

# A Methodology for Using Professional Knowledge in Corpus Annotation

A Dissertation

Presented to

The Faculty of the Graduate School of Arts and Sciences

Brandeis University

Computer Science

James Pustejovsky, Advisor

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Amber C. Stubbs

February, 2013

This dissertation, directed and approved by Amber C. Stubbs's committee, has been accepted and approved by the Graduate Faculty of Brandeis University in partial fulfillment of the requirements for the degree of:

**DOCTOR OF PHILOSOPHY**

Susan J. Birren, Dean of Arts and Sciences

Dissertation Committee:

James Pustejovsky, Chair

Guergana Savova

Marc Verhagen

Nianwen Xue

©Copyright by

Amber C. Stubbs

2013

For my husband, BJ

# Acknowledgments

This dissertation would never have been completed (or started, for that matter) without the help and support of many people. First, many thanks to my advisor, James Pustejovsky, whose encouragement, knowledge, and sense of humor made this dissertation (and our book!) not only possible, but more fun than I could have expected.

Thanks also to all the professors and researchers who provided me with guidance and listened to my ideas over the years: Rick Altman, Vincent Carey, George Hirsak, Guergana Savova, and Nianwen Xue. Special thanks to Marc Verhagen, who was always willing answer my questions and provide feedback on paper drafts and new ideas, and to Lotus Goldberg, for keeping me grounded with her excellent advice and sense of perspective.

Writing this dissertation was made easier by the people who shared the experience with me. Thanks to Anna Rumshisky and Roser Saurì, for their moral support and proving that there is life after grad school; to Jessica Moszkowicz for being my mentor and friend through my years at Brandeis; to Alex Plotnick for asking the hard questions, but also for sharing a sense of the absurd; and to Seohyun Im, my comrade-in-arms throughout the dissertation writing process. Thanks also to everyone who kept the Computer Science department running: Myrna, Jeanne, Julio, Chris, and all the CS Gurus.

My family and friends have been a constant source of love and support through my years in grad school. Thanks to my parents, for always being willing to lend an ear while I worked through a problem; to my brother Royce, for reminding me it's okay to relax sometimes; and to the rest of my siblings and their families for supporting me, even when it wasn't clear to anyone (including me) what exactly I was doing. Many thanks to my knitting group friends, who were an unbelievably excellent source of support, humor, advice, and (of course) yarn. You are all enablers in the best possible way.

Finally, there are not enough thanks in the world for my husband, BJ. I could never have completed this dissertation without his tireless encouragement, advice, patience, and love.

# Abstract

## A Methodology for Using Professional Knowledge in Corpus Annotation

A dissertation presented to the Faculty of  
the Graduate School of Arts and Sciences of  
Brandeis University, Waltham, Massachusetts

by Amber C. Stubbs

It is a well-known problem that performing linguistic annotation over a corpus can be an expensive and time-consuming task. The problem of annotation becomes even more difficult to solve when the task is based around a domain-specific corpus or specification. For example, extracting diagnosis information from clinical notes must be done by someone with sufficient medical training, who can understand all of the medical jargon and determine if a diagnosis can be made. However, hiring medical professionals to perform syntactic or semantic annotations can be extremely expensive, and few domain-expert annotators will have the time to create such an annotation.

This dissertation aims at finding a way to capture expert domain knowledge quickly and easily as annotations, and in a format where the information can then be used for more advanced natural language processing (NLP) tasks.

To that end, this dissertation proposes the use of *light annotation tasks*: linguistically under-specified, task- and domain-specific annotation models that can quickly capture expert knowledge in a corpus as it relates to a research question. The corpora created from light annotation tasks can then be augmented with additional, denser annotations (such as part-of-speech tagging), or used directly with an NLP system.

In addition to defining the light annotation task, this dissertation presents a set of principles that can be used to create annotation tasks for domain experts. These

principles are based on examining other “light” annotations, as well as the existing standards and methodologies used in more traditional annotation research. Software designed for light annotation projects is also presented.

Finally, in order to illustrate the utility of light annotations, a case study based around the medical research task of finding patients qualified to participate in a clinical study is presented. The medical settings that influence the case study’s design are discussed, and the light annotation task’s implementation is analyzed. The resulting corpus (called the Patient Evaluation Resource for Medical Information in Text (PERMIT) corpus) is then leveraged into a preliminary NLP system, which demonstrates the versatility of the light annotation methodology.



# Contents

<b>Abstract</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivations . . . . .	2
1.2 Goal of this dissertation . . . . .	3
1.3 Approach . . . . .	4
1.4 Related Work . . . . .	5
1.5 Overview . . . . .	17
<b>I Annotation Standards, Methodologies, and Tools</b>	<b>20</b>
<b>2 Desiderata of Annotation Tasks</b>	<b>21</b>
2.1 Corpus Creation and Selection . . . . .	21
2.2 General Annotation Desiderata . . . . .	24
2.3 Annotation Representation . . . . .	26
2.4 Annotation Guidelines and Reporting . . . . .	28
2.5 Annotation Tools . . . . .	31
2.6 Annotation Process . . . . .	34
<b>3 The MATTER Cycle</b>	<b>36</b>
3.1 Overview of MATTER . . . . .	37
3.2 MATTER and existing standards . . . . .	41
3.3 MATTER and light annotation tasks . . . . .	42
<b>4 Representing Expert Knowledge</b>	<b>44</b>
4.1 Approaches to Bioclinical Annotation . . . . .	46
4.2 Defining Light Annotation Tasks . . . . .	55
4.3 Principles of Light Annotation Tasks . . . . .	63
4.4 Methodology of Light Annotation Tasks . . . . .	68
4.5 Light annotation tasks and identified desiderata . . . . .	69

## CONTENTS

4.6	Overview of light annotation tasks . . . . .	72
<b>5</b>	<b>Tools for Light Annotations</b>	<b>74</b>
5.1	Existing annotation tools . . . . .	77
5.2	MAE - Multi-purpose Annotation Environment . . . . .	80
5.3	MAI - Multi-document Adjudication Interface . . . . .	85
5.4	Use and availability of MAE and MAI . . . . .	87
 <b>II The PERMIT Corpus: a case study in using light annotation</b>		 <b>88</b>
<b>6</b>	<b>Settings for a Domain Expert Clinical Task</b>	<b>89</b>
6.1	Goal of the case study . . . . .	90
6.2	Methods . . . . .	90
6.3	Research settings . . . . .	91
6.4	Overview of clinical trial settings . . . . .	101
<b>7</b>	<b>Domain Expert Annotation</b>	<b>104</b>
7.1	Annotation Task Settings . . . . .	105
7.2	The PERMIT annotation cycle . . . . .	109
7.3	Evaluation of the PERMIT corpus . . . . .	116
7.4	PERMIT corpus as a light annotation task . . . . .	126
7.5	Summary . . . . .	129
<b>8</b>	<b>Using Expert Annotation in an NLP system</b>	<b>131</b>
8.1	Data distribution in the PERMIT corpus . . . . .	132
8.2	Establishing a baseline accuracy with ML classifiers . . . . .	137
8.3	Keyword-based selection . . . . .	139
8.4	Using document structure for analyzing the PERMIT corpus . . . . .	144
8.5	Additional analyses . . . . .	150
8.6	Summary . . . . .	151
<b>9</b>	<b>Conclusions and Future Work</b>	<b>152</b>
9.1	Conclusions . . . . .	152
9.2	Future Work . . . . .	154
 <b>III Appendices and References</b>		 <b>157</b>
<b>A</b>	<b>Eligibility criteria analysis</b>	<b>158</b>

*CONTENTS*

<b>B Case-Control Annotation DTD</b>	<b>162</b>
<b>C Case-control Annotation Guidelines</b>	<b>164</b>
C.1 Overview . . . . .	164
C.2 Annotation . . . . .	165
<b>D Inter-annotator agreement table</b>	<b>168</b>
<b>E Selection criterion extent analysis</b>	<b>171</b>
<b>F Sample of augmented SecTag database</b>	<b>187</b>
<b>G NLP tools for the bioclinical and temporal domains</b>	<b>189</b>
<b>References</b>	<b>197</b>

# List of Tables

7.1	Keywords used to find initial annotation corpus . . . . .	108
7.2	Precision, recall, and f-measure using Annotator 2 as the gold standard	120
7.3	Precision, recall, and f-measure between Annotator 1 and the adjudicated gold standard . . . . .	125
7.4	Precision, recall, and f-measure between Annotator 2 and the adjudicated gold standard . . . . .	126
8.1	Relationship between initial corpus selection and case/control group membership . . . . .	134
8.2	Presence of keywords in the case and control groups . . . . .	134
8.3	Distribution of met criteria . . . . .	136
8.4	Distribution of Selection Criterion tags and attributes in corpus . . .	137
8.5	Baseline machine learning classification accuracy values per selection criterion . . . . .	138
8.6	Keyword-based ‘diabetes’ classifications compared to gold standard .	142
8.7	Keyword-based ‘recent cardiac event’ classifications compared to gold standard . . . . .	142
8.8	Keyword-based ‘no history of cardiac event’ classifications compared to gold standard . . . . .	143
8.9	Keyword- and section-based ‘diabetes’ classifications compared to gold standard . . . . .	148
8.10	Keyword-based ‘recent cardiac event’ classifications compared to gold standard . . . . .	149
8.11	Keyword-based ‘no history of cardiac event’ classifications compared to gold standard . . . . .	150
D.1	Abbreviations for tag and attribute combinations used in the confusion matrix table . . . . .	169
D.2	Confusion matrix annotations . . . . .	170

*LIST OF TABLES*

E.1	Modifiers and extents used to identify patients who met the diabetes criterion . . . . .	174
E.2	Modifiers and extents used to identify patients who did not meet the diabetes criterion . . . . .	175
E.3	Modifiers and extents used to identify patients who met the recent cardiac event criterion . . . . .	178
E.4	Modifiers and extents used to identify patients who did not meet the recent cardiac event criterion . . . . .	181
E.5	Modifiers and extents used to identify patients who met the no history of cardiac events criterion . . . . .	182
E.6	Modifiers and extents used to identify patients who did not meet the no history of cardiac events criterion . . . . .	186
F.1	Sample of the modified SecTag database . . . . .	188

# List of Figures

3.1	The MATTER cycle. . . . .	37
3.2	The Model-Annotate Cycle . . . . .	39
3.3	Corpus divisions for training and testing . . . . .	40
3.4	The Training-Evaluation Cycle . . . . .	40
4.1	A representation of a possible relationship between a light annotation model, $M_1$ , and a full annotation model, $M$ . . . . .	58
5.1	MAE: Multi-purpose Annotation Environment . . . . .	82
5.2	MAI - Multi-document Adjudication Interface . . . . .	86

# Chapter 1

## Introduction

It is a truism in the natural language processing community that corpus annotation is both an expensive and a time-consuming task. These two problems are compounded when the annotation being created requires knowledge that is outside of traditional linguistic or computational linguistic expertise, such as analysis of biomedical or clinical documentation, or other files that require a working knowledge of a specific domain.

This dissertation presents a novel methodology for capturing this type of professional knowledge, where domain experts are required in order to annotate the selected corpus. This methodology takes the form of a *light annotation task*, which is a linguistically under-specified annotation model designed to address a particular domain- and task-specific question. This model can be used to quickly capture expert knowledge in a corpus so that it can later be used as a first layer in a more semantically complete annotation tasks, or into natural language processing (NLP) systems.

In support of this new methodology, a set of annotation and adjudication software for light annotation tasks is also presented, and a case study based in the clinical

domain is used to demonstrate the potential uses of light annotation tasks.

## 1.1 Motivations

As computers become more adept at providing us with the information that we need, the information we request of them correspondingly increases in complexity. However, the more complex the information being requested, the more preparation needs to be done in order for computers to provide that information. Temporal queries are an excellent example of this problem: while humans are always asking questions such as “When was Wilson president?”, for a computer to provide an accurate answer it must understand the concept of “president”, that “Wilson” was a person, and that a time-related inquiry is being made.

Questions like this pose serious problems to a computer, and question answering systems struggle with time-sensitive queries. The standard approach to this problem is to use data annotated by humans—often linguistic researchers and students—with appropriate representations of the required knowledge, and to use that data to teach computers what cues to examine when presented with new queries or texts. The annotation process is complex and time-consuming, particularly when many aspects of the data must be represented, such as in the example of Wilson’s presidency.

The annotation problem becomes even more complex when the questions being asked are ones that require domain-specific<sup>1</sup> knowledge to understand and answer, such as questions related to medical studies. Consider the difficulties of determining whether a patient has had a heart attack within the past three years. For a computer

---

<sup>1</sup>While linguistic knowledge is also a specific domain of learning, standard practices for creating annotations are generally known to linguistic expert annotators. For the purposes of this dissertation, “domain expert” and “professional” are taken to refer to domains of knowledge other than linguistics.



## CHAPTER 1. INTRODUCTION

to make that determination, not only does it need to be able to compute the concept of “within the past three years”, it also needs to have a reference point—three years from when? On top of that, in a medical document the term “heart attack” may be used, but it is more likely to be called a “myocardial infarction”, or “MI”, or “STEMI” if it is identified as a particular type of heart attack.

The standard approach to data annotation fails here: a linguist is unlikely to be able to understand the jargon found in medical documents, but a medical expert is also unlikely to have the time or inclination to complete a detailed linguistic and temporal analysis of a full corpus of medical documents so that a dataset suitable for NLP can be created. An additional complication to the problem is cost: a domain expert who is qualified to answer complex questions on medical data is likely to have had years of training and experience in the field (for example, an M.D., R.N. or medical coder), and it is expensive to hire such qualified people for the length of time that would be required to fully annotate the needed corpus. Due to this overlap of three difficult problems (complexity, domain-specific knowledge, and cost), a methodology for creating meaningful annotation tasks that can be performed quickly is needed, particularly in domains that require expert knowledge to evaluate.

### 1.2 Goal of this dissertation

In order to address the problems inherent to domain-specific queries (specifically, queries that are not linguistic in nature) outlined above, this dissertation aims at defining a methodology that can be used to leverage domain-expert knowledge in the pursuit of processing complex data. This is done by defining the concept of a *light*

## CHAPTER 1. INTRODUCTION

*annotation task* and providing a set of principles<sup>2</sup> for creating these tasks that are simple enough to not be time-consuming or expensive when hiring experts as annotators, but complete enough that the expert annotations can be used in conjunction with other annotation layers in order to create a model that can be used in NLP systems, such as for creating analysis rules or for training machine learning (ML) algorithms. These ‘light’ annotation tasks are essentially task-specific annotation models that can be used to quickly capture expert knowledge in a corpus as it relates to a research question. Light annotations contrast with more traditional annotation tasks, such as complete part-of-speech tagging or full semantic role labeling, which are both linguistically complex and textually dense.

### 1.3 Approach

This dissertation approaches the goal of light annotation for domain experts by first examining other factors that must affect how such a task is defined, such as how the expert annotation will fit into existing annotation standards; how it can be used in conjunction with other annotations; other factors that may influence the success of a project, such as annotation tools; and finally, how the methodology can be practically applied to an actual annotation task.

The current standards and general desiderata for annotation research, as well as existing annotation tasks and tools that have been applied to all types of annotation tasks (including those that have been used in the bioclinical domain) are examined in Part I of this dissertation. Based on the results of this research, a methodology

---

<sup>2</sup>In a previous publication (Stubbs, 2012) these principles were referred to as ‘guidelines’, but because the term ‘guidelines’ is already used in the annotation community to describe the instructions for applying an annotation model to a corpus, ‘principles’ is used here instead.

## CHAPTER 1. INTRODUCTION

is proposed than can be used to create annotation tasks aimed at easily encoding expert knowledge. A set of annotation and adjudication tools designed to complement domain expert annotation tasks is also presented.

In order to demonstrate the application of a light annotation task, a case study is analyzed in Part II of this dissertation. Specifically, this case study involves the creation of the PERMIT (Patient Evaluation Resource for Medical Information in Text) corpus, which applies the principles and methodology outlined in Part I to a set of hospital discharge summaries. These summaries are analyzed by medical professionals for adherence to predetermined selection criteria. A light annotation specification is used to encode the relevant data points in the medical records, the inter-annotator agreement is assessed for the task, and a gold standard created from the annotations is used for preliminary research into using software to recreate the annotation results.

### 1.4 Related Work

Linguistic annotation finds its roots in corpus linguistics, a field which has grown steadily since the 1960s<sup>3</sup>, despite the warnings of linguists (most notably, Noam Chomsky (Chomsky, 1957)) who were wary of relying on corpora for linguistic insight<sup>4</sup>. While at the time these cautions were not without merit, advances in computational ability and the use of methodologies designed to ameliorate the potential

---

<sup>3</sup>Although some researchers place the beginning of the modern corpus linguistics era in the 1980s (McEnery and Wilson, 1996), as it was not until that time that the field was considered mainstream in linguistics, the Brown Corpus project (Kucera et al., 1967) began in the 1960s and is generally regarded as the first large-scale computerized corpus. As such it is reasonable to place the start of modern-day corpus linguistics there.

<sup>4</sup>Indeed, as recently as 2004, Noam Chomsky said in an interview that “Corpus linguistics is meaningless” (Andor, 2004).

## CHAPTER 1. INTRODUCTION

biases in collected corpora led to the eventual growth of corpus linguistics as its own field of research, which has directly influenced the creation of related fields such as computational linguistics and natural language processing.

The annotation of natural language plays a large role in many aspects of research into human language. However, as this dissertation seeks to provide a methodology for creating corpora that encode domain expert knowledge in a useful, easy to obtain manner, the examination of related work is limited to discussions of annotation methodologies and relevant annotation efforts, in particular those relating to domain-specific tasks. With this restriction in mind, the following sections examine methodologies as they exist in different corpus-related disciplines.

### 1.4.1 Annotation Methodologies in Corpus Linguistics

A survey of Corpus Linguistics textbooks and prominent papers reveals a surprising lack of methodology discussions as they relate to creating manually annotated corpora. While the corpus *building* process is widely analyzed, and the use of automated tagging software (such as part-of-speech recognition systems) is often discussed at length, the actual process of applying tags to a corpus by human annotators is often mentioned only in passing. The results of manual annotations are discussed, but the procedure by which such a resource is created has largely been ignored, at least until recently.

The first general methodology that has been applied to corpus annotation is the MATTER cycle (Pustejovsky, 2006; Pustejovsky and Stubbs, forthcoming 2012), which describes a system for creating annotation projects and applying them to machine learning algorithms. Other than that, the research most relevant to the goal of

## CHAPTER 1. INTRODUCTION

this dissertation involves other standards and criteria that can be used to inform the corpus annotation process, such as the Linguistic Annotation Framework (LAF) the ISO standard for representing annotation data (Ide and Romary, 2006) and Leech’s seven maxims for annotation tasks (Leech, 1993). These and other established and emerging standards in corpus annotation will be discussed more thoroughly in Chapter 2.

None of the existing standards or guidelines addresses the added complexity of capturing domain-expert knowledge in a corpus, although the concept of specific rather than general annotation tasks is not altogether new. The concept of ‘problem-oriented tagging’ was first addressed in 1984 by Pieter de Haan (de Haan, 1984), and annotation tasks that make use of expert-level knowledge are not uncommon, particularly in the bioclinical domains (Kim et al., 2008; Uzuner et al., 2007; Wilbur et al., 2006). The following sections will examine these annotation tasks in more detail.

### 1.4.2 Problem-oriented Tagging

‘Problem-oriented tagging’ is a phrase coined by Pieter de Haan (1984) to describe annotation tasks that “only provide information for the problem or structure being investigated”. In essence, this idea relies on creating a set of tags that is only meant to reflect a small aspect of the language—only that which is relevant to the phenomenon being studied. This provides an interesting counterpoint to the majority of annotation tasks, which rely on defining standard tagsets for entire linguistic sub-fields (such as part-of-speech tagging, syntactic parsing, etc.) and annotating a corpus with the entire set of tags. It is easy to see how the concept of problem-oriented tagging could

## CHAPTER 1. INTRODUCTION

be applied to corpus investigations requiring expert knowledge: by limiting the tags used to ones that apply only to the question being asked, the annotation project is automatically simplified and streamlined. However, no theoretical or methodological framework for creating problem-oriented tagsets is provided by de Haan, nor was one provided in later years by subsequent researchers, and in fact the concept is often dismissed or ignored by textbooks on the subject of corpus linguistics and annotation.

For example, in *Corpus Linguistics* (McEnery and Wilson, 1996), the authors devote a large section of Chapter 2 to discussing the current standards for annotation encoding, as well as various types of existing annotation (part of speech tagging, tree parsing, phonetic transcription, etc). However, no framework for creating tasks outside of those being discussed is given. In section 2.2.3 they discuss the concept of problem-oriented tagging, but rather than suggesting a way that such tasks could be created and used, McEnery and Wilson determine that the problem cannot be generalized in such a way so as to provide a methodology:

“ [...] problem-oriented tagging uses an annotation scheme which is selected not for its broad coverage and consensus-based theory-neutrality but for the relevance of the distinctions which it makes to the specific questions which each analyst wishes to ask of his or her data. In view of this dependence on individual research questions, it is not possible to generalize further about this form of corpus annotation, but it is a very important type to keep in mind in the context of practical research using corpora.” (pg. 57)

The phrase ‘problem-oriented tagging’ does not seem to have made a lasting impact on the literature of corpus linguistics: it is mentioned again in another book by McEnery (“Corpus-Based Language Studies” (McEnery et al., 2006)) and in one by Meyer (Meyer, 2002), but again, no suggestions are made for methodologies for using this approach. In other books on corpus linguistics and corpus annotation (McEnery

and Hardie, 2012; Gries et al., 2010; Wynne, 2005; Geoffrey Sampson, 2004; Garside et al., 1997; Meunier et al., 2011) , the phrase does not appear to be mentioned at all, nor do they discuss approaches that utilize a similar strategy under a different name.

### 1.4.3 Task-specific Annotations

The lack of specified methodologies for problem-oriented annotation has not, however, led to a lack of annotation tasks that bear similarities to the problem-oriented approach. Instead, the phrase “task-specific” has been substituted as a descriptor for that type of annotation, but still no set of guidelines for task-specific annotations has been presented.

Annotation projects that are task-specific are fairly common, particularly in conferences and workshops where researchers and labs are invited to participate in shared tasks and challenges. These shared tasks often involve an initial corpus and dataset that participants use to build or train systems to perform a particular linguistic task, such as word sense disambiguation or machine translation. Usually participants are given only a few weeks or months from the time the data is released to when they must report on their results or provide working systems to the task evaluators. Top-performing systems are then presented in the proceedings of the conference or workshop. While not all of the challenges discussed in this section are *only* task specific (some still involve broad part-of-speech tagging or semantic role labeling), most contain subtasks that can be considered task-specific, such as temporal processing, named entity recognition, sense disambiguation, and so on.

The first examples of this type of challenge were the Message Understanding Con-

## CHAPTER 1. INTRODUCTION

ferences (MUC), which were funded by the U.S. government and were focused primarily on named entity recognition and coreference resolution (Grishman and Sundheim, 1996). Seven MUCs were held between 1987 and 1997. While annotating named entities and coreference are still quite difficult and fairly broad, compared to full part-of-speech tagging, these annotations are clearly quite task-specific.

The MUCs were promptly followed by SENSEVAL, an “open, community-based evaluation exercise for WordSense Disambiguation programs” (Kilgarriff and Palmer, 2000). The first SENSEVAL focused on disambiguating words from English, Italian, and French texts, though in SENSEVAL 2 (2001) the languages studied were expanded to include Basque, Chinese, Czech, Danish, Dutch, English, Estonian, Italian, Japanese, Korean, Spanish, and Swedish. SENSEVAL 3 (2004), in addition to the word sense disambiguation tasks, also examined “semantic roles, multilingual annotations, logic forms, subcategorization acquisition” (Mihalcea, 2012).

SENSEVAL eventually became SemEval in 2007, which marked an expansion of the SENSEVAL tasks into a workshop that included not only word sense disambiguation in a variety of languages, but also metonymy resolution, semantic annotation, and temporal processing (SemEval, 2007). The second SemEval was held in 2010, and the third will be held in 2012 (Erk and Strapparava, 2010; SemEval, 2012). Each SemEval workshop hosts a wide range of semantic annotation tasks.

The Conference on Natural Language Learning (CoNLL) has also been hosting shared tasks as part of its proceedings since 1999. These tasks have included topics such as noun phrase bracketing, chunking, semantic role labeling, and syntactic dependencies (Sang, 2010).

Similarly, the Text REtrieval Conference (TREC) also hosts shared tasks (referred to as ‘Tracks’), focused on information retrieval from a variety of document types and



## CHAPTER 1. INTRODUCTION

for a variety of purposes. The TREC's started in 1992 and are still ongoing, with a variety of different tracks being offered each year (TREC, 2000).

However, while the MUC tasks initially focused on military transmissions, the majority of the CoNLL and SENS- and SemEval tasks used natural language texts that did not require extensive training to understand. The semantic content could be analyzed by the linguists and computer scientists participating in the challenges, and no domain-specific knowledge was required to create or interpret the annotations. Additionally, the corpora largely consisted of essays, newspaper articles, and other similar writing, rather than scientific, medical, or other domain-specific texts.

This is not true of the TREC Tracks, however, as from 2003 through 2007 a track focused on retrieving genomics data specifically from biomedical texts (Hersh and Voorhees, 2009). This track was then followed by the Medical Record Track in 2011 and 2012, which was another task specifically designed to examine biomedical texts. Tasks such as the Medical Record Track and Genomics Track are examples of domain-specific annotations, which will be discussed in the next section.

### 1.4.4 Domain-specific Annotations

Task-specific annotations for texts such as newspapers soon led to task-specific annotations for specialized domains, referred to here as domain-specific annotations. Of particular relevance to this dissertation are those annotation projects and challenges that have been done for the biomedical and clinical domains, though any type of annotation task requiring professional knowledge of a subject outside of linguistics or general language understanding could be considered a domain-specific task. Annotating legal contracts, for example, would be an excellent example of domain-specific

## CHAPTER 1. INTRODUCTION

annotation.

This section provides an overview of corpora and annotation tasks and challenges in these areas, with a primary focus on clinical resources, as those are more relevant to the case study presented later in this dissertation.

It is important to note that, as with the previously mentioned tasks and challenges, the participants are not generally required to annotate their own documents. While the process by which an annotated corpus was created for use in the task may be described in a later paper by the tasks organizers, these projects do not represent systematic *annotations* undertaken by disparate groups of researchers, but rather the building of systems for machine learning over the corpora provided. Therefore, the majority of papers resulting from these challenges do not discuss annotation methodologies, but rather the ML techniques used to approach the challenge.

Similarly, most other papers and articles relating information about NLP research projects undertaken in any area of natural language research, including the biomedical and clinical domains, do not usually discuss general methodology strategies, but instead provide information about their own annotation experience, such as the number of annotators, inter-annotator agreement scores, and other task-specific information. Papers that do discuss some aspects of methodologies in clinical and biomedical annotation will be discussed in Chapter 4. Because this dissertation utilizes a case study of clinical documents, the next sections focus primarily on resources and annotations in the biomedical and clinical domains.

### **Extant Biomedical and Clinical Corpora**

In order to perform domain-specific annotations, an appropriate domain-specific corpus must be assembled. The GENIA corpus (Kim et al., 2003) has been used for

## CHAPTER 1. INTRODUCTION

numerous biomedical annotation projects, including both private research and shared tasks and challenges (BioNLP2011, 2011; BioNLP2009, 2009; Farkas et al., 2010; Kim et al., 2008; Zhou et al., 2004). The GENIA corpus is comprised of roughly 2,000 MEDLINE abstracts and articles, which were selected by searching for the terms *human*, *blood cell*, and *transcription factor* (Kim et al., 2003).

In addition to the GENIA corpus, there are a number of other biomedical corpora, many of which are made up of MEDLINE abstracts, including the PDG (Protein Design Group) corpus (Blaschke et al., 1999), the YAPEX corpus (Franzén et al., 2002), and the University of Wisconsin corpus (Craven and Kumlein, 1999). A fairly up-to-date list of available biomedical corpora can be found here: <http://compbio.ucdenver.edu/ccp/corpora/obtaining.shtml>.

While biomedical corpora do require domain expert knowledge to interpret, the few existing corpora of clinical records are more germane to the case study presented in this dissertation. Unfortunately, one of the greatest hurdles facing researchers interested in clinical annotations and NLP is that it is extremely difficult to obtain permission to access patient medical record information for research purposes unless one is affiliated with a hospital. Chapman et al. (2011) suggest that due to “concerns regarding patient privacy and worry about revealing unfavorable institutional practices, hospitals and clinics have been extremely reluctant to allow access to clinical data for researchers from outside the associated institutions”. Even when an affiliation is present, the use of medical records will often only extend to the people directly involved in the research, and so any annotations or corpora used for research cannot be shared with others without the data being thoroughly de-identified, and sometimes even then it is not possible to obtain the proper permissions to distribute the annotated texts.

## CHAPTER 1. INTRODUCTION

As a result of this situation, very few publicly available clinical corpora exist, and often papers written about tools for parsing Electronic Health Records (EHRs—sometimes called Electronic Medical Records, or EMRs) describe results obtained on data that is not available to other researchers. For example, Pakhomov et al. (2006) developed a hand-annotated corpus of clinical notes for testing how accurate a Penn TreeBank-trained part-of-speech tagger would be in the medical domain. They discovered that having a domain-specific training set greatly increased the accuracy of the tagger, but it appears that this training set has not been made available for others. Similarly, the set of discharge summaries used by Cao et al. (2004) for summarization, the clinical notes used by Friedman et al. (2004) for automatically determining medical codes, the discharge summaries used by Long for diagnosis extraction (2005), the medical records used in the CLEF annotation project (Roberts et al., 2007), and the corpus of clinical documents from the VA used in an annotation project by South et al. (2009) are all private resources at the time of this writing.

Fortunately, there are some datasets of medical records that have been sufficiently de-identified to be made mostly public, and these are available to researchers who register for access to them. The University of Pittsburgh’s BLULab has a repository of de-identified medical records (<http://nlp.dbmi.pitt.edu/nlprepository.html>), which include almost 8,000 discharge summaries, as well as radiology reports, progress notes, and so on. Similarly, the BioScope corpus (Vincze et al., 2008) consists of clinical texts including radiology reports, as well as full papers and abstracts from the biomedical domain.

The largest resource of available medical records is the MIMIC II Clinical Database (MCD), a collection of de-identified medical records that includes nursing notes, discharge summaries, ICD9 codes, and more. The documents in the MCD have been

## CHAPTER 1. INTRODUCTION

de-identified by removing patient, doctor, and hospital names, and by systematically changing the dates in each patients files (the consistency of the temporal relationships within each patient’s records was maintained, so it is still possible to perform temporal reasoning over the data provided) (Clifford et al., 2010). The available portion of the MCD consists of 26,588 patient records from the Intensive Care Units of Beth Israel Deaconess Medical Center in Boston.

In terms of annotation methodology, it is vital that any annotation project have a suitable corpus of relevant and accurate data to analyze. The restrictions placed on the sharing of clinical data make this a more difficult hurdle to overcome than other domains of expert annotation, but the existing clinical corpora do provide starting points for research into the clinical domain, even for researchers who are not affiliated with hospitals or other medical facilities.

### **Overview of existing biomedical and clinical annotations and shared tasks**

Some of the available biomedical and clinical corpora were created as a result of domain-specific challenges and shared tasks that were affiliated with biomedical NLP conferences and workshops. In addition to the aforementioned TREC genomics and medical records tracks (Hersh and Voorhees, 2009; TREC, 2000), the BioCreAtIvE (Critical Assessment for Information Extraction in Biology) workshops have been hosting tasks ranging from gene mention tagging, gene normalization, interactor annotation to the more recent (2011) tasks related to biocuration workflow (BioCreAtivE, 2006). Similarly, in 2009 and 2001 the BioNLP workshop hosted main shared tasks related to bio-event extraction and domain recognition, as well as supporting tasks such as identifying co-reference, entity relations, and negation (BioNLP2011, 2011; BioNLP2009, 2009; Kim et al., 2009).

## CHAPTER 1. INTRODUCTION

Of special relevance to this dissertation are the i2b2 (Informatics for Integrating Biology and the Bedside) Center’s shared tasks, which have focused exclusively on extracting medical information from clinical documents. The i2b2 challenges usually have a gold standard corpus associated with them for training, and previous tasks have involved identifying obesities and co-morbidities, coreference labeling patient smoking status, extracting medication names and dosages and finding relations between events and entities (Uzuner, 2008; Uzuner et al., 2012; i2b2 team, 2011; Uzuner et al., 2010a; Uzuner et al., 2010b; Uzuner et al., 2007).

In addition to annotations and corpora associated with shared tasks and challenges, individual groups have pursued their own clinical annotation projects. For example, the CLinical E-science Framework (CLEF) collaboration from Sheffield University resulted in semantic annotation tags and guidelines specifically for patient medical records (Roberts et al., 2008; Roberts et al., 2007). The tags and guidelines are currently available to anyone from <http://nlp.shef.ac.uk/clef/TheGuidelines/TheGuidelines.html>, though the corpus is not. There have been a plethora of other annotation efforts over medical documents, including temporal annotations for ordering events (Zhou et al., 2007; Bramsen et al., 2006), medication information extraction (Gold et al., 2008), relating EHRs to ICD9 codes (Friedman et al., 2004), phenotypic information related to Inflammatory Bowel Disease (South et al., 2009), uncertainty and negation (Vincze et al., 2008), anaphoric relations and coreference (Savova et al., 2011; Cohen et al., 2010), medical disorders (Ogren et al., 2006), document structure (Denny et al., 2008) and part-of-speech tagging (Pakhomov et al., 2006). Naturally, this is not a complete list of all annotation tasks performed over clinical text, but it is a roughly representative sample of existing biomedical and clinical annotations.

## CHAPTER 1. INTRODUCTION

As with the majority of other publications on annotation projects, these papers do not contain discussions of general methodology, though they do often contain information about specific aspects of their tasks, such as number of annotators, the annotation goal and specification, and other relevant information about their corpus. In fact, some of these annotation tasks can be considered to be *light* in the “under-specified” sense used in this dissertation. While again, these light annotations do not discuss general methodologies for those types of annotations, an analysis of those tasks in conjunction with the case study in Part II of this dissertation led to the development of the light annotation principles discussed in Chapter 4.

### 1.5 Overview

This dissertation proposes a novel methodology for taking advantage of domain expert knowledge in the form of light annotations tasks by examining existing annotation standards and best practices as well as existing domain expert annotation tasks. It also presents annotation and adjudication tools made for light annotations tasks, and examines a case study involving an annotation in the clinical domain to demonstrate the applicability of light annotation tasks. To that end, the rest of the dissertation is organized as follows:

#### **PART I: Annotation Standards, Methodologies, and Tools**

- **Chapter 2** examines the desiderata of linguistics corpus annotation, including existing and emerging standards in the annotation community, particularly annotation representations, tools, and processes;

## CHAPTER 1. INTRODUCTION

- **Chapter 3** describes the MATTER cycle, the first methodology for linguistic annotation that can be applied to all levels of linguistic annotation;
- **Chapter 4** explores the problems with utilizing domain expert knowledge in annotation projects, defines the concept of “light annotation”, and explores existing light annotation tasks, particularly in the biomedical and clinical domains. Finally, it presents principles and a methodology for creating new light annotation tasks;
- **Chapter 5** describes annotation and adjudication tools created for capturing expert knowledge using light annotations;

### **PART II: The PERMIT Corpus: a case study in using light annotation**

- **Chapter 6** presents the medical research settings for the case study, including an overview of epidemiological studies, examinations of selection and matching criteria, the format of discharge summaries and information about temporal expressions in both eligibility criteria and medical records;
- **Chapter 7** discusses how the methodology from Chapter 4 was applied to the creation of the PERMIT corpus, as well as how the corpus was selected, considerations for future annotation tasks, and the adjudication of the gold standard;
- **Chapter 8** analyzes the information in the PERMIT corpus and shows how the light annotation can be leveraged into a basic NLP system. Performance of the system is compared to baseline scores generated by machine learning algorithms;



## *CHAPTER 1. INTRODUCTION*

- **Chapter 9** summarizes the contributions of this dissertation, and discusses future improvements and applications of this research.

# Part I

## Annotation Standards, Methodologies, and Tools

## Chapter 2

# Desiderata of Annotation Tasks

Any methodology for domain expert annotation tasks must be created in accordance with current accepted standards (and identified standards that are still under development) for linguistic annotation tasks. Because light annotations can only exist in the context of more traditional annotation tasks, the literature review provided in this chapter examines the different aspects of natural language annotation. This allows the proposed methodology and principles to be grounded in current best practice as described in the literatures of corpus linguistics, as well as biomedical and clinical NLP.

### 2.1 Corpus Creation and Selection

As noted in Chapter 1.4, natural language processing has its roots in corpus linguistics, and it is there that we find some of the most in-depth discussions regarding corpus selection, particularly with regards to creating corpora that are *representative* and *balanced*. At this time, there are no existing all-purpose guidelines that can be

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

used to determine exactly when a corpus can be deemed appropriate for analysis in general or for one task in particular, and this dissertation does not attempt to create such guidelines. Below is an overview of what has been written on this topic to date, both in corpus linguistics in general and the biomedical domain in particular.

McEnery et al. (2006) paraphrase Leech's (1991) definition as follows: "a corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety" (pg. 13). However, it is no mean feat to determine if a corpus can, in fact, be used to generalize about its subject matter. If a corpus is being collected to provide an overview of, for example, American English then it will need to provide broad coverage of all different genres, styles, registers, and so on. Meyer (2002) provides an excellent overview of all the potential factors that can affect whether a corpus can be determined to be sufficiently representative.

Much like the distinction between general and task-based annotations, corpora that are assembled for particular purposes—such as to examine a particular linguistic construction (e.g., passive voice) or to provide a resource for annotation and ML testing of some type of reasoning (e.g., temporal analysis)—will have different criteria for what it means to be "representative". For a corpus exploring passive voice, is it the genres and styles of the text that the examples come from that is important, or is it the form of the sentences that the passive constructions are in? Biber (1993) suggests that "Representativeness refers to the extent to which a sample includes the full range of variability in a population" (where 'population' refers to, essentially, the scope of the corpus being explored). However, he notes that the theoretical considerations the corpus is being collected to examine will determine what population should be explored.

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

Similarly, ‘balance’ can be defined in different ways as well. In some cases, a corpus is considered to be balanced when it contains roughly equal numbers of examples of all types of the text it is meant to represent. On the other hand, some corpora require that the different types of examples be proportional to how each type is represented in the real world. This can be particularly important in corpora that will be used for machine learning purposes, as a disproportionate number of examples of a particular linguistic construction can cause an algorithm to apply a particular label more frequently than would be seen in actual natural language.

Clearly, representativeness and balance are not easy concepts to define, as they are quite context-dependent. In general, a good rule of thumb for corpus creation is that ‘bigger is better’, however this can be difficult to apply to corpora that will be annotated, as the annotation process is time-consuming and can be quite expensive, particularly when annotators need to be experts in a particular field. The idea of always having more data is also complicated when a corpus is being compiled in a domain-specific field, where large amounts of data may simply not be available (as is the case with clinical data; see Section 1.4.4).

On the topic of domain-specific corpora, Cohen et al (2005) examined corpus design for biomedical processing. Their study primarily focuses on the attributes that an annotated corpus should have in order to be widely used/useful, rather than solely on the factors that should be considered when finding documents for the corpus. The factors they identified as significant are: 1) recoverability of original text and annotations, 2) availability of guidelines and documentation, 3) balance and representativeness, 4) presence of annotation<sup>1</sup>.

---

<sup>1</sup>Cohen et al. make a distinction between a corpus, which they define as a set of annotated data, and a ‘text collection’, which they define as “textual data sets that [...] do not contain mark-up of the document contents”. As this is a non-standard usage of the term ‘corpus’, it will not be used

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

While not being strictly about the corpus building process, Cohen et al.'s paper does address other factors that should be considered when creating any annotated corpus, not just ones in the biomedical domain. Indeed, these suggestions are echoes of other guidelines that have been suggested in the past (as the authors themselves note), and ones that are still being implemented today, as will be discussed later in this chapter.

In terms of domain-expert annotation tasks, the size of the corpus and the need for representativeness and balance must be weighed against the potential cost of the annotated resource being created, as well as the ultimate goal of the annotation task. The “bigger is better” corpus rule is one that conflicts sharply with the restraints that often restrict domain-expert annotation tasks, where time and money can limit how much data can be annotated by a domain-expert consultant. However, this very conflict highlights the need for a light annotation layer that can capture the information that requires an expert to interpret without demanding an excessive amount of time be dedicated to the task.

### 2.2 General Annotation Desiderata

One of the sources that Cohen et al. (2005) cite in their analysis of successful corpora was Leech's seven maxims for annotation schemes (1993). These maxims are some of the first guidelines published for determining how to create an annotation task. Though they are generic enough in scope to not truly be guidelines for an annotation methodology, they have had a lasting impact on the field of corpus annotation and must be considered in any review of the literature. These maxims are paraphrased further in this dissertation.

---

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

below:

1. The annotation should be easily separated from the corpus– “the raw corpus should be recoverable”;
2. Similarly, the annotations should be able to stand alone, away from the corpus as well;
3. The scheme should be based on symbols, definitions, and guidelines that are available to the users of the corpus;
4. The information about how the annotation was created and who performed the annotation should also be available;
5. The scheme cannot claim to be, or be presented as, the only way of representing the task being undertaken;
6. Annotation schemes should be based on theory-neutral analysis of data;
7. No single annotation scheme should be held up to be the ‘standard’, though emergence of data standards should be encouraged.

Most of these maxims have been incorporated into the canon of what is considered best practice for corpus building: maxims 1 and 2 define how the annotation data should be separable from the main corpus, an idea that has been encoded into the Linguistic Annotation Framework/Graphical Annotation Framework, which will be discussed further in Section 2.3; maxims 3 and 4 have been incorporated into annotation guidelines and reporting suggestions, which will be discussed in Section 2.4, and the remaining maxims refer to the way that the annotations should be regarded by the community.

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

In what at first appear to be a violation of maxims 5-7, there has been a movement to standardize some aspects of language annotation. For example, the ISO (International Organization for Standardization) has standards for annotation representation (the aforementioned LAF/GrAF), as well as standards for some semantic annotation tasks, temporal representation (Pustejovsky et al., 2010), spatial representation (Pustejovsky et al., 2011), word segmentation (ISO-24614, 2011), and other standards, managed under the ISO group TC 37/SC 4. However, while efforts have been made to standardize the specifications and guidelines for these tasks, these remain *de facto* rather than *de jure* standards in the annotation community, and many offshoots and variations of all types of annotations exist, so the spirit of the last few maxims is still preserved. These aspects of annotations are discussed further in Section 2.3.

### 2.3 Annotation Representation

As previously mentioned, the Linguistic Annotation Framework (LAF) (Ide and Romary, 2006) exists as an ISO standard for representing annotation data for a corpus. Other suggestions for standards have been made over the years (such as those put forth by the Text Encoding Initiative (TEI, 1987) and the Corpus Encoding Standard (CES, 1996)), and LAF provides models for encoding data that aim to “provide a standard infrastructure for representing language resources and their annotations that can serve as a basis for harmonizing existing resources as well as developing new ones” (Ide and Romary, 2007).

Rather than dictate the format that annotations should take, LAF simply requires that any format used be mappable to an *abstract data model* that is defined by a rigid “dump” format. The creators of an annotation scheme are asked to provide a schema



## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

that maps their corpus annotation to the dump format, which allows all annotation to be convertible from one form to another (Ide and Romary, 2007). LAF also embraces *stand-off annotation* as a standard for annotation representation (Ide et al., 2003), where the annotations are kept separate from the data being annotated to maintain the integrity of the corpus.

LAF also specifies the creation of a Data Category Registry (DCR), which will be used to harmonize the content of different annotations by providing a resource for finding “pre-defined data elements and schemas” (Ide and Romary, 2007). As of the writing of this dissertation, a preliminary version of the DCR exists as <http://www.isocat.org/>.

The Graphical Annotation Framework (GrAF) is an extension of LAF that provides a platform for merging, analyzing, and visualizing disparate annotation structures by creating an XML serialization of LAF’s dump format (Ide and Suderman, 2007). GrAF has been used effectively to transmute output from various annotation systems, such as Apache’s UIMA (Unstructured Information Management Applications) (Ide and Suderman, 2012).

While the LAF and GrAF standards are still primarily *de facto* standards, despite their ISO affiliations, they provide a solid base for representing linguistic annotations that can be used to maximize interoperability between different annotation schemas or specifications.

The focus of LAF and GrAF on interoperability is supremely important to the idea of light annotation tasks. As these tasks are designed to be augmented with other annotations in order to maximize output from automated systems, it is imperative that any light annotation task be created in a format that will not conflict with tags or labels that are applied to the data later on. In particular, the use of stand-off an-

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

notation to represent light annotations is an effective way to address this issue. While some have objected to the use of stand-off annotation on the grounds that it leads to a loss in performance when being queried compared to more ‘traditional’ in-line annotations (Dipper et al., 2007), these objections are based on the software available to process those formats. Not only has that software improved as more people use stand-off formats for annotated data, but the importance of interoperability trumps any lags in performance that may be experienced.

Clearly, any light annotation task should still adhere to existing guidelines for representation, particularly ones that help ensure compatibility between annotation schemes. Because light annotations are meant to capture only expert knowledge, not linguistic knowledge, it is necessary that the light annotation representation be one that can be augmented with other annotation schemes, so as to maximize the ability to leverage the encoded expert knowledge.

### 2.4 Annotation Guidelines and Reporting

In the annotation community, “annotation guidelines” refer to the instructions given to the annotators to create the annotated corpus. While there is no set standard for what information about an annotation task should be made available to users of the corpus, or how the guidelines should be written or presented, there has been discussion of those topics that help inform annotation creators as to what aspects of their work others find useful. For example, the following list of guideline users is paraphrased/restructured from Dipper et al. (2004a):

- *The annotator*: Annotators apply the annotation specification (the tags and attributes) to the chosen corpus by following the annotation guideline instruc-

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

tions. Interests: the goal of the annotation, and how to apply the annotation specification to the corpus.

- *The corpus explorer*: Explorers are interested in using the corpus to examine linguistic theories. Interests: how to find examples of linguistic phenomena and interpret the annotation, information about the corpus itself.
- *The language engineer*: Engineers want to use automatic methods to explore the corpus and annotations; for example: through extracting linguistic information through evaluation scripts, or using the corpus to train machine learning algorithms. Interests: tagsets, corpus creation methods.
- *The guideline explorer*: The annotation guidelines themselves are of interest to linguists who want to understand the theory behind the annotation, or people who want to create their own guidelines for a different annotation task. Interests: guideline creation and underlying theory.
- *The guideline author*: Because guidelines often have to be revised multiple times before a task, the guideline authors themselves will often need to refer back to their own work to ensure the coverage of the guidelines is complete. Interests: clear organization of instructions.

It should be noted that Dipper et al. (2004a) use the term 'guidelines' to refer to all the information relevant to researchers, including information about the corpus, the linguistic theory, the annotation, and so on, while this dissertation uses the traditional definition of 'guidelines', as described at the beginning of this section. However, despite this terminological difference, Dipper et al. make excellent points about the

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

information that should be made available when an annotated corpus is released to the public based on the different interests of various types of researchers.

Another recent study by Bayerl and Paul (2011) examined what factors in an annotation task may influence inter-annotator agreement (IAA) scores. They examined 96 annotation tasks over three categories of study (word-sense disambiguation, prosodic transcriptions, and phonetic transcriptions), and were able to identify seven aspects of annotation tasks that influenced IAA scores: “annotation domain, number of categories in a coding scheme, number of annotators in a project, whether the annotators received training, the intensity of annotator training, the annotation purpose, and the method used for calculation of percentage agreements” (Bayerl and Paul, 2011). For the purposes of this dissertation, however, the factors influencing IAA scores are less interesting than the fact that the authors wanted to analyze more factors that could affect annotation, but couldn’t find enough consistent reporting on those factors in the studies that they looked at to determine statistical significance.

As a result of this lack of information, Bayerl and Paul developed a list of factors that they suggest should be included in all reports of annotation efforts. These are:

1. Number of annotators;
2. Type and amount of material annotated;
3. Number of categories in the scheme;
4. Criteria for selecting annotators;
5. Annotator’s expert status (novices, domain experts, schema developers, native speakers, etc.);

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

6. Type and intensity of training;
7. Type and computation of the agreement index;
8. Purpose for calculating the agreement index (including whether the goal was to reach a certain threshold or achieve “highest-possible” agreement).

The very fact that these fairly basic pieces of information are not always included in reports on annotation tasks shows that there are no established or adhered to standards for reporting on annotation tasks, either in papers or guidelines. However, these two papers provide excellent suggestions for what information should be included in reporting about an annotation task, and how it should be presented.

In terms of light annotations tasks, while it is clearly important to report on the knowledge and training of annotators, particularly for domain expert tasks, it is an equally pressing issue that the annotations guidelines fully show the annotators what their task is. In keeping with the idea of a light annotation task, the guidelines should be equally ‘light’—that is not to suggest that they should not fully explain the annotation the experts are being asked to create, but rather that they should not contain extraneous information that could slow down the annotation of the corpus.

### **2.5 Annotation Tools**

Any annotation task that is being undertaken by human annotators must have accompanying software that allows for the selected corpus to be marked up with the chosen annotation scheme. However, opinions differ on whether annotation tools should be specialized—providing support for only one type of annotation task in order to speed

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

up the annotation process—or generalized—providing support for a diverse set of annotation tasks so that annotators and researchers can learn to use a single piece of software and not require repeated training sessions.

Dipper et al. (2004b) developed a list of requirements for annotation tools based on their analysis of a set of 12 research projects from a variety of disciplines. Their requirements (paraphrased) are:

- **Diversity of data:** Support of different modalities (written or spoken), different character sets, and annotation units (sentence, discourse, etc.);
- **Multi-level annotation:** Support of different levels of annotation, such as syntactic, morphological, etc.;
- **Diversity of annotation:** Support for pairs of relations (directed and non-directed), and cross-level relation annotation;
- **Simplicity:** Tools must be simple to use and create tasks for; time learning how to use the tool should be minimal;
- **Customizability:** Support for creation of new tagsets; tags and attributes can be easily modified;
- **Quality assurance:** Support for making annotations consistent and complete, as well as compliance with encoding standards;
- **Convertibility:** Support of converting data from one format to another, by either providing standardized output, or providing built-in conversion tools.

Here, Dipper et al. are supporting the idea of general-purpose tools: their list of requirements promote tools that can be used for many different types of tasks, though

## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

they acknowledge that “individual requirements might be of different relevance to different annotation projects” (ibid.).

On the other hand, Reidsma et al. (2005), who performed a meta-analysis of papers critiquing annotation tools (including that of Dipper et al. (2004b)), reach the conclusion that “...to meet the annotation requirements for very large corpora, it may be necessary to develop annotation tools that are specialized to reduce the time and effort for creating the annotations.” Somewhere between these two views, Dybkjaer and Bernsen (2004), in their evaluation of multi-modal annotation tools reach the conclusion that “Comparing ... tools is clearly a multi-dimensional exercise. No tools is just simply better or poorer than another”, but still advocate for the creation of a general-purpose multi-modal annotation tool.

In some ways, however, the debate over whether tools should be specialized or generalized is moot: a variety of annotations tools of all types exist, from the very general-purpose GATE (Cunningham et al., 2010) to task-specific tools such as SAPIENT (Liakata et al., 2009), a web browser plug-in specifically for creating sentence-level annotations. The range between those two extremes is not small: a search of the LRE Resource Map<sup>2</sup> for “annotation tools” returns over 200 results. While there are many instances of repetition in those results (the Map is not currently curated), there is clearly a variety of annotation tools available.

Of all the desiderata discussed in this chapter, the annotation tool is the one that will have the most direct impact on a domain expert annotator, particularly when an annotation task has a limited amount of time or money. In that case, it is best for the tool to be easy to use and appropriate for the annotation task. This aspect of domain-expert annotations will be discussed further in Chapter 5.

---

<sup>2</sup><http://www.languagelibrary.eu/lremap/>

## 2.6 Annotation Process

As Section 1.4.1 illustrates, books on corpus linguistics do not provide discussions of the process required to create an annotated corpus. It was not until recently that this topic has been examined, beginning with Pustejovsky’s description annotation methodology in 2006 (Pustejovsky, 2006).

Palmer and Xue (2010) also addressed the issue of linguistic annotation, and they noted that the “development of an annotation scheme requires addressing at a minimum the following issues [...]:

- target phenomena definition;
- corpus selection;
- annotation efficiency and consistency;
- annotation infrastructure;
- Annotation evaluation;
- Use of machine learning for pre-processing and sampling.”

The authors go on to provides details for aspects of each item that need to be addressed when defining an annotation scheme, such as deciding whether the annotation can be done automatically, or if the task can be narrowed down to a smaller set of tags; whether the corpus that is chosen should be full articles or could be isolated sentences; how the annotation guidelines can be defined and the need for testing the annotations before finalizing the scheme; how the tool chosen will affect the annotation process, through data management and/or ease of use; the appropriate way of



## CHAPTER 2. DESIDERATA OF ANNOTATION TASKS

finding inter-tagger agreement scores for the task; and what types of pre-processing can be used in conjunction with the annotated corpus.

In terms of a light annotation, the process used has less of an impact on the annotators, and more on the researchers organizing the annotation and its uses. The items and discussion provided by Palmer and Xue provide an excellent way of viewing the annotation process, and some similarities exist between these items and the stages of the MATTER cycle—however, the work described later in this dissertation was done within the paradigm of the MATTER cycle, and so a fuller description of that system is given in Chapter 3.

# Chapter 3

## The MATTER Cycle

The MATTER cycle provides the first general methodology for an annotation and machine learning task. It was conceptualized by Pustejovsky in 2006 as the *Annotate, Train, Test* model for annotation (Pustejovsky, 2006), and has recently been expanded upon in the book *Natural Language Annotation for Machine Learning* by James Pustejovsky and Amber Stubbs (forthcoming 2012). This chapter provides an overview of the MATTER cycle as it is presented in *Natural Language Annotation for Machine Learning*.

This material is presented here in order to provide a platform from which the methodology of light annotations tasks can be discussed. The end of the chapter describes the relationship between the full MATTER cycle as it is used for natural language processing and machine learning and the principles of light annotation tasks.

### 3.1 Overview of MATTER

MATTER stands for *Model, Annotate, Test, Train, Evaluate, Revise*. Figure 3.1 provides a visualization of the cycle. These steps describe a general methodology for creating annotation and machine learning tasks of all different types, from part-of-speech tagging to detailed semantic or discourse analysis.

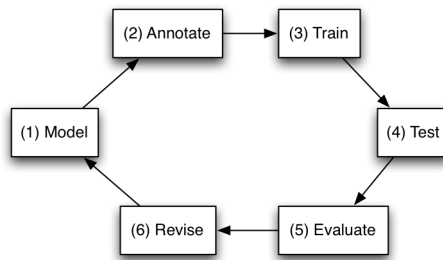


Figure 3.1: The MATTER cycle.

The rest of this section describes the MATTER cycle in more detail.

#### 3.1.1 Goal, corpus and annotators

Prior to the beginning of the MATTER cycle, it is important to define the goal of the annotation task, and determine the source and method for collecting an appropriately balanced and representative corpus. Naturally, as the MATTER cycle progresses it is possible that the goal (and therefore the metrics by which it is determined if the corpus is sufficiently balanced/representative) may change, but establishing these two aspects of an annotation task early in the process helps keep the project on track.

While it is not a requirement that specific annotators be chosen this early in the cycle, it is recommended that the researcher have an idea of what sort of background

## CHAPTER 3. THE MATTER CYCLE

and experience the annotators will have, as this information can affect the model chosen and how the guidelines are written.

### 3.1.2 M - Model

The Model of an annotation task is the specification or schema that describes what the annotation will be—the tags, attributes, and other features that will be added to the corpus being annotated. The Model can be described as  $M = \langle T, R, I \rangle$ , where  $M$  is the model,  $T$  is the set of terms being used,  $R$  is the relations between those terms, and  $I$  is the interpretation of the terms and relations (Pustejovsky and Stubbs, forthcoming 2012).

Generally speaking, only a single model is used during an annotation task, and that model represents both the information that is going to be collected from the corpus during the course of the annotation, and the information that will be recreated later using machine learning or other NLP systems.

### 3.1.3 A - Annotate

The Annotate step in the MATTER cycle actually represents a number of different steps, from writing the annotation guidelines, finding annotators, selecting annotation software, and testing the Model on the corpus through practice annotations. In fact, it is useful to envision the Model and Annotation phases of MATTER as a smaller cycle—the MAMA (Model-Annotate-Model-Annotate) cycle, or the “babbling” phase of the annotation development process (see Figure 3.2). This is the part of the process where the problems with the model and annotation system are worked out, and the final versions of both are determined so the full annotation task can be completed.

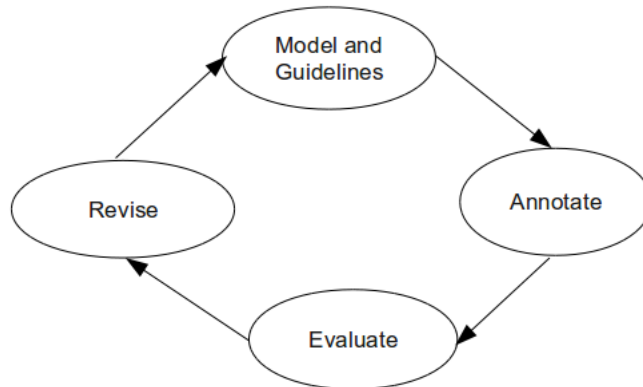


Figure 3.2: The Model-Annotate Cycle

Once the MAMA cycle is complete—when there are no more changes being made to the specifications or guidelines, and they have been applied to the entire corpus—and there is an annotated and adjudicated Gold Standard corpus with the complete Model represented, the machine learning part of the process can begin.

### 3.1.4 T - Train

Training is the process by which a machine learning (ML) algorithm is taught to look for and recognize features that will be used for creating the desired output from the system. Training is only performed on a part of the corpus that has been set aside for that purpose—generally the corpus is split into three parts: development training (training, roughly 44% of the full corpus), development testing (dev-test, 22%), and final testing (testing, 33%) Figure 3.3 shows a graph of the distribution proportions.

The training set is used to supply the chosen algorithm with the features it will use later on, in the testing phase of the MATTER cycle. The features that it is trained to look for will generally be annotation-based (reliant on the annotations created during

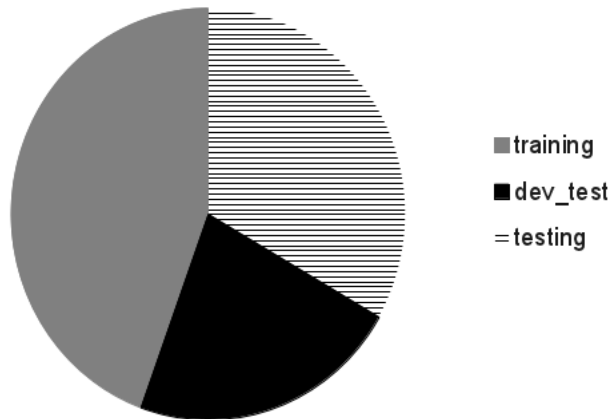


Figure 3.3: Corpus divisions for training and testing

the MAMA cycle), or structure-based (reliant on the format of the document), though other sources of features can also be used (for example, dictionaries or ontologies).

### 3.1.5 TE - Test and Evaluate

Much like the Model and Annotation stages, the Training, Testing, and Evaluation stages are also involved in a smaller cycle within MATTER: the Training-Evaluation cycle (see Figure 3.4).

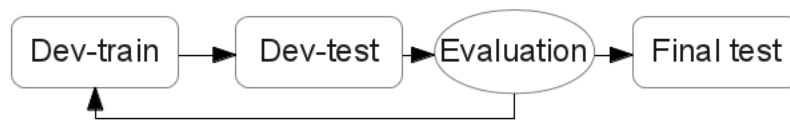


Figure 3.4: The Training-Evaluation Cycle

Once the algorithm has been trained on the training data, it is run on the dev-test data, then evaluated for accuracy (usually with precision and recall scores and F-measures). If the scores are not good, the features are changed, the algorithm re-trained, and the output evaluated again until a satisfactory level of accuracy is

## CHAPTER 3. THE MATTER CYCLE

reached, at which point the algorithm is run over the test dataset, which is then used to calculate the accuracy scores that are reported in papers or presentations.

### 3.1.6 R - Revise

The Revision state of the MATTER cycle is the point at which the entire project, from corpus selection to ML evaluation results, is reviewed. Topics for revision include: aspects of the task that may have contributed to poor performance, changes to the task that could result in improved performance later on, and new applications of the task that could be done successfully in the future, such as expanding the task to a new language or domain.

## 3.2 MATTER and existing standards

Because the MATTER cycle is a general description of the process used for creating annotated data and training machine learning algorithms, there is no conflict between it and the existing (or proposed) annotation standards described in Chapter 2. The MATTER cycle is agnostic to the decisions made regarding corpus selection, annotation tools, representation formats, etc. It does, however, provide a set of guidelines for the process of creating an annotated corpus and using that corpus for machine learning techniques, and provides standard reference points for the process as it is usually performed. As was discussed in Section 1.4, until the MATTER cycle was developed there was no set methodology for annotation tasks, so this cycle provided a stable platform for the development of annotation tasks, and therefore provides a way for light annotation tasks to be added to the toolkit of natural language researchers.

### 3.3 MATTER and light annotation tasks

As explained in Section 3.1.3, in a standard iteration of the MATTER cycle, or specifically the MAMA cycle, the entire Model of the annotation task is applied to the chosen corpus by the annotators. However, this paradigm can cause problems when domain expert knowledge is needed for a task. There are many examples of annotation tasks where it would be impractical to ask a medical professional to annotate all the aspects of the Model in the corpus due to the time it would take to add them and the cost of hiring domain experts (doctors, nurses, biologists, etc.) as consultants or annotators. While complex annotation tasks using domain experts have certainly been done in the past (Kim et al., 2008; Roberts et al., 2008), such endeavors are expensive in both time and money.

Therefore, this dissertation proposes, when possible, the creation of a *light annotation task*, which is essentially a linguistically under-specified, task- and domain-specific Model that overlaps with the full annotation task and is used to quickly capture expert knowledge in a corpus as it relates to a research question, but does not require the experts to perform intensive annotation tasks such as part-of-speech tagging or semantic role labeling. Instead, the domain expert is given an annotation task that can represent the information that they are being asked to discover, which can then be augmented later with the rest of the Model by other annotators. This light annotation can then be augmented with other annotation layers, created by other annotators or automatically generated with software. While the concept of layered and light annotations is not new, prior to this dissertation there did not exist an analysis of existing light annotations, or what is an effective way to encode domain expert knowledge in an annotation task.



### *CHAPTER 3. THE MATTER CYCLE*

The next chapter discusses existing light annotation tasks and other strategies for annotations in the bioclinical domain, and presents principles for the creation of light annotations that are compatible with the standards described in Chapter 2 and the methodology of the MATTER cycle.

# Chapter 4

## Representing Expert Knowledge

As discussed in Section 1.4.4, while a number of biomedical and clinical corpora and corpus annotations exist, there is no standardized methodology for capturing domain expert knowledge in an annotation task. Traditionally, the basic structure of the MATTER cycle has been used for developing corpora using domain expert annotators, but this practice has recently come under criticism by some of the researchers attempting to create these annotated resources due to the difficulty in asking domain experts to perform full semantic or syntactic annotations. Papers exploring the difficulties of domain expert annotations have recently been published, and the case study presented in Part II of this dissertation provides an in-depth look at some of the problems domain-specific tasks can present.

Xia and Yetisgen-Yildiz (2012) recount their own experiences in managing three different clinical annotation tasks requiring domain expert knowledge (identifying critical recommendations in radiology reports, diagnosing disease from chest x-ray reports, and diagnosing pneumonia from ICU reports), and identify several areas of the annotation process where the addition of domain experts causes additional

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

challenges to the annotation task. In addition to the usual troubles that plague any clinical annotation task (obtaining Internal Review Board (IRB) approval, the legal problems surrounding trying to release a corpus of clinical notes, and the expense of hiring domain experts as annotators), they note that for domain expert annotation tasks, much of the work in the traditional annotation cycle falls on the shoulders of the domain experts, because the NLP researchers are not necessarily qualified to perform tasks such as writing guidelines and finding annotators in the clinical domain. However, because domain experts are not likely to be familiar with standard practices for writing annotation guidelines and training annotators, the annotation process is affected by lack of agreement between annotators (which, they note, is sometimes caused by different medical backgrounds as well—something that is critical to a radiologist may not be read the same way by a general practitioner). Additionally, many domain experts do not realize how much time and effort must go in to creating a gold standard corpus, which can become problematic when the project takes longer than anticipated.

Similarly, Scott et al. (2012) also observe that it can be difficult to find qualified domain experts who are willing to spend the time required to create an annotated corpus, and further discuss the general problem of annotation tasks not generally being viewed as scientific tasks, leading to a dearth of wholly accepted standards for creating and evaluating annotated corpora.

The problems outlined above need to be addressed in order to create quality bio-clinical annotations. Existing approaches to these problems (including those proposed by Xia and Yetisgen-Yildiz and Scott et al.) will be discussed in Section 4.1, and a novel set of guidelines for creating light annotation tasks in the biomedical domain<sup>1</sup>

---

<sup>1</sup>These guidelines (and some of the themes of this chapter) were first presented in Stubbs (2012).

are presented in Section 4.3.

## 4.1 Approaches to Bioclinical Annotation

A variety of different tactics have been used in order to create annotated corpora in the bioclinical domains. Some of these focus solely on the annotation of the corpus, while others take a more holistic view and attempt to re-frame the entire corpus-creation process. This section provides an overview of some of the different ways bioclinical annotation has been approached. It should be noted that some of these processes (and others not discussed here) have been attempted for other, non-domain expert annotation tasks, but these are not germane to the problem at hand and are therefore not discussed further.

The approaches discussed here can be divided into three general types: changes to the annotators and specifications; changes to the annotation system; and alterations in the entire annotation task structure.

### 4.1.1 Annotators and Annotation Specifications

One problem with annotation tasks for clinical documents is that it can be difficult for researchers of differing backgrounds (for example, linguists and physicians) to see eye to eye on what should be included in the specification for a clinical annotation task. In order to create an annotation specification for a task focused on annotating clinical conditions, Chapman and Dowling (2006) worked together in an iterative process (essentially, using the MAMA cycle) to create a specification for the task that both authors/annotators (one a bioinformaticist with a background in linguistics, the other a physician) could apply to emergency department reports to obtain high

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

inter-annotator agreement scores. The experiment was successful, and the f-measure for annotations over 20 documents was 93% using a schema with 45 variables and 10 exceptions (an exception was a place in the text where it would seem that a particular tag applied, but the guidelines specifically stated that it should not be—for example, the ‘thorazine’ in “allergic to thorazine” would not be annotated as a medication, but the whole phrase would be annotated as a clinical condition).

In later research, the specification (referred to as the ‘Annotation Schema’) developed by Chapman and Dowling was expanded and used in an experiment to see if physicians agreed more often when performing the annotation task, or if laypeople (in this case, bioinformatics masters students) would be able to perform the annotation equally well (Chapman et al., 2008).

The experiment was again iterative, with the annotators first being trained for one hour on a Baseline Schema that listed only the medical concepts that were to be annotated, and the annotators were asked to annotate a set of seven documents. Later, each annotator was trained for an hour on the Annotation Schema (a version of the specification from Chapman and Dowling (2006) which was expanded to contain 57 variables divided into three medical concepts) and asked to annotate three documents, after which detailed feedback was provided for each of the annotations. This annotate-feedback process was repeated twice more. Finally, after three months of performing no further annotations, the annotators were asked to annotate a set of seven reports, with no additional training or feedback.

Unsurprisingly, annotation accuracy and inter-annotator agreement rose quickly once the annotation-feedback cycle began, even with the more complicated annotation specification. Somewhat more surprisingly, the lay annotators performed nearly as well at the Annotation Schema task as the physicians, although the scores were not

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

quite as high for the laypeople, and they forgot more of the specification during the three-month break (Chapman et al., 2008).

This research does not necessarily indicate that someone with no medical training will be able to perform complex diagnoses over medical records. The specification did not require that the annotators be able to understand a patient’s condition, merely that they be able to recognize a phrase as a clinical condition, test result, or other high-level concept. Were the annotation task to diagnose whether a patient had, or was at risk for, a particular disease, asking a person with no medical background to make that determination based a medical record would be a poor use of that person’s time and the researcher’s resources (and would, presumably, not present a positive outcome for any patients relying on the research).

However, this idea of using laypeople for annotation tasks is one that has been researched in other areas of study, not just the bioclinical domain. Another way that some researchers have been looking to make the process of domain-specific annotations cheaper and faster is by *crowdsourcing*—that is, rather than asking two or three people to provide hundreds of annotation; using the Internet to ask hundreds of people to do a few annotations each. The most commonly used platform for crowdsourcing at the moment is Amazon Mechanical Turk<sup>2</sup>(AMT), a website where people can create Human Intelligence Tasks (HITs), which are then performed by non-expert annotators (both in the sense that the annotators are not domain experts, nor are they linguistic experts ) all over the world. Crowdsourcing has been used fairly effectively for some types of linguistic annotation tasks such as event recognition, word sense disambiguation, word similarity and textual entailment (Snow et al., 2008) and not only in the bioclinical domain. However, it has also been criticized for producing

---

<sup>2</sup><https://www.mturk.com/mturk/welcome>

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

poor quality annotations for other types of tasks, and there is some concern that the HIT paradigm is exploitative (Fort et al., 2011). Due to the limitations placed on sharing even de-identified medical data, it is difficult to obtain permission to use AMT and other crowdsourcing sites as annotation resources for clinical corpora. Additionally, the lack of guaranteed expert knowledge makes bioclinical annotations through generic crowdsourcing particularly difficult, though work being done to use expert crowds is discussed in Section 4.1.3.

### 4.1.2 Annotation Approaches

Given that one of the problems with creating an annotated bioclinical corpus is that domain experts are not always familiar with what it means to annotate a document according to a specification and set of guidelines, one logical approach to making the process easier is to modify the way that the annotation is created. While medical professionals are familiar with chart reviews and evaluating those on a set of criteria, the processes for text annotation used in computational linguistics is not always familiar territory for domain experts.

One modification to the system of annotation was used during the addition of event markups to the GENIA corpus, where events in text were annotated and mapped to an event ontology by domain experts (Kim et al., 2008). However, due to the complexity of the task, the researchers developed two policies for the annotation process in order to create more accurate annotations: *Text-bound Annotation* and *Single-facet Annotation*.

The concept behind *Text-bound Annotation* is relatively simple: “Associate all annotation with actual expression in text” (ibid.). Essentially, this means that anno-

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

tators should not infer the existence of information in the text; rather, they should only annotate the information that can be verified through assertions in the data. The practice of requiring annotators to mark only what is explicit in the text is one that appears fairly regularly, particularly in domain-specific annotations.

Another annotation concept that was used during the GENIA event annotation is that of *Single-facet Annotation*, described as: “Keep the view point for annotation as simple and focused as possible” (ibid.). For the purposes of the GENIA annotation, this means that the annotators were asked to read the texts while focused solely on identifying events in the text and connecting them to the ontology: the researchers describe this as ‘defining one aspect of the text as the focus of annotation’ (ibid).

While the GENIA annotation task had only one focus (event annotation), the concept of Single-facet Annotation is one that can be applied to annotations that use multiple tags as well: instead of giving an annotator the entire specification and guidelines for all of the tags at once, the annotation can be split into different tasks based on each tag, and the annotators can focus on each facet of the annotation one at a time. This approach can reduce errors in an annotation task by lowering the cognitive load on the annotators at any given stage of the process. A similar approach to annotation is used in the Brandeis Annotation Tool (BAT), which splits annotation tasks into layers that are annotated one at a time (Verhagen, 2010), an approach that was very effective during the TempEval-2 annotation task (Verhagen et al., 2010) as well as for word sense disambiguation in the bioclinical domain (Savova et al., 2008).

Another approach to the process of annotation that has been used in the bioclinical domain is that of “accelerated annotation”, which is based on the active learning framework of training machine learning algorithms, and was created to speed up the annotation process for sparse corpora (Tsuruoka et al., 2008). Active learning is



## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

a system for training statistical classifiers that is based on the idea that “machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns” (Settles, 2010), and has been used for other tasks in the bioclinical domain, such as coreference resolution (Miller et al., 2012). Essentially, this means that the system queries a human annotator about a piece of data that it selects for itself<sup>3</sup>, the human’s annotation is added to the pool of data, the classifier is retrained and a new query is selected based on the improved model. Generally speaking, the full annotation specification is not presented to the human annotator, only the parts that the system needs information about.

The “accelerated annotation” process described in Tsuruoka et al. (2008) is based on the active learning approach, but differs slightly in goal. While active learning seeks to train a classifier by annotating only a subset of the given corpus, the goal of Tsuruoka et al. was to create a complete corpus of annotated named entities that could be used by other researchers (ibid). The primary difference in the accelerated annotation system from active learning methods is that the accelerated annotation framework sought to train the classifier only to detect what section of the text was most likely to have an entity that needed to be annotated; the framework did not attempt to automate the actual annotation process, which is the purpose of most active learning systems.

Each sentence that was identified as being of interest was then sent to a human annotator to be marked up. This system therefore allows domain expert annotators to add their knowledge to the corpus much more rapidly than if they had to read the entire dataset themselves. The authors note, however, that this method for

---

<sup>3</sup>There are a variety of ways that this piece of data is chosen; for an overview of all the different methods, see Settles (2010).

speeding up annotation is best used in a corpus where the objects being annotated are sparse, but they find that the system allowed the number of sentences that had to be annotated reduced by almost half (ibid).

### 4.1.3 Annotation Cycles

Finally, some researchers have chosen to consider the entire MAMA cycle, and have looked at ways to improve that process for the clinical domain. While the cycle itself remains unchanged, better ways to incorporate the knowledge of NLP researchers and domain experts have been suggested.

Xia and Yetisgen-Yildiz (2012), based on their own experiences and problems with domain-specific annotations (discussed at the beginning of this chapter), made the following suggestions for improving the annotation experience for domain experts (paraphrased):

- Have the annotators work together to create the annotation guidelines and review each other’s work as part of the training. Doing so significantly increased annotator agreement scores.
- Have annotators provide additional information in their annotations—instead of simply having them apply a label to the text (such as whether or not the person described has a particular disease), have the annotators indicate what parts of the text led to their conclusion.
- Be sure that the domain experts (in this case, physicians) are aware of how much time it can take to create a gold standard annotated corpus and that they are able to make the appropriate time commitment.

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

- Have the NLP researchers involved early in the creation of the annotation corpus and task, as they will be more familiar with the annotation process and common pitfalls.

Although they are not explicitly stated, there are similarities between these suggestions and some of the other techniques that have been used for bioclinical annotations. The suggestion of having annotators work together to create the guidelines is similar to the technique used to create the specification in Chapman and Downing (2006), as well as the technique used to train the annotators in Chapman et al. (2008). Additionally, the suggestion to have the annotators indicate the place in the text that they are basing their judgments on does help locate features for machine learning algorithms, as Xia and Yetisgen-Yildiz (2012) suggest, but it also means that the annotations are more likely to be based on what is in the text rather than what can be inferred, which echoes back to the Text-bound Annotation technique from Kim et al. (2008).

On the other hand, Scott et al. (2012) focus on amending the annotation process to be closer to the research techniques used in psycholinguistics and experimental psychology for collecting judgments from annotators in order to make the annotation process more scientific and repeatable.

Specifically, Scott et al. use their own annotation task—determining the certainty that a patient has a particular condition by interpreting the hedge statements used to preface the diagnosis (i.e., “may have”, “probably does not have”, etc.)—in clinical texts as a basis for this scientific approach to annotation. Rather than simply find some annotators and ask them to mark and label hedge statements in the texts, they set up a system where the annotators were presented with individual statements

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

containing hedges and given a scale from 0% to 100% to mark the certainty level of the statement being made. Each annotator was forced to provide a judgment for each sentence, and the system contained redundant sentences in order to control for participants who contradicted themselves.

The annotations were collected through the website SurveyMonkey<sup>4</sup>, and annotators were recruited through “professional newsgroups in medicine and biomedicine and to colleagues in medical schools”. The annotation effort was quite successful, with low standard deviations in hedge assessments for native speakers of American and British English.

Here, Scott et al. (2012) provide a way to crowdsource domain-expert annotations, as an alternative to crowdsourcing with Amazon’s Mechanical Turk system, which cannot guarantee domain expertise in the annotators. However, it should be noted that it is difficult to break down every type of annotation task into one that will be effective in a crowdsourcing system, and any task requiring a medical diagnosis may not be a good candidate for this type of annotation.

### 4.1.4 Overview of Bioclinical Annotations

Overall, the techniques for modifying tasks for the bioclinical domain that were reviewed here are all effective for the tasks they were targeted to handle. Unfortunately, none of them fully address the problems faced by NLP researchers who do not have easy access to domain expert trained annotators, but rather have to hire such experts as consultants for annotation tasks. In order to maximize the use of the resulting annotation but still minimize the time that the annotation takes to create, the an-

---

<sup>4</sup><http://www.surveymonkey.com/>

notation task must require a relatively low cognitive load on the annotator, but still be comprehensive enough that the information can be used later in NLP tasks. To that end, *light annotation tasks* are an ideal solution to the problems presented by annotations requiring domain expert knowledge.

## 4.2 Defining Light Annotation Tasks

Clearly, many of the approaches to annotation tasks outlined in Section 4.1 have been effective, but none of them fully address all the problems associated with domain expert annotations, particularly those of the time and money required to create annotated corpora. To address this problem, this dissertation proposes the use of *light annotation tasks*, a framework for annotation that is based on the MATTER cycle and is compatible with the established standards and desiderata of annotation projects (as described in Chapter 2). The purpose of a light annotation task is to create a dataset that represents complex information, but that is itself not complex (Stubbs, 2012).

A light annotation is an annotation task that uses linguistically underspecified, task-specific tags, i.e., tags that do not supply full syntactic or semantic content, but instead can be used to indicate more broadly the areas of interest in a document without being anchored to strict linguistic concepts. Here, “underspecified” does not mean that the tags are not specific to the task—indeed, their very specificity to, for example, a clinical question can be what keeps them from being clearly defined semantically, as will be shown in Part II of this dissertation. An example of a light annotation task can include top-level document classification tasks, where a single label is applied to an entire document.

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

In an annotation task, the Model for the task (as described in Section 3.1.2) can be described as  $M = \langle T, R, I \rangle$ , where  $M$  is the model,  $T$  is the set of terms being used,  $R$  is the relations between those terms, and  $I$  is the interpretation of the terms and relations. Traditionally in annotation tasks, a single Model is used to represent all of the tags and attributes that will be used to annotate a document. However, this approach is problematic when it comes to domain expert bioclinical annotations, because in many cases it would be impractical for NLP researchers to hire consultants to perform a full syntactic and/or semantic annotation for all of the information in a document that may be required to build an NLP system.

For example, if a researcher was interested in building an NLP system to determine whether a patient was diagnosed with a condition (such as asthma) based on their hospital discharge summary, there are a large number of aspects of the document that need to be accounted for in the system's feature set in order to accurately make a diagnosis. Some of these features include the section of the document (if asthma is mentioned in the 'Family History' section, it is much less likely to be related to the patient; more information about the structure of EHRs can be found in 6.3.3), whether there are hedge phrases or negations surrounding the key phrases, any possible coreferences that could be resolved to give a positive or negative diagnosis, and so on. In addition to these issues, NLP systems often use sentence markers, part-of-speech tags, syntactic chunks, and other linguistic features to help train statistical systems, and to write rules for rule-based systems.

It would clearly be a waste of time and money to ask that two or more domain experts annotate all of that information so that it can be incorporated into an automated system. Aside from the fact that a bioclinical domain expert is probably not familiar with syntactic theory, there is simply too much information that needs to

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

be annotated. However, the overall goal of the research effort can be described in a single question: *Does this patient have asthma?*, which is certainly a question that a pair of clinicians could answer. The simple classification task created by that question could even be augmented by asking *What parts of this text provide that information?* without disproportionately adding to the annotator’s cognitive load: since they would have to be reading the document anyway to make that determination, it would not be hard to have them mark the relevant passages.

Still, any additional information asked outside of the classification should not require linking to semantic classes, disambiguating word senses, or any other annotation task that is too grounded in linguistic theory. The domain experts should not be asked to create the full Model,  $M$ , that will be used for any future NLP systems, but should rather be asked to annotate with  $M_1$ , a complementary light annotation model that contains task-specific tags.

It should be noted that  $M_1$  is not necessarily a subset of  $M$ , just as  $\langle T_1, R_1, I_1 \rangle$  are not necessarily subsets of  $\langle T, R, I \rangle$ . Indeed, the relationship between  $M_1$  and  $M$  will depend greatly on the specific annotation task being attempted; and the relationship between the two could instead be that of supersets (should  $M_1$  use umbrella terms for the concepts in  $M$ ), or some other, less standardized relationship. The interpretation ( $I_1$ ) of the tags in  $M_1$  will generally also be semantically underspecified. Figure 4.1 shows a sample relationship between  $M$  and  $M_1$ , where the tags in  $M_1$  are broader in scope than those in  $M$ .

In effect,  $M_1$  provides a model for the task-specific question being asked, such as whether a patient has a particular condition based on their records, or whether a certain type of reaction is discussed in a biology paper. On the other hand,  $M$  is the model that will be used in the NLP system that will ultimately attempt to emulate

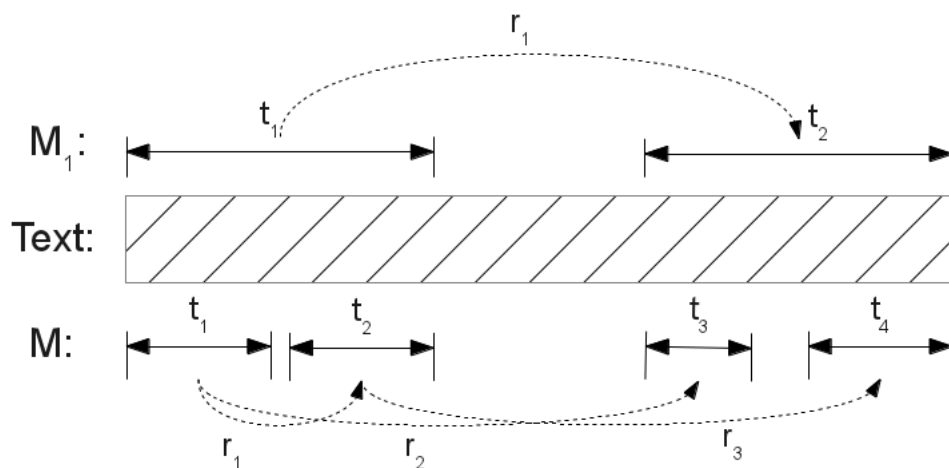


Figure 4.1: A representation of a possible relationship between a light annotation model,  $M_1$ , and a full annotation model,  $M$

the results of the  $M_1$  annotation. The exact relationship between these two models will depend on the the specific task being attempted, and what aspects of the text  $M_1$  reveals to be most relevant for training the NLP system.

While the above description of the relationship between a light annotation task and a traditional annotation model is new, light annotation tasks are already being used in the bioclinical domain. However, there is currently no set of guidelines for how to create effective light annotation tasks. Section 4.2.1 provides examples of how light annotations have been used already for clinical tasks, and Section 4.3 presents a novel set of guidelines for creating light annotation tasks.

### 4.2.1 Light annotations in the bioclinical domain

Many annotation studies in the biomedical and clinical domains cite the time and money required to create full semantic or syntactic annotations of clinical texts as barriers to good corpus creation, and many have sought ways to find the needed



## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

information without going over budget or running out of time (though they do not refer to their chosen process as being a light annotation). Using a light annotation specification is done mostly when the information needed about a document is a domain-specific question (rather than a purely linguistic inquiry) and requires domain expert knowledge to answer. For example, having a doctor, research nurse or medical coder indicate whether a document suggests that the patient described has a particular diagnosis is a good candidate for a light annotation task, and the annotation could still be light if the annotators were asked to tag the portions of the document that pertained to their diagnosis. However, asking the domain experts to provide deep linguistic annotations, such as semantic classes or part of speech tags, would take the task out of the realm of domain-specific light annotations and into that of more traditional linguistic annotation.

South et al. (2009) looked at identifying patients with IBD (Inflammatory Bowel Disease), and asked their annotators (two clinicians) to annotate the textual extents in a corpus of clinical notes that were related to the IBD concepts and to indicate the semantic class and attributes of those extents.

While South et al.'s full annotation task is not one that would be considered 'light' by the definition given here, aspects of the analysis were reflective of light annotation tasks and their applications. The researchers examined agreement at the document, concept, and attribute levels and found that agreement over which documents contained IBD-related information (93%), agreement for concept annotation was lower (.72 on average across all the concepts), and attribute agreement varied widely (.06 to .67). Unfortunately, an analysis of whether the disagreement at the concept level was over tag placement or semantic class was not provided. That is, if the annotators agreed that an instance of "ulcerative colitis" was annotated by both annotators, but

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

labeled a “sign or symptom” by one and a “diagnosis” by the other, that is a very different kind of disagreement than if the phrase was annotated by one but not annotated at all by the other. However, even without that information about the concept level annotation, it is clear that the higher-level the annotation (whether the document contains indications of IBD) the more accurate the annotation was compared to the levels of annotation requiring assignment of a semantic class or other, more detailed attributes.

There are a number of studies that use annotation models that can be considered “light”, but for the sake of brevity this dissertation will focus on a few representative studies. One example of a study designed to use a light annotation task is Yetisgen-Yildiz et al. (2011), which sought to identify patients who could be identified as having pneumonia. The researchers asked that an annotator “[...] with 6 years of experience as a research study nurse manually [classify] a patient as ‘positive’ if the patient had pneumonia within the first 48 hours of ICU admission and as ‘negative’ if the patient did not have pneumonia or the pneumonia was detected after the first 48 hours of ICU admission[...]”.

Simply using “positive” and “negative” as document-level labels made the task relatively easy for the annotator to perform, and his or her time could then be spent focusing on reading the document to determine the diagnosis, though the use of only one annotator is unusual for an annotation task. However, the researchers were able to use these single labels to create an NLP system that was able to replicate those diagnoses with 58.3% precision and 42.4% recall, which are impressive figures considering the complexity of the question asked about the patients, which required both analysis of physical state as well as temporal analysis of the document. This indicates that light annotation tasks can be useful starting places for NLP in the

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

bioclinical domain; in fact, the ARC<sup>5</sup> (Automated Retrieval Console) system for identifying patients with diseases is built around the idea that users will indicate areas of the text that are of interest, rather than perform deep annotations of the text (D’Avolio et al., 2010; D’Avolio et al., 2011).

Finally, the 2007 i2b2 (Informatics for Integrating Biology and the Bedside) NLP challenge task was that of identifying the smoking status of a patient from his or her medical discharge records (Uzuner et al., 2007). The participants in the challenge were provided with an annotated dataset, and attempted to recreate the labels given to the documents in the training and testing datasets. The annotation scheme used was again a single label classification task using the following labels:

- *past smoker*: someone who quit smoking a year or more ago;
- *current smoker*: someone who smoked within the past year;
- *smoker*: someone who was either a current or past smoker, but it could not be determined when or if they had quit;
- *non-smoker*: someone who never smoked;
- *unknown*: someone whose record contained no information about smoking status.

The definition of “smoker” used here may be unintuitive to someone who is not a clinician, but it is founded on medical practices: because the effects of smoking linger even after has someone quit for a short amount of time, a person isn’t generally grouped as a ‘past smoker’ until they have not smoked for a sufficient length of time.

---

<sup>5</sup><http://arc.4thparadigm.org/>

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

The annotation for the i2b2 smoking status dataset was performed by two pulmonologists, and was in fact performed twice: first, the annotators were asked to use their intuition when creating labels for each document, and second they were asked to label the documents based only on what was stated in the text. Unsurprisingly, the intuition-based annotations obtained significantly lower agreement scores (average Cohen’s kappa of .45) compared to the text-based annotations (average Cohen’s kappa of .84).

Given this dataset for training and a second for testing, challenge participants were able to build systems that performed well, and 12 system runs were able to obtain F-scores of over .84 (Uzuner et al., 2007). Despite the lack of semantic and syntactic information in the annotated corpus, some of the top-performing systems were able to augment the data with that information themselves (Clark et al., 2008; Cohen, 2008; Szarvas et al., 2006).

While there are certainly other ‘light’ bioclinical annotations that have been performed (for example, the BioNLP workshop tasks have also benefited from simplifying annotation tasks used for other purposes (Kim et al., 2009; Kim et al., 2011)), the ones discussed here are sufficiently representative of the state of light annotation tasks in the bioclinical domain for the purposes of discussion in this dissertation.

It is clear that light annotation tasks are relevant and useful for the biomedical domain in general and NLP research specifically, but until now there has been no methodology for creating annotation tasks for domain experts. The next section explores generalizations about the annotation tasks discussed in this and previous sections, and presents a set of principles that can be applied to the creation of light annotation tasks.

### 4.3 Principles of Light Annotation Tasks

Just as how, until recently, there was no established methodology for creating and performing annotation and machine learning tasks (c.f. Chapter 3), despite corpus annotations being a staple of corpus linguistics and computational linguistics research for decades, there are no established standards for designing light annotation tasks for the purpose of capturing domain-expert knowledge, despite the fact that such tasks have been used in biomedical and clinical domains for years.

Because the majority of existing related research has been done in the bioclinical domain, and the case study presented at the end of this dissertation uses clinical data, the discussion of light annotations tasks for this dissertation have been primarily in the bioclinical domain. However, it should be noted that light annotation tasks could be used for any type of annotation requiring the capturing of professional knowledge (legal texts, for example, would be excellent candidates for light annotation tasks). It should be noted that complex linguistic annotations are not considered part of the “professional” annotations being discussed in this dissertation. While significant study is often necessary to perform detailed linguistic-based annotations, the MATTER cycle is generally sufficient for capturing that type of professional knowledge. This dissertation seeks to find a reliable way to encode professional knowledge of fields where textual annotation is not already standard practice.

Based on the examples presented in the previous section, as well as the other approaches to bioclinical annotations discussed in Section 4.1, and the lessons learned from the case study presented in Part II, this dissertation presents a novel set of guidelines aimed at designing effective ways for leveraging expert knowledge in domain-specific annotation tasks; these guidelines are also compatible with the general desider-

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

ata for annotation tasks described in Chapter 2, as well as the methodology described in the MATTER cycle (Chapter 3).

By synthesizing all of the above standards, projects, and desiderata, maxims for creating good light annotation tasks can be established. The following principles are therefore proposed<sup>6</sup>:

- The annotations are performed by experts in the field;
- The task is divided into as few classification questions as possible;
- The classifications used in the model are based on current best theories and practices for the chosen domain;
- Annotation should be done based only on what is in the text, not on expert's intuitions about the text;
- If possible, the annotations should be applied to sentence- or phrase-level sections of the document, in support of document-level classifications;
- Additional layers of annotation can be provided before or after the light annotation is performed without conflicting with the given classifications.

The following section examines each of these guidelines in turn:

**Expert annotators:** Put simply, if the purpose of the annotation task is to obtain complex information about the data, the annotations should be done by people who are qualified to make those determinations. On the surface this is obvious, but it is

---

<sup>6</sup>A preliminary version of these principles and discussion were originally presented in Stubbs (2012), but have been expanded and revised in this dissertation.

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

a departure from more traditional linguistic annotations where linguists and domain experts have shown roughly equal ability to apply part-of-speech tags, tree structures, and coreference markers (Tateisi and Tsujii, 2004; Tateisi et al., 2005; Cohen et al., 2010). Resolving to use expert annotators allows the  $M_1$  annotation model to capture information about the text that only a domain expert would be able to determine, such as whether a patient has a particular clinical condition, or is at risk for one, without getting bogged down in deep syntactic or semantic annotations.

**Minimal classifications:** By breaking down the needed information into a small set of classification tasks (or even a single task, as is seen in the Smoking Status corpus), the annotation can be done much more quickly and accurately. This is particularly helpful for research groups who may not have a domain professional in-house, but instead need to hire domain expert annotators as consultants: a process that can be costly and time-consuming. This approach was used, to one extent or another, in all of the light annotation tasks discussed previously (South et al., 2009; Uzuner et al., 2007; D’Avolio et al., 2011; Yetisgen-Yildiz et al., 2011), as well as by Wilbur et al. (2006) in which five aspects of scientific papers that can be used generally in text mining were identified: focus, polarity, certainty, evidence, and directionality.

**Based on current theories, techniques and resources:** Beyond simply suggesting that annotations should not be intrinsically unscientific, the point of this principle is to say that the domain expert’s understanding of the text should take *precedence* over strictly linguistic analyses. For the Smoking Status corpus (Uzuner et al., 2007), for instance, a textual reading of ‘quit smoking 3 months ago’ by a layperson would indicate a status of ‘Past Smoker’, but that would be incorrect according to the med-

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

ical interpretation. The annotation must therefore reflect medical standards, and not be subordinated to easier or more obvious interpretations.

Additionally, any existing resources that are considered standards in the domain being examined should also be used as a starting point for any domain-specific light annotation task. Standard definitions, dictionaries, ontologies, and so on should be used where possible, in order to ensure that the light annotation is compatible with existing work in the field. For example, in the biomedical community, the resources contained in the UMLS<sup>7</sup> are repositories for standardized terminology.

**Evidence-based annotations:** It seems reasonable to suggest that, if supplied with an expert’s knowledge in a field, making use of the intuitions that go along with that knowledge would be a great boon to interpreting biomedical texts. However, both the Smoking Status challenge and the GENIA event annotations found that using expert intuitions resulted in greater discrepancies in inter-annotator agreement (Uzuner et al., 2007; Kim et al., 2008). Kim et al. relied instead on what they referred to as *Text-bound annotation*: annotations that required the annotators to “indicate clues in the text for every annotation they made”. This resulted in higher inter-annotator agreement and more useful annotations. There is a key difference between making use of expert *knowledge* and relying on expert *intuition*. Relying on intuition may result in annotators trying to read between the lines of a text, or past experience that tells them, ‘If a patient says this, it’s usually actually that’. Limiting annotations and classifications to what is said in the text will result in annotations that are both more agreed upon between annotators and more useful for machine learning or other NLP techniques.

---

<sup>7</sup><http://www.nlm.nih.gov/research/umls/>



**Sentence- or phrase-level annotations to support document classifications:**

While the simplest possible annotation task is to have domain experts assign a label to an entire document (for example “positive” if a patient meets the criterion being examined, “negative” if they do not), if the ultimate purpose of the dataset is for use in training and testing an NLP system, then it is advisable to have the annotators show what aspects of the text are leading them to their conclusions. This helps ensure that the annotations are evidence-based, and also provides suggestions for where an NLP system should start looking for features, and where the full  $M$  annotation model may be applied. This principle is directly supported by Xia and Yetisgen-Yildiz (2012), who also suggest that annotators provide additional information for their conclusions.

However, it is important that these supporting annotations not become too dense. One way to keep an annotation ‘light’ is to not require that the set of terms,  $T_1$ , be applied specifically to any particular type of syntactic structure or semantic class, but rather that it be a general marker that the annotator can use to indicate that a phrase or sentence contains a relevant piece of information. This allows the task to be performed much more quickly, but also ensures that the NLP researchers will have a solid foundation for their own work, when determining where and how to focus their own efforts.

**No conflict with additional annotations:** This guideline applies to the practical matter of the actual encoding of the annotation, and the current standards for corpus annotation in the computational linguistics community. The annotation task should not rely on tools or outputs that will not be compatible with other layers of annotation. The easiest way to ensure this is to use tools that are LAF-compliant (c.f.

Section 2.3, and to represent annotations in stand-off XML or a similar scheme that does not change the text being annotated. This will make it easier to add layers of other annotations later in the annotation process for use in machine learning, either by hand (other annotators) or automatically (with software). This is particularly important if the existing annotation will be later augmented, either by hand (by linguistics or other non-domain experts) or using automated systems, with other, more dense, annotations, as was done by many of the groups in the i2b2 smoking status challenge (Uzuner et al., 2007).

## 4.4 Methodology of Light Annotation Tasks

Just as the definition of light annotation tasks is based on the Model step of the MATTER cycle, the method by which a light annotation is applied is based on the Annotation step of MATTER. More specifically, the MAMA annotation cycle that is used to apply traditional annotation tasks to corpora is the same one that is used for creating annotated datasets with a light annotation model.

The primary difference in the annotation process is in intent rather than execution. While a traditional MAMA cycle in the MATTER cycle aims to encode all the linguistic annotation necessary to represent the Model, the light annotation Model and MAMA cycle aim to represent the domain expert knowledge. This will affect the choice of tags, attributes, software, and annotators, but it does not change the fundamental methodology of Model, Annotate, Evaluate, Revise. Chapter 7 uses the case study presented in Part II of this dissertation to demonstrate how a light annotation task is designed and performed for a task requiring professional knowledge, in this case of the clinical domain.

Once the light annotation is done, the light Model,  $M_1$  can be augmented with further annotations, also within the MATTER cycle. The specific implementation of the stages between the light Model,  $M_1$  and the complete, MATTER-ready model,  $M$  will vary based on the domain and task being addressed; some examples of this process could include: incorporating word-sense disambiguations based on existing domain-specific dictionaries; performing part-of-speech analysis; or adding separate layers of semantic annotations, such as incorporating temporal or spatial information as a separate layer of the text.

## 4.5 Light annotation tasks and identified desiderata

Chapter 2 identified the existing and emerging standards and desiderata in the annotation community as important considerations for the creation of light annotation tasks. In this section, those desiderata are revisited, and the relationship between those standards and the light annotation task principles presented here are discussed.

**Corpus creation and selection** (Chapter 2.1): The light annotation methodology does not conflict with the principles of representativeness and balance required for good results in corpus linguistics, nor does it pose a problem for the mantra “bigger is better” when it comes to training machine learning algorithms. In fact, the use of light annotation tasks makes it more likely that a larger domain-specific dataset can be annotated, since any domain expert annotators that are hired as consultants will be able to complete more annotations in the same (or even a shorter) period of time.

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

**General annotation desiderata** (Chapter 2.2): Leech’s seven maxims for annotation schemes have been so thoroughly integrated into the de facto standards of the annotation community that it is almost nonsensical to address them individually here. Suffice it to say that the light annotation principles do not directly contradict any of the maxims, and for the most part actively support them.

**Annotation representation** (Chapter 2.3): As one of the principles of the light annotation task is that it should not conflict with other annotations, and specifically mentions the LAF standard for annotation encoding, it is clear that light annotations are, if the principles are followed, fully compliant with current standards for representing annotated data.

**Annotation guidelines and reporting** (Chapter 2.4: The established and emerging de facto standards for the creation of annotation guidelines and what should be reported about the annotation process are in some ways entirely separate from the light annotation process. While it is absolutely important for explaining a light annotation task to others that information about the domain experts, their training, and the model used for the light annotation task be made available, those are all considerations for after the annotation process is over. In terms of the annotation guidelines that are provided to the annotators, they should be as clear as possible for the domain experts to refer to quickly when they are working on creating the annotated corpus. Again, however, this suggestion is not encoded in the principles or description of the light annotation task; the execution is left up to the researchers.

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

**Annotation tools** (Chapter 2.5): The environment in which an annotated corpus is created can have a large impact on the quality of the created corpus. While the principles of light annotation tasks do not specify a particular type of annotation software be used, it does recognize the barriers that unsuitable tools can create, particularly for domain experts who are unlikely to be familiar with the annotation process. Therefore, this dissertation presents a set of annotation and adjudication tools designed for light annotation tasks in Chapter 5.

**Annotation process** (Chapter 2.6, Chapter 3): The relationship between the light annotation tasks proposed in this chapter and the MATTER development cycle for traditional annotation tasks should be quite clear based on the description of light annotation tasks provided in Chapter 4.2. The light annotation process is intended to utilize an underspecified annotation model,  $M_1$ , which can later be transformed into the fully specified model,  $M$ . In terms of the MATTER cycle, it would be entirely possible to perform the full process with only a light annotation model, but it is likely that the results from training and testing machine learning algorithms would not be as good as if  $M_1$  had been augmented with other annotations designed to leverage the professional knowledge captured by the light annotation task.

Overall, the light annotation model and principles provided here do not conflict with the established standards and desiderata described in Chapter 2. Just as the MATTER cycle is agnostic to most of the established and emerging standards, so too is the model and principles of light annotation tasks. This is not a weakness in the theory behind these tasks, but rather the fact that the light annotation principles do not conflict with or explicitly support the stated desiderata means that as stan-

standards change, light annotation tasks can change with them. Even the standard of representational compatibility, which is described here as being LAF-specific, can be applied to any representational standards that may emerge in the future. Therefore, the light annotation methodology does not conflict with existing standards, and is open to changes in standards that may occur in the future, while still providing a platform from which domain expert knowledge can be collected.

## 4.6 Overview of light annotation tasks

The purpose of the light annotation task is not to create a complete representation of all the relevant data in a domain-specific text. It can, however, create a highly accurate layer of annotation that will be used in conjunction with other linguistic information, as was the case with the Smoking Status challenge. In terms of the MATTER cycle, the light annotation is not the full representation of the Model ( $M = \langle T, R, I \rangle$ ). Rather, the light annotation Model,  $M_1$ , is a top-level set of annotation that is used to indicate portions of the document relevant to the classification, or to apply a label to a document as a whole. It does not represent the entire set of features necessary to create an algorithm (during the Training and Testing phases of MATTER) that is able to generate the desired classifications.

The methodology for the creation of a light annotation task does, however, fit neatly into the MAMA cycle of the MATTER process, as it will undergo the same formative and refining steps that any annotation task, regardless of how dense or light it is, must go through to be vetted. Indeed, because of the domain expertise required for many bioclinical tasks, it is imperative that at least one iteration of the MAMA cycle be observed, so that both the NLP researchers and the domain experts can be

## CHAPTER 4. REPRESENTING EXPERT KNOWLEDGE

satisfied with the way that the information is being collected and represented.

Because the light annotation guidelines only make suggestions for the Model of the annotation task, there will be no conflicts should an NLP/biomedical researcher wish to take advantage of some of the other methods used in biomedical annotations, such as the Single-facet Annotation used in the GENIA event corpus (Kim et al., 2008), domain-expert crowdsourcing as described by Scott et al. (2012), or an active learning-based system such as the accelerated annotation program used by Tsuruoka et al. (2008).

Naturally, if domain-expert annotators are asked to create a light annotation, it should be the case that they are provided with an annotation tool that allows them to label the documents without confusing the task by providing too many options or requiring a long time to learn to use. A set of tools for annotating and adjudicating light annotation tasks is presented in the next chapter.

# Chapter 5

## Tools for Light Annotations

In addition to the guidelines for light annotation tasks presented in the previous chapter, this dissertation also provides software for annotating and adjudicating light annotation tasks<sup>1</sup>.

As described in Section 2.5, the software is an important part of any annotation task, and much thought has been put into analyzing what makes a good annotation tool.

While no outside study has been done on the specific problem of annotation tools for domain expert annotators or light annotation tasks, Dipper et al. (2004b) examined what attributes an annotation tool should have for it to be most generally useful, and created the following list of practical requirements:

- diversity of data;
- multi-level annotation;
- diversity of annotation;

---

<sup>1</sup>The basic information in this chapter was first presented in Stubbs (2011), but has been revised and expanded for this dissertation.



## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

- simplicity;
- customizability;
- quality assurance;
- convertibility.

These are all excellent general goals for an annotation task, but as Dipper et al. discovered in their study, there are often trade-offs between the different criteria. For example, tools that were more “ready-to-use” (that is, they did not require work to be done before hand, such as pre-defining annotation specifications) were more user-friendly, but those applications performed less well in the quality assurance aspect of the evaluation than tools that took in pre-defined tagsets.

Ultimately, Dipper et al. determine that “...it is clear that the annotation scenario determines which tools are suitable and which are not” (2004b)). While they do not explore the scenario of performing light or domain-expert annotation tasks, it is clear that an appropriate annotation tool may be just as important as an appropriate annotation model when it comes to best using a domain expert’s time and expertise.

While the specifications of annotation tools for domain-experts and light annotation tasks have not yet been fully studied outside of this dissertation, there are a few common-sense desiderata for an expert annotation task that can be easily identified.

A tool for domain expert annotation should:

- be easy to install;
- provide a simple way to modify the annotation specification to accommodate changes to the task;

## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

- be easy to learn to use;
- clearly display the created annotations and their attributes;
- have all the necessary capabilities for the annotation task.

The last two items on the list will, of course, vary by the annotation task being performed. The specifics of the annotation task undertaken for this dissertation are described in Chapter 7, but can be generalized into the following items:

- mark text extents with user-defined tags and attributes;
- annotate partial as well as whole words;
- allow the creation of tags that can be applied to the entire document (non-consuming extent tags);
- create links between extent tags (including non-consuming extent tags);
- generate LAF-compliant output.

Here, ‘non-consuming extent tag’ means a tag that acts as though it is anchored to an extent in the text, but can be used to indicate the entire document. While it would generally be possible in most annotation tools to simply annotate all of the text of a document using a single tag, this usually results in all of the text being displayed as belonging to that one tag type (usually by changing background or text color), which can be very distracting for the annotator. A non-consuming extent tag has the interpretation of being applied to the whole document, but without the visual distractions (this is similar to a metadata tag, but a “non-consuming extent tag” can be linked to other tags in the annotation should the need arise).

## 5.1 Existing annotation tools

While the lists of general and task-specific desiderata listed in the previous section seem fairly straightforward, at the time the annotation work for this dissertation began (summer 2010), it was surprisingly difficult to find an annotation tool that met all of them. While it is true that a recent (summer 2012) search of the LRE Map returns over 200 results for “annotation tools”, the Map was not available at the time this research began. Therefore, the search for tools was generally limited to those that were well-known and readily available. The tools that were considered and tested for this dissertations light annotation task are described below.

**GATE** - The General Architecture for Text Engineering<sup>2</sup> is an open source platform for text annotation and processing developed at Sheffield University (Cunningham et al., 2010). GATE is widely used for annotation tasks, and works on all major operating systems. In addition to providing annotation support, GATE provides a full set of plug-ins for automatic text processing, such as part-of-speech tagging, tokenization, sentence splitting, etc., as well as other features such as annotation merging and inter-annotator agreement scoring.

However, due to the large number of plug-ins and other features, the learning curve for GATE is quite steep, and while creating new extent tags was relatively easy, developing original link tags was much more difficult, as GATE seemed geared towards creating links automatically, then adjudicating them by hand. This setup was not feasible given the novelty of the light annotation scheme being used, as there were no existing systems suitable for the task, and one could not be built and tested

---

<sup>2</sup><http://gate.ac.uk/>

## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

until the annotation task was complete.

Additionally, the version of GATE that was available in 2010 did not create stand-off annotation, but rather inserted nodes into the text where tags were meant to start and end, then had the tag information at the end of the document, as shown below:

```
<Node id="1092"/>Monitor<Node id="1099"/> in NICU til
<Node id="1112"/>CXBC<Node id="1116"/> returns.

<Annotation Id="91" Type="Intervention"
  StartNode="1092" EndNode="1099"/>
<Annotation Id="92" Type="Investigation"
  StartNode="1112" EndNode="1116"/>
```

The nodes that GATE added to the text made the documents extremely difficult to analyze outside of the GATE architecture, and the difficulty in creating links made the platform unusable for the domain-expert annotation task.

**Callisto** - Callisto<sup>3</sup> was developed at the MITRE corporation and comes with many annotation specifications already installed, and it is possible for users to create their own extent-based specifications. However, while some pre-loaded tasks in Callisto do contain their own sets of link tags, at the time this program was being tested users who wished to define their own annotation specifications using link tags were required to not simply create their own schema, but write their own plug-in for the program in Java. While the resulting system would have been relatively easy for the annotator to use, having to revise the plug-in each time a change was made to the specification would have slowed down the MAMA cycle for the light annotation task.

---

<sup>3</sup><http://callisto.mitre.org/>

## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

**The Brandeis Annotation Tool (BAT)**- BAT has been used for a variety of annotation tasks, including creating the gold standard for SemEval tasks and evaluations (SemEval 2010), and provides a task-specific web interface for authorized users (Verhagen, 2010). While BAT provides excellent support for adjudicating the work of multiple annotators, it was designed for layered annotation, where each piece of an annotation task is done individually from the others by all annotators (similar to the idea of Single-facet Annotation described in Chapter 4.1.2), then adjudication is performed on each layer before moving on to the next.

This system is extremely useful for tested annotation tasks with multiple annotators and judges. However, for any annotation specification that is still being revised in the MAMA cycle, a layered annotation format makes it much more difficult to find problems in the specification, as some will take a long time to appear. Additionally, at the time BAT could only annotate complete whitespace-separated tokens, which is potentially problematic for certain bioclinical annotations. For example, if a patient is recorded as “HIV-” it maybe be necessary to annotate “HIV” separately from “-”, a functionality that did not exist in BAT at the time.

Other tools were also examined as options for creating a light annotation task, such as Knowtator (Ogren, 2006), Protégé<sup>4</sup>, and SLAT 2.0 (2010). However, many of the same problems described above continued to reappear: steep learning curves, difficult to create and modify annotation schemas, or simply unable to create all the necessary aspects of the task being designed. This is not to suggest that these tools are poorly designed or otherwise should not be used for annotating corpora—indeed, their popularity proves that they are all well-suited for corpus annotations. However,

---

<sup>4</sup><http://protege.stanford.edu/>

the combination of requirements for domain-expert annotation tasks in general and the specific light annotation task undertaken for this dissertation required that a new annotation tool (and an accompanying adjudication tool) be created.

## 5.2 MAE - Multi-purpose Annotation Environment

In order to provide a suitable environment for a light annotation task for domain experts, the Multi-purpose Annotation Environment (MAE) was built. MAE is a lightweight annotation tool written in Java with an SQLite backend database<sup>5</sup>. It has been tested on a variety of operating systems, including Windows Vista and XP, Mac OS X, Ubuntu, Linux Mint 12, and Red Hat. It should be compatible with any operating system that can run Java 6 or higher.

In terms of the desiderata outlined above for a domain-expert annotation tool, MAE meets all of the specifications:

**Easy to install:** MAE is written in Java and distributed as a .jar file, so on most platforms running MAE simply requires double-clicking on the application. Aside from ensuring that Java is installed and up-to-date (which most operating systems do automatically), no setup or installation is required.

**Simple task creation and modification:** In order to define an annotation tagset, users are asked to create what is essentially a Document Type Definition (DTD) file, with a few modifications to account for some of MAE's features. As was previously discussed, Dipper et al. (2004b) noticed in their survey of annotation tools that

---

<sup>5</sup>SQLiteJDBC driver created by David Crawshaw <http://www.zentus.com/sqlitejdbc/>

## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

annotation software often has to make a trade-off between reliable output and ease of beginning to annotate. MAE attempts to bridge this gap by requiring that the user have some idea of what they want to use as tags, but by using a DTD rather than a more complicated XML schema, the barrier to starting an annotation project is still rather low. The DTD for the annotation task described later in this dissertation can be found in Appendix B, and a sample DTD for a different task is shown here:

```
<!ENTITY name "NounVerbTask">

<!ELEMENT NOUN ( #PCDATA ) >
<!ATTLIST NOUN start #IMPLIED >
<!ATTLIST NOUN type ( person | place | thing | other ) >
<!ATTLIST NOUN comment CDATA >

<!ELEMENT VERB ( #PCDATA ) >
<!ATTLIST VERB tense ( past | present | future | none ) >
<!ATTLIST VERB aspect ( simple | progressive |
    perfect | perfect progressive ) >

<!ELEMENT ACTION EMPTY >
<!ATTLIST ACTION relationship ( performs | performed_by ) >
```

The ENTITY tag provides the name of the annotation, and each !ELEMENT tag defines either an extent tag (tags specifying “#PCDATA” or link tags (tags specifying “EMPTY”. “!ATTLIST” entries associated with each tag define an attribute for that tag, which can either be character data or a list of options. To create a non-consuming tag, the attribute “start” is declared and set to “#IMPLIED”, which allows the tag to be created but not associated with any of the text in the file.

Using these DTD features, it is quite easy to create new tags and attributes for any light annotation task, and DTD files can be quickly modified should a task specification be changed.

## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

**Easy to use:** Very little effort is needed to begin annotating in MAE. The annotator loads the DTD into MAE in order to create the database structures needed to store the information about the tagset being used, and then loads the next text file that will be annotated. No preprocessing needs to be done to the files before they are loaded (though they do need to be UTF-8 encoded, especially if the language being annotated is not the operating system default). Figure 5.1 shows MAE with the annotation scheme shown above, with a sample file loaded.

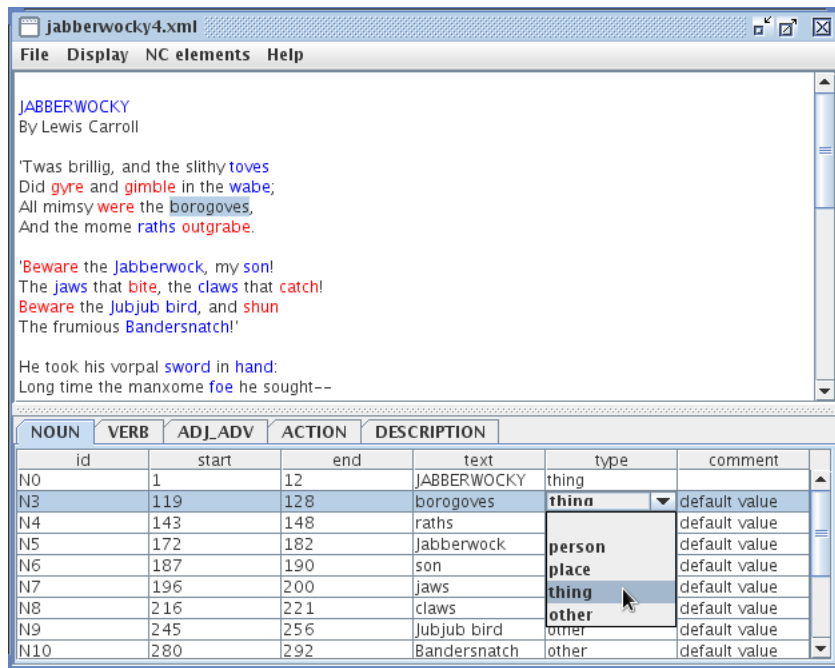


Figure 5.1: MAE: Multi-purpose Annotation Environment

The text of the file being annotated is shown in the upper portion of the window, and each tag defined in the DTD gets its own table/tab in the bottom of the window. Each tag is also given its own color, so it is easy to tell by looking at the text how each word has been annotated.



## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

New extent tags are easily created by highlighting the text and right-clicking, then selecting the appropriate tag from the menu that appears. That tag is then added to the related table, where the tag’s attributes can be filled in. Document-unique IDs for the tags are automatically generated by MAE for each tag created, for easy reference in link tags.

Links are easily created by holding the control key (command key on Macs) and left-clicking on each of the entities that will be included in the link. It is possible for users to link to non-consuming tags as well.

Tags can be deleted both from the text window by highlighting all or part of the tag and right-clicking, as well as from the tables by highlighting the row describing the tag being removed. If an extent tag is deleted, any link tags that use the deleted extent tag as an anchor will also be removed, in order to maintain the consistency of the annotation.

**Clear display:** Because MAE is designed for light annotation tasks, the display is equally light and easy to interpret. The text is highlighted where extent tags have been placed, and the tables display all of a tag’s attributes for easy modification. Additionally, MAE contains some features that make finding the correct tag in the table easier: selecting an annotated extent in the text window will highlight all the rows in the tables where there is a tag involving that extent, including link tags. Newer versions of MAE include functionality that will auto-select the associated text in the upper window when a tag’s ID is double-clicked.

**All necessary capabilities for task:** In addition to creating extent tags, link tags, and non-consuming tags (these are created from the “NC Elements” menu at the

## CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

top of the screen), MAE has other functionalities that are useful for a variety of annotation tasks. It supports overlapping tags, links between annotated text as well as non-consuming tags and annotation of partial words.

MAE outputs character-based stand-off XML, which allows the user-generated annotation to be easily merged with other annotations, either those created by hand or those generated by NLP systems. A sample of the output format is shown here, based on the DTD provided above:

```
<?xml version="1.0" encoding="UTF-8" ?>
<NounVerbTask>
<TEXT><![CDATA[
JABBERWOCKY
By Lewis Carroll

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.
]]></TEXT>
<TAGS>
<NOUN id="N0" start="-1" end="-1" text="" type="person"
  comment="author = Lewis Carroll" />
<NOUN id="N1" start="61" end="66" text="toves" type="thing"
  comment="" />
<VERB id="V1" start="80" end="86" text="gimble"
  tense="past" aspect="" />
<ACTION id="A1" fromID="V1" fromText="gimble" toID="N1"
  toText="toves" relationship="performed_by" />
</TAGS>
</NounVerbTask>
```

XML is always a bit hard to simply read, but the selection above shows a non-consuming tag that is being used to indicate the author of the piece, a noun tag for “toves” and a verb tag for “gimble”. The noun and verb tags are then connected by

an action link tag using the ID numbers of the extent tags that were automatically generated by MAE.

### 5.3 MAI - Multi-document Adjudication Interface

The task of adjudication—determining which annotator tags are correct, and creating a Gold Standard corpus from them, as well as adding any tags that might have been left out—can be a difficult and time-consuming process. While it is possible to do through the use of Python scripts, an adjudication tool that was designed to work with MAE’s output and could display the annotations and discrepancies to the adjudicator seemed a preferable way to solve the problem.

Therefore, the Multi-document Adjudication Interface (MAI) was written as a complement to MAE in order to ease the process of Gold Standard creation. MAI is built on the same basic code base as MAE, and provides a simple interface that allows users to load the annotations from different annotations over the same document, then compare the annotations one tag at a time and determine which tags should be included in the gold standard. Figure 5.2 shows a version of the MAI interface using the same annotation scheme shown previously.

When a tag is selected from the menu on the left, each text extent that was annotated with that tag by the annotators is highlighted in the text window: text in blue was annotated by all of the annotators, and text in red was annotated by some but not all of the annotators. When an extent is selected, the tags and attributes from each annotator at the selected location is displayed in the table at the bottom of the screen, and the adjudicator can select which one to copy to the gold standard. Once a tag is in the gold standard the text at that location in the top window turns

CHAPTER 5. TOOLS FOR LIGHT ANNOTATIONS

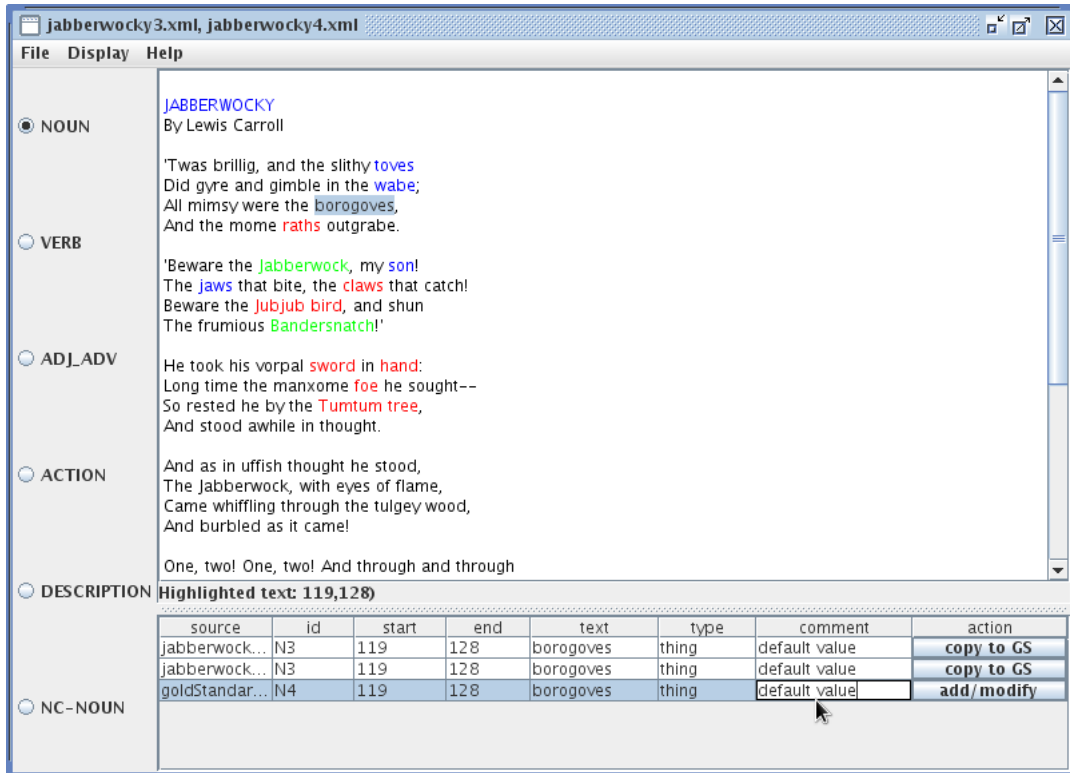


Figure 5.2: MAI - Multi-document Adjudication Interface

green. The attributes of gold standard can be modified, so it doesn't matter if all the annotators were incorrect. The adjudicator can also add tags that were left out of all of the annotations.

## 5.4 Use and availability of MAE and MAI

MAE and MAI, while initially built for the light annotation task described in this dissertation, have since been distributed and used for a variety of annotation tasks, including Chinese verb and aspect annotation, Russian morphemes, and spatial annotation tasks, as well as the 2012 i2b2 NLP challenge corpus.

MAE and MAI are both open source under the GNU GPL v3 license, and are available for download through Google Code:

MAE: <http://code.google.com/p/mae-annotation/>

MAI: <http://code.google.com/p/mai-adjudication/>

## Part II

**The PERMIT Corpus: a case  
study in using light annotation**

## Chapter 6

# Settings for a Domain Expert

## Clinical Task

In order to explore the potential uses of light annotation tasks, Part II of this dissertation presents the PERMIT (Patient Evaluation Resource for Medical Information in Text) corpus, an annotated dataset of hospital discharge summaries built around a light annotation task for the clinical domain. The PERMIT corpus implements a prototype of the light annotation methodology, as the lessons learned from creating the dataset influenced the principles for light annotation tasks provided in Section 4.3. The annotations were created using the software described in Chapter 5, which was built initially for the PERMIT corpus.

This chapter describes the medical settings that informed the creation of the annotation task: patient selection for clinical trials. Chapter 7 describes the MAMA cycle of the PERMIT corpus and analyzes the light annotation methodology and its applications, and Chapter 8 explores some potential ways for the corpus to be used in NLP systems in the clinical domain.

## 6.1 Goal of the case study

As Section 1.4.4 discussed, a variety of annotation tasks have already been undertaken in the clinical domain, from part-of-speech tagging to medicine recognition to medical event identification. One particularly interesting challenge in medical research is identifying patients who are eligible to participate in clinical trials of medications and other treatments for conditions and diseases.

Under NIH grant number 5R21LM009633-02 (PI: James Pustejovsky), researchers at Brandeis University and the Channing Laboratory at Brigham and Women’s Hospital (BWH) collaborated to explore the ways that the patient selection process could be aided by NLP techniques, particularly those involving temporal analysis of the data. Specifically, the researchers undertook a project to mimic the patient selection process for a mock retrospective case-control study, and to create an annotation scheme to capture the information necessary to identify qualified patients, and then use the resulting annotated corpus to study how an NLP system could make use of the annotated information.

## 6.2 Methods

The researchers at BWH provided insight into the patient selection process, suggested eligibility criteria for the mock-study, arranged for medical researchers to perform the annotation, and provided feedback on the annotation schemes. The annotation schema and process was headed by the author of this dissertation, with support from the grant’s PI, James Pustejovsky.

Due to difficulties in obtaining permission to share medical data across institu-



## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

tions, it was determined that the annotation would be done using discharge summaries from the MIMIC II Clinical Database (physionet.org, 2010). Eligibility criteria for the mock-study would be based on the types of criteria used in actual clinical research, and the annotators would be professionals qualified to evaluate medical records. More information about the specifics of the annotation will be provided in Chapter 7; the remainder of this chapter will discuss the settings in which retrospective case-controls studies traditionally take place.

### 6.3 Research settings

In order to create an annotation scheme that is appropriate to the project being undertaken, it is important to have an thorough understanding of the settings in which the project is being done. In the case of the annotation project described here, it is necessary to understand what a retrospective case-control study is, why they are performed, and how they fit into the panoply of clinical research. In the rest of this chapter, settings for current practices in large-scale medical studies (epidemiology) will be discussed (Section 6.3.1, in addition to eligibility criteria for medical studies (Section 6.3.2), and the format of the type of data being examined in this annotation project (Section 6.3.3).

#### 6.3.1 Current epidemiology practices

In order to fully set the stage for the annotation and NLP task described in the next chapters, the following descriptions of epidemiological studies are provided; they are paraphrased from Chapter 6 of *Modern Epidemiology* by Kenneth Rothman, Sander Greenland, and Timothy Lash (2008).

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

**Experimental and nonexperimental studies** – Experimental studies are, put simply, studies involving test conditions and groups of people deliberately given treatments (or placebos) to test for outcomes. In contrast, nonexperimental studies rely on patients being exposed to different situations by their own actions or circumstances. There are a variety of reasons why nonexperimental studies are run in lieu of experimental ones—it is, naturally, unethical to expose study participants to cancer-causing agents on purpose, but observing the effects of such agents on patients who were exposed to them through other means is not only ethical, but vital to expanding medical knowledge.

**Cohort and case-control studies** – Both of these can be types of nonexperimental studies. Cohort studies have “two or more groups of people that are free of disease and that differ according to the extent of their exposure to a potential cause of the disease” (pg. 94). The researcher then observes how frequently each cohort is affected by disease (the authors give an example of people working with chemicals, with each cohort being exposed to a different chemical).

Case-control studies, on the other hand, are conducted by choosing a population (such as chemical workers) and a particular disease to be studied—cases are the members of the population who are afflicted with the disease, while controls are unaffiliated. This allows the researcher to compare the cases and controls by matching them demographically (or by other related factors) in order to determine how each factor might affect whether a person will contract the disease.

**Prospective and Retrospective Studies** – Prospective and retrospective studies

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

are also both forms of nonexperimental studies. The authors note that previously the literature would use the term “cohort” interchangeably with “prospective” and “case-control” interchangeably with “retrospective” but that this conflation of terms is false and misleading: a cohort study can be prospective or retrospective, and the same is true for case-control studies. The difference between prospective and retrospective studies (as the terms are used now) is based primarily on the relational timing of events: studies where patients are not recruited until after they show signs of a disease are retrospective, while studies where patients are recruited on the basis that they may become afflicted are prospective<sup>1</sup>.

Given these distinctions between different types of clinical studies, the rationale behind using a nonexperimental retrospective case-control paradigm for the case study presented in this dissertation can be explained:

**Nonexperimental:** Because the purpose of the grant was to explore the use of annotation and NLP techniques in clinical research, rather than actually perform a medical study, it was not necessary or feasible to adopt an experimental study design. However, nonexperimental studies are commonly used for medical research, and as they rely on examining existing medical records, the application of annotation and NLP techniques was much more immediate, and still very relevant to the medical research field.

**Retrospective:** From speaking with the consultants at BWH, it became clear that

---

<sup>1</sup>This description is somewhat simplified, and studies may not be purely prospective or retrospective, but can be a mixture of both. However, this definition is sufficient for the purposes of this discussion and dissertation.

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

retrospective studies rely more heavily on patient records than do prospective studies, where ongoing patient interviews for data collection are much more common. Again, as the purpose of this research was to examine the use of NLP for practical contributions to medical research (but not actually perform medical research or interviews, as that was outside of the permissions of the grant), a retrospective paradigm was adopted.

**Case-control:** The case-control paradigm for medical studies requires an additional level of analysis that is quite intriguing—not only do the cases have to meet particular eligibility criteria to be included in a study, they must be matched with people from the general population that have similar demographic characteristics so that more general trends about the condition being studied can be analyzed. This additional requirement of matching criteria (an auxiliary set of selection criteria that ensure the control group is matched to the case group to help remove other possible influences on the outcome of the study) is one that seemed to be a good blend of NLP and statistics (a specialty of the BWH researchers included in the grant) and so it was made a part of the annotation/NLP project as a whole.

From a medical research perspective, there are also distinct advantages to the retrospective case-control study paradigm. Retrospective studies are often used for rare diseases because they allow researchers to examine data collected at different points in time (Coggon et al., 1997). Additionally, retrospective studies are much less expensive to run than traditional experimental cohort studies because they do not require the presence of the people being studied (Clark and Doughty, 2008).

Naturally, the retrospective paradigm is not without drawbacks. Retrospective

case-control studies are traditionally more susceptible to observational and sampling bias (Mann, 2003), as well as selection bias (Geneletti et al., 2009), particularly when it comes to matching controls to cases on information that is not commonly found in medical records. For example, income, location, education, etc. can all have effects on patients, and can therefore affect study outcomes in ways that may not be obvious. However, even with these problems retrospective case-control studies are widely used in research, and it seems possible that use of NLP techniques may in fact alleviate some of these problems, as computers can make analyzing medical records significantly faster, and the more records that are examined, the more likely it is that people who fully meet the eligibility and matching requirements can be found.

### 6.3.2 Eligibility criteria

The requirements used to determine who is eligible to be included in a medical study are referred to as ‘eligibility criteria’<sup>2</sup> In order to fully understand how to use these criteria in a mock case-control retrospective study, it is important to fully understand their role in the study and the types of criteria that are frequently used.

The website `clinicaltrials.gov`, which contains information about clinical trials that are NIH-funded, as well as privately funded studies, defines these criteria as “The medical or social standards determining whether a person may or may not be allowed to enter a clinical trial. These criteria are based on such factors as age, gender, the type and stage of a disease, previous treatment history, and other medical conditions.” Eligibility criteria are almost always written in natural language, as they are intended to be read and understood by human researchers searching for qualified

---

<sup>2</sup>They are also sometimes referred to as “inclusion/exclusion criteria”, “selection criteria”, or “enrollment criteria”: this dissertation uses all of these terms interchangeably.

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

study participants. They are also divided into two types: *inclusion criteria*, which are criteria that must be met in order for a patient to qualify, and *exclusion criteria*, which are criteria that disqualify a person from participating in the study if they are met.

Essentially, eligibility criteria help to determine the cohorts or the case/control groups for a study by selecting the groups of people who are appropriate for the treatments being tested or diseases being examined. For example, an ongoing study examining the effect of pre-natal Vitamin D on asthma rates in infants (the VDAART study<sup>3</sup>) requires (inclusion criteria) that expectant mothers entering the study be between 18 and 39 years old, be between 10 and 18 weeks pregnant when they enter the study, and that they have a “personal history of asthma, eczema, allergic rhinitis” (Weiss, 2009).

There are also some restrictions (exclusion criteria) on who can participate: women who have chronic medical conditions, who already take more than 2,000iu of Vitamin D a day, who have a “multiple gestation pregnancy” (i.e., are pregnant with twins, triplets, etc), or who used IVF or other reproduction techniques to become pregnant are all unable to participate in the study (ibid.). These restrictions help to remove confounding variables that could muddy the clarity of the study outcomes by keeping the study population heterogeneous enough to make any positive (or negative) results clearly correlate with the two factors being studied (in this case, Vitamin D and asthma).

The VDAART eligibility criteria are fairly straightforward—the investigators want to know whether Vitamin D can reduce the prevalence of asthma in the children of parents with asthma (or related diseases), and so they are looking at people who

---

<sup>3</sup><http://www.vdaart.com/>

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

meet those conditions, while removing people from the pool whose circumstances may interfere with the test (i.e., smokers, mothers who had difficulty getting pregnant, and those with other chronic illnesses that may affect the child).

While the VDAART study is a prospective study that was actively recruiting participants (a setup that was not going to be used for the NLP research described here), the eligibility criteria are a good example of the types of information that can be used to determine who can be included in any type of clinical trial, including retrospective studies.

The types of information needed to properly balance a medical study can vary widely, and not all questions asked of potential patients are necessarily ones that can be answered by examining existing medical records such as discharge summaries—in another study, this one about home pregnancy tests (Nettleman, 2006), one of the criteria was that the women entering the study do not wish to become pregnant, which is not a piece of information that is likely to appear in most clinical records. However, that study also has age restrictions for participation, and age is certainly information that can be found in medical records, either directly (“34 y/o”) or indirectly (using the date of birth). Therefore, even in studies where the majority of criteria deal with information not commonly found in medical records, a system that uses NLP techniques for examining medical records could still be useful for narrowing the field of candidates based on existing records.

Many eligibility criteria are complex, in terms of semantics, syntax, and/or computation. The challenges inherent in parsing and evaluating the criteria themselves are many, but they are not the focus of the work described in this dissertation. Significant work has been done in this area by researchers at Columbia University, including a literature review of eligibility criteria representations (Weng et al., 2010),

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

using e-screening to help identify patients for studies (Li et al., 2008; Botsis et al., 2010), and the development of ELiXR, a UMLS-based system for representing eligibility criteria in studies (Weng et al., 2011). This work is extremely valuable, and its existence allows other researchers to focus on other aspects of clinical research, such as the analysis of discharge summaries and the use of annotation schemes and NLP techniques presented here.

However, simply because this research does not focus specifically on the difficulties of representing complex eligibility criteria does not mean that it will ignore them entirely. One focus of the research was to examine temporal relations in clinical discharge summaries, and so temporal constraints in eligibility criteria must also be examined.

### **Temporal modifiers in Selection Criteria**

Temporal modifiers are often important parts of study selection criteria. For example, they can be used to ensure that all patients included in the study were in similar stages of their diseases or healing process (e.g., cardiac event within the past 2 years) at the time they were recruited or their records added to the analysis. Also, because medical care standards change frequently, placing limits on when events must have occurred (e.g., between 2005 and 2010), researchers can avoid analysis errors that would be caused by conflating the results of completely different treatments.

Temporal constraints in eligibility criteria are not limited to particular types of clinical studies, and are used in many different fields of medicine. A recent study of the selection criteria listed on `clinicaltrials.gov` showed that 11.07% of the corpus examined contained temporal concepts—the largest group of concepts found in the selection criteria (Luo et al., 2010).



## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

A PubMed search for case-control studies (performed in 2010) revealed a plethora of eligibility criteria using temporal constraints. Some examples of these criteria are:

- “650 patients who underwent elective or urgent CABG with cardio pulmonary bypass ... between 1 January 2001 and 31 December 2004” (Badreldin et al., 2010)
- “... patients with a radiological confirmed diagnosis of a first hip fracture within the past three years.” (Jha et al., 2010)
- “... a history of low back pain and/or low back related leg pain over the previous 6 months ...” (Moloney et al., 2010)
- “... at least 6 months of isoniazid or 4 months of rifampin...” (Xu and Schwartzman, 2010)
- “... patients who had a minimum number of 2 visits for diabetes and who had a diagnosis of diabetes mellitus for more than 6 months ...” (Hueston, 2010)

Not only were these studies conducted by researchers in very different medical research areas, but this list also describes criteria regarding medications, diseases, conditions, and actions taken by the patient, all with time-related modifiers. The existence of these temporal modifiers elevates patient selection from a simple dictionary-based search to one requiring more sophisticated information extraction techniques, including more sophisticated temporal reasoning.

An analysis of the eligibility criteria on <http://clinicaltrials.gov> as of October 5, 2010 using regular expressions to locate and count time-related temporal

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

expressions in the studies listed at that time; a total of over 96,000. Table A.1 contains a breakdown of the different temporal expressions found in the eligibility criteria, and can be found in Appendix A.

While this table doesn't represent all temporal expressions found in clinical trial criteria (it includes no dates, and the analyzed text does not include the age restrictions on the studies), it is clear that time is an important factor in selection for medical studies. On average, each clinical trial has 2.27 time-related expressions in the selection criteria. Since the list of analyzed expressions is small, we can assume that the actual number of time sensitive queries in trial selection criteria is actually larger.

### 6.3.3 Structure and Language of Discharge Summaries

In addition to analyzing the general setup of medical studies, it is important to be aware of attributes of the type of data that will be processed; in this case, medical discharge summaries.

It is common for medical records, particularly discharge summaries, to be divided into sections containing different types of information about the patient. The headers show whether the information in that section is related to the patient's medical status ("Allergies", "Past Medical History"), about the patient's family ("Family Medical History"), about actions taken while the patient was at the hospital ("Course of Treatment"), or recommendations for the future ("Discharge Instructions"). In order to take advantage of the information these headers provide, Denny et al. (2008) developed SecTag, a system for identifying and classifying header information in clinical notes. While the code for this system is not available, they do provide a database of

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

section headers that have been categorized into super-types, a resource that will be included in the NLP work described in Chapter 8

In an analysis of six different genres of medical records including discharge summaries, Mowery et al (2009) determined that “. . . the following sections alone can be used to predict a condition as historical: *Past Medical History*, *Allergies* and *Social History*. Clearly, section headers provide important context for anchoring medical events in time.

Within each section, the text could consist entirely of narrative, lists (usually of medications), or a mixture of both. Discharge summaries also contain multiple forms of temporal expressions, in relatively high frequency: Mowery et al. also found that Discharge Summaries contain more temporal expressions than any other genre of medical record (Mowery et al., 2009).

However, the presence of temporal expression in discharge summaries does not necessarily imply that the temporal information is correct. In a study that compared the stated temporal expressions in clinical reports to the times those events actually took place (as verified by other hospital records), Hripcsak et al. discovered that the reported times showed significant deviation (roughly 20%) from the time the event actually took place (Hripcsak et al., 2009). This finding provides a fascinating insight into the way that medical events are reported in free-text records, and will need to be accounted for (or knowingly discounted) in any NLP analysis of discharge summaries.

### 6.4 Overview of clinical trial settings

In order to create a case study that can be considered representative of “real” medical studies, it is important to determine what generalizations can be made about clinical

## CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK

studies. Overall, the settings presented here can be divided into two categories: those that determined the type of case study that would be performed and presented in this dissertation, and those that will need to be accounted for in the case study itself.

Of the first type, it is clear that any NLP-focused case study (as opposed to one that is focused on medical research or an actual clinical trial) will be best approached by using a nonexperimental, retrospective, case-control paradigm, as those settings are ones that are actually used in medical research (and therefore any findings will be applicable to other studies), but do not require that actual medical research be done in order to approximate the conditions such studies are done in (and so frees the NLP researchers from having to obtain permission to do actual medical studies for which they are not qualified).

Of the second type, we have study conditions that the NLP research will be trying to emulate in the annotation and NLP task. These can be itemized as follows:

- The study must have identified eligibility criteria;
- The study must also have criteria for matching case and control groups;
- The eligibility criteria should be representative of those used in existing clinical trials;
- The eligibility criteria should contain at least one form of temporal modifier;
- The discrepancies in reported and actual times of events needs to be accounted for;
- The annotation and/or NLP system needs to account for the structure of the discharge summaries.

*CHAPTER 6. SETTINGS FOR A DOMAIN EXPERT CLINICAL TASK*

How these conditions are accounted for in the case-control study are presented in the next chapter.

# Chapter 7

## Domain Expert Annotation

One of the outcomes of the research described here is the creation of a corpus annotated by domain experts in a light annotation task. However, as with most corpora the creation process of the annotation was not a simple, one-step process. The MATTER cycle contains within it the MAMA (model-annotate-model-annotate) cycle specifically because corpus annotation tasks nearly always require multiple iterations before reaching their final forms. Because one of the goals of this dissertation is to explore the application of the light annotation methodology for domain experts, this chapter will work through the MAMA cycle that was used to create the final annotation of the Patient Evaluation Resource for Medical Information in Text (PERMIT) corpus.

Section 7.1 of this chapter fully explains the specific criteria used for the mock-study described in the previous chapter and how the corpus used for annotation was chosen. Section 7.2 explores the iterations of the MAMA cycle used in order to create the PERMIT corpus, the annotated, gold standard corpus created through the MAMA process. Finally, Section 7.3 evaluates the inter-annotator agreement for the corpus, and examines how the light annotation methodology affected the annotation

task<sup>1</sup>.

## 7.1 Annotation Task Settings

As Chapter 3 described, the first steps that must be completed before entering the MATTER cycle is to define the goal of the dataset and the corpus that will be used. This section describes the eligibility and matching criteria that were used to define the goal of the annotation task, as well as how a suitable corpus was chosen and the backgrounds of the annotators.

### 7.1.1 Defining the goal: criteria selection

As discussed in Section 6.1, the overall goal of the research undertaken by Brandeis University and the Channing Laboratory was to explore the use of NLP techniques in clinical data by mimicking the properties of retrospective, case-control studies and the procedures used in them to find qualified participants. In order to accomplish this goal a set of eligibility and matching criteria had to be developed, and these criteria had to then be applied to a corpus of clinical documents. The patient eligibility criteria that were used as the basis for the annotation task are:

**Selection criteria:**

**General criterion 1:** must be under 55 years old at time of admission

**General criterion 2:** must have diabetes

**Case criterion 1:** must have had a cardiac event within 2 years of admission

date

---

<sup>1</sup>Some of the information in this chapter was first presented in Stubbs and Pustejovsky (2011) and Stubbs (2012).

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

**Control criterion 1:** no history of cardiac events

**Matching criteria:**

**Matching Criterion 1:** race

**Matching Criterion 2:** sex

**Matching Criterion 3:** lipid measurement w/in 6 months of admission

**Matching Criterion 4:** information on diabetic treatment

**Matching Criterion 5:** lipid medications

These criteria were not selected randomly, but were modeled after medical trials that have been and are currently being run, in order to ensure that the different criteria would be relatively likely to appear in the same documents. Specifically, the relationship between cardiac events and diabetes is one that exists but is being studied (Sumner, 1999; Armitage et al., 2007; Zürn, 2011), and so provided a reasonable base for the mock-study being created for the annotation project.

In order to mimic the case-control nature of some retrospective studies, some of the selection criteria apply to all candidates (age and diabetes status), one applies only to the study’s “cases” (recent cardiac event), and one applies to the “controls” (no history of cardiac events; i.e., the absence of the case criterion).

The matching criteria were based on general factors that the researchers at Channing suggested as plausible confounding factors—since no actual medical study was being performed, these criteria were not vetted by an Internal Review Board or placed under any sort of medical review; their presence is primarily to mimic the case-control study paradigm, rather than to provide a basis for actual medical analysis.



### 7.1.2 Corpus selection

Initially, the corpus for this research was to be provided by Brigham and Women’s Hospital medical records, but it was not possible to obtain permission from the organization’s Internal Review Board for those records to be released to a different institution. Therefore, the corpus was instead collected from that MIMIC II Clinical Database ([physionet.org](http://physionet.org), 2010) version 2.4. The MIMIC database contains over 26,000 de-identified medical records, most of which include hospital discharge summaries from Intensive Care Units of Beth Israel Deaconess Medical Center in Boston. These records have the names of all patients and doctors changed, and the times have been shifted systematically throughout the documents in order to maintain temporal consistency as well as anonymity (Clifford et al., 2010).

Unfortunately, in order to de-identify the data as much as possible, information about the patients’ race was removed from the records, so Matching Criterion 1 did not represent a significant portion of the resulting annotation. Fortunately, as the corpus was not intended to be used for actual medical research, this did not have an adverse effect on the annotation process, though the annotations themselves could not be used for analysis of matching procedures later on.

In order to create a corpus that was both representative in terms of the different ICUs that the records came from, and balanced with respect to the criteria the annotation task was based on, the documents in the corpus were a mix of randomly selected files, and files that were selected based on their containing keywords related to the selection criteria. Some files were randomly selected both to make the PERMIT corpus representative of the the larger MIMIC corpus, but also to ensure that any NLP system built later on would be able to disregard completely unrelated doc-

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

uments. Only discharge summaries were used in the corpus, as they provide the clearest description of what a patient experienced while in the hospital. Table 7.1 contains an overview of what keywords were used in the selection process, and how many documents containing each keyword were randomly selected from the database.

keyword	number
“diabetes”	16
“DMII”	8
“heart attack”	9
“insulin”	14
“LDL”	7
“myocardial infarction”	16
randomly selected	16
	total: 86

Table 7.1: Keywords used to find initial annotation corpus

In addition to the 86 keyword-based and randomly selected files, the corpus contained another 14 discharge summaries that were randomly chosen when the clinical trial model was first being developed (described later in Section 7.2.1), and different annotation ideas were being explored. In total, the PERMIT corpus contained 100 discharge summaries that were a mix of randomly selected files (to help create a representative corpus) and keyword-selected files (to help maintain balance towards the goal of the annotation).

### 7.1.3 Annotators

The intention for this annotation was always that it would be a domain expert task: any inference about whether a patient has a particular condition or medical history based on clinical documents requires that the document annotator be qualified to

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

interpret the documents. Therefore, two medical researchers were asked to perform the annotations: one is a Registered Nurse who is regularly involved in medical research, and the other is involved in patient selection and data collection for medical studies; both are employees of Harvard Medical and Brigham and Women’s Hospital in Boston, MA.

A third annotator with experience in medical billing was also used to test initial versions of the annotation guidelines for the light annotation task, but this person’s medical billing experience proved to not be sufficient to perform an annotation of the full corpus<sup>2</sup>. This annotator’s input was useful, but ultimately these annotations were not included in the PERMIT corpus.

### 7.2 The PERMIT annotation cycle

The methodology for creating light annotation tasks for leveraging domain expert knowledge described in Chapter 4 was not the original model for the PERMIT annotation. In fact, it was by working through the MAMA cycle with the PERMIT corpus that inspired the idea of the light annotation, and eventually informed the creation of the principles and methodology for annotation tasks of that type. In this section, the full MAMA cycle for the PERMIT corpus is described, from initial plans for a more semantically “dense” model to the final light annotation model and guidelines.

---

<sup>2</sup>This should not be taken to indicate that medical coders cannot be used for clinical annotations; in fact, coders have made excellent annotators for other clinical annotation tasks.

### 7.2.1 Initial Model: CLEF

As discussed in Section 1.4.4, other projects have examined the problem of annotating medical documents. Therefore, rather than build a model from scratch, original plans for this research called for the annotation to be based on an existing specification from a different corpus.

Initially this project was going to use the Clinical E-Science Framework (CLEF) (Roberts et al., 2007) annotation schema and guidelines (working group, 2007) for the PERMIT corpus. CLEF has two extent tags, *Entities* and *Signals*, and two link tags, *Coreference* and *Relationships*. Each of these tags has subcategories that are used to further classify the text being annotated; for example, *Entities* is further subdivided into Condition, Intervention, Investigation, Result, Drug or Device, and Locus.

However, an initial annotation effort using only the different *Entities* tags quickly made it apparent that using the CLEF schema and guidelines would be extremely time-consuming, and would also require substantial changes. The existing CLEF guidelines were unclear in terms of defining what made something a ‘condition’ rather than a ‘result’, or an ‘intervention’ instead of an ‘investigation’. Additionally, CLEF did not have a way of capturing demographic information, an important factor in the matching criteria

### 7.2.2 Initial Annotation: CLEF Entities

Two documents were chosen randomly from the MIMIC database, and the R.N. annotator was asked to apply only the *Entities* tags to the files based on the available CLEF guidelines (working group, 2007). Unfortunately, the guidelines proved to be unclear in terms of distinguishing sufficiently between the different types of Entities,

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

and much time was spent between the researchers and the annotators in discussing the different possible applications of the different tags. Additionally, the annotated CLEF corpus was not (and still isn't) available for download due to problems obtaining permission to de-identify and distribute the corpus (R. Gaizauskas, private communication, December 17, 2009).

Eventually, however, the annotation of the sample discharge summaries using the CLEF extent guidelines was completed, and the results were evaluated.

### 7.2.3 Initial Evaluation

The Entity annotation with the CLEF guidelines required that the annotator (and everyone else involved in the project) be involved in long, detailed discussions attempting to determine the semantic distinctions between the different classes of entities, a process that was difficult, time-consuming and had little to do with the goal of the corpus as a resource for evaluating patient records in relation to the selection and matching criteria.

While having a corpus annotated with a specification as detailed as CLEF would certainly be a useful resource for general NLP purposes, the time it would have taken to create and adjudicate the annotations would have completely consumed the entire length and budget of the grant—each document took two hours to annotate with the CLEF entity guidelines. Additionally, requiring domain experts to make judgments about semantic roles was not an efficient use of their expertise.

Eventually, it was decided that a new annotation model would be developed and applied, one that adhered more closely to the goal of the corpus, and any semantic or syntactic annotation that were required for NLP processing would be applied auto-

matically with existing resources, or provided by annotators who were not bioclinical domain experts (the second set of annotators could, of course, be linguistic experts).

#### 7.2.4 Second Model – Light annotation

The light annotation specification that was developed for the PERMIT corpus used only four tags: three extent tags used to identify sections of text relevant to each criterion and to identify significant modifying phrases, and one linking tag used to associate different extents, generally between criterion markers and their modifiers. Non-consuming tags showed where annotators did not find mentions of any relevant text for a particular eligibility criterion. The full DTD for this task can be found in Appendix B.

##### Extent tags

The bulk of the annotation in the PERMIT corpus is contained in extent tags. There are three types: `selection_criterion`, `matching_criterion`, and `modifies`.

The `selection_criterion` and `matching_criterion` tags were used to mark text that was relevant to the criteria outlined in Section 7.1.1. The `selection_criterion` and `matching_criterion` tags both have an attribute called “`criterion`”, which annotators used to indicate which criterion the text they were marking was relevant to; for example: ‘age’, ‘diabetic’, ‘recent card. event’, ‘no card. event’. Another attribute was used to indicate whether the annotated text showed that the criterion was met or not (or present or not, in the case of matching criteria).

The `modifies` tag was used to annotate context that would change the interpretation of the criterion-related text. The use of this tag varied widely: in some cases

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

it was used to mark dates related to time-dependent criteria, in others it was used to indicate if the criterion-related text was about a family member rather than the patient, or was in some way negated or theorized about (“may be at risk for...”). Applications of this tag will be discussed further in Section 7.4.

### **Link tags**

This annotation task uses only one link tag, called `modifies`, which is used to connect text tagged as `modifier` to the criterion-related task it provided context for.

### **Non-consuming tags**

Both the `selection_criterion` and `matching_criterion` tags have the option of being “non-consuming”. For the purposes of this annotation task, a non-consuming criterion tag is used to indicate that there was no text in the document related to the indicated criterion. This annotation task makes the assumption that if a disease or condition is not mentioned in a discharge summary, then the person in question does not have that disease or condition. Therefore, if a discharge summary contains no mention of diabetes, then the annotator creates a non-consuming `selection_criterion` tag, sets the “criterion” attribute to “diabetic” and sets the “meets” criterion to “DOES NOT MEET”. This way it is clear from the annotation that the annotator found no mention of diabetes-related text in the document and is making a positive claim that this person most likely does not have diabetes.

### 7.2.5 Second Annotation

The two annotators were provided with the corpus of 100 documents, the DTD for the case-control annotation task described above, a short set of annotation guidelines, and MAE, the software for light annotation tasks described in Chapter 5.

The annotation guidelines as they were presented to the annotators are provided in Appendix C; this section gives a brief overview of how the guidelines indicated the tags in the specification were to be applied.

As training, each annotator was asked to annotated a sample set of documents according to the new specification and guidelines, and were encouraged to ask questions about the annotation task. As a result of these questions some modifications and clarifications were added to the annotation guidelines (these were technically a few additional iterations of the MAMA cycle, though the specifics of what were changed were minor and don't need to be discussed), and once the annotators seemed to understand the task and the annotation, they were allowed to annotate the remainder of the corpus.

#### Light Annotation Procedure

Essentially, the procedure outlined by the annotation guidelines was intended to imitate the actual process that would be used by a medical researcher when trying to determine if a patient met a set of eligibility criteria: the annotators were asked to read the entire document, and indicate what keywords stood out as being relevant to the “study” being recruited for. When a set of relevant text was identified, it was annotated as either `selection_criterion` or `matching_criterion`. Then the annotator checked the document for nearby information that indicated whether this



## CHAPTER 7. DOMAIN EXPERT ANNOTATION

keyword is being used in a way that classifies the criterion as being met or not (for selection criteria) or present or not (for matching criteria). Information used to determine whether a keyword is related to a criterion was annotated as a “modifier”, and linked to the related keyword with a “modifies” tag.

For example, if the matching criteria are related to whether a person has diabetes, a mention of a person being admitted to the hospital because of diabetic shock will be annotated as a positive indication that the patient is diabetic. However, if the phrase “father w/ DMII” appears in the “Family History” section of the document, then “DMII” will be annotated as a `Selection_criterion`, but the “father” will be marked as a modifier and linked to “DMII” and the “DMII” mention will be marked as one that does not meet the requirements for selection criteria. The resulting annotation will look something like this:

```
<Selection_criterion id="SC16" text="DMII"
  criterion="diabetes" meets="NO"/>
<Modifier id="M2" text="father"/>
<Modifies id="ML26" from="M2" to="SC16"/>
```

This method of annotation lessens the work done by the domain experts—rather than having to do a full linguistic annotation on top of the complex medical texts they have to interpret, they were able to focus on their specialized knowledge, and linguistic knowledge can be added later where needed.

### **Light annotation guidelines**

In addition to the procedure outlined above, the annotators were asked to evaluate each mention of relevant text separately from the rest: that is, each mention of ‘diabetes’ should be evaluated in the context of whether that particular instance

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

indicated whether the criterion requiring a patient be diagnosed with diabetes was met, not whether the document as a whole indicated that the patient had diabetes.

Annotators were also asked to use clues other than simple mentions of the disease names to determine if a patient had had a disease, but only if those clues were unambiguous. For example, a blood test revealing that the patient had a high blood sugar level may not specifically indicate that the patient has diabetes (they could only be at risk, or something else could have affected the test), but a mention of insulin being taken was unambiguous because no other disease uses insulin as a treatment. The ‘unambiguous’ indication does not appear in the guidelines in Appendix C, however; this was added during the phase where the annotators were testing their understanding of the guidelines.

Finally, as described in the previous section, annotators were used to mark up context that indicated whether a criterion was being met, such as temporal markers or negations. However, this request did not include indicators such as document section headers, as those were determined to be too numerous to annotate and they slowed the annotation down too much.

### **7.3 Evaluation of the PERMIT corpus**

As one of the primary concerns surrounding the abandonment of the CLEF annotation specification was the length of time it would take to annotate the entire PERMIT corpus, a key factor in evaluating the light annotation task is whether or not the new specification helped solve this problem. In fact, the annotation process was greatly sped up by the adoption of the light annotation task. While the CLEF Entity annotation took the annotator roughly two hours per document, the criteria-specific

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

annotation was done much faster, at an average rate of 3.72 documents per hour (averaged across both annotators) (Stubbs, 2012).

As a consequence of saving time during the annotation of the PERMIT corpus, money was also saved: based on the time each document took to annotate under the CLEF Entity annotation, the actual cost of the light annotation was roughly 90% less than the projected cost of the CLEF annotation (ibid.). While the light annotation method is not appropriate for all tasks (certainly the CLEF model does serve a purpose in bioclinical annotation tasks), when it can be used it does provide a reduction in both time and expense.

Because of the cost of hiring domain expert annotators, being able to reduce the amount of time and money spent on annotations for a research project are an important outcome of the light annotation task methodology. However, this is not the only metric by which an annotated corpus is evaluated, and the rest of this chapter examines the more standard conventions for evaluating an annotation task, such as inter-annotator agreement and the results of the corpus adjudication.

In an annotation task, the inter-annotator agreement coefficient and precision and recall scores of each annotator compared to the gold standard are used as indicators for how clearly the task was defined, and how reproducible the results would be. High scores indicate good levels of agreement and a more reproducible task.

### 7.3.1 Inter-annotator agreement scores

While in some areas of corpus linguistics it has become standard to measure inter-annotator agreement (IAA) with Cohen's kappa (Cohen, 1960) (or Fleiss' kappa, depending on the number of annotators (Fleiss, 1971)), these measures rely on being

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

able to determine the number of true negatives found in the corpus, a metric that can be difficult to quantify. Hripscak and Rothschild (2005) observed that “In text markup studies, computer systems mark relevant phrases in documents. Negative cases correspond to nonrelevant phrases. Their number is poorly defined because phrases can overlap and vary in length”. An additional problem with kappa statistics is that in cases where positive results are sparse, such as medical informatics, kappa scores will be artificially lower even when the annotators agree, due to the large number of negative responses (Hripscak and Heitjan, 2002).

This same problem of data sparsity and high numbers of negative results applies to evaluating the agreement between annotators in sparse annotations such as the PERMIT corpus. However, Hripscak and Rothschild also determined that the unweighted f-measure can be used to evaluate such instances. They note that if the number of negative examples “...is at least known to be large, however, the probability of chance agreement on positive cases approaches zero; [...]  $k$  approaches the positive specific agreement. Therefore, for experiments with large but unknown [true negatives], the average positive specific agreement, which equals the average F-measure among the raters, approaches the  $k$  that would be calculated if [the number of true negatives] were known.”

Because the PERMIT corpus annotation is sparse, Cohen’s kappa will not provide a true representation of the accuracy of the annotation process. While it might be argued that the agreement could be evaluated based on whitespace-separated tokens, because the different tags could be applied equally to whole phrases instead of single tokens, this compromise would require that most tags be counted multiple times in an evaluation, which will lead to inaccurate results.

The F-measure metric is the harmonic mean of the precision and recall of the

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

documents being evaluated, as shown:

$$Precision = \frac{true\_positives}{true\_positives + false\_positives}$$

$$Recall = \frac{true\_positives}{true\_positives + false\_negatives}$$

$$Fmeasure(F_B) = \frac{(1 + B)(precision * recall)}{precision + recall}$$

The  $B$  coefficient can be used to weight the equation to favor precision or recall, but here there is no need to do so, and so 1 will be substituted for  $B$  in the equation, making it:

$$Fmeasure(F) = \frac{(2)(precision * recall)}{precision + recall}$$

By using one annotator as the ‘gold standard’ prior to the actual adjudication, an f-measure can be calculated that reflects the IAA score of the annotation task. Because the f-measure is the harmonic mean of the two scores, it does not matter which annotator is used as the ‘gold standard’.

For Table 7.2, the inter-annotator agreement was calculated by using Annotator 2 as the gold standard. In cases where no true positives were found, the precision, recall, and f-measure were counted as being 0. In some annotation tasks, this would necessarily be an accurate portrayal of the state of agreement between the annotators because it would be possible that the “gold standard” annotator simply did not use the tag being evaluated, which could mean that agreement was perfect if the other annotator also did not use the tag. However, because this annotation task required that the annotators use every tag at least once, this scenario does not apply. While

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

this could be interpreted as having the effect of artificially lowering the agreement scores, it is preferable to artificially raising them by removing all those cases from the analysis altogether. Scores were calculated by tag across all documents that both annotators annotated (Annotator 1 was missing a document, and that one was left out of the evaluation), and the table shows the average precision, recall, and f-measure for each tag across all documents.

Prior to this analysis being done, some cleanup of the data was performed. First, a Python script was run over the annotated files to remove whitespace from the beginnings and ends of annotated extents, in order to more accurately assess strict agreement of extent tags. Additionally, Annotator 2 misread the age restriction as “55 or over” rather than “under 55”, and so the “MEETS/DOES NOT MEET” attribute for those tags were the opposite of what they were meant to be. As this was a simple misreading with a binary interpretation, rather than ask the annotator to fix those documents by hand, a script was used to fix the error automatically.

	strict			overlap		
	prec.	recall	f-meas.	prec.	recall	f-meas.
Selection criteria	0.531	0.592	0.541	0.617	0.68	0.626
age	0.516	0.864	0.635	0.523	0.871	0.642
diabetic	0.688	0.578	0.604	0.853	0.698	0.735
recent card. event	0.305	0.457	0.34	0.34	0.512	0.379
no card. event	0.48	0.468	0.47	0.48	0.468	0.47
Matching criteria	0.719	0.722	0.711	0.877	0.883	0.867
sex	0.44	0.445	0.439	0.91	0.917	0.893
race	0.897	0.904	0.897	0.897	0.904	0.897
recent lipid test	0.891	0.887	0.888	0.899	0.895	0.896
diabetic treatment	0.689	0.681	0.666	0.73	0.727	0.705
lipid medications	0.813	0.814	0.797	0.866	0.884	0.857
Modifier	0.182	0.265	0.204	0.246	0.376	0.279

Table 7.2: Precision, recall, and f-measure using Annotator 2 as the gold standard

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

The agreement metrics were calculated twice: first using strict rules for counting whether tags matched: a true positive was only found if both annotators had a tag of the same type and attribute with the exact same start and end locations; and second with relaxed rules for extent agreement, where a true positive was found so long as the tag and attributes agreed and the annotators tags shared at least some characters in common.

Unsurprisingly, agreement was much higher using the overlap metric. As the annotation guidelines did not have strict rules about where the beginning and end of a criterion tag should be located, the annotators used their judgment, which meant that they used different rules for including determiners, possessives, and descriptive phrases. However, the fact that they tended to agree on general location is a positive result for a light annotation task.

The agreement scores are still a bit lower than might generally be expected for what appears to be a relatively simple task. The reasons for this will become clearer in the next section, where the creation of the gold standard is discussed, and an error analysis of the entire dataset is performed, including an overview of the agreement between the annotators and the adjudicated gold standard.

### **7.3.2 Adjudication of the gold standard and annotation error analysis**

The gold standard was created by adjudicating the annotations of the two annotators using the software MAI described in Section 5.3. The adjudication was performed by the author of this dissertation, in consultation with the annotators themselves and the researchers at the Channing Lab.

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

The adjudication process revealed the most common sources of error in the annotated documents, both in terms of problems with the guidelines that needed to be reconciled, and problems with applying the annotation model to the documents. A confusion matrix of the annotator’s tags is in Appendix D.

In addition to the tags described in Section 7.2.5, an additional tag was added to the annotation, called `Overall`. This tag was used to indicate whether the patient being described met each of the criteria individually, and whether they could therefore be categorized as a case, control, or neither. See Appendix B for the DTD description of the tag.

### **Cardiac events**

One primary source of disagreement between the annotators was that they were not using the same definition of “cardiac event”. Although “cardiac event” is a term that is used in the medical community, its use varies depending on the context of its use, such as the goal of the study being performed. As a result, there was some confusion between annotators over whether the term’s use in the criteria should apply to only medical events that were caused by something going wrong with the heart itself, or if it should be taken to include events such as surgeries on the heart.

Because the case-control study was not intended to be used for actual medical research, the determination on this point was somewhat arbitrary, as there was no medical reason to use one definition over the other. Eventually it was decided that only internal conditions (and therefore not surgeries or other external interventions) would be considered “cardiac events”. While this rule was made clear to the annotators during the annotation process, there was additional confusion as to whether events that more indirectly affected the heart (such as arterial blockage) were con-



## CHAPTER 7. DOMAIN EXPERT ANNOTATION

sidered cardiac events; unfortunately, because the annotators worked independently from each other and not at the same time, this discrepancy was not discovered until adjudication began.

For the purposes of the adjudication, a list of qualifying cardiac events was compiled and used to determine which selection criteria tags would be used in the gold standard. Based on the advice of the consultants at the Channing Lab, the list of accepted cardiac events included acute coronary syndrome, congestive heart failure, coronary artery disease, heart attack, heart disease, heart failure, myocardial infarction and vessel disease; but it did not include events such as av block, cardiogenic shock, chest pain/discomfort/tightness, heart block, pericardial effusion, pulmonary edema or tamponade.

Because of the confusion over what constituted a cardiac event, the agreement scores for both the ‘recent cardiac event’ and ‘no history of cardiac events’ was lower than it might have been otherwise. However, this could easily be fixed in future annotation projects by specifying qualifying events in the guidelines.

### **Modifier and modifies tags**

Another source of low agreement scores was confusion over how large a span of text should be annotated, and what should be considered a modifier. Given the string “type II diabetes”, one annotator tended to annotate “diabetes” as a selection criterion and “type II” as a modifier, while the other would label the longest string as a selection criterion. The ‘longest string’ approach is what was used in the gold standard. This aspect of the annotation/adjudication effort is discussed further in Section 7.4.

### Locating relevant extents

The biggest source of inter-annotator disagreement was not technically a disagreement at all. In many cases, both annotators were, in fact, annotating the same types of events and medications, but a large amount of time an extent that was correctly annotated by one annotator was simply missed by another. The most likely cause of this problem is that the discharge summaries in MIMIC are often quite long and repetitive, and so it was easy to miss relevant data. Future annotation efforts would likely benefit from splitting of the files into more manageable pieces. The gold standard version of the PERMIT corpus contains all the correct annotations from both annotators. While a human would be able to tell from a single example if a criterion was met, it was thought that a computer would need more input data than a single mention per document, so the gold standard has “meets”/“does not meet” annotations for all mentions of relevant terms in a document.

### 7.3.3 Annotator agreement with the gold standard

Once the gold standard was adjudicated, precision, recall, and f-measure were again calculated, this time by comparing each annotator’s tags per file to the tags in the corresponding gold standard file. Again, extent matches were calculated using both strict and overlapping metrics. Table 7.3 shows the relationship between Annotator 1 and the gold standard annotations, and Table 7.4 shows the same statistics for Annotator 2.

These tables clearly show just how large a problem the annotator’s missing relevant extents was. With the exception of the modifier tag for Annotator 1, the precision, recall, and f-measure scores were higher for both annotators across the board when

CHAPTER 7. DOMAIN EXPERT ANNOTATION

	strict			overlap		
	prec.	recall	f-meas.	prec.	recall	f-meas.
Selection criteria	0.697	0.664	0.661	0.764	0.727	0.725
age	0.941	0.92	0.916	0.948	0.927	0.923
diabetic	0.807	0.656	0.699	0.952	0.749	0.808
recent card. event	0.371	0.551	0.416	0.399	0.603	0.449
no card. event	0.52	0.443	0.456	0.52	0.443	0.456
Matching criteria	0.956	0.878	0.91	0.974	0.897	0.929
sex	0.93	0.902	0.904	0.941	0.913	0.914
race	0.914	0.914	0.912	0.914	0.914	0.912
recent lipid test	0.921	0.915	0.918	0.927	0.921	0.924
diabetic treatment	0.902	0.806	0.839	0.926	0.83	0.862
lipid medications	0.931	0.902	0.903	0.926	0.902	0.902
Modifier	0.225	0.332	0.252	0.277	0.403	0.307

Table 7.3: Precision, recall, and f-measure between Annotator 1 and the adjudicated gold standard

compared to the gold standard, as opposed to when they were compared to each other. If one annotator had simply not followed the guidelines at all, that annotator’s scores compared to the gold standard would be just as low as when compared to the other annotator. However, we see that this is not the case, and the fact that precision tends to be higher than recall backs up this assessment of the annotation task. Clearly, the density and repetitiveness of the discharge summaries was a challenge for this annotation task, though other domain expert tasks may not experience the same difficulties with their annotation data.

Overall, the agreement even between the annotators and the gold standard is not quite as high as it might be expected to be, but this case-control annotation task provides an excellent source of data for evaluating the creation of light annotation tasks. The next section provides an analysis of how the PERMIT corpus compares to the stated principles for light annotations, and what lessons were learned from this

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

	strict			overlap		
	prec.	recall	f-meas.	prec.	recall	f-meas.
Selection criteria	0.938	0.806	0.863	0.967	0.843	0.896
age	0.96	0.599	0.726	0.96	0.599	0.726
diabetic	0.925	0.889	0.899	0.949	0.908	0.921
recent card. event	0.861	0.838	0.837	0.881	0.861	0.858
no card. event	0.833	0.789	0.802	0.858	0.804	0.82
Matching criteria	0.834	0.768	0.794	0.971	0.892	0.924
sex	0.506	0.493	0.497	0.948	0.92	0.924
race	0.965	0.958	0.958	0.965	0.958	0.958
recent lipid test	0.967	0.97	0.968	0.97	0.972	0.97
diabetic treatment	0.873	0.776	0.803	0.901	0.8	0.829
lipid medications	0.871	0.845	0.852	0.938	0.896	0.909
Modifier	0.605	0.622	0.596	0.657	0.668	0.643

Table 7.4: Precision, recall, and f-measure between Annotator 2 and the adjudicated gold standard

annotation process.

### 7.4 PERMIT corpus as a light annotation task

The principles of light annotation presented in Section 4.3 were not fully formed prior to this case study; rather, this case study helped refine the principles by highlighting areas of the annotation task that had the most problems or resulted in the most errors. The rest of this section examines the PERMIT annotation task in light of the established principles.

**The annotations are performed by experts in the field:** Both of the annotators had the experience to evaluate the documents presented; the third annotator (with medical billing experience) had difficulty identifying relevant medications and other

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

markers, and was therefore asked to not continue the annotation process after a few documents.

Working with the annotator with medical billing experience made it clear that for the given task, it was important that the annotators be able to understand the text of the discharge summaries in a very complete way, as well as be able to recognize what medications, test results, etc. were related to various conditions. However, this should not be taken to mean that no medical coders are qualified to perform clinical annotations; simply that any who do should have sufficient experience and confidence to perform well at their task (something that should be true of any annotator).

**The task is divided into as few classification questions as possible:** This principle was implemented when the task was re-defined as pertaining only to the criteria established for the mock case-control study. Rather than perform an annotation of all the medical events and related objects in the text, the annotation was limited to extents directly pertaining to the established criteria.

The selection criteria and matching criteria were given separate tags, and each of those tags had an attribute that indicated which criterion the associated extent was referring to, and an attribute that indicated whether the extent suggested the criterion was met or not.

**The classifications used in the model are based on current best theories and practices for the chose domain:** The use of insulin as an unambiguous indicator for diabetes was vetted by the collaborators at the Channing Laboratory. However, as was discussed in Chapter 7.3.2, the definition of “cardiac event” proved much more difficult to define without simply providing a list of qualified events.

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

This principle was heavily influenced by the experience with the cardiac event criteria. While any research project that is primarily medical research and secondarily NLP research would presumably have the definitions of all the terms worked out well before any annotation task is undertaken, it is still vital that the medical priorities and definitions be made clear to all the participants, and that medical expertise is the deciding factor in these definitions.

**Annotation should be done based only on what is in the text, not on expert’s intuitions about the text:** This principle was adhered to entirely: by asking the annotators to indicate the areas of text that were relevant to the criteria (as well as the areas that might appear to be positive indicators but were, in fact, not), the annotation task was grounded in the existing text of the discharge summaries.

While annotators were encouraged to use medications as evidence of certain conditions, this was only allowed when the medication use was unambiguous (such as “insulin” as an indicator for diabetes) and still had extents in the text associated with them. In addition, if the text contained no mention of anything that was relevant to a criterion, the annotators were asked to use a non-consuming tag to indicate that lack of evidence. This requirement, because it meant that the annotators were actively making a claim about the document rather than simply not annotating anything, helped ensure that they were more careful about checking the document for text related to each criterion.

**If possible, the annotations should be applied to sentence- or phrase-level sections of the document, in support of document-level classifications:** In retrospect, based on the error analysis in Section 7.3.2, the `modifier` extent tag and

`modifies` link tag should not have been included in the annotation task. Instead, annotators should have simply annotated the entire relevant phrase or sentence, without trying to break up the phrase into constituent parts. While annotating an entire extent may have ended up including information that was not relevant to the stated criteria, that information could have been weeded out later in the MATTER cycle when a fuller Model was being constructed. As it stands, the `modifier` tag was the source of the greatest amount of disagreement and confusion for the annotators. Future light annotation tasks should more closely adhere to the principle of sentence- or phrase-level annotations.

**Additional layers of annotation can be provided before or after the light annotation is performed without conflicting with the given classifications:**

Because the annotation environment generated stand-off annotation based on character offset, the final version of the annotations are fully augmentable with other annotation layers and information. As will be shown in Chapter 8, this makes leveraging the expert annotations into an NLP system much easier.

## 7.5 Summary

The case-control annotation performed over the PERMIT corpus provided an excellent way to test and refine the principles of light annotation tasks. While not all the principles were strictly adhered to, due to the fact that they were still being determined at the time the corpus was annotated, the PERMIT corpus is a good example of a light annotation task, and examining the places where the case-control annotation was less than ideal reveals the motivations behind the principles of light

## CHAPTER 7. DOMAIN EXPERT ANNOTATION

annotation tasks.

Even with some of the difficulties encountered with the PERMIT corpus annotation, the use of a light annotation over the more semantically dense CLEF specification led to a decrease in time and money spent on the annotation project, and these aspects of research are always concerns when hiring consultants for professional-level work. Additionally, the resulting annotation of the PERMIT corpus is sufficiently grounded that it can easily be used for building and testing NLP systems, as will be shown in the next chapter.



## Chapter 8

# Using Expert Annotation in an NLP system

The purpose of light annotation tasks is to encode domain-specific knowledge in a way that allows the expert-level understanding of the text to be used in other annotation models and natural language processing systems. While the light annotation allows the domain expert knowledge to be saved, it does not provide a complete annotation of all the aspects of the text that would be used for a machine learning or rule-based NLP system. However, the principles underlying the light annotation provide insight into the aspects of the text that can be turned into features, and the format allows the annotated gold corpus to be augmented with other annotated data, either by hand or by machine.

This chapter examines how the light annotation of the PERMIT corpus can be used to set the foundation of a system designed to mimic the process of finding patients who meet the eligibility criteria. The system described here does not provide a start-to-finish process for fully automating the patient selection process; rather, it

## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

contains some preliminary research into using the domain expert knowledge encoded by the light annotation task.

Section 8.1 examines the distribution of data in the PERMIT corpus, in terms of the number patients who meet each criterion, the number of cases and controls, and how many times each tag was used. Section 8.2 uses the corpus to test and train machine learning algorithms in order to set a baseline for evaluation of the corpus in an NLP system. Section 8.3 demonstrates the use of domain-expert identified keywords for recognizing met criteria, and Section 8.4 augments the keyword-based system with some course information about the document structure of the discharge summaries. The remainder of the chapter looks at some of the different ways the PERMIT corpus could be tested using other, existing NLP systems for clinical texts.

The work in this chapter focuses primarily on the selection criteria rather than the matching criteria; the techniques for leveraging the two types of tags would be largely the same, so discussion is focused on the selection criteria in order to keep the chapter narrative focused on the strategies used for assessing and leveraging the light annotation.

### 8.1 Data distribution in the PERMIT corpus

In order to fully utilize the domain expert annotation, it is important to have an understanding of how the annotations are distributed in the corpus, and to look for trends in the annotated texts and patient diagnoses. This section examines the distribution of cases and controls in the PERMIT corpus, as well as provides analysis of the use of the different tags for each of the selection criteria.

### 8.1.1 Cases and Controls

The motivation for the PERMIT corpus was to mimic the circumstances of a mock retrospective case-control study by generating selection and matching criteria and asking domain experts to create annotations based on the goals of the mock medical study. For the purposes of the PERMIT corpus, a ‘case’ is someone who meets the following criteria: is under 55 years old at the time of hospital admission, is diagnosed with diabetes, and has had a cardiac event within two years of the date of hospital admission. A ‘control’ in the corpus must also meet the age and diabetic criteria, but have no history of cardiac events. While these criteria are relatively simple compared to some of the more complex criteria used for more ambitious medical studies, the methodology of the light annotation should easily scale to criteria requiring more detailed patient evaluation.

Based on these definitions, the PERMIT corpus contains 5 cases and 5 controls, out of the full set of 100 patient records, despite many of the files having been chosen based on their containing keywords relevant to the study criteria (c.f. Section 7.1.2). The other 90 files met either none or only one or two of the stated criteria, and were therefore neither cases nor controls; the distribution of individual criteria in the corpus will be examined later in this chapter. The rest of this section investigates how the selected documents broke down into files that met the different case/control criteria.

Nine of the ten files in the set of cases and controls came from the documents that were added to the corpus because they contained one of the keywords that was being selected for. Table 8.1 shows how many files from each keyword selection met the case or control criteria.

CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

keyword (total documents set)	cases	controls
“diabetes” (16)	3	3
“DMII” (8)	0	0
“heart attack” (9)	0	0
“insulin” (14)	0	2
“LDL” (7)	0	0
“myocardial infarction” (16)	1	0
randomly selected (16)	0	0
initial document set (14)	1	0

Table 8.1: Relationship between initial corpus selection and case/control group membership

Table 8.1 does indicate that the “diabetes” keyword did comparatively well as a selector for relevant discharge summaries, though it does not tell the whole story about what keywords were contained in the qualifying files. Table 8.2 shows the relationship between the keywords used to selected files to be part of the corpus, and how many of the cases and controls contained each of those keywords, regardless of why each file was included in the PERMIT dataset.

keyword	cases	controls
diabetes	5	5
DMII	1	0
heart attack	0	0
insulin	4	5
LDL	0	0
myocardial infarction	3	0

Table 8.2: Presence of keywords in the case and control groups

There does appear to be a strong correlation between some of the keywords and whether the files meet the selection criteria, with all 5 cases and controls containing the word ‘diabetes’ and 3 of the 5 cases containing ‘myocardial infarction’. However, further inspection of the files in the full corpus reveals that 20 documents in the

## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

corpus contain both “diabetes” and “myocardial infarction”, and only three of those files qualified as cases.

Narrowing down the field of potential candidates through keywords is clearly helpful, but it is also clear that keywords alone will not suffice to accurately select patients who qualify for studies. In particular, finding patients who meet all the criteria for the control group would be difficult in this scenario, as the controls were identified by their *lack* of cardiac events.

While the stated goal of the study is to identify cases and controls from the patient population, focusing on the ‘case’ and ‘control’ labels misses the underlying problem: identifying patients who meet (or don’t meet) each of the individual criteria, then looking at all the criteria at once to determine membership in the case or control group. The next section examines the distribution of the different selection criteria in the corpus.

### 8.1.2 Distribution of selection criterion tags

In order to begin making suggestions about how to leverage the expert annotations in order to help find patients who meet each of the individual criteria, it is helpful to understand the distribution of which criteria were met or not met in the PERMIT corpus. Table 8.3 shows the number of files that meet each criterion. At the criterion level, files tended to not meet the age or cardiac event-related criteria (both having a recent cardiac event and having no history of such events), but the patients did tend to be diabetic.

This trend is particularly interesting, because while 38 of the files in the PERMIT corpus were included for diabetes-related keywords (“diabetes”, “DMII”, “insulin”),

CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

Criterion	file count
Age	
Meets	32
Does not meet	68
Diabetic	
Meets	60
Does not meet	40
Recent cardiac event	
Meets	38
Does not meet	62
No history of cardiac events	
Meets	33
Does not meet	67

Table 8.3: Distribution of met criteria

12 additional files met that criteria from the other file-inclusion groups. Similarly, though 25 files were included for containing cardiac-event related terms (“heart attack” and “myocardial infarction”), an additional 13 files met the recent cardiac event criterion.

The individual meets/does not meet attributes on the tags themselves roughly followed the distribution of those attributes over the files, as shown in Table 8.4. Again, tags skewed towards meeting the diabetic criterion, and not meeting the criterion of not having a previous cardiac event. Whether or not a cardiac event occurred within the past 2 years was often difficult to determine based on the given text, making the “recent cardiac event” criterion less heavily biased. The age restriction is often not met, likely because diabetes and cardiac events tend to affect people over 55, rather than under. On average, there were 123 selection tags per discharge summary.

tag	tag count
Age	
Meets	84
Does not meet	162
Diabetic	
Meets	363
Does not meet	66
Recent cardiac event	
Meets	137
Does not meet	130
No history of cardiac event	
Meets	43
Does not meet	250
Total tags	1235

Table 8.4: Distribution of Selection Criterion tags and attributes in corpus

## 8.2 Establishing a baseline accuracy with ML classifiers

The use of machine learning (ML) algorithms for all types of NLP tasks, not only those in the bioclinical domain, is an area of research that has been expanding rapidly for years. For this dissertation, a series of ML classifiers from both the NLTK (Natural Language Toolkit (Bird et al., 2009)) and WEKA (Waikato Environment for Knowledge Analysis (Hall et al., 2009)) were used to establish a performance baseline for accuracy of classifying the PERMIT corpus files.

Because there were so few cases and controls (five each) in the PERMIT corpus for the selected criteria, the overall classification problem was instead divided into four classifications problems: one for each of the selection criteria used for the mock study.

As the purpose of this part of the experiment was to generate a representative

CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

baseline for accuracy, all types of classifiers were used to test this dataset, rather than just one. The classifiers were each run 100 times using a random sampling cross-validation methodology that selected 65 training documents and 35 testing documents each time the classifier was run. The aggregate accuracy data for all of those runs is shown in Table 8.5. Accuracy was calculated as percentage of documents that were correctly labeled in each run.

Criterion (# tags met/not met)	NLTK			WEKA			
	Naïve Bayes	Dec. Tree	Max. Ent.	Dec. Tree (C4.5)	Naïve Bayes	Ripper	SVM
Age (84/162)							
Average	0.704	0.561	0.532	0.561	0.672	0.612	0.673
Min	0.486	0.343	0.171	0.4	0.429	0.343	0.514
Max	0.857	0.771	0.829	0.771	0.829	0.829	0.829
Med	0.686	0.571	0.6	0.571	0.686	0.6	0.657
Diabetes (363/66)							
Average	0.655	0.687	0.531	0.683	0.663	0.725	0.696
Min	0.486	0.371	0.286	0.343	0.457	0.457	0.514
Max	0.829	0.857	0.743	0.829	0.857	0.886	0.857
Med	0.657	0.714	0.543	0.714	0.657	0.743	0.686
Rec. Card. (137/130)							
Average	0.689	0.608	0.519	0.621	0.679	0.627	0.701
Min	0.514	0.343	0.257	0.429	0.457	0.4	0.571
Max	0.829	0.771	0.8	0.8	0.829	0.829	0.857
Med	0.686	0.629	0.543	0.629	0.686	0.629	0.714
No Card. (43/250)							
Average	0.739	0.643	0.531	0.617	0.745	0.72	0.748
Min	0.543	0.486	0.2	0.371	0.6	0.514	0.543
Max	0.971	0.8	0.829	0.857	0.886	0.914	0.914
Med	0.743	0.657	0.6	0.629	0.743	0.743	0.771

Table 8.5: Baseline machine learning classification accuracy values per selection criterion



## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

For features, the classifiers were given the label that indicated whether or not each criterion was met and the “bag of words” in each document. The classifiers performed fairly well, though the rest of the chapter will provide suggestions and methods for improving these scores, primarily using rule-based analysis. It should be noted that the accuracies reported here are percentage accuracies (number correct divided by total number), not precision, recall, or f-measure.

### 8.3 Keyword-based selection

When attempting to leverage the domain expert annotations, one of the first aspects of the corpus that should be examined are the words and phrases that were marked by the annotators, and how often those text extents appear in the entire corpus. Some preliminary phrase analysis was done in Section 8.1, but that was done based on prospective keywords: this section examines the extents that were actually marked as relevant to the goals of the corpus as part of the annotation task.

Appendix E shows the extents that the annotators used for each of the selection criterion<sup>1</sup>. These tables provide an overview of all the phrases used to mark whether a criterion was met or not met, and they are ordered by the frequency that each phrase appears in the annotated corpus for each criterion.

For the most part, the basic terms that are annotated for each of the criteria are in line with what might be expected if a medical dictionary was used, the annotations do reveal some surprising results. For example, there is a much wider range of phrases

---

<sup>1</sup>The extents labeled for the “age under 55 at time of admission” criterion are not included, as those extents are simply a list of dates of birth and phrases such as “44 year old”, which did not add useful information to the analysis process.

used to indicate diabetes than is found in the SNOMED-CT browser<sup>2</sup>. While a SNOMED-CT search for “diabetes mellitus” returns a set of terms and child terms that do reflect some of what the expert annotators marked up in the PERMIT corpus, variations such as “dmii”, “t2dm”, “iddm” and similar variations which would likely need to be added manually to any system built to identify patients with diabetes. The annotators also indicated insulin (and brand-names for insulin) as an indicator of diabetes, another set of diabetes-indicators that would need to be built into an NLP system.

While there is much more variation in the lists of phrases that met the recent cardiac event criterion, the overall theme of the annotations is much the same. Many of the annotated phrases do appear in SNOMED-CT, but many of the annotated variations do not. For example, a search for “myocardial infarction” does reveal “heart attack” and “Myocardial infarct” as variations of the term, it does not list “non-st elevation myocardial infarction” as a related term, let alone the abbreviation “nSTEMI”. While these phrases do appear further down the concept tree in SNOMED-CT from the higher-level term “myocardial infarction”, an annotator who is not an expert in the domain would not know at what point the relationship between “cardiac event” and the conceptual children of “myocardial infarction” might stop, which necessitates the inclusion of domain professionals in the annotation task.

### 8.3.1 Using keywords with the PERMIT corpus

Despite the small variations in the annotations that were created for each criterion, a core list of relevant phrases can be used to help identify documents that fall into

---

<sup>2</sup><https://uts.nlm.nih.gov/snomedctBrowser.html>, accessed July 19, 2012

## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

the “meets” and “does not meet” category for each criterion.

One of the most obvious trends that can be gleaned from the annotations in Appendix E is that for the diabetes and recent cardiac event criteria, positive instances of a disease (that is, instances where the person meets the criterion being examined) generally don’t have modifiers, while those that do not meet the criterion do. This trend is reversed with the criterion pertaining to not having a history of cardiac events. Additionally, documents that do not meet a criterion often have no text related to that criterion, a fact that is relayed by the number of times a non-consuming tag was used to indicate that a criterion was not met.

The rule that, generally, a criterion is met if there is a related phrase but not met if there is no relevant text can easily be applied to each of the diagnosis-related criteria. For each criterion a list of relevant phrases was created, based on the annotated extents in the gold standard corpus. Each text file was searched for the list of keywords, and classified as “meets” or “does not meet” based on the rules outlined above. The results of this rule and keyword combination are discussed below.

**Diabetes:** The list of diabetes-related keywords generated from the gold standard as as follows: *diabetes, diabetic, dm, dm1, dm2, dmi, dmii, t2dm, iddm, glargine, humalog, glyburide, metformin, lantus, glucophage*. A Python script to perform a regular expression search, looking for places where these expressions appear only as stand-alone tokens (so ‘dm’ will not count if it is part of another word) and ignoring the case of the expressions searched. Files that contained any of the diabetes-related words were counted as meeting the criterion, and those that did not include any diabetes-related words were counted as not meeting the criterion. The lists of files were then compared to the list of files from the gold standard that were classified as

meeting and not meeting the diabetes criterion. The results for this test are shown in Table 8.6.

	gold standard/script	true pos.	false pos.	false neg.
<b>Meets</b>	60/69	59	10	1
<b>Does not meet</b>	40/31	30	1	10

Table 8.6: Keyword-based ‘diabetes’ classifications compared to gold standard

While there was a tendency for this method to over-label patients as meeting the diabetes criterion, the overall accuracy for the ‘diabetes’ criterion is .89, a number that compares favorably to the classifier results shown in Table 8.5, where the average accuracy across all the classifiers for the diabetes criterion was .663, and the best-performing classifier had an average accuracy of .696.

**Recent cardiac event:** Similar to the ‘diabetes’ criterion, a determination for the ‘recent cardiac event’ criterion was also based on whether or not a file contained any of a list of cardiac event-related keywords: *stemi*, *myocardial infarction*, *congestive heart failure*, *chf*, *nstemi*, *mi*, *coronary artery disease*, *artery disease*, *cad*, *heart attack*, *imi*, *heart failure*, *vessel disease*. The results of this keyword and rule-based classification are shown in Table 8.7.

	gold standard/script	true pos.	false pos.	false neg.
<b>Meets</b>	38/78	38	40	0
<b>Does not meet</b>	62/22	22	0	40

Table 8.7: Keyword-based ‘recent cardiac event’ classifications compared to gold standard

Similar to the ‘diabetes’ criterion, there was a tendency here for files to be over-classified as meeting the criterion compared to not meeting it. Overall accuracy for

## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

the ‘recent cardiac event’ criterion using this method is .60. This compares well to the classifiers, which overall performed slightly better, with an average accuracy of 0.635 and the highest-performing classifier obtaining an average accuracy of .701.

**No history of cardiac events:** This criterion used the same keyword list as the ‘recent cardiac event’ criterion, but the rule used for classifying documents was flipped: if a document contained one of the keywords, it was put into the ‘does not meet’ category, and if it did not contain any of the keywords it was classified as ‘meets’. The results of this system are shown in Table 8.8.

	gold standard/script	true pos.	false pos.	false neg.
<b>Meets</b>	33/22	20	2	13
<b>Does not meet</b>	67/78	65	13	2

Table 8.8: Keyword-based ‘no history of cardiac event’ classifications compared to gold standard

This method obtained an overall accuracy of .85, which is significantly better than both the classifiers’ average performance of .678, and the highest-performing classifier’s average of .748.

Overall, simply checking for keywords found in the annotations provided fairly good results for an initial attempt at document classification according to each diagnostic criterion. Two of the three criteria out-performed the baseline performance metric created by the classifier systems, and the third was not far off from matching the classifiers’ average performance.

Naturally, these performance results cannot necessarily be extrapolated to new sets of documents, but the purpose of this chapter is to explore some of the ways

that light annotations tasks can be leveraged into NLP systems for processing clinical data, not to build a fully-functioning system for identifying patients with particular diagnoses.

## 8.4 Using document structure for analyzing the PERMIT corpus

In order to further improve the results of the test-system that was discussed in the previous system, the next step was to examine the modifiers that the gold standard identifies as being important for correctly interpreting the documents.

In many cases, these modifiers are used to indicate that the document is describing the patient's family or medical history, rather than current events. Section 6.3.3 discussed how patient discharge summaries often use section headers to distinguish different parts of the patient's medical record, including identifying sections about the patient's history and his or her family's medical history as well. By determining what section of the document an identified keyword is in, it may be possible to further improve the accuracy of the keyword-based system without resorting to fully syntactic and/or dependency parsing.

### 8.4.1 Section headers and narrative containers

Using section headers as a way to gather more information about a clinical text has been proven an effective method of analyzing clinical documents (Mowery et al., 2009), and the SecTag system was built to utilize this type of information (Denny et al., 2008). While the SecTag system is not freely available, the database of section

## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

headers and relational concepts can be downloaded from Vanderbilt University<sup>3</sup>.

The SecTag database and ontology provides an excellent resource for identifying section headers and classifying them into types, but the database does not include information most relevant to the modifiers that were annotated in the PERMIT corpus, such as whether or not each of the section headers refers to the patient, and when the events described in that section likely took place.

The idea of determining when an event was likely to have taken place comes from recent research into the TimeBank corpus examining temporal constraints on events. Recent analysis of news texts suggests that temporal constraints on reported-on events may be inferred by their source and are inherently understood by readers (Pustejovsky and Stubbs, 2011). This constraint, called a “narrative container”, supplies temporal structure to news reports by limiting the time frame in which an unanchored event (that is, an event mentioned in the past tense without any temporal modifiers) can take place.

For example, an article from a daily newspaper that simply reports “The White House said...” without indicating when the ‘said’ even took place will still be interpreted by most readers as having taken place within the past day, due to the assumptions that can be made about the time frame in which an article for such a publication is written. The one-day window is the narrative container, which provides constraints regarding when an event can be assumed to have taken place. The left-most constraint (the one farthest in the past) in this example is 24 hours before the news story was published, and the right-most constraint (the one closest to the time the article is being read) is the time just prior to the publication of the

---

<sup>3</sup><http://knowledgemap.mc.vanderbilt.edu/research/content/sectag-tagging-clinical-note-section-headers>

## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

newspaper.

This concept of “narrative containers” can be applied in a general way to the section headers in discharge summaries in order to help determine when the discussed events (including diagnoses) are most likely to have taken place.

The SecTag database does not inherently contain information about section headers, or whether each section pertains to the patient, his or her family, or some other aspect of a discharge summary such as hospital-related metadata (for example, the attending physician or room number). The addition of this information to the SecTag database is described in the next section.

### **Augmentations to the SecTag database**

In order to take advantage of the information in section headers with regards to narrative containers and subject, the SecTag database needed to be augmented with some additional information about the section headers it contained. This was done by creating an additional table with information about the top-level concepts in the SecTag ontology.

Specifically, each of the section headers was given one of three labels that described its type/function in the document. These three labels were:

**Metadata** - This is information that is important to understanding the record, such as the admission date, discharge date, and the patient’s date of birth, as well as information that is meant for the hospital but is not important for this annotation task, such as the names of the doctors, who created the file, etc.

**Document Substructure Component** - These provide context for what the fol-



## CHAPTER 8. USING EXPERT ANNOTATION IN AN NLP SYSTEM

lowing block of text refers to in the discharge summary. These are phrases such as “Allergies”, “Past Medical History”, “Discharge Disposition”, etc. The substructure components are especially important because they often provide temporal context, and are the most directly relevant to the PERMIT corpus.

**Component Topics** - these are sections within the Substructure Components that provide topical information. For example, in the Substructure Component “Hospital Course”, the description of actions may be divided by each medical problem that was addressed while the patient was in the hospital, such as “Diabetes”, “Cardio”, etc. Knowing the general topic of the text segment can be useful when trying to disambiguate an abbreviation, though this information is generally more detailed than the PERMIT corpus currently requires.

In addition to the labels describing the function of the section header data in the document, each top-level concept was assigned given an “about” marker, used to indicate whether the information referred to the hospital, patient, family, or some combination of those, as well as the likely start and end date for that concept’s narrative container. The narrative container start and end times were generally limited to fairly broad times, such as the patient’s date of birth, the hospital admission date, and the hospital discharge date. It should be noted that these additions to the SecTag database are meant to represent a quick, general way to garner additional information from identified section headers.

### 8.4.2 Results of keywords + section header processing

Making use of the section header terminology was done in a two-step process. First, potential section headers in each file were identified with a regular expression search. Each identified segment was then matched against the SecTag database to see if a match could be found so that the identified string could be connected to a top-level concept and provided with the extra information, including level, narrative container information, and who the section was most likely about.

Once the section headers in each document were identified, the text of the documents was searched for the keywords identified in the previous section. If a keyword was found, the document was checked to see what type of section it was in, and whether the topic was the patient or his or her family, as well as whether the narrative container might have information relevant to the criterion being examined. The results of this experiment per criterion are described below.

**Diabetes:** When augmenting the keyword list with section header information for this criterion, only the header information that determined if the section was about the patient was used. Narrative container information was not relevant, as diabetes is generally an ongoing state. The results for this analysis are shown in Table 8.9.

	gold standard/script	true pos.	false pos.	false neg.
<b>Meets</b>	60/65	59	6	1
<b>Does not meet</b>	40/35	34	1	6

Table 8.9: Keyword- and section-based ‘diabetes’ classifications compared to gold standard

The accuracy for this criterion increased from .89 to .93, a moderate improvement for performance here, but still significantly better than the classifier baseline shown

in Table 8.5.

**Recent cardiac event:** Because this criterion has a temporal component, header information about whether the section pertained to the patient was used in the analysis, as well as information about the narrative container. Patients were only placed in the “meets” category if the section a match was found in had a narrative container start time of the hospital admission date. Table 8.10 shows the results of this analysis.

	gold standard/script	true pos.	false pos.	false neg.
<b>Meets</b>	38/53	32	21	6
<b>Does not meet</b>	62/47	41	6	21

Table 8.10: Keyword-based ‘recent cardiac event’ classifications compared to gold standard

Accurate classifications using this method rose to .73, a significant increase from the keyword-only score of .60, which also raises the accuracy higher than both the average classifier score (0.635) and the highest-performing classifier (.701).

**No history of cardiac events:** Like the ‘diabetes’ criterion, this criterion was evaluated using only the header information that determined whether the relevant section was about the patient. As any cardiac event that the patient experience would remove them from meeting this criteria, no matter when it occurred, narrative container information was not a relevant factor in this analysis. Table 8.11 shows the results of adding header information to the ‘no history’ criterion.

This method obtained an overall accuracy of .89—not a large increase from using only keywords (.85), but still much better than both the classifiers’ average performance of .678, and the highest-performing classifier’s average of .748.

	gold standard/script	true pos.	false pos.	false neg.
<b>Meets</b>	33/28	25	3	8
<b>Does not meet</b>	67/72	64	8	3

Table 8.11: Keyword-based ‘no history of cardiac event’ classifications compared to gold standard

## 8.5 Additional analyses

At this point in the processing of the PERMIT corpus, it would be necessary to augment the light annotations with more detailed or deep annotations, such as syntactic parsing, dependency trees, temporal processing, etc. Appendix G provides an overview of some of the existing systems used for analyzing clinical and temporal data. While some of these systems are not available for public use, they provide insight into the types of analysis that could be done on the PERMIT corpus.

Based on the modifiers used in the light annotation, the next steps for an NLP system based on the PERMIT corpus would be to address the contextual negations (“ruled out for”) and the time-sensitive aspects of the cardiac event criteria. Systems such as cTAKES (Savova et al., 2010), ConText (Chapman et al., 2007; Harkema et al., 2009), the TARSQI Tool Kit (Verhagen and Pustejovsky, 2008; Verhagen and Pustejovsky, 2012), and TimeText (Zhou et al., 2007) would be the most likely candidates to be brought into an NLP system for the corpus being analyzed here, but this dissertation does not seek to build an entire automated analysis system around the light case-control annotation task.

## 8.6 Summary

This chapter provides an introductory approach to ways that a light annotation can be leveraged in an NLP system by using keywords and document structure as starting points for analysis, and discusses how further analysis could be done by taking advantage of existing systems, even without acquiring a deeper annotation.

The light annotation for the case-control ‘study’ provided a platform from which the clinical documents could be analyzed and evaluated, and that it also encoded information about relevant phrases that may not have been identified if the annotators had not been domain experts and had instead relied on medical dictionaries such as SNOMED-CT. While this dissertation does not seek to build a full NLP system that would mimic the results of the annotation task, the foundation for such a system are embedded in the light annotation model.

# Chapter 9

## Conclusions and Future Work

### 9.1 Conclusions

This dissertation examines the problem of capturing professional (specifically, non-linguistic) knowledge in annotated corpora, and proposes a solution to this problem in the form of *light annotation tasks*: linguistically under-specified, task- and domain-specific annotation models that can be used to quickly capture expert knowledge in a corpus as it relates to a research question. The resulting annotation is one that can become a layer in a more detailed annotation model, or directly into an NLP system. The findings and contributions of this dissertation are summarized below.

**Evaluation of established standards:** A thorough overview of the established standards and general desiderata for annotation tasks, as they have been described in the corpus and computational linguistic communities is provided in this dissertation. Chapter 2 examines these standards, which include corpus selection, representation of annotated data, annotation guidelines, reporting on annotation tasks, and require-

## CHAPTER 9. CONCLUSIONS AND FUTURE WORK

ments for annotation software. Chapter 3 describes the MATTER cycle, the first general methodology for all types of annotation and machine learning tasks. By pulling together the disparate desiderata of the different factors that influence annotation tasks, this dissertation not only builds a solid base on which to ground light annotation tasks, but also produces a resource that other researchers involved in computational linguistics may find valuable when designing their own annotation tasks.

**Definition and principles of light annotation:** Representing domain expert knowledge in an annotation task for later use in an NLP system is a difficult task, due to the different research interests and styles of, for example, medical doctors and computational linguistics. While the concept of a light annotation task is not new, this dissertation provides the first definition of what a light annotation task is, and how such a task fits into the established annotation standards. Additionally, until this dissertation no research had been done into what makes a good light annotation task, particularly with regards to domain-specific research. Chapter 4 provides a thorough analysis of light, domain expert annotations that have been performed in the past, and uses that information to establish a set of principles that can be used to create effective light annotation tasks.

**Software for light annotation:** Because light annotation tasks are designed to be performed by domain experts rather than linguistic researchers, it is important that the annotation environment be easy to install, set up, and use. In order to provide such an environment for domain experts, this dissertation also presents MAE and MAI, annotation and adjudication software designed for light annotation tasks. This software, described in Chapter 5, was not only used in the light annotation case study

## CHAPTER 9. CONCLUSIONS AND FUTURE WORK

presented in Part II of this dissertation, it has also been used by other annotation tasks (not all of them light), including spatial and temporal annotations, and annotations in a variety of other languages.

**A case study for clinical light annotation:** Part II of this dissertation presents a case study examining how the light annotation methodology and principles were applied to the clinical domain in the form of a mock retrospective case-control study. The resulting annotation, in the form of the PERMIT corpus, was both easier and faster to create than other, more fully specified annotations that have been performed over clinical data. Additionally, the light annotation provided a solid basis for preliminary experiments in automating the selection of patients for the mock study.

In sum, this dissertation establishes a standard for the newly-defined *light annotation task*, which is grounded on established standards in the corpus and computational linguistic communities. In addition to providing software that is designed for use with light annotation tasks, this dissertation also describes and analyzes a case study that applies the concept of light annotation to the clinical domain, and shows preliminary research into leveraging that annotation into an NLP system.

### 9.2 Future Work

The light annotation methodology provides many different areas in which further research can be done. Some of these research areas are described in this section.

**Application to other annotation approaches:** Section 4.1 described some of



## CHAPTER 9. CONCLUSIONS AND FUTURE WORK

the different approaches to bioclinical annotations that have been used for other annotation tasks. While some of these have already been incorporated into the light annotation methodology (such as text-bound annotation (Kim et al., 2008)) others have not. A light annotation task could, for example, be paired effectively with an active learning (Settles, 2010) or accelerated annotation framework such as the one used by Tsuruoka et al. (2008).

**Application to other domains:** The case study in this dissertation focuses on the bioclinical domain, but the light annotation principles and methodology described here could certainly be applied to any annotation task requiring domain expert knowledge, such as evaluations of legal documents, math or computer science papers, and so on. Future research regarding light annotation tasks should examine any factors that might influence the creation of these tasks in other domains.

**Application to other modalities:** The research presented here focuses on text annotation, but applying the light annotation principles to other modalities, such as speech or video, could also provide an interesting platform for future research. Expert analysis of games such as Go or chess, or even sporting events, could provide valuable insight to laypeople as to what makes good moves or plays, or what separates a perfect gymnastic routine from one that is subtly flawed.

**Further applications in bioclinical research:** The PERMIT corpus annotation task was designed around a retrospective case-control study, and used the finding of qualified patients as a motivation for the annotation and preliminary NLP research. While the main theoretical contribution of this dissertation is the description of the

## CHAPTER 9. CONCLUSIONS AND FUTURE WORK

light annotation principles and methodology, the case study presented does provide a potential avenue for the use of this methodology in actual clinical research. However, the uses for light clinical annotations are not limited to retrospective case-control studies, or even to patient selection in general. Other possible uses for light annotation in the clinical domain should be explored.

**More specific principles:** The purpose of this dissertation was to provide a methodology and set of principles that could be applied to any type of light annotation task, but research into whether more specific guidelines should be developed for different domains, modalities, or research goals is another potentially useful avenue for future research.

**Addition of other annotation layers:** While a defining principle of light annotation tasks is that they can (and most likely will) be augmented with additional layers of annotation, the specific process of turning the light annotation Model,  $M_1$  into the full model ready for the MATTER cycle,  $M$ , can be further investigated. This process was briefly discussed in Section 4.4, and Chapter 8 provides a beginning point for how the process can be done. However, the particulars of how the  $M_1$  created for a specific task and/or domain will affect the creation of a fuller  $M$  is an area that should be explored in order to even more fully exploit the uses of light annotation tasks.

This list of possible future research is not exhaustive, but it does show that there is still more to be learned about the application of light annotation tasks, and how they can be used in to aide research in a variety of different disciplines and modalities.

## Part III

# Appendices and References

# Appendix A

## Eligibility criteria analysis

This appendix provides an overview of temporal expressions in clinical trial eligibility criteria, based on the trials listed on `clinicaltrials.gov` on October 5, 2010. Overall, there were 3,960 studies containing the terms “retrospective” or “case-control” and 96,673 studies in total.

The ‘#’ in the left-most column stands for a match with a number, or any of these words: *a, an, the, one, two, three, four, five, many, few*. The ‘\_\_\_\_\_’ represents *hour, day, week, month, year*, with the frequency of each represented in that cell of the table. The last cell of the table shows the frequency of words commonly associated with temporal ordering.

The “Retrospective and case-control” column in the table is not meant to represent all retrospective and case-control studies in the `clinicaltrials.gov` database: a full analysis of all 96,673 studies was not performed to determine which ones were, in fact, retrospective and/or case-control studies. The column is present simply to provide some perspective on potential comparisons, not to make thoroughly researched statements on disparities on study types.

APPENDIX A. ELIGIBILITY CRITERIA ANALYSIS

Terms	Retrospective and case-control	All studies
within/for (the past) # -----	636	80,483
hour(s)	76	4,226
days(s)	141	27,145
week(s)	74	16,581
month(s)	245	24,849
year(s)	100	7,681
# ----- ago/before	154	13,562
hour(s)	13	856
days(s)	21	4,231
weeks(s)	23	3,683
month(s)	52	3,704
year(s)	45	1,090
after # -----	22	558
hour(s)	4	47
day(s)	2	81
week(s)	7	213
month(s)	5	126
year(s)	4	91
# ----- (any time) in the past	6	103
hour(s)	0	1
day(s)	1	29
week(s)	1	25
month(s)	4	44
year(s)	0	4
at least # -----	454	26,607
hour(s)	11	811
day(s)	23	3,070
week(s)	56	6,523
Continued on next page		

APPENDIX A. ELIGIBILITY CRITERIA ANALYSIS

Table A.1 – continued from previous page

Terms	Retrospective/case-control	All studies
month(s)	177	9,951
year(s)	187	6,252
between # and # -----	92	2,806
hour(s)	0	30
day(s)	2	80
week(s)	8	130
month(s)	1	67
year(s)	81	2,499
between # ----- and # -----	7	177
hour(s) and hour(s)	0	4
hour(s) and day(s)	0	2
hour(s) and week(s)?	0	1
hour(s) and month(s)	0	3
hour(s) and year(s)	0	0
day(s) and day(s)	0	5
day(s) and week(s)	0	1
day(s) and month(s)	0	3
day(s) and year(s)	1	5
week(s) and week(s)	1	5
week(s) and month(s)	0	11
week(s) and year(s)	0	3
month(s) and month(s)	1	14
month(s) and year(s)	0	59
year(s) and year(s)?	4	54
other	0	6
# -----	4593	259,569
hour(s)	245	10,972
Continued on next page		

APPENDIX A. ELIGIBILITY CRITERIA ANALYSIS

Table A.1 – continued from previous page

<b>Terms</b>	<b>Retrospective/case-control</b>	<b>All studies</b>
day(s)	381	49,186
week(s)	489	52,939
month(s)	1094	73,306
year(s)	2384	73,165
previously	147	6,845
subsequently	10	225
after	473	24,036
before	359	19,899
history of	868	47,164

# Appendix B

## Case-Control Annotation DTD

```
<!ENTITY name "CCml">

<!ELEMENT Selection_criteria ( #PCDATA ) >
<!ATTLIST Selection_criteria id ID prefix="SC" >
<!ATTLIST Selection_criteria start #IMPLIED >
<!ATTLIST Selection_criteria criterion
    ( age | diabetic | recent card. event |
      no card. event ) >
<!ATTLIST Selection_criteria meets
    ( MEETS | DOES NOT MEET ) >
<!ATTLIST Selection_criteria comment CDATA >

<!ELEMENT Matching_criteria ( #PCDATA ) >
<!ATTLIST Matching_criteria id ID prefix="MC" >
<!ATTLIST Matching_criteria start #IMPLIED >
```



*APPENDIX B. CASE-CONTROL ANNOTATION DTD*

```
<!ATTLIST Matching_criteria criterion
  ( sex | race | recent lipid test |
    diabetic treatment | lipid medications ) >
<!ATTLIST Matching_criteria present
  ( PRESENT | NOT PRESENT ) >
<!ATTLIST Matching_criteria comment CDATA >

<!ELEMENT Modifier ( #PCDATA ) >
<!ATTLIST Modifier id ID prefix="M" >
<!ATTLIST Modifier comment CDATA >

<!ELEMENT Modifies #EMPTY >
<!ATTLIST Modifies id ID prefix = "ML" >
<!ATTLIST Modifies comment CDATA >

<!-- The Overall tag was used only during adjudication -->

<!ELEMENT Overall ( #PCDATA ) >
<!ATTLIST Overall start #IMPLIED >
<!ATTLIST Overall age ( MEETS | DOES NOT MEET ) >
<!ATTLIST Overall diabetic ( MEETS | DOES NOT MEET ) >
<!ATTLIST Overall recent_card ( MEETS | DOES NOT MEET ) >
<!ATTLIST Overall no_hist ( MEETS | DOES NOT MEET ) >
<!ATTLIST Overall determination ( CASE | CONTROL | NEITHER) >
<!ATTLIST Overall comment CDATA >
```

# Appendix C

## Case-control Annotation Guidelines

The guidelines provided below are the instructions that were given to the annotators of the Case-Control annotation task. They were modified only for formatting to fit into this dissertation.

### C.1 Overview

Annotation for the case-control task (CCT) will be done in MAE. Before beginning this task, please read the user guide included in the documents folder that MAE comes with, and familiarize yourself with the controls. There is a sample annotation task in the samples directory if you would like to experiment with that data first.

For the CCT, we will be examining two types of information: selection criteria, and matching criteria. Selection criteria are used to determine if a person is eligible for a study, and for this task are composed of the following:

## APPENDIX C. CASE-CONTROL ANNOTATION GUIDELINES

**General criteria 1:** must be under 55 years old at time of admission

**General criteria 2:** must have diabetes

**Case criteria 1:** must have had a cardiac event within 2 years of admission date

**Control criteria 1:** no history of cardiac events

The matching criteria are used to determine which patients share similar characteristics in order to make data analysis as accurate as possible. For this task, the matching criteria are:

**Matching Criteria 1:** race

**Matching Criteria 2:** sex

**Matching Criteria 3:** lipid measurement w/in 6 months of admission

**Matching Criteria 4:** information on diabetic treatment

**Matching Criteria 5:** lipid medications

In general, when selecting patients for eligibility in a study, it would be the case that once a criterion is met, the person doing the assessing would move on to the next one. However, the purpose of this task is to create a set of annotations that can be used to determine which mentions of events, drugs, etc would indicate meeting a criteria, and which mentions would not. Therefore, it is important that all mentions of items that could be used to assess a criterion be annotated.

### C.2 Annotation

We suggest that you go through each document at least twice using a keyword-based strategy. For example, look for words that describe a cardiac event, and tag them

## APPENDIX C. CASE-CONTROL ANNOTATION GUIDELINES

as a selection\_criteria with type “recent cardiac event”. Then, check the context around the words to determine whether this cardiac event meets the requirements for inclusion in the study. Words providing context should be marked as “modifier”, and linked to the text of the event with the “modifies” tag.

In most cases, modifiers will be words that indicate: whether or not an event actually happened (ex: “admitted for” versus “ruled out for”); whether or not the event happened to the patient (ex: if a criterion is mentioned in relation to a family member); dates that indicate if the event happened within the timeframe (when there is a timeframe specified).

Section headers should not be marked as modifiers, but any other information should be annotated. Not all annotated extents will have modifiers.

### C.2.1 Notes

- It is fine to extrapolate from the text whether a criterion is met or not. For example, if it is not specifically mentioned that a patient has diabetes, but it is mentioned that they are taking insulin daily, you can annotate the insulin as both a selection criteria and a matching criteria.
- The ‘Meets’/‘does not meet’ (or ‘present’/‘not present’) attribute should be set for each annotated extent based on the context *for that event*. Each extent should be evaluated individually as to whether that particular mention meets stated requirements. You are not, for this task, evaluating the person as a whole.
- Please note that if a criterion is met by certain things *not* being mentioned (for example, no history of cardiac events), then this should be annotated by

### *APPENDIX C. CASE-CONTROL ANNOTATION GUIDELINES*

creating a non-consuming tag and setting the appropriate “meets” or “does not meet” flag. For each file there should be at least one tag per criteria. If a person meets the criteria for having a recent cardiac event, then there should also be a tag in their file showing that they do *\*not\** meet the criteria for no cardiac events. Non-consuming tags are created from the “NC elements” option in the top menu bar of MAE.

- It is also possible to not meet either of the cardiac event criteria—for example, if a person’s only cardiac event occurred over 2 years ago. In this case, the cardiac event in question should be annotated along with the relevant date as “does not meet”, and a non-consuming tag for “no history of cardiac events” should also be created and labeled as “does not meet” as well.
- As previously mentioned, any information related to any of the criteria should be annotated. However, for the matching criteria pertaining to sex, it is not necessary to annotate every use of “he” or “she” to refer to the patient.

# Appendix D

## Inter-annotator agreement table

Table D.2 is the confusion matrix for the PERMIT corpus annotation, which was generated by checking each whitespace-separated token in the corpus against the annotations of both annotators, and entering them into the table based on what tags each annotator used for that token. Because many multi-word extents were annotated by each annotator, most tags are counted more than once. This table is not used for inter-annotator agreement calculations (such as Cohen's kappa), but it does provide a general overview of what parts of the annotation the annotators were most likely to disagree on. Table D.1 shows the abbreviations used for the tags in the confusion matrix.

APPENDIX D. INTER-ANNOTATOR AGREEMENT TABLE

Abbreviation	tag and attribute
S_c age	Selection_criteria: age
S_c dia	Selection_criteria: diabetic
S_c r_c	Selection_criteria: recent cardiac event
S_c n_h	Selection_criteria: no history of cardiac events
M_c sex	Matching_criteria: sex
M_c race	Matching_criteria: race
M_c l_t	Matching_criteria: recent lipid test
M_c d_t	Matching_criteria: diabetic treatment
M_c l_m	Matching_criteria: lipid medications
Mod.	Modifier
S_c&	Selection_criteria (no criterion given)
M_c&	Matching_criteria (no criterion given)
None	the annotator did not use a tag

Table D.1: Abbreviations for tag and attribute combinations used in the confusion matrix table

APPENDIX D. INTER-ANNOTATOR AGREEMENT TABLE

Annotator 1	Annotator 2													
	S_c age	S_c dia.	S_c r_c	S_c n_h	M_c sex	M_c race	M_c l_t	M_c d_t	M_c l_m	Mod.	S_c&	M_c&	None	
S_c&age	265	0	0	0	0	0	0	0	0	0	0	0	111	
S_c&diabetic	0	225	1	2	0	0	0	5	0	1	1	0	57	
S_c&recent card	0	3	223	168	0	0	0	0	0	0	0	0	608	
S_c&no hist.	0	0	6	0	0	0	0	0	0	0	0	0	5	
M_c&sex	2	0	0	0	210	0	0	1	0	0	0	0	21	
M_c&race	0	0	0	0	3	7	0	0	0	0	0	0	3	
M_c&lipid test	0	0	0	0	0	0	70	0	1	0	0	1	5	
M_c&diab. treat	0	3	0	0	0	0	0	133	0	4	0	0	107	
M_c&lipid meds	0	2	0	0	0	0	1	0	109	0	0	1	26	
Modifier	0	98	6	1	0	0	0	10	0	218	0	0	455	
S_c&	2	1	2	0	0	0	0	0	0	0	0	0	1	
M_c&	0	1	0	0	0	0	0	2	1	0	0	0	1	
None	37	192	249	249	116	1	26	78	21	400	0	0	140368	

Table D.2: Confusion matrix annotations



# Appendix E

## Selection criterion extent analysis

This appendix provides all of the extents that were included in the gold standard of the PERMIT corpus selection criterion tag. The annotations for the age criteria (being under 55 years old at the time of admission) are not included because they were simply annotations of dates of birth and phrases such as “44 year old”, and therefore do not need to be reproduced here.

For each table, the extents and any accompanying modifiers were all turned into lower case, and are ordered based on how frequently they appeared in the gold standard. These tables give a complete overview of what phrases were identified as relevant text for each criterion, as well as which modifiers were used to determine whether a criterion was met or not met. Lines in each table that are blank represent the times that no text in the document was related to a criterion, and therefore a non-consuming tag was created to reflect that lack of relevant information.

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

<b>Diabetes: meets</b>	<b>Count</b>
insulin	93
diabetes	31
dm	21
diabetes mellitus	21
glargine	14
humalog	13
dka	13
dmii	12
dm2	11
glyburide	10
diabetic	9
metformin	7
diabetic ketoacidosis	7
dm ii	6
type 2 diabetes mellitus	5
type ii diabetes mellitus	4
type 1 diabetes	4
glucotrol	4
type ii diabetes	3
t2dm	3
lispro	3
iddm	3
glipizide	3
dm type 1	3
diabetes mellitus type ii	3
type i diabetes mellitus	2
type 2 dm	2
type 1 dm	2

Continued on next page

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.1 – continued from previous page

Diabetes: meets	Count
lantus	2
glucophage	2
dm type i	2
diabetic neuropathy	2
diabetes type 2	2
diabetes type 1	2
diabetes mellitus type 2	2
diabetes mellitus type ii	2
diabetes mellitus	2
type ii dm	1
type i dm	1
type i diabetic	1
type 2 diabetes	1
type ii diabetes mellitus	1
traglitazone	1
rosiglitazone	1
nph	1
non-insulin-dependent diabetes mellitus	1
non-insulin dependent diabetic	1
niddm	1
medical history + dm	1
insulin-dependent + insulin-dependent + diabetes mellitus	1
insulin-dependent + insulin-dependent + diabetes	1
insulin sliding scale	1
insulin lispro	1
insulin	1
humulin	1

Continued on next page

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.1 – continued from previous page

Diabetes: meets	Count
humilog	1
h/o + dm	1
glipizide	1
dm2	1
dm-ii	1
dm ii	1
diabetes, type ii	1
diabetes type ii	1
diabetes type i	1
diabetes type i	1
diabetes ketoacidosis	1
avandia	1
actos	1
insulin dependent diabetes mellitus	1
insulin	1

Table E.1: Modifiers and extents used to identify patients who met the diabetes criterion

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Diabetes: does not meet	Count
	38
mother + diabetes	3
mother + dm	2
diabetes mellitus	2
diabetes	2
specialists + diabetes	1
son and daughter have + dm	1
sisters + dm	1
relatives + not related to + diabetes	1
paternal + diabetes mellitus	1
no other hx + dm	1
no history of + diabetes mellitus	1
no h/o + dm	1
no history of + dm	1
grandmother + diabetes	1
gm + gm + type i diabetes	1
father + father w + dmii	1
family history of + diabetes	1
family history + diabetes	1
early signs of + early signs of + diabetes	1
determined not to be + diabetic	1
cousin + type ii dm	1
borderline + diabetes mellitus	1
at risk for developing + at risk for + diabetes	1

Table E.2: Modifiers and extents used to identify patients who did not meet the diabetes criterion

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Recent cardiac event: meets	Count
STEMI	12
myocardial infarction	10
congestive heart failure	10
NSTEMI	7
MI	7
coronary artery disease	5
CHF	5
heart attack	3
anterior myocardial infarction	3
single vessel coronary artery disease	2
one vessel coronary artery disease	2
NSTEMI	2
non ST segment elevation myocardial infarction	2
found to have + 3 VD	2
3 vessel disease	2
this year + April of this year + myocardial infarction	1
status post + coronary artery disease	1
status post + anterior myocardial infarction	1
ST-segment elevation myocardial infarction	1
ST segment elevations myocardial infarction	1
ST elevations myocardial infarction	1
ST elevation myocardial infarction	1
ST elevation inferior myocardial infarction	1
showed + 3 vessel coronary artery disease	1
should be repeated in 4 weeks post + MI	1
ruled out for MI + inferior MI	1
ruled in for + ruled in for + NSTEMI	1
ruled in for + myocardial infarction	1

Continued on next page

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.3 – continued from previous page

Recent cardiac event: meets	Count
ruled in + mi	1
revealed + three vessel coronary artery disease	1
revealed + single vessel disease	1
revealed + 3 vessel disease	1
presents with + stemi	1
premature coronary artery disease	1
one vessel non-flow limiting disease	1
non-stemi	1
non-st-elevation mi/cad/	1
non-st elevation myocardial infarction	1
non st elevation myocardial infarction	1
non st elevation mi	1
most recent about 2 years ago + most recent about 2 years ago + mi	1
may result in another + heart attack	1
left heart failure	1
inferior mi	1
infarct	1
imi	1
heart failure	1
he was felt to be in + congestive heart failure	1
ekg changes consistent with + anteroseptal infarction	1
ekg and cardiac enzymes were consistent with + consistent with + imi	1
decompensated heart failure	1
coronary artery with diffuse disease	1
coronary artery disease	1
cad	1
apical infarction	1

Continued on next page

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.3 – continued from previous page

Recent cardiac event: meets	Count
anterior/ inferior mi	1
anterior st elevation myocardial infarction	1
admitted with + admitted with + imi	1
admitted to cardiology with + nstemi	1
2016-01-21 + coronary artery disease	1
2015-09-07 + two vessel disease	1
2015-08-21 + shortly after + nstemi	1
2015-05-26 + myocardial infarction	1
2015-05-26 + coronary artery disease	1
2015-05-26 + 2015-05-26 + myocardial infarction	1
2013-06-25 + cad	1
2012-07-06 + myocardial infarction	1
2012-07-06 + coronary artery disease	1
2012-05-06 + 2012-05-06 + myocardial infarction	1
2009-11-13 + st segment elevation myocardial infarction	1
2009-10-09 + nstemi	1
2005 + nstemi	1
12-01 + 2 vessel disease	1
08-22 + nstemi	1
07-29 + coronary artery disease	1
07-29 + cad	1
05-28 + status post + mi	1
coronary artery with diffuse disease	1

Table E.3: Modifiers and extents used to identify patients who met the recent cardiac event criterion



APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Recent card. event: does not meet	Count
	33
father + mi	6
rule out + myocardial infarction	5
father + myocardial infarction	5
ruled out for + myocardial infarction	4
ruled out + myocardial infarction	3
chf	3
ruled out for + mi	2
mother + mi	2
coronary artery disease	2
without angiographic evidence of + coronary artery disease	1
without angiographic evidence of + coronary artery disease	1
without angiographic evidence of + coronary artery disease	1
without angiographic evidence + coronary artery disease	1
son + myocardial infarction	1
sister + sister + dm	1
sister + mi	1
sister + congestive heart failure	1
sister + cad	1
ruled out for + patient ruled out + mi	1
ruled out for + negative cardiac enzymes + mi	1
ruled out for + mi	1
ruled out for + acute myocardial infarction	1
ruled out + heart attack	1
rule out + mi	1
rule out + congestive heart failure	1
prior + inferior myocardial infarction	1
presenting with signs and symptoms of + chf	1
Continued on next page	

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.4 – continued from previous page

Recent cardiac event: does not meet	Count
post + 1987 + mi	1
possible + cp/nstemi	1
patient's mother + myocardial infarctions	1
paternal father + mi	1
no signs of + myocardial infarction	1
no history of premature + cad	1
no history of + mi	1
no history of + cad	1
no h/o + cad	1
no family history of + coronary artery disease	1
no family history + coronary artery disease	1
no evidence of any + coronary artery disease	1
no evidence of + chf	1
no evidence + congestive heart failure	1
no family history of + cad	1
never + congestive heart failure	1
negative cardiac enzymes + r/o + mi	1
mother + myocardial infarction	1
mother + heart attack	1
mother + cad	1
mi	1
maternal uncles and aunts + cad	1
however then became apparent that + chf	1
father w/ + cad	1
father + father + chf	1
father + coronary artery disease	1
father + chf	1
Continued on next page	

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.4 – continued from previous page

Recent cardiac event: does not meet	Count
father + cad	1
family history + coronary artery disease	1
family history + cad	1
f + mi	1
doubt the presence of + doubt the presence + chf	1
did not show evidence of + severe heart disease	1
did not reveal + coronary artery disease	1
demonstrated no clinically significant + coronary artery disease	1
coronary arteries are normal	1
concerning for + stemi	1
concern for acute + myocardial infarction	1
concern for + chf	1
cardiac enzymes were negative for + myocardial infarction	1
cad	1
but cannot rule out + myocardial infarction	1
brother + myocardial infarction	1
brother + mi	1
2009 + myocardial infarction	1
2006 + 2006 + myocardial infarction	1
2000 + inferior myocardial infarction	1

Table E.4: Modifiers and extents used to identify patients who did not meet the recent cardiac event criterion

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

<b>No cardiac event history: meets</b>	<b>Count</b>
	36
no + congestive heart failure	2
no prior history + coronary artery disease	1
no previous history of + heart disease	1
no previous + mi	1
no history of + cardiac disease	1
no + myocardial infarction	1

Table E.5: Modifiers and extents used to identify patients who met the no history of cardiac events criterion

<b>No cardiac event history: does not meet</b>	<b>Count</b>
chf	35
coronary artery disease	29
congestive heart failure	24
cad	23
	19
three vessel disease	6
two vessel coronary artery disease	5
nstemi	4
mi	4
history of + chf	4
myocardial infarction	3
history of + coronary artery disease	3
h/o + chf	3
3vd	3
Continued on next page	

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.6 – continued from previous page

No cardiac event history: does not meet	Count
patient has a history of + coronary artery disease	2
hx of + chf	2
history + cad	2
congestive heart failure	2
2 vessel coronary artery disease	2
1vd	2
two-vessel coronary artery disease	1
three-vessel coronary artery disease	1
three vessel native disease	1
single (1) vessel coronary artery disease	1
s/p + mi	1
pvd	1
prior to admission + 3vd	1
prior + myocardial infarction	1
post coronary artery bypass grafting + coronary artery disease	1
post + post + myocardial infarction	1
post + myocardial infarction	1
post + coronary artery disease	1
pmh signif for + cad	1
patient's history of + coronary artery disease	1
patient has a history of + two vessel disease	1
past medical history + three vessel coronary artery disease	1
past medical history + 2003 + coronary artery disease	1
one vessel disease	1
one vessel coronary artery disease	1
omi	1
non-st elevation myocardial infarction	1
Continued on next page	

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.6 – continued from previous page

No cardiac event history: does not meet	Count
multi-vessel cad	1
mi + coronary artery disease	1
medical history + cad	1
long history of + coronary artery disease	1
long history of + congestive heart failure	1
likely + chf	1
known history of + coronary artery disease	1
inf. infarct	1
history of five + myocardial infarctions	1
history of + history of + congestive heart failure	1
history of + congestive heart failure	1
history of + cad	1
history of + coronary artery disease	1
history of + cad	1
history + congestive heart failure	1
history + chf	1
heart attacks	1
heart attack	1
h/o + cad	1
exacerbation + chf	1
dm	1
diastolic heart failure	1
cororary artery disease	1
coronary artery disease	1
consistent with + congestive heart failure	1
class ii + chf	1
cardiac disease	1
Continued on next page	

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

Table E.6 – continued from previous page

No cardiac event history: does not meet	Count
age indeterminate + anteroseptal myocardial infarct	1
6/04 + cad	1
3 vessel disease	1
3 vessel coronary disease	1
3 vessel cad	1
2015-05-26 + coronary artery disease	1
2013-08-19 + coronary artery disease	1
2009 + nstemi	1
2008 + coronary artery disease	1
2008 + ast medical history of + cad	1
2005 + three vessel coronary artery disease	1
2005 + coronary artery disease	1
2003 + coronary artery disease	1
2003 + congestive heart failure	1
2 vessel disease	1
2 vessel disease	1
1997 + mi	1
1997 + coronary artery disease	1
1987 + cad	1
1987 + 1987 + mi	1
08-27 + chf	1
05-28 + coronary artery disease	1
status post + coronary artery disease + 2005 + mi	1
pt did not have any syptoms or signs of + chf	1
past medical history significant for + coronary artery disease	1
Continued on next page	

APPENDIX E. SELECTION CRITERION EXTENT ANALYSIS

**Table E.6 – continued from previous page**

<b>No cardiac event history: does not meet</b>	<b>Count</b>
--	--------------

Table E.6: Modifiers and extents used to identify patients who did not meet the no history of cardiac events criterion

---



# Appendix F

## Sample of augmented SecTag database

Table F.1 shows a sample of the SecTag database augmented with information about the type of concept that each line represents (metadata, document substructure component and component topics), as well as the beginning and ending points of the narrative container associated with that concept and whether the content of the concept's section is most likely to refer to the patient, the patient's family, the hospital itself, etc.

APPENDIX F. SAMPLE OF AUGMENTED SECTAG DATABASE

concept id	concept name	label	NC start	NC end	about
5	general_history_and_physical	compTop	DOB	ADMIT	patient
11	progress_note	compTop	ADMIT	DISCHARGE	patient
20	physician	metaData			staff
63	patient_history	compTop	DOB	ADMIT	patient
64	demographics	metaData	ongoing	ongoing	patient
96	date_of_birth	metaData	timex		DOB
148	principal_procedures	compTop	ADMIT	DISCHARGE	patient
158	history_present_illness	compTop	DOB	ADMIT	patient
315	family_medical_history	subComp	DOB	ADMIT	family
351	hematology_family_history	subComp	DOB	ADMIT	family
352	coagulation_disorders_family_history	subComp	DOB	ADMIT	family
353	cardiovascular_family_history	subComp	DOB	ADMIT	family
375	chronic_illnesses	compTop	DOB	ongoing	patient
435	admission_medications	compTop	ADMIT	ADMIT	patient
437	outpatient_medications	compTop	DISCHARGE	ongoing	patient
603	visual_field_exam	subComp	ADMIT	DISCHARGE	patient
648	cardiovascular_exam	subComp	ADMIT	DISCHARGE	patient
707	neurological_exam	subComp	ADMIT	DISCHARGE	patient
963	discharge_instructions	compTop	DISCHARGE	ongoing	patient
986	health_maintenance_plan	subComp	DISCHARGE	ongoing	patient
1089	author	metaData			records

Table F.1: Sample of the modified SecTag database

# Appendix G

## NLP tools for the bioclinical and temporal domains

While this dissertation does not seek to build a full NLP system for processing clinical data for a mock case-control study, a fundamental principle of light annotation tasks for domain experts is that the any resulting annotations should be in a format that can be augmented by other sources. Naturally, this principle is useless if no other sources for analyzing these types of files exist; fortunately there are a plethora of available systems for parsing all types of texts, including bioclinical text. Because this dissertation focuses on clinical data, and uses as a case study both diagnoses and temporal analysis, this appendix provides an overview of some of the available systems that could be used to augment the case-control annotation task described in Part II. Naturally this is not a comprehensive listing of all the available clinical and temporal tools that have been built, but it does provide a sense of what types of other annotations and analyses could be added to the PERMIT corpus.

## APPENDIX G. NLP TOOLS FOR THE BIOCLINICAL AND TEMPORAL DOMAINS

**AMBIT** - Uses CLEF and myGrid to provide a tool for mining biomedical and clinical text (Gaizauskas et al., 2003; Harkema et al., 2005). AMBIT is a multi-stage processing engine for clinical and biomedical texts, using multiple components including an *information extraction engine* for processing terminology, syntactic and semantic information, and discourse; a *terminology engine* which maps to the UMLS; a *database* of the texts and annotations produced by the other components; an *interface layer* that allows web-based user access; and a *query engine* which provides an interface for users to extract information from the database (Harkema et al., 2005). While AMBIT is not currently available for use outside of the project it was created for, the existence of the program helps illustrate how many approaches are being taken in the processing of clinical documents and some of the successful methods being used in this domain.

**ConText** - ConText is based on the NegEx (Chapman et al., 2001) algorithm for identifying negations in clinical texts. ConText extends the negation detection algorithm to determine whether contextual clue indicate if a conical condition is negated or affirmed, recent, historical or hypothetical, and whether the experiencer is the patient or someone else (Chapman et al., 2007). Both NegEx and ConText rely on regular expressions for assigning labels to conditions, though ConText uses a more extensive list of terms and more intricate rules for scope than NegEx. ConText performs best at identifying the negation and hypothetical attributes for conditions, with more errors in identifying the historical status (Harkema et al., 2009). ConText is currently available for download, and was built to be easily integrated into other applications.

## *APPENDIX G. NLP TOOLS FOR THE BIOCLINICAL AND TEMPORAL DOMAINS*

**cTAKES** - The cTAKES (clinical Text Analysis and Knowledge Extraction System) is an open-source natural language processing system developed for clinical narrative data. It is a “modular system of pipelined components combining rule-based and machine learning techniques” for the purposes of sentence boundary detection, tokenization normalization, part-of-speech tagging, shallow parsing, named entity recognition, and annotation of the status and negation of clinical conditions (which, like ConText, is based on the NegEx algorithm) (Savova et al., 2010). The cTAKES performs well on sentence boundary detection, tokenization, part-of-speech tagging and shallow parsing, with each of those achieving accuracy scores between 0.924 and 0.949. The dataset that the cTAKES was trained and tested on has not been made available, but cTAKES as a whole is available for download.

**ELiXR** - EliXR was developed at Columbia University for the purposes of parsing eligibility criteria and representing the information in them in a standardized way (Weng et al., 2011). EliXR is made up of a pipeline of processes: lexical marking, semantic annotation, dependency parsing, pattern mining, grammar induction, and criteria representation. Specifically, once the researchers had dependency trees for all of the criteria they examined (1,000 different criteria from various studies), they mined the trees for patterns and developed a set of 175 frequent patterns that represent 81% of the test set (Weng et al., 2011). They suggest that the extracted templates can be used to fill the gap between criteria representations and what attributes medical researchers are looking for in their patients, and a preliminary version of this system has been used help identify patients for studies (Li et al., 2008; Botsis et al., 2010).

**Lancet** - one of many programs developed for the i2b2 challenges (Uzuner et al.,

## *APPENDIX G. NLP TOOLS FOR THE BIOCLINICAL AND TEMPORAL DOMAINS*

2010a). Lancet was designed to extract information about medications from narrative text in EHRs by using three supervised machine learning models.(Li et al., 2010). Annotation was done by the Lancet researchers on a portion of the dataset released for the i2b2 challenge, then trained the following components: medications, dosages, and other information related to the taking of medicine were identified with a conditional random fields model, medication names were linked to the other information (dosages, etc.) with the AdaBoost classification model, and sections of narratives and lists were identified with a support vector machine model supplied by the WEKA toolkit. Overall, Lancet performed well in the i2b2 challenge, though the research team has since gone on to improve accuracy (Li et al., 2010).

**MedLEE** - MedLEE is the most comprehensive and widely-applied NLP system for medical records that currently exists. It was originally made for processing radiological reports but has since been extended to a variety of other fields and applications, including mapping medical problems in patient records to ICD9 codes (Carlo et al., 2010), assessing quality control measures for cardiovascular care (Chiang et al., 2010), de-identifying medical records (Morrison et al., 2009), parsing discharge summaries into XML and assigning UMLS codes (Wang et al., 2008), patient smoking status (McCormick et al., 2008), identifying patients for medical studies (Li et al., 2008), for encoding clinical records (Friedman et al., 2004), and so on. The MedLEE program as described by Friedman (2000) is a modular system consisting of: a preprocessor that identifies sentences, abbreviations, and “categorizes words and phrases”; a parser used to identify sentence structure; a compositional regularizer for finding multi-word phrases; an encoder for mapping words and phrases into UMLS and other coding systems; and a recovery component that uses backup methods for extracting infor-

mation where the previous components might have failed. MedLEE has, of course, been expanded and modified over time. Currently it is available for commercial use (<http://www.nlpapplications.com/index.html>).

**SecTag** - SecTag is a program used to identify section headers in patient H&P (history and physical) records. SecTag uses a sequential set of algorithms to process the records, including: 1) a variation of the KnowledgeMap sentence identifier to identify sentences and lists; 2) regular expressions, spell checking and stop word removal to locate potential section headers; 3) use Bayesian probabilities to what sections each sentence belongs to; 4) use Bayesian probabilities to disambiguate section header classifications; and 5) identify the end of each section (Denny et al., 2009). The section header classification is based on an ontology of headers that map to LOINC and RxNorm data types (Denny et al., 2008). While the SecTag algorithm is not available for download, the database is available for use, however, and its use in analyzing the PERMIT corpus is discussed in Section 8.4.

**The TARSQI Toolkit** - The TARSQI Toolkit (TTK) is a “modular system for automatic temporal and event annotation of natural language” (Verhagen and Pustejovsky, 2008), and was developed at Brandeis University. It takes text as input, and outputs a TimeML annotation of the times and events in the document. The different modules of the TTK are:

- *Preprocessing* - tokenizer, chunking, part of speech tagging
- *EVITA* - event recognizer (Saurí et al., 2005)
- *BTime* - temporal expression parser (previously GUTime)

## APPENDIX G. NLP TOOLS FOR THE BIOCLINICAL AND TEMPORAL DOMAINS

- *Slinket* - parser for modal expressions
- *Temporal processing* - set of modules used for creating and cleaning TLinks (temporal relations):
  - *Blinker* - a rule-based system using parts of speech
  - *S2T* - a system that turns subordinating links into TLinks
  - *Classifier* - MaxEnt classifier trained on TimeBank
  - *Sputlink/merger* - runs closure and removes extraneous/conflicting links

The TTK has been tested against Time Bank, a corpus of 183 news articles from various sources that have been annotated with TimeML (Pustejovsky et al., 2003). It is available for download from <http://timeml.org>. While not initially designed for processing clinical documents, a preliminary experiment aimed at determining whether a set of patients were on a particular type of medication at the time they were admitted to the hospital showed promising results for being able to reconfigure the TTK to work with clinical documents (Stubbs and Harshfield, 2010).

**TEXT2TABLE** - A medical text summarization program that “extracts medical events and date times from a text. It then converts them into a table structure” (Aramaki et al., 2009). Specifically, TEXT2TABLE focuses on recognizing what medical events are mentioned in a discharge summary, identifying whether or not they happened (as opposed to being negated or hypothetical), and attempting to identify when the events occurred so that a summary of all events can be generated. In order to do this, TEXT2TABLE uses a 4-step process: event identification using conditional random fields; normalization of dates, times, and events; time-event link-



ing (in this case, events are linked to latest time/date); and identification of negative events. In general, TEXT2TABLE performed well, obtaining an 85.8% average accuracy in identifying negative events and other modalities.

**TimeText** - TimeText was developed “to represent, extract, and reason about temporal information clinical text”, discharge summaries in particular (Zhou et al., 2007). It uses a four-stage process for information extraction: a Temporal Constraint Structure and Temporal Constraint Tagger for representing temporal expressions; the MedLEE system (Friedman, 2000) for processing the narrative information in the discharge summaries; a subsystem that uses medical and linguistic knowledge for handling uncertainties in text (Zhou et al., 2006b); and a formal temporal model based on a simple temporal constraint satisfaction problem. TimeText was used to generate connections between medical events and times in discharge summaries, and the output was compared against human annotators. The researchers found that the temporal information in discharge summaries was extremely difficult to interpret consistently, even among humans. TimeText was later extended to incorporate “fuzzy times” into the constraint satisfaction problem (Lai et al., 2008).

**TN-TIES** - TN-TIES (Triage Note Temporal Information Extraction System) is used to generate human-readable timelines from the triage notes that are recorded when a patient comes to the Emergency Room of a hospital (Irvine et al., 2008). It is based on work done by Zhou et al. (2006a) on modeling temporal relations in discharge summaries, the same work that was the basis for TimeText. TN-TIES uses a three-step process for analyzing the triage notes: first, documents are broken into chunks to identify content phrases; then the chunks are sent to a classifier that determines

## *APPENDIX G. NLP TOOLS FOR THE BIOCLINICAL AND TEMPORAL DOMAINS*

what temporal class each chunk belongs to (here, temporal classes are ‘relative date and time’, ‘duration’, ‘key event’, etc.); finally the classified chunks are sent to an interpreter, where the program attempts to determine what order events occurred in and when they occurred. Output from TN-TIES was compared to a manually annotated corpus of triage notes. The system performed well on chunking, with 91% accuracy, and accurately identified the relative date and time classes, but less well on other, less frequent classes. Numbers for the interpreted data were not provided. The work done with TN-TIES provides a useful datapoint for examining temporal information in clinical texts.

## References

- [Andor2004] J. Andor. 2004. The Master and His Performace: An Interview with Noam Chomsky. *Intercultural Pragmatics* 1.
- [Aramaki et al.2009] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. TEXT2TABLE: medical text summarization system based on named entity recognition and modality identification. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 185–192, Morristown, NJ, USA. Association for Computational Linguistics.
- [Armitage et al.2007] Jane Armitage, Colin Baigent, Zhengming Chen, and Martin Landray. 2007. Treatment of HDL to Reduce the Incidence of Vascular Events HPS2-THRIVE. <http://clinicaltrials.gov/ct2/show/NCT00461630>. clinicaltrials.gov ID: NCT00461630; last accessed July 2012.
- [Badreldin et al.2010] Akmal Badreldin, Axel Kroener, Hiroyuki Kamiya, Artur Lichtenberg, and Khosro Hekmat. 2010. Effect of clopidogrel on perioperative blood loss and transfusion in coronary artery bypass graft surgery. *Interact Cardiovasc Thorac Surg*, 10(1):48–52.
- [Bayerl and Paul2011] Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37:243–257.
- [Biber1993] Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.
- [BioCreAtivE2006] BioCreAtivE. 2006. BioCreAtivE homepage. <http://www.biocreative.org/>. Last visited May 2012.
- [BioNLP20092009] BioNLP2009. 2009. BioNLP Shared Task on Event Extraction. <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/>.
- [BioNLP20112011] BioNLP2011. 2011. BioNLP Shared Task. <http://sites.google.com/site/bionlpst/>.
- [Bird et al.2009] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc, Sebastopol, CA, first edition edition.
- [Blaschke et al.1999] Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. In *ISMB-99 Proceedings*.

## References

- [Botsis et al.2010] Taxiarchis Botsis, Valsamo K Anagnostou, Gunnar Hartvigsen, George Hripcsak, and Chunhua Weng. 2010. Modeling prognostic factors in resectable pancreatic adenocarcinomas. *Cancer Informatics*, 7:281–91.
- [Bramsen et al.2006] Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Finding temporal order in discharge summaries. *AMIA Annual Symposium Proceedings*, pages 81–5.
- [Cao et al.2004] Hui Cao, Michael Chiang, James J. Cimino, Carol Friedman, and George Hripcsak. 2004. Automatic Summarization of Patient Discharge Summaries to Create Problem Lists using Medical Language Processing. In Marius Fieschi, Enrico Coiera, and Yu-Chan Jack Li, editors, *Proceedings of the 11th World Congress on Medical Informatics*, page 1540, The Netherlands. IOS Press.
- [Carlo et al.2010] Lorena Carlo, Herbert S Chase, and Chunhua Weng. 2010. Aligning Structured and Unstructured Medical Problems Using UMLS. *AMIA Annu Symp Proc*, 2010:91–5.
- [CES1996] CES. 1996. Corpus Encoding Standard. <http://www.cs.vassar.edu/CES/>. Last visited June 2012.
- [Chapman and Dowling2006] Wendy W Chapman and John N Dowling. 2006. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *J Biomed Inform*, 39(2):196–208, Apr.
- [Chapman et al.2001] W W Chapman, W Bridewell, P Hanbury, G F Cooper, and B G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, 34(5):301–10.
- [Chapman et al.2007] Wendy W. Chapman, David Chu, and John N. Dowling. 2007. ConText: an algorithm for identifying contextual features from clinical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Chapman et al.2008] Wendy W Chapman, John N Dowling, and George Hripcsak. 2008. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform*, 77(2):107–13.
- [Chapman et al.2011] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D’Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal*, 18 no. 5.

## References

- [Chiang et al.2010] Jung-Hsien Chiang, Jou-Wei Lin, and Chen-Wei Yang. 2010. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). *J Am Med Inform Assoc*, 17(3):245–52.
- [Chomsky1957] Noam Chomsky. 1957. *Syntactic Structures*. Mouton.
- [Clark and Doughty2008] Michael Clark and Dorothy Doughty. 2008. Retrospective Versus Prospective Cohort Study Designs for Evaluating Treatment of Pressure Ulcers A Comparison of 2 Studies. *Journal of Wound, Ostomy and Continence Nursing*, Volume 35 Number 4(July/August).
- [Clark et al.2008] Cheryl Clark, Kathleen Good, Lesley Jezierny, Melissa Macpherson, Brian Wilson, and Urszula Chajewska. 2008. Identifying Smokers with a Medical Extraction System. *Journal of American Medical Informatics Association*, 15:36–39.
- [Clifford et al.2010] G. Clifford, D. Scott, and M. Villarroel. 2010. User Guide and Documentation for the MIMIC II Database. <http://mimic.physionet.org/UserGuide/UserGuide.html>, August. accessed Nov. 23, 2010.
- [Coggon et al.1997] D. Coggon, Geoffrey Rose, and DJP Parker. 1997. Epidemiology for the Uninitiated, Chapter 8. Internet, <http://www.bmj.com/epidem/epid.html>.
- [Cohen et al.2005] K. Bretonnel Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus Design for Biomedical Natural Language Processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 38–45, Detroit, June. Association for Computational Linguistics.
- [Cohen et al.2010] K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. *BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining*, pages 37–41.
- [Cohen1960] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.
- [Cohen2008] Aaron M. Cohen. 2008. Five-way Smoking Status Classification Using Text Hot-Spot Identification and Error-correcting Output Codes. *Journal of American Medical Informatics Association*, 15:32–35.

## References

- [Craven and Kumlein1999] Mark Craven and Johan Kumlein. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of ISMB-99*.
- [Cunningham et al.2010] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, and Ian Roberts, 2010. *Developing Language Processing Components with GATE*, 5 edition, July.
- [D’Avolio et al.2010] Leonard W D’Avolio, Thien M Nguyen, Wildon R Farwell, Yongming Chen, Felicia Fitzmeyer, Owen M Harris, and Louis D Fiore. 2010. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *Journal of the American Medical Informatics Association*, 17(4):375–82.
- [D’Avolio et al.2011] Leonard W D’Avolio, Thien M Nguyen, Sergey Goryachev, and Louis D Fiore. 2011. Automated concept-level information extraction to reduce the need for custom software and rules development. *Journal of the American Medical Informatics Association*, 18(5):607–13.
- [de Haan1984] Pieter de Haan, 1984. *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*, chapter Problem-oriented Tagging of English Corpus Data, pages 123–139. Rodopi: Amsterdamn.
- [Denny et al.2008] Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard. 2008. Development and evaluation of a clinical note section header terminology. *AMIA Annual Symposium proceedings*, pages 156–60.
- [Denny et al.2009] Joshua C Denny, Anderson Spickard, Kevin B Johnson, Neeraja B Peterson, Josh F Peterson, and Randolph A Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc*, 16(6):806–15.
- [Dipper et al.2004a] Stefanie Dipper, Michael Götze, and Stavros Skopeteas. 2004a. Towards User-Adaptive Annotation Guidelines. In *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora LINC-2004*, pages 23–30, Geneva, Switzerland.
- [Dipper et al.2004b] Stefanie Dipper, Michael Götze, and Manfred Stede. 2004b. Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. In *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, pages 54–62, Lisbon, Portugal.

## References

- [Dipper et al.2007] Stefanie Dipper, Michael Götze, Uwe Kssner, and Manfred Stede. 2007. Representing and Querying Standoff XML. In *In Proceedings of the GLDV-Frhjahrstagung*.
- [Dybkjr and Bernsen2004] Laila Dybkjr and Niels Ole Bernsen. 2004. Towards GeneralPurpose Annotation Tools - How far are we today. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC'2004*, pages 197–200.
- [Erk and Strapparava2010] Katrin Erk and Carlo Strapparava. 2010. SemEval-2. In *ACL 2010*, <http://semeval2.fbk.eu/semeval2.php>.
- [Farkas et al.2010] Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning -- Shared Task, CoNLL '10*, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Fleiss1971] J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- [Fort et al.2011] Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37:2:413–420.
- [Franzén et al.2002] Kristofer Franzén, Gunnar Eriksson, Fredrid Olsson, Lars Asker, Per Lidén, and Joakim C oster. 2002. Protein Names and How To Find Them. *International Journal of Medical Informatics*, 67:49–61.
- [Friedman et al.2004] Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. 2004. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402.
- [Friedman2000] C Friedman. 2000. A broad-coverage natural language processing system. *Proc AMIA Symp*, pages 270–4.
- [Gaizauskas et al.2003] Robert Gaizauskas, Mark Hepple, Neil Davis, Yikun Guo, Henk Harkema, Angus Roberts, and Ian Roberts. 2003. AMBIT Acquiring Medical and Biological Information from Text. In Simon Cox, editor, *Proceedings of UK e-Science All Hands Meeting 2003*, pages 370–373, Nottingham, UK, September.
- [Garside et al.1997] Roger Garside, Geoffrey N Leech, and Tony McEney. 1997. *Corpus Annotation: Linguistic Information From Computer Text Corpora*. London: Longman.

## References

- [Geneletti et al.2009] Sara Geneletti, Sylvia Richardson, and Nicky Best. 2009. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*, 10(1):17–31.
- [Geoffrey Sampson2004] Diana McCarthy Geoffrey Sampson, editor. 2004. *Corpus Linguistics: Readings in a Widening Discipline*. Continuum:London, New York.
- [Gold et al.2008] Sigfried Gold, Noémie Elhadad, Xinxin Zhu, James J Cimino, and George Hripcsak. 2008. Extracting structured medication event information from discharge summaries. *AMIA Annual Symposium Proceedings*, pages 237–41.
- [Gries et al.2010] Stefan Th. Gries, Sefanie Wulff, and Mark Davies, editors. 2010. *Corpus-linguistic Applications: Current studies, new directions*. Rodopi: Amsterdam.
- [Grishman and Sundheim1996] Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 466–471.
- [Hall et al.2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).
- [Harkema et al.2005] H. Harkema, I. Roberts, R. Gaizauskas, and M. Hepple. 2005. Information Extraction from Clinical Records. In S. J. Cox, editor, *Proceedings of the 4th UK e-Science All Hands Meeting*, Nottingham, UK. Available at: <http://www.allhands.org.uk/2005/proceedings/papers/477.pdf>.
- [Harkema et al.2009] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J. of Biomedical Informatics*, 42(5):839–851, October.
- [Hersh and Voorhees2009] William Hersh and Ellen Voorhees. 2009. TREC genomics special issue overview. *Information Retrieval*, 12:1–15.
- [Hripcsak and Heitjan2002] George Hripcsak and Daniel F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. *Journal of Biomedical Informatics*, 35:99–110.
- [Hripcsak and Rothschild2005] George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12:296–298.



## References

- [Hripcsak et al.2009] George Hripcsak, Noémie Elhadad, Yueh-Hsia Chen, Li Zhou, and Frances P Morrison. 2009. Using empiric semantic correlation to interpret temporal assertions in clinical texts. *J Am Med Inform Assoc*, 16(2):220–7.
- [Hueston2010] William J Hueston. 2010. Does having a personal physician improve quality of care in diabetes? *J Am Board Fam Med*, 23(1):82–7.
- [i2b2 team2011] i2b2 team. 2011. i2b2 Previous Challenges. <https://www.i2b2.org/NLP/Coreference/PreviousChallenges.php>. last accessed November 2011.
- [Ide and Romary2006] Nancy Ide and Laurent Romary. 2006. Representing Linguistic Corpora and Their Annotations. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*.
- [Ide and Romary2007] Nancy Ide and Laurent Romary, 2007. *Evaluation of Text and Speech Systems*, chapter Towards International Standards for Language Resources. Springer.
- [Ide and Suderman2007] Nancy Ide and Keith Suderman. 2007. Graf: A Graph-based Format for Linguistic Annotation. In *Proceedings of the Linguistic Annotation Workshop*.
- [Ide and Suderman2012] Nancy Ide and Keith Suderman. 2012. Bridging the gaps: interoperability for language engineering architectures using GrAF. *Language Resource Evaluation*, 46(1):75–89, March.
- [Ide et al.2003] Nancy Ide, Laurent Romary, and Eric de la Clergerie. 2003. International Standard for a Linguistic Annotation Framework. In *Proceedings of HLT-NAACL’03 Workshop on The Software Engineering and Architecture of Language Technology*.
- [Irvine et al.2008] Ann K Irvine, Stephanie W Haas, and Tessa Sullivan. 2008. TN-TIES: A system for extracting temporal information from emergency department triage notes. *AMIA Annual Symposium proceedings*, pages 328–32.
- [ISO-246142011] ISO-24614. 2011. Language resource management – Word segmentation of written texts. [http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=41666](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41666). Visited June 15, 2012.
- [Jha et al.2010] Ruchira M Jha, Ambrish Mithal, Nidhi Malhotra, and Edward M Brown. 2010. Pilot case-control investigation of risk factors for hip fractures in the urban Indian population. *BMC Musculoskelet Disord*, 11:49.

## References

- [Kaplan et al.2010] Dain Kaplan, Ryu Iida, and Takenobu Tokunaga. 2010. SLAT 2.0: Corpus construction and annotation process management. In *Proceedings of the 16th Annual Meeting of The Association for Natural Language Processing*, pages 510 – 513.
- [Kilgarriff and Palmer2000] A. Kilgarriff and M. Palmer. 2000. Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, 34:1–13.
- [Kim et al.2003] J-D Kim, T Ohta, Y Tateisi, and J Tsujii. 2003. GENIA corpus—semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–2.
- [Kim et al.2008] Jin-Dong Kim, Tomoko Ohta, and Jun’ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.
- [Kim et al.2009] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun’ichi Tsujii. 2009. Overview of BioNLP’09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP ’09, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Kim et al.2011] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Junichi Tsujii. 2011. EXTRACTING BIO-MOLECULAR EVENTS FROM LITERATURE THE BIONLP09 SHARED TASK. *Computational Intelligence*, 27(4):513–540.
- [Kucera et al.1967] Henry Kucera, W. Nelson Francis, and John B. Carroll. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.
- [Lai et al.2008] Albert M Lai, Simon Parsons, and George Hripcsak. 2008. Fuzzy temporal constraint networks for clinical information. *AMIA Annu Symp Proc*, pages 374–8.
- [Leech1991] Geoffrey Leech, 1991. *English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik*, chapter The State of the Art in Corpus Linguistics, pages 8–29. London: Longman.
- [Leech1993] Geoffrey Leech. 1993. Corpus Annotation Schemes. *Lit Linguist Computing*, 8(4):275–281.
- [Li et al.2008] Li Li, Herbert S Chase, Chintan O Patel, Carol Friedman, and Chunhua Weng. 2008. Comparing ICD9-encoded diagnoses and NLP-processed discharge

## References

- summaries for clinical trials pre-screening: a case study. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 404–8.
- [Li et al.2010] Zuofeng Li, Feifan Liu, Lamont Antieau, Yonggang Cao, and Hong Yu. 2010. Lancet: a high precision medication event extraction system for clinical text. *J Am Med Inform Assoc*, 17(5):563–7.
- [Liakata et al.2009] Maria Liakata, Claire Q, and Larisa N. Soldatova. 2009. Semantic Annotation of Papers: Interface & Enrichment Tool (SAPIENT). In *Proceedings of the BioNLP 2009 Workshop*, pages 193–200, Boulder, Colorado, June. Association for Computational Linguistics.
- [Long2005] William Long. 2005. Extracting diagnoses from discharge summaries. *AMIA Annu Symp Proc*, pages 470–4.
- [Luo et al.2010] Zhihui Luo, Robert Duffy, Stephen Johnson, and Chunhua Weng. 2010. Corpus-based Approach to Creating a Semantic Lexicon for Clinical Research Eligibility Criteria from UMLS. In *2010 AMIA Clinical Research Informatics Summit*.
- [Mann2003] C. J. Mann. 2003. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medical Journal*, 20:54–60.
- [McCormick et al.2008] Patrick J McCormick, Noémie Elhadad, and Peter D Stetson. 2008. Use of semantic features to classify patient smoking status. *AMIA Annu Symp Proc*, pages 450–4.
- [McEnery and Hardie2012] Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics*. Cambridge University Press.
- [McEnery and Wilson1996] Tony McEnery and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press.
- [McEnery et al.2006] Tony McEnery, Richard Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: an advanced resource book*. Routledge: London and New York.
- [Meunier et al.2011] Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, and Magali Paquot, editors. 2011. *A Taste for Corpora*. John Benjamins.
- [Meyer2002] Charles F. Meyer. 2002. *English corpus linguistics:an introduction*. Cambridge University Press.
- [Mihalcea2012] Rada Mihalcea. 2012. Senseval website. [www.senseval.org](http://www.senseval.org).

## References

- [Miller et al.2012] Timothy A. Miller, Dmitriy Dligach, and Guergana K. Savova. 2012. Active Learning for Coreference Resolution. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*.
- [Moloney et al.2010] Niamh Moloney, Toby Hall, and Catherine Doody. 2010. An investigation of somatosensory profiles in work related upper limb disorders: a case-control observational study protocol. *BMC Musculoskelet Disord*, 11:22.
- [Morrison et al.2009] Frances P Morrison, Soumitra Sengupta, and George Hripcsak. 2009. Using a pipeline to improve de-identification performance. *AMIA Annu Symp Proc*, 2009:447–51.
- [Mowery et al.2009] Danielle L. Mowery, Henk Harkema, John N. Dowling, Jonathan L. Lustgarten, and Wendy W. Chapman. 2009. Distinguishing historical from current problems in clinical reports: which textual features help? In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.
- [Nettleman2006] Mary Nettleman. 2006. Usefulness of Home Pregnancy Testing. <http://clinicaltrials.gov/ct2/show/NCT00390754>, October. last visited July 2012.
- [Ogren et al.2006] Philip V Ogren, Guergana Savova, James D Buntrock, and Christopher G Chute. 2006. Building and evaluating annotated corpora for medical NLP systems. *AMIA Annual Symposium Proceedings*, page 1050.
- [Ogren2006] Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.
- [Pakhomov et al.2006] Serguei V. Pakhomov, Anni Coden, and Christopher G. Chute. 2006. Developing a corpus of clinical notes manually annotated for part-of-speech. *Int J Med Inform*, 75(6):418–29.
- [Palmer and Xue2010] Martha Palmer and Nianwen Xue, 2010. *The Handbook of computational Linguistics and Natural Language Processing*, chapter Linguistic Annotation, pages 238–270. Wiley-Blackwell.
- [physionet.org2010] physionet.org. 2010. MIMIC II Clinical Database. website, March.

## References

- [Pustejovsky and Stubbs2011] James Pustejovsky and Amber Stubbs. 2011. Increasing Informativeness in Temporal Annotation. In *Proceedings of the Linguistic Annotation Workshop V*.
- [Pustejovsky and Stubbsforthcoming 2012] James Pustejovsky and Amber Stubbs. forthcoming, 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media Inc.
- [Pustejovsky et al.2003] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK Corpus. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference*, pages 647–656, Lancaster University. UCREL.
- [Pustejovsky et al.2010] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In Nicoletta Calzolari Conference Chair, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and DanielEditors Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, volume 2010, pages 394–397. European Language Resources Association (ELRA).
- [Pustejovsky et al.2011] James Pustejovsky, Jessica Moszkowicz, and Marc Verhagen. 2011. ISO-Space: The Annotation of Spatial Information in Language. In *Proceedings of ISA-6: ACL-ISO International Workshop on Semantic Annotation*.
- [Pustejovsky2006] James Pustejovsky. 2006. Unifying Semantic Annotations for Linguistic Description. In *Invited talk at the Proceedings of Text, Speech, and Dialogue Conference*. <http://video.google.com/videoplay?docid=-3733053928905745721>.
- [Reidsma et al.2005] Dennis Reidsma, Dennis Hofs, and Natasa Jovanovic. 2005. Designing Focused and Efficient Annotation Tools. In L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens, and P.H. Zimmerman, editors, *5th International Conference on Methods and Techniques in Behavioral Research, Measuring Behavior 2005*, pages 149–152, Wageningen. Noldus Information Technology.
- [Roberts et al.2007] Angus Roberts, Robert Gaizauskas, and Mark et al Hepple. 2007. The CLEF corpus: semantic annotation of clinical text. *AMIA Annual Symposium proceedings*, pages 625–9.
- [Roberts et al.2008] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, and A. Setzer. 2008. Semantic Annotation of Clinical Text: The CLEF Corpus.

## References

- In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*.
- [Rothman et al.2008] Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. 2008. *Modern Epidemiology*. Lippincott, Williams and Wilkins, Philadelphia, PA, third edition.
- [Sang2010] Erik Tjong Kim Sang. 2010. CoNLL website. <http://ifarm.nl/signll/conll/>. visited Jan. 2012.
- [Saurí et al.2005] Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. 2005. Evita: A Robust Event Recognizer For QA Systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 700–707.
- [Savova et al.2008] Guergana K. Savova, Anni R. Coden, Igor L. Sominsky, Rie Johnson, Philip V. Ogren, Piet C. de Groen, and Christopher G. Chute. 2008. Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics*, 41 issue 6:10881100.
- [Savova et al.2010] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–13.
- [Savova et al.2011] Guergana K Savova, Wendy W Chapman, Jiaping Zheng, and Rebecca S Crowley. 2011. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18(4):459–465.
- [Scott et al.2012] Donia Scott, Rossano Barone, and Rob Koeling. 2012. Corpus Annotation as a Scientific Task. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*.
- [SemEval2007] SemEval. 2007. SemEval-2007. <http://nlp.cs.swarthmore.edu/semEval/tasks/index.php>.
- [SemEval2012] SemEval. 2012. SemEval-3. <http://www.cs.york.ac.uk/semEval-2012/>.
- [Settles2010] Burr Settles. 2010. Active Learning Literature Survey. Technical report, University of Wisconsin-Madison.

## References

- [Snow et al.2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [South et al.2009] Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics*, 10 Suppl 9:S12.
- [Stubbs and Harshfield2010] Amber Stubbs and Benjamin Harshfield. 2010. Applying the TARSQI Toolkit to Augment Text Mining of EHRs. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Uppsala, Sweden, July. Association for Computational Linguistics.
- [Stubbs and Pustejovsky2011] Amber Stubbs and James Pustejovsky. 2011. Annotation of Discharge Summaries Based on Selection Criteria. In *Proceedings of Clinical Research Informatics Summit*, page 125, San Francisco, CA.
- [Stubbs2011] Amber Stubbs. 2011. MAE and MAI: Lightweight Annotation and Adjudication Tools. In *Proceedings of the Linguistic Annotation Workshop*, Portland, OR.
- [Stubbs2012] Amber Stubbs. 2012. Developing Specifications for Light Annotation Tasks in the Biomedical Domain. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTextM)*, Istanbul, Turkey.
- [Sumner1999] Anne E. Sumner. 1999. Diabetes and Heart Disease Risk in Blacks. <http://clinicaltrials.gov/ct2/show/NCT00001853>. Clinicaltrials.gov ID: NCT00001853; last accessed July 2012.
- [Szarvas et al.2006] G. Szarvas, R. Farkas, S. Ivn, A. Kocsor, and R. Busa Fekete. 2006. Automatic Extraction of Semantic Content from Medical Discharge Records. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.
- [Tateisi and Tsujii2004] Yuka Tateisi and Jun'ichi Tsujii. 2004. Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of the 4th International Conference on Language Resource and Evaluation*.

## References

- [Tateisi et al.2005] Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax Annotation for the GENIA Corpus. In *Proceedings of the IJCNLP, companion volume*.
- [TEI1987] TEI. 1987. Text Encoding Initiative. <http://www.tei-c.org/index.xml>. last visited June 23, 2012.
- [TREC2000] TREC. 2000. TREC website. <http://trec.nist.gov/>, August. accessed March 20, 2012.
- [Tsuruoka et al.2008] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2008. Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC Bioinformatics*, 9.
- [Uzuner et al.2007] Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2007. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.
- [Uzuner et al.2010a] Ozlem Uzuner, Imre Solti, and Eithon Cadag. 2010a. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–8.
- [Uzuner et al.2010b] Ozlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010b. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–23.
- [Uzuner et al.2012] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*.
- [Uzuner2008] Ozlem Uzuner. 2008. Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA Annual Symposium Proceedings*, pages 1252–3.
- [Verhagen and Pustejovsky2008] Marc Verhagen and James Pustejovsky. 2008. Temporal Processing with the TARSQI Toolkit. In *Coling 2008: Companion volume: Demonstrations*, pages 189–192, Manchester, UK. Coling 2008 Organizing Committee.
- [Verhagen and Pustejovsky2012] Marc Verhagen and James Pustejovsky. 2012. The TARSQI Toolkit. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).



## References

- [Verhagen et al.2010] Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.
- [Verhagen2010] Marc Verhagen. 2010. The Brandeis Annotation Tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- [Vincze et al.2008] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9 Suppl 11:S9.
- [Wang et al.2008] Xiaoyan Wang, Amy Chused, Noémie Elhadad, Carol Friedman, and Marianthi Markatou. 2008. Automated knowledge acquisition from clinical narrative reports. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 783–7.
- [Weiss2009] Scott Weiss. 2009. Randomized Trial: Maternal Vitamin D Supplementation to Prevent Childhood Asthma (VDAART). <http://clinicaltrials.gov/ct2/show/NCT00920621>, June. last visited July 2012.
- [Weng et al.2010] Chunhua Weng, Samson W Tu, Ida Sim, and Rachel Richesson. 2010. Formal representation of eligibility criteria: a literature review. *J Biomed Inform*, 43(3):451–67.
- [Weng et al.2011] Chunhua Weng, Zhihui Luo, and Steven B. Johnson. 2011. EliXR: An Approach to Eligibility Criteria Extraction and Representations. In *2011 CRI Summit Proceedings*. accessed Nov. 22, 2010.
- [Wilbur et al.2006] W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC B*.
- [working group2007] CLEF working group, 2007. *CLEF Annotation Guidelines*. <http://nlp.shef.ac.uk/clef/TheGuidelines/TheGuidelines.html>, May. accessed March 2010.
- [Wynne2005] Martin Wynne, editor. 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Available online from <http://ahds.ac.uk/linguistic-corpora/>.

## References

- [Xia and Yetisgen-Yildiz2012] Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical Corpus Annotation: Challenges and Strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTextM)*.
- [Xu and Schwartzman2010] Yining Xu and Kevin Schwartzman. 2010. Referrals for positive tuberculin tests in new health care workers and students: a retrospective cohort study. *BMC Public Health*, 10:28.
- [Yetisgen-Yildiz et al.2011] Meliha Yetisgen-Yildiz, Bradford Glavan, Fei Xia, Lucy Vanderwende, and Mark Wurfel. 2011. Identifying Patients with Pneumonia from Free-Text Intensive Care Unit Reports. In *In Proc. of the ICML workshop on Learning from Unstructured Clinical Text*.
- [Zhou et al.2004] GuoDong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–90.
- [Zhou et al.2006a] Li Zhou, Genevieve B Melton, Simon Parsons, and George Hripcsak. 2006a. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform*, 39(4):424–39.
- [Zhou et al.2006b] Li Zhou, Simon Parsons, and George Hripcsak. 2006b. Handling implicit and uncertain temporal information in medical text. *AMIA Annu Symp Proc*, page 1158.
- [Zhou et al.2007] Li Zhou, Simon Parsons, and George Hripcsak. 2007. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *J Am Med Inform Assoc*, 15(1):99–106.
- [Zürn2011] Christine Zürn. 2011. Risk Prediction in Type II Diabetics With Ischemic Heart Disease. <http://clinicaltrials.gov/ct2/show/NCT01422057>. clinicaltrials.gov ID: NCT01422057; last accessed July 2012.