Developing Specifications for Light Annotation Tasks in the Biomedical Domain

Amber Stubbs

Laboratory for Linguistics and Computation
Department of Computer Science
Brandeis University, Waltham, MA 02453
astubbs@cs.brandeis.edu

Abstract

Biomedical texts pose an interesting challenge in natural language processing tasks. While the information contained in them is important to people of all backgrounds, often they are stylistically complex with specialized vocabularies, and require advanced degrees or other special training to interpret correctly. Because of this, researchers in Natural Language Processing are often at a disadvantage when it comes to extracting task-specific information from these texts: the experts who are best able to understand them may not have the time or interest in completing complicated and time-consuming annotations for use in corpus analysis and machine learning. This paper proposes a methodology for creating light annotation tasks for biomedical corpora that can be used to create useful annotations without requiring extensive training or exceptionally long annotation periods. The utility of the proposed methodology is examined in light of existing annotation projects, as well as through the lens of a case study using hospital discharge summaries for patient selection based on eligibility criteria.

Keywords: annotation, biomedical, methodology

1. Introduction

Text mining of biomedical corpora is a field that has been growing rapidly over the past decade. However, complex biomedical texts offer a unique challenge to computational linguists, who may not always have the domain-specific knowledge required to fully understand and interpret the texts from which they wish to mine information. At the same time, the people who do have this knowledge (doctors, biologists, chemists, etc.) may not have the time or inclination to provide sufficient professional information and linguistic insight to help researchers create useful datasets for training machine learning algorithms. Additionally, hiring such experts as consultants in order for them to perform annotations can be prohibitively expensive.

Naturally, not every query into biomedical texts requires the help of an MD or biologist— part-of-speech tagging, for instance, can generally be done by native speakers of the language even when the vocabulary is unfamiliar. Similarly, named entity and event annotations also do not always require domain knowledge, unless a terminologically rich annotation system is being used. However, as the field of biomedical text mining and information extraction expands, the questions being asked about the data begin to move from "Which of these words are nouns?" and "Which of these are events?" to "Which of these indicate disease X?" and "What are the temporal relationships between these events?" These questions are less easily answered by computational linguists, and more often require domain-specific knowledge and/or training to be properly addressed.

Light annotation tasks[1] are, in theory, an ideal way of solving this dilemma of linguistic complexity versus expert understanding of the literature, as they can exploit information about the chosen corpus without requiring full linguistic annotation. However, it is not easy to create an annotation task that is light (in terms of work required to obtain the annotation, both physically and mentally) and contentful (in terms of later utility). Ideally, a light annotation is acquired for a particular question or corpus, and then additional relevant information (part-of-speech tagging, semantic roles, document structure) is added later as a way of providing more features to the machine learning algorithms.

While light annotation tasks themselves are not new in the biomedical domain—some parts of past i2b2 (Informatics for Integrating Biology and the Bedside) challenges (Uzuner et al., 2007) have relied on them for creating datasets that are augmented by challenge participants, BioNLP tasks have benefited from simplifying annotation tasks used for other purposes (Kim et al., 2009; Kim et al., 2011), and systems like the Automated Retrieval Consol (ARC) (D'Avolio et al., 2010; D'Avolio et al., 2011) use them for data mining, for instance—no methodology or desiderata has been proposed to date for creating meaningful light annotation tasks.

This paper introduces such a methodology, which can be used in conjunction with current standards in corpus and computational linguistics. It is meant to be used in relation to the MATTER (Model, Annotate, Train, Test, Evaluate, Revise) cycle, a development cycle for annotation tasks (Pustejovsky, 2006; Pustejovsky and Stubbs, 2012). At the end of the paper, a case study using this methodology is examined, which focuses on using expert knowledge to create an annotation that represents patients who meet selection criteria for a medical study based on their hospital discharge summaries.

[1] For the purposes of this paper, a 'light annotation' is a textual markup that uses tags that are under-specified in terms of linguistic content, generally for the purpose of creating a task that requires less work to complete. This is in contrast to shallow annotation, such as when a shallow syntactic parse is performed over sentence structures.

2. Related Work

There are existing examples of light annotation tasks in the biomedical domain. The 2007 i2b2 NLP challenge task of identifying the smoking status of patients is a perfect example of a light annotation for a biomedical task, and one that will be discussed later in this paper. Similarly, the Automated Retrieval Console (ARC) system seems to be designed around the idea of asking only for light annotations from users.

The existence of these tasks proves that light annotation projects can be undertaken to yield datasets that represent complex information, but are themselves not complex, and can also later be useful for machine learning projects. However, so far no guidelines or methodology has been established for generalizing these types of tasks.

While investigating useful annotations in biomedical texts, Wilbur et al. (2006) identified five aspects of scientific papers that can be used generally in text mining: focus, polarity, certainty, evidence, and directionality. They reported that the inter-annotator agreements resulted in scores between 70 and 80 percent, which are good indications of an accessible and useful annotation task.

Another example of biomedical annotation is the semantic annotation done by Kim et al. (2008) over the GENIA corpus. In light of the complexity of the task, they employed what they call Single-Facet annotation, a system of presenting annotation tasks to the annotator in order to reduce the cognitive load on the annotators by "defining one aspect of the text as the focus of annotation". This is similar to the annotation approach used in the Brandeis Annotation Tool (BAT), which reduces error in an annotation project by reformulating an annotation task to be performed one layer at a time (Verhagen, 2010).

More generally, the fields of Corpus and Computational Linguistics have yielded specific criteria and methodologies for creating annotation tasks: the MATTER cycle provides a generalized system for developing annotated corpora (Pustejovsky, 2006), the Linguistic Annotation Framework (LAF) is part of an ISO standard for representing annotations in ways that ensure compatibility with other projects (Ide and Romary, 2006), and the seven maxims for annotation tasks identified by Leech (Leech, 1993) have been largely unchallenged over the years.

3. The MATTER Cycle

MATTER is a development cycle for natural language processing tasks involving annotation and machine learning. The steps are: *Model, Annotate, Train, Test, Evaluate, Revise* (Pustejovsky, 2006; Pustejovsky and Stubbs, 2012) (See Figure 1). MATTER represents a general methodology of standard development for all types of annotation tasks.

Within the MATTER cycle there is a smaller development cycle related specifically to the Model and Annotation phases—often when creating an annotation task, the model and annotation are re-evaluated and modified multiple times before the algorithm training is even attempted (see Figure 2). This is referred to as the MAMA (Model-Annotate-Model-Annotate), or the "babbling" phase of the development cycle, as it is the part of the process where the model and annotation become fully formed as a representation of the task (Pustejovsky and Moszkowicz, 2012).

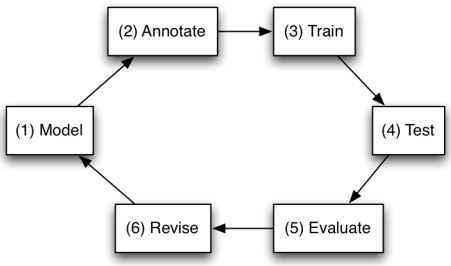

Figure 1: The MATTER cycle.

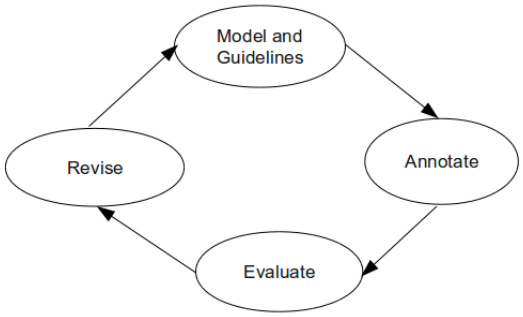

Figure 2: The Model-Annotate Cycle

For most annotation tasks, the Model is the specification used to describe the features of the annotation being applied to the corpus. It defines the tags, attributes, and metadata that will be represented by the annotators over the data being analyzed. The model, M, can be represented as a tuple: M = <T, R, I> where T is the vocabulary of terms, R is the relations between the terms, and I is their interpretation (Pustejovsky and Stubbs, 2012). In most tasks, a single model is used to represent all the information that is needed for the corpus being annotated; that is to say that, even in cases where a task may be divided into steps (for example, event tagging is done first, relation tagging is done later), the annotation output is a unified, complete representation of the desired model.

While this approach has worked well for many annotation tasks, it is easy to find examples of data mining questions where it would be impractical to ask a medical professional to provide all relevant aspects of annotation. Consider again the i2b2 challenge regarding smoking status (Uzuner et al., 2007). While the primary task—determining which of 5 categories (past smoker, current smoker, smoker, non-smoker, and unknown)—is straightforward, and can be represented simply as a single label applied to an entire document, there are many other factors involved in making that classification. In a document, problematic issues can include: which person is being described (some medical documents provide family history as well as patient history), the age of the document, the presence of negations, the scope of a modifying clause, ambiguities introduced by

coreference, etc., not to mention linguistic features such as part-of-speech tags, chunking, tokenization, and so on. These features are not trivial to encode, but the results of the challenge show that at least some of them were used in top-performing systems (Clark et al., 2008; Cohen, 2008; Szarvas et al., 2006).

The Smoking Status dataset is an excellent example of a successful light annotation task specifically because it does *not* include any of those linguistic features. In the next section I will discuss what characteristics make it a good example of light annotation, and how those characteristics can be generalized to other annotation tasks.

4. Creating Light Annotation Tasks

The Smoking Status dataset has a number of attributes which individually may not make it a noteworthy task, but taken together provide an excellent example of a light annotation task. These are (Uzuner et al., 2007):

- The annotation was done by professionals in a field directly related to the area of study (pulmonologists);
- The task used only 5 categories of classification, and two of those contained degrees of uncertainty. For the specific classifications, Past Smoker was someone who quit a year or more ago, Current Smoker is someone who smoked within the past year, and a Non-Smoker is someone who never smoked. Less specifically, Smoker was used to classify someone who was either a current or past smoker, but the temporal reference was vague, and Unknown was used for files where no reference to smoking was made at all;
- The categories used were based on current medical practices and understanding. A layperson would be inclined to label someone who quit smoking 3 months ago as a past smoker, but someone in the medical profession would know that, because the effects of smoking are long-lasting, even people who recently quit are considered smokers for up to a year afterwards;
- Information about both textual and intuitive classifications were collected, though only textual information was used for the challenge due to disagreements over intuition.

Within this task there are some points that can be generalized to other tasks that intend to examine complex biomedical questions. By merging these points with existing linguistic annotation standards, we can establish maxims for creating good light annotation tasks. The following guidelines for creating light annotation tasks in the biomedical domain are therefore proposed:

- 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 biomedical theories and practices;
- 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, or applied as labels to the entire text;
- Additional layers of annotation can be provided before or after the light annotation is performed without conflicting with the given classifications.

These guidelines are primarily targeted at projects that are looking to extract domain expert knowledge from texts (this paper focuses on biomedical examples, but this could be applied to other forms of expertise as well). That is, the Smoking Status was determined by pulmonologists because it is a subject which is directly related to their professional knowledge, and may not be easily interpreted by a layperson. Projects looking to add linguistic information, such as part-of-speech tags, to a text probably do not need to take this approach.

Let us examine each of these guidelines in turn:

Expert annotators: If the purpose of the task being performed is to learn something complex 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 a departure from more traditional linguistic annotations, where linguists and doctors 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).

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 biomedical professional in house, but instead need to hire domain expert annotators as consultants: a process that can be costly and time-consuming. Wilbur et al. (2006) used a similar approach in their text classification task to great success. This is also similar to the Single-facet Annotation as explored by Kim et al. (2008).

Based on current theories and techniques: Beyond simply suggesting that annotations should not be intrinsically unscientific, the point of this guideline is to say that the medical or biomedical understanding of the text should take *precedence* over strictly linguistic analyses. For the Smoking Status corpus, 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 medical interpretation. The annotation must thus reflect medical standards, and not be subordinated to easier or more obvious interpretations.

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 Kim et al. found that leveraging expert knowledge 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, if that is your goal.

Sentence- or phrase-level annotations: Once the annotation task has been cast as one of simple classification, it becomes much easier to instruct domain expert annotators to find sentences or phrases that are used to determine what classification a document or document section should be given. This task can be done much more quickly if the annotators are not asked to create careful markups of the entire document, but rather just to highlight the relevant portions, add a classification label, and then move on.

No conflict with additional annotations: This guideline applies to the practical matter of the actual encoding of the annotation. 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 (Ide and Romary, 2006), 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 process for use in machine learning.

Overall, the purpose of the light annotation task using this methodology is not necessarily to create a complete representation of all the relevant data in a biomedical 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 = <T, R, I>). Rather, the light annotation Model 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 light annotation can and should still be done in the context of the MATTER and MAMA cycles, as they represent established guidelines for text annotation tasks. The next section discusses a corpus of medical documents and examines how the MATTER and light annotation guidelines were applied to an annotation task using that data.

5. Case Study: Finding Patients who Match Selection Criteria

Finding patients who are eligible for participation in medical studies is not a trivial task, even when hospital billing codes can be used to help narrow the field of candidates. At some point medical records need to be examined, and that process is time-consuming and error-prone due to the complexity of the documents being reviewed.

Under a grant from the NIH (NIHR21LM009633-02, PI: James Pustejovsky), this problem was explored in collaboration with the Channing Laboratory at Brigham and Women's Hospital and Harvard Medical.

In order to explore the possibility of automating at least part of the selection process, a test set of selection criteria for a mock case-control study was created, as well as a set of matching criteria in the interest of exploring the information required to create matched case-control groups. A set of 100 discharge summaries was selected from the MIMIC II Clinical Database (Clifford et al., 2010) for review. Documents were chosen based on keywords relevant to the chosen criteria; for a full discussion of the corpus selection process see Stubbs (forthcoming).

The complexities of representing eligibility criteria have been and are still being explored (Weng et al., 2010; Weng et al., 2011). However, rather than focusing on that aspect of the eligibility problem, this annotation effort looked specifically at what would be required for information extraction from the discharge summaries themselves.

The selection and matching criteria used to identify patients for the study were:

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
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

It was immediately clear that a great deal of information would be required to automate this task with any degree of accuracy: document structure, temporal processing, and event recognition would likely be necessary, and possibly other information as well. However, given the complexity and domain-specific vocabulary of the discharge summaries, it was obvious that the document analysis would have to be done, at least in part, by someone working in medical research.

5.1. Annotation Task

Initially this project was going to use the Clinical E-Science Framework (CLEF) (Roberts et al., 2007; Roberts et al., 2008) annotation schema and guidelines (working group, 2007). CLEF has two extent annotations, *Entities* and *Signals*, and two link annotations, *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 such an approach would be extremely time-consuming, and would also require substantial effort to be feasible. The existing CLEF guidelines were found to be unclear in terms of defining what made something a 'condition' rather than a 'result', or an 'intervention' instead of an 'investigation'. Unfortunately, the CLEF corpus is not available to the public and so it could not be used as a resource for making these distinctions.

While an underspecified annotation guideline is not an insurmountable problem, the deciding factor in moving away from CLEF was the amount of time it would take to perform the annotation. For each document that was annotated by a Registered Nurse (i.e., someone who was familiar with the terminology and structure of the files in the corpus), annotating only the entities took several hours. It was clear that the budget for the grant could not support such an intensive annotation project, and a different system would have to be used.

Therefore, it was agreed that the document annotation would be broken down into parts: the linguistic processing (part-of-speech, temporal processing, dependency parsing) could be done by the computational linguistics researchers as needed, while the determination of who met what criteria would be completed as a light annotation task by the medical researchers.

The annotation was done by two medical researchers: one is a Registered Nurse, and the other is involved in patient selection and data collection for medical studies. Because the discharge summaries being examined were so dense with information, rather than have the annotators give a single label per criterion to each document, they were asked to indicate which parts of the document were relevant to each criteria.

The annotation scheme used only four tags: three extent tags used to identify sections of text relevant to each criterion, and one linking tag used to associate different extents where necessary. More specifically:

The **selection_criterion** and **matching_criterion** tags were used to mark text that was relevant to the criteria described above. Both **criterion** tags have an attribute called "criterion", which annotators used to indicate which criterion the text they were marking was relevant to, and 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 **modifier** tag was used to annotate contexts (such as adjectival phrases) that would change the interpretation of the criterion-related text. The use of this tag varied widely: in some cases 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 (e.g., "may be at risk for..."). In order to create a connection between criterion-related text and the modifying extents, a **modifies** link tag was used to connect the two spans where needed.

For the phrase "father with DMII" the resulting annotation would look like this:

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

The annotators did not have to give a document-level classification to each discharge summary; rather, the status of each file in relation to the established criteria was determined automatically after adjudication was performed. Annotations were done in MAE (Multi-purpose Annotation Environment), a intuitive light-weight annotation tool that did not require the annotators to be trained in using a complex software package or understand the underlying XML representation (Stubbs, 2011).

Because the annotators were asked to mark only the parts of the document that were relevant to the criteria, the annotation process was able to go much faster than when the CLEF annotation was attempted. Using CLEF, a single document took hours for the annotator to generate, while with this scheme an average of three documents an hour could be marked up.

This annotation scheme adheres to the guidelines for light annotation outlined above: the annotators had expert knowledge of the field; the task was reduced to a small set of classification tasks at the phrase level; the classifications were based on selection criteria modeled after existing studies; the use of the **modifier** tag required them to provide support for their claims when needed; and the annotations were encoded in stand-off XML so they can be distributed separately from the discharge summaries themselves, and later merged with other annotation layers.

5.2. Annotation Results

The annotation effort over the discharge summaries benefited from the concepts outlined in the guidelines for light annotation tasks. From the outset, the light annotation was much faster for the annotators to complete than the CLEF annotation: under CLEF, each document took roughly 2 hours to annotate, while under the light annotation an average of 3.72 documents could be annotated per hour. Because of the time saved by using the light annotation scheme, the actual cost of the annotation project was roughly 90% less than the projected cost of the CLEF annotation. This improvement in speed reflected the reduced cognitive load that the task placed on the annotators; both annotators felt that the light annotation task was much more tractable and easy to both understand and perform. By using the light annotation task, the data was encoded with expert opinions much more quickly and cheaply than if we had continued to use CLEF. Additionally, the format of the annotation is such that the textual evidence for their opinions can be analyzed later without consulting the annotators, and the format of the annotation is compatible with a variety of existing tools.

Aannotation tasks, however, are always iterative, and the results reported on here are from only the second itera-

tion of the MAMA cycle over this data. As expected under the MATTER and MAMA methodologies, the analysis presented here revealed some problems with the model that can be corrected in later annotations. Specifically, the level at which the annotators were asked to evaluate the text will be expanded—rather than use the **modifier** and **modifies** tags, only the **criterion** tags will be used, and the annotators would be instructed to annotate the entire section relevant to the criteria, including any modifying phrases. This would both cut down on the confusion over how to use the **modifier** tag, as well as speed up the annotation process even more. In future work, the definition of 'cardiac event' should also be more clearly defined in the guidelines.

Due to the complexity of the discharge summaries being annotated, agreement scores based on extent markup between the two annotators were not as high as can be achieved in later iterations. They both generally agreed on what aspects of the text were relevant to the criteria, and which patients met and did not meet the different requirements. However, because of the density of the texts and the amount of repetition in each record, the exact extents that they used to make those determinations did not always overlap, though they were often complementary. For example, Annotator 1 may have spotted a mention of "type2dm+" in the "patient medical history" section of the record, but missed the "patient has diabetes" phrase in the "hospital course" portion of the document, while Annotator 2 did the opposite. Therefore, while Cohen's Kappa (Cohen, 1960) for all the extents marked by each annotator is .505, when compared to the Gold Standard corpus each annotator had high precision (an average of .92) and lower recall (an average of .84) for the **criterion** tags, indicating that both annotators had a higher percentage of false negatives, an analysis that backs up the interpretation of discharge summaries being particularly difficult to read closely for content.

The task of applying the Gold Standard to machine learning algorithms is still in progress, though it is benefiting greatly from the research done for SecTag (Denny et al., 2008; Denny et al., 2009) and cTAKES (Savova et al., 2010), and the temporal analyses of discharge summaries done by Hripcsak et al. (Hripcsak et al., 2009), Zhou et al.(Zhou et al., 2007) and Mowery et al. (Mowery et al., 2009).

6. Conclusions and Future work

This paper presents a methodology for creating light annotation tasks, specifically in the biomedical domain. The guidelines presented represent suggestions extracted from other recognized endeavors, and are based on solid theoretical and practical foundations. While a full application of the light annotation methodology to the case-control annotation task has not yet been completed, it is clear that even these preliminary results show improvements over what would have been achieved with a 'heavier', more complex annotation task.

Using this methodology for extracting light annotation tasks from more complicated endeavors is a viable way for researchers who want to process biomedical texts to approach the problem, without being expert knowledge of the chosen fields themselves. These guidelines enable researchers to obtain contentful, evidence-based expert analyses of domain-specific texts without excessive cost or time investments. This approach could also be used in conjunction with other systems that have been designed for enhancing annotation systems, such as an accelerated annotation framework (Tsuruoka et al., 2008).

While not all tasks are necessarily going to be able to be converted into this format, those that are may benefit from using this approach, particularly in labs where access to domain experts is limited. Admittedly, there is potential for data loss in using a light annotation framework—if the task is not sufficiently well-defined in relation to the goal of the annotation task, there is potential for wasted effort. While this is true of any annotation task, because the focus of this effort is to assist programs where access to domain experts is limited and therefore more costly, it is imperative that any light annotation task undertaken with limited resources be carefully considered in terms of utility.

The guidelines presented here are not limited to use with annotation efforts in the biomedical domain; they can be used for any light annotation task requiring expert knowledge. Applying the methods described here to other domains should be explored for future research.

7. Acknowledgements

I would like to BJ Harshfield, Cheryl Keenan, and Vincent Carey for their assistance with the annotation task, and my advisor James Pustejovsky for his help clarifying the theories presented here. Thanks also to Lotus Goldberg, Jessica Moszkowicz, and Marc Verhagen for proofreading and providing comments on this paper.

Funding for this research was provided by NIH grant NIHR21LM009633-02, PI: James Pustejovsky.

8. References

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.

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.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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.

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). *J Am Med Inform Assoc*, 17(4):375–82.

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. *J Am Med Inform Assoc*, 18(5):607–13.

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.

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.

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.

Nancy Ide and Laurent Romary. 2006. Representing linguistic corpora and their annotations. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

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.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Junichi Tsujii. 2011. Extracting biomolecular events from literaturethe bionlp09 shared task. *Computational Intelligence*, 27(4):513–540.

Geoffrey Leech. 1993. Corpus annotation schemes. *Lit Linguist Computing*, 8(4):275–281.

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.

James Pustejovsky and Jessica L. Moszkowicz. 2012. The role of model testing in standards development: The case of iso-space. In *In the Proceedings of the Eight International Conference on Language Resources and Evalutation (LREC)*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media Inc.

James Pustejovsky. 2006. Unifying linguistic annotations: A timeml case study. In *Proceedings of Text, Speech, and Dialogue Conference*.

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.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, and A. Setzer. 2008. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*.

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.

Amber Stubbs. 2011. Mae and mai: Lightweight annotation and adjudication tools. In *Proceedings of the Linguistic Annotation Workshop*, Portland, OR.

Amber Stubbs. forthcoming. *A Methodology for Leveraging Professional Knowledge in Corpus Annotation*. Ph.D. thesis, Brandeis University.

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*.

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*.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotationfor the genia corpus. In *Proceedings of the IJCNLP, companion volume*.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2008. Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC Bioinformatics*, 9.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2007. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24.

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).

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.

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.

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC B*, 7.

CLEF working group, 2007. *CLEF Annotation Guidelines*. `http://nlp.shef.ac.uk/clef/TheGuidelines/TheGuidelines.html`, May. accessed March 2010.

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.