BioTagger-GM for Detecting Clinical Concepts in Electronic Medical Reports

Manabu Torii, PhD, Hongfang Liu, PhD
Lab of Text Intelligence in Biomedicine
Georgetown University Medical Center Washington, DC

Abstract

Two initial steps towards information extraction from free text are i) concept mention identification and ii) concept mention normalization. As for the concept mention identification task, reasonable performance can be achieved using sequential labeling techniques, such as Conditional Random Field (CRF). In this paper, we report our participation in the concept mention identification task at I2B2 NLP challenge. We adapted an existing gene/protein name identification system, BioTagger-GM, for clinical concepts identification.

Introduction

Concept identification, also known as named entity recognition, is a critical component in biomedical text mining systems that extract biomedical/clinical knowledge from free text or semi-structured text, such as biology research literature and electronic medical records (EMRs). Approaches to tackling this task can be categorized into two main types: i) dictionary lookup methods requiring a list of target entity names and a mechanism for word sense disambiguation [1-9], and ii) machine learning methods requiring text annotated with named entities as a training corpus for building a tagging model[10-12]. In natural language processing (NLP) shared-task workshops in the biomedical domain as well as in the generic domain[11, 13-17], machine learning methods such as Conditional Random Field (CRF)[18] and variants of Support Vector machine (SVM)[19] have achieved promising performance, provided with a large annotated corpus. The availability of machine learning software packages, such as SVM^{struct}[19], YamCha[20] and MALLET[21] among other resources, has boosted the baseline performance of concept recognition systems.

In the past, we have participated gene/protein name tagging and normalization tasks in BioCreAtive workshops[22, 23] and developed a tagging system called BioTagger-GM. Realizing that the clinical concept identification task in the I2B2 NLP Challenge 2010 is essentially similar to gene/protein name tagging task, we rapidly adapted BioTagger-GM for clinical concept identification[24]. In the following, we first describe the background and the original design of BioTagger-GM developed for gene/protein name recognition. Description of the adapted system for clinical concept identification is described next.

Background of BioTagger-GM

A distinctive property of BioTagger-GM is the use of an extensive gene/protein thesaurus, BioThesaurus, and also a large terminology resource in the life science domain, UMLS Metathesaurus, in a machine learning framework for sequential labeling. Initially, phrases in text that are recorded in BioThesaurus and Metathesaurus are first marked regardless of their ambiguity in the context and then they are considered as one type of features for a machine learning algorithm. The details of dictionary lookup and machine learning are provided in the following:

Dictionary lookup - BioTagger-GM uses a normalized dictionary lookup method, where both input text and name entries in the dictionaries, BioThesaurus and Metathesaurus, are normalized. Nonsensical terms and lengthy names in BioThesaurus are ignored during dictionary lookup. The normalization step includes a) converting tokens to base forms according to the SPECIALIST lexicon, b) changing letters to lower case, c) ignoring punctuation marks, and d) convert digit sequences and Greek alphabet to 9 and G, respectively. All phrase occurrences, including overlapping phrases, are recorded during dictionary lookup.

Machine learning – BioTagger-GM uses the CRF frameworks implemented in MALLET to detect gene/protein names. Sentences were tokenized and specified with features characterizing tokens and token occurrences in their contexts. BioTagger-GM uses a simple tokenization strategy where sentences are tokenized according to the following punctuation marks: “:”, “/”, “-” as well as parentheses, brackets, and white spaces. Given tokenized texts, name recognition tasks can be modeled as sequential labeling by associating each token (e.g., word) with an appropriate label (B, I, and O) to demarcate gene/protein names. Here, the label B indicates the token is the beginning of a gene/protein name, I the middle of a gene/protein name, and O the tokens not part of gene/protein names. Each token is represented by features, which include the token itself as one type of features. Besides widely-used features, such as nearby words and part of speech tags within a window size, BioTagger-GM incorporates dictionary lookup results as additional features. If a phrase in text (a sequence of tokens) can be mapped to a dictionary, the phrase is assigned with labels "*L_SemT*", where *L* is one of the three labels, B, I and O, and *SemT* is the type of the phrase in the designated dictionary, e.g., UMLS semantic type. Note that it is possible that multiple labels are assigned in case of overlapping mapping.

Ensemble – A second-order CRF model described above is a base system in BioTagger-GM, which in our experiments outperformed similarly trained taggers using other machine learning algorithms, such as Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM)[25]. To further enhance the performance of CRF taggers, ensemble (combination) of multiple taggers has been implemented in BioTagger-GM. Besides the CRF tagger, an MEMM tagger using the same feature set as the CRF tagger, a first-order CRF tagger trained with ABNER[26], and an HMM tagger trained with LingPipe[27] have been used for gene/protein name recognition. Outputs from different taggers are combined through at-least-n voting[28, 29], which selected entity names detected by n or more taggers.

Adaptation of BioTagger-GM

Figure 1 outlines clinical concept identification using BioTagger-GM. We split the identification of three entity categories ("problem", "treatment", and "test") into three separate subtasks, where each subtask handles one of the three categories. In this manner, we could rapidly adapt a framework to train named entity taggers for a single category to that for multiple categories.

BioTagger-GM is critically based on BioThesaurus, an extensive thesaurus of the target concept, which is gene/protein. In the current task, instead of using BioThesaurus, we used a collection of clinical terms. These terms were extracted from discharge summaries used to load in the vocabulary viewer [30]. Metathesaurus was used just as the original implementation of BioTagger-GM. We employed the first-order CRF model for this task, instead of the second-order model, since first-order models yielded superior performance in our preliminary tests on the training corpus. We did not use a post-processing error-correction module in the original BioTagger-GM, which is a set of hand-coded rules specifically for gene/protein names and for biological research literature. The resulting system was regarded as a base system for the current tagging task.

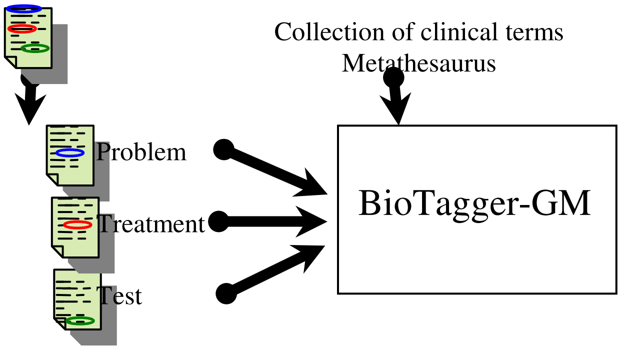

Figure 1. The implementation process.

Besides the base system of the adapted BioTagger-GM, two slight variations of the base tagger were considered in order to derive an ensemble system. As one variant, we supplement the section header, e.g., "problem", "procedure", or "indication", as additional feature for each token. As another variant, we adjusted the context window size to derive nearby words as features. The window size has previously been tuned for gene/protein name recognition in biological literature, where asymmetric window was used, but we considered a larger, symmetric context window, assuming that the asymmetric window might be overly tuned to gene/protein names in biological literature. Finally, outputs from the three variants (base, base + section information, base + wide window) were combined through at-least-n voting with n=2, i.e., given all phrases tagged as a certain category by the three taggers, the ensemble system filters out phrases tagged by only one tagger.

Results and Discussion

Performance of the three single taggers and the ensemble tagger combining them is shown in Table 1. We evaluated using the exact matching criteria (exact boundaries). Among the single taggers, the base system appeared to perform better than the other two taggers in terms of the micro-averaged F-score, although the difference was small. The ensemble tagger consistently performed better than the constituent single taggers in terms of F-scores for individual categories and also of the micro-averaged F-score. The improvement, however, was rather limited possibly because the diversity among the constituent taggers was limited and, thus, the outputs from the three taggers were similar.

We have demonstrated that the rapidly adapted BioTagger-GM system could yield reasonable

	Problem			Treatment			Test			Micro Avg.		
	Pre	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**
Base	.868	.804	**.834**	.869	.787	**.826**	.866	.772	.816	.868	.790	**.827**
Base w/o dictionaries	.868	.781	.822	.864	.751	.803	.870	.757	.809	.867	.765	.813
Base using MEMM	.821	.768	.793	.839	.748	.791	.843	.740	.788	.932	.754	.791
Base w/ section info	.864	.803	.832	.869	.787	**.826**	.864	.770	.814	.866	.790	.826
Base w/ wide window	.866	.800	**.834**	.865	.785	.823	.869	.773	**.818**	.866	.788	.826

Table 1. Performance of the taggers in 5-fold cross-validation tests on the training corpus. Except for the tagger using MEMM ("Base using MEMM"), the 1st order CRF were used as the tagging algorithm. We could observe that (i) the base tagger ("Base") performed better than the one without using dictionary-based features ("Base w/o dictionaries"), and (ii) the base tagger using CRF ('Base") performed better than the one using MEMM, instead of CRF ("Base using MEMM").

	Problem			Treatment			Test			Micro Avg.		
	Pre	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**
√ Base	.856	.786	.820	.853	.760	.804	.856	.763	.807	.855	.772	.811
√ Base w/ section info	.855	.780	.816	.858	.754	.803	.855	.747	.797	.856	.763	.807
Base w/ wide window	.857	.781	817	.856	.753	.801	.858	.760	.806	.857	.767	.809
√ Ensemble	.861	.770	**.821**	.859	.758	**.806**	.863	.761	**.809**	.861	.770	**.813**

Table 2. Performance of the single and ensemble taggers in BioTagger-GM on the evaluation corpus. The checked sign (√) indicates the configurations we selection for our three submissions in the i2b2 challenge. The two variants of the base taggers, "Base w/ section info" and "Base w/ wider window", yielded comparable performance, and thus we arbitrarily selected the one using the section information.

performance in the clinical domain, but we have not explored different opportunities to further improve the tagger performance. Potential topics to explore include the followings:

1. Different machine learning algorithms and diverse feature configurations should be tested especially for the ensemble purpose. For example, SVM-based taggers, such as SVM^{struct}[19] and YamCha[20], and a bi-directional MEMM[31], among others, are competitive with CRF. These algorithms are of great interest for the ensemble purpose.

2. In practice, it is also important to consider different implementations of the machine learning algorithms. For example, it has been reported that CRF++ implementation of CRF yielded better performance than (the default setting of) the MALLET implementation of CRF[32]. In addition, speed and memory requirement become important consideration[33] not only for practical applications, but also for conducting numerous cross-validation tests to fine-tune system.

3. For gene/protein name recognition, BioTagger-GM exploited domain-specific properties through a large terminology resource, BioThesaurus, a domain-specific part-of-speech tagger, GENIA tagger[34], and heuristic-based pre- and post-processing modules, such as an acronym detector, among others. We believe that in clinical domain extensive terminology resources/lexicons, domain-specific syntactic analyzers, and/or hand-coded pre- and post-processing modules can enhance the recognition performance.

4. It is of interest to compare a set of single-category taggers (i.e., three separate taggers for "problem", "treatment", and "test") and a multi-category tagger (i.e., one tagger to detect the three categories). We adopted the first approach to rapidly adapt BioTagger-GM to the current task, but given that the three categories are mutually exclusive, the second approach may help avoid confusions among the categories and/or take advantage of possible interactions among entities belonging to different categories.

5. Another consideration in building generic concept identification systems would be the adaptation of a trained tagger to different institutions. In our preliminary experiments, performance of taggers was apparently degraded when they were trained and evaluated on records from different institutions, compared to the case of the same institutions.

Conclusion

Through our participation in the challenge workshop, we have demonstrated that a gene/protein name tagger, BioTagger-GM, could be rapidly adapted to the clinical domain with reasonable recognition performance. We identified several challenges and research topics in further improving concept identification systems in the clinical domain. Provided with the large corpus in this domain, we plan to investigate these topics further.

Acknowledgement

The work was supported by ABI: 0845523 from National Science Foundation and R01LM009959 from National Institute of Health.

References

1. Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform.* Dec 2004;37(6):461-470.
2. Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics.* Oct 15 2007;23(20):2768-2774.
3. Koike A, Niwa Y, Takagi T. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics.* Apr 1 2005;21(7):1227-1236.
4. Kou Z, Cohen WW, Murphy RF. High-recall protein entity recognition using a dictionary. *Bioinformatics.* Jun 2005;21 Suppl 1:i266-273.
5. Egorov S, Yuryev A, Daraselia N. A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc.* May-Jun 2004;11(3):174-178.
6. Mika S, Rost B. NLProt: extracting protein names and sequences from papers. *Nucleic Acids Res.* Jul 1 2004;32(Web Server issue):W634-637.
7. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Medinfo.* 2004;11(Pt 1):268-272.
8. Aronson AR, Mork JG, Neveol A, Shooshan SE, Demner-Fushman D. Methodology for creating UMLS content views appropriate for biomedical natural language processing. *AMIA Annu Symp Proc.* 2008:21-25.
9. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc.* Sep-Oct 2004;11(5):392-402.
10. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics.* 2005;6 Suppl 1:S2.
11. Wilbur JW SLaTL. BioCreative 2. Gene Mention Task. *Second BioCreative Challenge Evaluation Workshop*. Madrid; 2007:pp. 7-16.
12. Kim J-D, Ohta T, Tsuruoka Y, Tateisi Y, Collier N. Introduction to the bioentity recognition task at JNLPBA. *InternationalWorkshop on Natural Language Processing in Biomedicine and its Application*. Geneva, Switzerland; 2004.
13. Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA Annu Symp Proc.* 2008:1252-1253.
14. Gainer V, Hackett K, Mendis M, et al. Using the i2b2 hive for clinical discovery: an example. *AMIA Annu Symp Proc.* 2007:959.
15. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. *AMIA Annu Symp Proc.* 2006:931.
16. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L. Evaluation of text mining systems for biology: overview of the Second BioCreAtIve community challenge. *Genome Biology.* 2008;9((Suppl 2): S1).
17. Morgan AA, Lu Z, Wang X, et al. Overview of BioCreative II gene normalization. *Genome Biol.* 2008;9 Suppl 2:S3.
18. Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Paper presented at: ICML, 2001; Williamstown, MA.
19. Tsochantaridis I, Hofmann T, Joachims T, Altun Y. Support Vector Learning for Interdependent and Structured Output Spaces. Paper presented at: ICML, 2004; Banff, Alberta, Canada.
20. Kudo T, Matsumoto Y. Chunking with support vector machines. Paper presented at: In Proc. of the Annual Meeting of the Association for Computational Linguistics, 2001
21. McCallum AK. Mallet: A machine learning for language toolkit. *Unpublished. http://mallet. cs. umass. edu.* 2002.

22. Liu H, Torii M, Hu Z, Wu CH. Gene Mention and Gene Normalization Based on Machine Learning and Online Resources. Paper presented at: Proceeding of BioCreAtIve II workshop, 2007.

23. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics.* 2005;6 Suppl 1:S11.

24. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc.* Mar-Apr 2009;16(2):247-255.

25. McCallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation. Paper presented at: Proc. ICML 2000, 2000; Stanford, California.

26. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics.* Jul 15 2005;21(14):3191-3192.

27. Alias-i. LingPipe http://alias-i.com/lingpipe.

28. Torii M, Liu H. At-least-N voting over biomedical named entity recognition systems. Paper presented at: BioLINK, 2008.

29. Kambhatla N. Minority vote: at-least-N voting improves recall for extracting relations. Paper presented at: COLING/ACL on Main conference poster sessions, 2006; Sydney, Australia.

30. Friedman C, Liu H, Shagina L. A vocabulary development and visualization tool based on natural language processing and the mining of textual patient reports. *J Biomed Inform.* Jun 2003;36(3):189-201.

31. Tsuruoka Y, Tsujii Ji. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. Paper presented at: HLT/EMNLP, 2005.

32. Hsu C, Chang Y, Kuo C, Lin Y, Huang H, IF C. Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics* 2008 24:i286-i294.

33. Okazaki N. CRFsuite: a fast implementation of Conditional Random Fields (CRFs) http://www.chokkan.org/software/crfsuite/; 2007.

34. Tsuruoka Y, Tateishi Y, Kim J-D, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics. Paper presented at: 10th Panhellenic Conference on Informatics, 2005.