Online Biomedical Concept Annotation Using Language Model Mapping[1]

Lawrence H. Reeve
IBM Corporation, USA,
larry.reeve@us.ibm.com

Hyoil Han
College of Information Science and Technology, Drexel University, USA,
hyoil.han@acm.org

Ari D. Brooks
College of Medicine, Drexel University, USA,
Ari.Brooks@DrexelMed.edu

Abstract

We report the results of applying language technology to the bioinformatics problem of online concept annotation of biomedical text. We extend our concept annotator, CONANN, to find biomedical concepts in using concept language models. The goal of CONANN is to improve annotation speed without losing annotation accuracy as compared to offline systems, facilitating the use of concept annotation in online environments. Intrinsic and extrinsic evaluations show accuracy competitive with a state-of-the-art biomedical text concept annotator with a speed improvement of more than four times.

1. Introduction

The task of a biomedical concept annotator is to map each unit of source text, such as a phrase, into one or more domain-specific concepts. Biomedical concepts are defined in resources such as the Unified Medical Language System (UMLS) [1] and National Cancer Institute (NCI) Thesaurus [2]. Among the benefits of using domain-specific concepts, rather than terms, is 1) synonym merging, where synonymous phrases are merged to a single concept, and 2) the use of a domain-specific language for querying.

In this paper, we present the results of extending CONANN[7] to use concept language models to find the best matching domain-specific concept for a biomedical text phrase. We are not aware of any biomedical concept annotators using a language model approach.

2. Biomedical Concept Resource - Background

The UMLS Metathesaurus contains concepts and real-world instances of the concepts, including a concept name and its synonyms, lexical variants, and translations [9]. The Metathesaurus is derived from over 100 different vocabulary sources resulting in over one million biomedical concepts. A *concept instance* is a phrase belonging to a UMLS concept. Each UMLS concept is associated with one or more synonymous concept instances. A single UMLS concept may have multiple instances. The UMLS Metathesaurus organizes concept instances into concepts. A *concept name* is the name given to a particular UMLS concept.

3. Language Model Concept Mapping

A unigram language model is built for each of the concepts in UMLS. A concept language model contains word identifiers, frequencies of words in a concept, and the probability of the word occurring across all concept instances within the concept, that is, $P(w) = \frac{|w|}{N}$, where w is a word in a particular concept's concept instances, $|w|$ is a count of the number of times the word appears in all of the concept instances of a concept, and N is the total number of words in all of the particular concept's concept instances. In the UMLS resource we used, there were 797,152 concepts defined resulting in the generation of 797,152 corresponding unigram language models. All of the UMLS concepts are assumed to be independent.

The overall annotation strategy of a single biomedical source phrase with a language model extension is shown in Figure 2 and is as follows:

1. *Candidate Phrase Generation*: Construct a list of candidate phrases based on the common words between all UMLS concept instances and the source

[1] This work was conducted when the first author was at Drexel University.

phrase. If only one candidate phrase remains, return its associated concept name.

2. *Candidate Phrase Filtering*: The list of candidate phrases is trimmed using a coverage filter based on Involvement scoring[10], which first computes the normalized word weights of words in common between a source phrase and a candidate phrase in both directions (phrase involvement), and then averages the two phrase involvement values to determine the candidate phrase score.

The standard deviation of the involvement phrase scores for the set of candidate phrases is calculated as the mean candidate involvement score plus two standard deviations. All candidate phrases having involvement phrase scores greater than or equal to the threshold value are passed to the language model concept mapper in order to capture the top 5% [11] of candidate phrases. If no involvement phrase scores are greater than or equal to the threshold value, candidate phrases with the highest involvement phrase score value are used. Two possible scenarios are possible after coverage filtering. 1) A single candidate phrase remains and is returned. 2) More than one candidate phrase remains and they are passed to the language model concept mapper for final mapping.

3. *Final Mapping*: Each candidate concept is then assigned a score based on its probability of generating the source phrase. The probability score is calculated using a standard unigram language mixture model [8] combining the source word (*w*) probability within the concept language model ($M_{concept}$ in Figure 1) with the probability of the word occurring across the entire UMLS phrase collection ($M_{conceptCollection}$ in Figure 1).

$$P(concept) = \prod_{w \in SrcPhrase} ((1-\lambda)P(w \mid M_{concept} + \lambda P(w \mid M_{conceptCol\,lection}))$$

Figure 1: Multinomial unigram language mixture model for concept mapping

We initially set λ=0.5 to balance the concept language model with the collection model. We expanded the source phrase words to include their variants provided by UMLS. The highest-scoring candidate concept, or multiple concepts in the case of a tie, is then output as the best-matching concept for the source phrase.

4. Related Work

Our work is most closely related to MetaMap [3], KnowledgeMap [5], and SAPHIRE [4]. SAPHIRE uses simple and partial mapping, and for candidate phrase scoring combines measures of term overlap, term proximity, and length of term matches. KnowledgeMap uses simple and partial matching, and for candidate phrase scoring uses an exact match approach and if no matches are found, performs iterative variant-word-generation and re-matching. MetaMap scores candidate phrases using a mixture of four different scores: *Centrality*, *Variation*, *Coverage*, and *Coherence*. CONANN uses simple and partial matching, but does not score every candidate phrase for final mapping. Like KnowledgeMap and MetaMap, we incorporate word variants of the source phrase, but we do not incorporate disambiguation or exact matching as KnowledgeMap does or extensive word variant generation as MetaMap does. CONANN defers complex scoring until after most candidate phrases have been eliminated. In addition, we build a language model of each concept's phrases, whereas existing systems consider each candidate phrase as independent of one another, even when from the same concept.

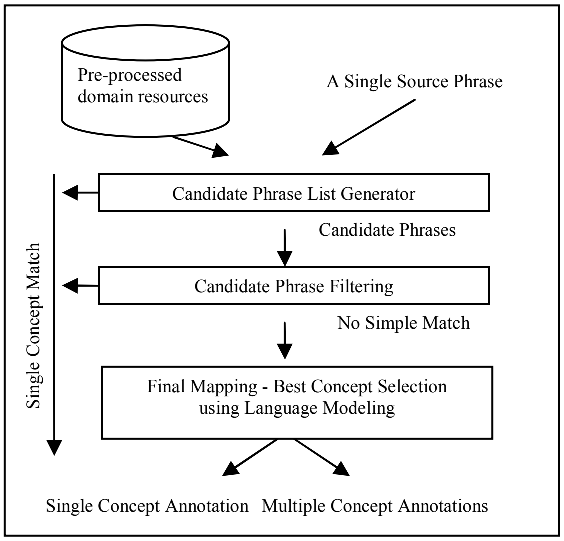
Pre-processed domain resources
A Single Source Phrase
Candidate Phrase List Generator
Candidate Phrases
Candidate Phrase Filtering
No Simple Match
Final Mapping - Best Concept Selection using Language Modeling
Single Concept Match
Single Concept Annotation
Multiple Concept Annotations

Figure 2: Concept mapping process in CONANN with a language model extension
(Adaptation of the old CONANN's architecture in Reeve & Han [7])

Other related systems include SENSE [12], which translates source and concept instance to low-level semantic factors, then performs exact matching of the semantic factors; Concept Locator [13] which simply sub-divides a phrase and looks for exact matches;

PhraseX [14] which focuses on phrase identification and performs an exact match with candidate phrases; and IndexFinder [15] which treats the source text as a bag of words and finds all matching words, regardless of their location.

5. Evaluation Methods

Evaluation of CONANN is done using both intrinsic and extrinsic methods. The intrinsic evaluation is intended to evaluate the speed and accuracy of CONANN against an existing state-of-the-art biomedical concept annotator. The extrinsic evaluation is designed to measure the effect of the annotator's output on a task. We chose concept-based text summarization as the task. A corpus of unique phrases and a subset of texts for summarization was constructed from a citation database of approximately 1,200 oncology clinical trial papers physicians feel are important to the field [16]. The final corpus consists of 1,628 unique phrases for the intrinsic evaluation and 24 texts each with three summaries generated by third-year medical students for use with the extrinsic evaluation.

5.1 Intrinsic Evaluation

We use the MetaMap system [3] provided by the United States National Library of Medicine as the baseline system. MetaMap and CONANN were executed using 1,628 unique phrases from our evaluation corpus and their execution times compared. Accuracy of CONNAN against was compared using precision [17], where the number of matching was divided by the number of phrases. Average time to annotate a phrase is calculated by taking the total annotation time of all phrases divided by the total number of phrases annotated.

The average time to annotate a phrase using MetaMap was 208 milliseconds, while for CONANN it was 47 milliseconds, an improvement of speed of nearly 4.5 times. The total time to annotate all phrases in the corpus was 5.67 minutes using MetaMap and 1.28 minutes for CONANN. We assume that the precision of MetaMap is 1 (i.e. 100%). The precision of CONNAN compared to MetaMap was 0.85. If we relax the concept matching so that the top five concepts produced by CONANN for a phrase were compared to MetaMap, the precision rises to 0.93.

5.2 Extrinsic Evaluation

Two probabilistic summarizers, FreqDist [18] and a version of SumBasic [19] modified to use concepts rather than terms were used. Both summarizers use concept frequency as the sole feature to select salient sentences.

The FreqDist and SumBasic summarizers were used to generate a summary of each of the 24 texts using the concept output from both MetaMap and CONANN. The system-generated summaries were evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [20][21] summary evaluation tool. ROUGE provides automated comparison of system-generated summaries with the three human summaries using n-gram co-occurrence. ROUGE-2 measures bigram co-occurrence while ROUGE-SU4 measures skip-bigrams with a maximum distance of 4 words.

Table 1 shows the ROUGE-2 and ROUGE-SU4 scores using the FreqDist and SumBasic summarizers with both CONANN and MetaMap annotation output. The use of the CONANN concept annotator with both the FreqDist and SumBasic summarizers provides the best text summarization performance. CONANN with FreqDist shows an improvement of 7% for ROUGE-2 and 2% for ROUGE-SU4. CONANN with SumBasic shows an improvement of 7.8% for ROUGE-2 and 5.6% for ROUGE-SU4.

Table 1: Text summarization performance

Summarizer	***ROUGE-2 Score***	***ROUGE-SU4 Score***
FreqDist with MetaMap	0.12080	0.21864
FreqDist with CONANN	0.12897	0.22292
SumBasic with MetaMap	0.10920	0.19868
SumBasic with CONANN	0.11839	0.21053

6. Conclusion

We presented a language model version of our online biomedical concept annotator, CONANN. Online annotation is useful for physician and biomedical research tasks such as personalized text summarization, question-answering, information extraction, data mining and concept-based indexing and retrieval. An intrinsic evaluation was performed to compare the precision of CONANN's concept output and annotation speed to MetaMap, a state-of-the-art concept annotator. Precision was measured at 0.85 for exact match and 0.93 for relaxed match, with a speed improvement of 4.5 times. An extrinsic evaluation showed summaries generated from two different types of summarizers using concepts generated by CONANN slightly outperformed the same summarizers using MetaMap concept output.

The intrinsic and extrinsic evaluations show the use of filtering candidate phrases using word coverage and then mapping source phrases to biomedical concepts using language models is effective and faster than current systems. Future work includes integrating synonymous word variants into the concept language models and incorporating concept disambiguation.

6. References

[1] United States National Library of Medicine, "Unified Medical Language System (UMLS)," 2005.

[2] Center for Bioinformatics, United States National Cancer Institute. 2006, National cancer institute thesaurus.

[3] A. R. Aronson, "Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program," in *Proceedings of the AMIA Symposium 2001,* 2001, pp. 17-21.

[4] W. R. Hersh and R. A. Greenes, "SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships," *Comput. Biomed. Res.,* vol. 23, pp. 410-425, Oct. 1990.

[5] J. C. Denny, P. R. Irani, F. H. Wehbe, J. D. Smithers and A. Spickard 3rd, "The KnowledgeMap project: development of a concept-based medical school curriculum database," *Proceedings of the Annual AMIA Symposium,* pp. 195-199, 2003.

[6] L. Reeve, H. Han and A. D. Brooks, "BioChain: Using lexical chaining methods for biomedical text summarization," in *Proceedings of the 21st Annual ACM Symposium on Applied Computing, Bioinformatics Track,* 2006, pp. 180-184.

[7] L. H. Reeve and H. Han, "CONANN: An online biomedical concept annotator," in *Proceedings of the 2007 Data Integration in the Life Sciences Conference (DILS'07),* 2007.

[8] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval.* ,1st ed.Cambridge, England: Cambridge University Press, 2007.

[9] X. Liu and W. B. Croft, "Statistical language modeling for information retrieval," in *Annual Review of Information Science and Technology* , vol. 39, B. Cronin, Ed. American Society for Information Science and Technology, 2005, pp. 1-31.

[10] A. R. Aronson, "MetaMap Evaluation," pp. 1-12, 2001.

[11] H. O. Kiess, *Statistical Concepts for the Behavioral Sciences.* ,Third ed., vol. 1, Boston, MA: Allyn and Bacon, 2002, pp. 568.

[12] Y. L. Zieman and H. L. Bleich, "Conceptual mapping of user's queries to medical subject headings," *Proc. AMIA. Annu. Fall. Symp.,* pp. 519-522, 1997.

[13] P. M. Nadkarni, "Concept locator: a client-server application for retrieval of UMLS metathesaurus concepts through complex boolean query," *Comput. Biomed. Res.,* vol. 30, pp. 323-336, Aug. 1997.

[14] S. Srinivasan, T. C. Rindflesch, W. T. Hole, A. R. Aronson and J. G. Mork, "Finding UMLS Metathesaurus concepts in MEDLINE," *Proc. AMIA. Symp.,* pp. 727-731, 2002.

[15] Q. Zou, W. W. Chu, C. Morioka, G. H. Leazer and H. Kangarloo, "IndexFinder: A method of extracting key concepts from clinical texts for indexing," in *Proceedings of the AMIA Annual Symposium,* 2003, pp. 763-767.

[16] A. D. Brooks and I. Sulimanoff, "Evidence-based oncology project," in *Surgical Oncology Clinics of North America* , vol. 11, Anonymous 2002, pp. 3-10.

[17] W. R. Hersh, M. Mailhot, C. Arnott-Smith and H. J. Lowe, "Selective Automated Indexing of Findings and Diagnoses in Radiology Reports," *J. Biomed. Inform.,* vol. 34, pp. 262-273, 2001.

[18] L. Reeve, H. Han, S. V. Nagori, J. Yang, T. Schwimmer and A. D. Brooks, "Concept frequency distribution in biomedical text summarization," in *Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management (CIKM'06),* 2006, pp. 604-611.

[19] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization," Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101, 2005.

[20] C. Lin and E. H. Hovy, "Automatic evaluation of summaries using N-gram co-occurrence statistics," in *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003),* 2003, pp. 71-78.

[21] National Institute of Standards and Technology (NIST), "Document Understanding Conferences," vol. 2005, 2005.