Can Distributional Statistics Aid Clinical Concept Extraction?

Siddhartha Jonnalagadda, Graciela Gonzalez
Department of Biomedical Informatics, Arizona State University, Phoenix, AZ, USA

Abstract

Extracting concepts and relationships among them from clinical narratives constitute the basic enabling technology that will unlock the knowledge contained in them and drive more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment, to name a few secondary uses of the data. However, the training data cannot possibly contain all concepts and their synonyms. We propose novel natural language processing approaches to transcend this limitation in relation to automatically extracting concepts from clinical notes using distributional statistics. Until now[1]*, "It is not well-understood what settings are appropriate to induce distributional word representations for structured prediction tasks (like parsing and MT) and sequence labeling tasks (like chunking and NER)".* ***We observed that distributional statistics of terms across large unlabeled corpus significantly aids concept extraction****. We were also successful at implementing a competitive baseline relationship extraction system. Our systems submitted to the task achieved 80.9% f-score for class exact span concept extraction and 69.7% f-score for correctly matching relationships.*

Introduction

We propose a novel approach to the task of extracting concepts and relationships from clinical corpus that applies distributional semantics. Distributional semantics is an emerging area of research arising from the notion that the semantics of a piece of text (discourse) can be inferred from the distribution of the elements of that discourse in relation to their surroundings.

Recent research by Sahlgren et al.[2] in distributional semantics has explored the differences between relations extracted depending on the type of context used to construct a model. There are two types of relationships between terms – syntagmatic and paradigmatic[3]. If two terms co-occur significantly in passages or sentences, they are said to be in syntagmatic relationship. Examples include terms that occur frequently in succession such as p53 and tumor, APOE and AD, and poliomyelitis and leg. If two terms can substitute for each other in the sentences while maintaining the integrity of the syntactic structure of a sentence, they are said to be in a paradigmatic relationship. Examples include: p53 and gata1, AD and SDAT, and poliomyelitis and polio. Since terms in paradigmatic relationship don't occur together in the same context, extracting such a relationship typically requires 2nd order analysis, while a 1st order analysis is sufficient to extract syntagmatic relationships. Sahlgren et al. argue that using a small sliding-window rather than an entire document as a context is better suited to extracting paradigmatic relations, and supports this argument with empirical results.

At the core of deriving the meaning While earlier attempts were almost all dictionary or rule-based systems, most of the modern systems use supervised machine-learning where a system is trained to recognize named entity mentions in text based on specific (and numerous) features associated with the mentions that the system learns from annotated corpora. In this latter category, generative models (Naïve Bayes Classifier and Hidden Markov Models) and instance-based classifiers (Logistic Regression and Naïve Bayes Classifier) are theoretically shown to be less accurate for extracting concepts or named entities from text than sequence-based discriminative models like Conditional Random Fields[4,5]. The high computational cost associated with using deep syntactic and semantic features largely restricted the concept extraction systems to orthographic, morphological and shallow syntactic features. The first application of information extraction in clinical domain can be attributed to Hirschman et. al[6] for converting a corpus of x-ray reports on patients with breast cancer into a structures database using a theory of sublanguage of grammars[7]. Medical Language Extraction and Encoding System (MEDLEE[8]) which automatically generates coded information from general clinical notes using rule-based approach in addition to finding modifiers is an important step towards generic reusable clinical NER systems. MetaMap[9], an initiative of NLM to map text to UMLS metathesarus that uses large lexicon is one of the next notable advancements in state of the art since it is still being actively used. A direct application of MetaMap for detecting medical problems[10] shows that it has an f-score accuracy of 75% for that task and a recent open-source tool HITEX[11] uses

MetaMap to map concepts to UMLS strings. Recently, we saw the emergence of cTakes[12] that uses naïve bayes classifier with lexical and syntactic features for identifying clinical named entities such as diseases, signs/symptoms, anatomical sites and procedures and it achieves 56% strict f-score accuracy. Other systems have been developed for specialized genres of clinical text. As pointed out by Meystre et al.[13], clinical NLP is currently lagging behind biomedical NLP mainly because of lack of experience of NLP researchers with clinical text and the rarity of annotated corpora.

Concept extraction is a fundamental building block that enables more advanced applications like sentiment analysis and relationship extraction. We define relationship extraction as finding n-tuples of entities satisfying a defined function. In clinical domain, relationship extraction often involves finding associations between medical problems, tests and treatments. Establishing a relationship among concepts is often more than a binary classification problem and requires further classification of the kind of association. The 4th i2b2/VA NLP shared task is requires classification of relationships between tests and problems as whether: 1) test reveals the medical problem, and 2) test conducted for the medical problem; and relationships between treatment and medical problem as whether: 1) treatment improves medical problem, 2) treatment worsens medical problem, 3) treatment causes medical problem, 4) treatment is administered for medical problem, or 5) treatment is not administered for medical problem. Unlike in biomedical domain, relationship extraction is still in beginning stages in the clinical domain with just a couple of the systems planned for general use[14,15]. More often we see specialized systems such as ONYX[16] that uses probabilistic context free grammar to extract surface-of-part relationship and location-of-condition relationship with an accuracy of 30% (they reported ERROR of 70%). Most existing systems for finding relationships between clinical concepts use measures of simple co-occurrence, or variations of it like chi-square[17] and pointwise mutual information[18]. Some of the recent work involves use of machine learning techniques[14] and dependency trees[19] to extract relationship between an entity and the modifiers around it (not necessarily a specific entity-type).

Methods

For concept extraction, the main challenge is the scarcity of annotated examples and the fact that no such large corpus can practically be created without raising privacy concerns. We hypothesize that the distributional information of terms in the unannotated corpora can be used to compensate for the limited vocabulary present in a small annotated corpus and allow more accurate concept recognition. Existing state of the art machine-learning systems in biomedical domain (since biomedical domain is considered to be ahead in concept extraction and we feel that the features should be adaptable) will be extended by adding distributional semantics features, respectively to extract medical problems and treatments from clinical narratives. Improvement in accuracy after adding distributional semantics features (using i2b2/VA NLP shared task corpus as gold standard) would validate the **utility of distributional semantics features**.

Because of privacy concerns, large corpora of clinical text are not available for research purposes, and thus the annotated sets that are publicly available are even smaller in size. We recently proved that distributional semantic features in a corpus can be successfully applied for extracting concepts from text through our SimFind system[20]. We used unannotated data from clinical text and propose a semi-supervised machine-learning approach for the purpose of extracting concepts and relationships. While there is an earlier system in biomedical domain from IBM[21] that used a large amount of unannotated data for winning the BioCreative II shared task for detecting protein entity names, it uses the computationally expensive machine-learning algorithm called Alternating Structure Optimization (ASO). Liu et al.[22] who applied ASO to Semantic Role Labeling, state, "Some of our experiments are limited by the extensive computing resources required for a fuller exploration. However, we have been unable to use unlabeled data to improve the accuracy." We are proposing the use of unlabeled data through construction of vector-based similarity model using random indexing which is much faster than previous methods (processing the entire Medline corpus takes around 30 min using an octa-core Xeon server and 16GB RAM). Our approach is thus scalable to huge unannotated corpora and will promote widespread use of unannotated data for the task of clinical concept extraction.

The concept extraction task can be seen as finding terms that could conceivably replace the token we want to label without disturbing syntactic structure, i.e., finding terms that are paradigmatically related. To model paradigmatic relationships the above vector coordinate permutations model was chosen, as it has been observed that the relations captured by this method tend to emphasize terms of a similar semantic class[23,24]. Since distributional semantics models constructed based on millions of documents and millions of terms have terabytes of cells, we need to reduce dimensionality. Traditional dimensionality reduction techniques such as Singular Value Decomposition (SVD) are cubic in complexity. Recently, Random Indexing[25] emerged as promising alternative to the use of SVD for the dimension reduction step in the generation of term-by-context vectors. Random Indexing and other similar methods are motivated by the Johnson–Lindenstrauss Lemma[26] which states that the distance between points in a vector space will be approximately preserved if they are projected into a reduced-dimensional subspace of sufficient dimensionality. Using Random Indexing for building a paradigmatic space using involves:

a) constructing random term vectors of pre-determined dimension N and seed S, where N-2*S dimensions are zeroes, S dimensions are +1s, and S dimensions are -1s. To preserve the applicability of Johnson–Lindenstrauss Lemma, S<<N.

b) based on random vector representations of each term, the semantic vector representation of the term is computed by a linear function (Sahlgren et al., 2008) of the all terms surrounding each occurrence of the term.

As discussed before, random indexing is only a computationally cheaper alternative to SVD (Singular Value Decomposition) with almost the same accuracy. Hence, for a small size of corpus like i2b2/VA corpus, there are supervised dimensionality reduction techniques such as LDA (Linear Discriminant Analysis) that use SVD computation and could replace random indexing for designing kernels based on training set annotations. LDA, while having the limitation of not being applicable for reducing dimensionality in unlabeled data, is widely applied in NER before and it will not be surprising if kernels built using LDA perform better than the kernels built using Random Indexing that is an unsupervised dimensionality reduction method and does not exploit the labels of the data. Random indexing is more suitable when applied to a huge unlabeled corpus such as hundreds and thousands of clinical notes. As the initial step, we used all the documents (349 annotated clinical notes and 827 unannotated clinical notes) and in the subsequent experiment, we used all the MEDLINE abstracts that are indexed as "clinical trials". There are more than 500k MEDLINE abstracts indexed as "clinical trials".

An inverted index of the chosen unlabeled data is constructed using Lucene Index (lucene.apache.org) and Semantic Vectors (code.google.com/p/semanticvectors) software is used to find the vector representation of each term in the abstracts that appears at least twice. We then constructed the kernel K over the terms in the i2b2/VA corpus, where

K(w1, w2)=	cosine of the semantic vector representations of w1 and w2, if both terms exist in the inverted index
	zero, if both terms don't exist in the lucene index

This kernel is used to automatically build a thesaurus of terms. Each entry in the thesaurus consists of a token from the i2b2/VA NLP corpus and N most similar terms based on the distributional semantics. Computing the kernel (K) scales linear in the number of dimensions of each vector and quadratic in the number of term vectors. Computing the thesaurus with a pre-determined number of similar terms from the kernel scales linear to the number of terms. Overall, it scales as linear in the number of dimensions of each vector and quadratic in the number of term vectors. Using the kernel instead of directly computing thesaurus saves computing the cosines (O(N) time complexity) more than once during the construction of present thesaurus and more importantly for multiple values of parameter N.

We are using 1st order Conditional Random Fields (CRF) algorithm as implemented by MALLET (http://mallet.cs.umass.edu).. Since the time complexity of CRF algorithm is $O(L^2*N*M*F*I)$, where L is the number of labels, N is the number sequences (sentences), M is the average length of the sequences, and F is the average number of the features, and I the number of iterations. It is observed[28] that the accuracy is almost the same for all label types such as – IO, IOB and IOBEW. We chose IO notation for labeling for minimizing computational time; thus we use 4 labels Iproblem, Itest and Itreatment. As proposed before, we used all the features used in a state of the art Biomedical NER system known as BANNER28 and try three additional feature types based on: 1) thesaurus, 2) vector representation of the token, and 3) dictionaries.

<u>Semantic features</u>

Dictionary: We used 4 dictionaries compiled from multiple sources for the 3 concepts we are annotating.

SIMFIND: 1) the entries in the thesaurus of each token, and 2) values of the dimensions in the vector representation of the token, also known as word embeddings1.

<u>Pragmatic features</u>

SECTION: the section in which the sentence appears

<u>Syntactic features</u>

POS: the part of speech of the token

DEP: type of dependency/ies with the closed set words

<u>Lexical features</u>

lower case token; lemma; prefixes; suffixes; n-grams; patterns such as beginning with a capital.

Relationship Extraction: We used machine-learning features proposed by Uzuner et al.[14] for a similar task and used MaxEnt classifier as implemented by MALLET.

Results

The results for the system trained on the competition training corpus and tested on the testing corpus, but without any distributional semantics features and that of the system with distributional semantics features are in the below table.

Table 1: The accuracy of baseline (subscript 1) vs. system with distributional semantic features (subscript 2)

Type	R_1	R_2	P_1	P_2	F_1	F_2
Concept exact span	80.4	**81.9**	85.1	**85.9**	82.7	**83.9**
Class exact span	78.1	**79.9**	82.7	**83.7**	80.3	**81.7**
Problem exact span	80.0	**81.7**	83.6	**84.8**	81.8	**83.2**
Treatment exact span	80.2	**81.3**	85.8	**86.4**	82.9	**83.7**
Test exact span	81.3	**82.8**	86.6	**87.0**	83.4	**84.8**
Problem matching class	78.6	**80.4**	81.6	**82.9**	80.0	**81.7**
Treatment matching class	77.0	**78.7**	82.9	**83.9**	79.8	**81.2**
Test matching class	78.6	**80.2**	84.1	**84.8**	81.3	**82.5**
Concept inexact span	89.0	**90.1**	94.2	**94.5**	91.6	**92.3**
Class inexact span	88.9	**89.8**	90.8	**91.6**	89.7	**90.7**
Problem inexact span	89.7	**90.9**	93.8	**94.6**	91.7	**92.7**
Treatment inexact span	88.3	**89.1**	**94.4**	94.3	91.2	**91.7**
Test inexact span	88.8	**90.0**	94.7	**94.7**	91.6	**92.3**
Problem inexact span matching class	87.6	**89.1**	90.9	**91.8**	89.2	**90.4**
Treatment inexact span matching class	84.2	**85.7**	90.6	**91.4**	87.3	**88.4**
Test inexact span matching class	85.1	**86.7**	91.0	**91.5**	87.9	**89.0**

R=Recall, P=Precision, F=F-score.

It is encouraging to see that addition of distributional semantics features increases both the recall and precision. We experimented with different types of distributional semantics based features and all the experiments improved accuracy with the increase highest in the above case where both the local thesaurus (constructed using i2b2/VA NLP corpus) and that constructed using clinical trials are used. We chose 20 similar words for the thesaurus and the model was built using 2000 dimensions and 5 seeds of +1s and -1s. We infer from Bootstrap Resampling [citation] with 1000 repetitions on the test corpus that the improvement because of adding the distributional semantic features is 100.0% significant. However, the improvement because of adding local distributional semantic based features after adding distributional semantic features from clinical trials was insignificant (confidence: 56.6%). Hence, it might be concluded that addition of distributional semantic features using a large unannotated corpus are sufficient and needn’t be supplemented by distributional semantics features from smaller corpora. The system submitted to the competition only had local distributional semantic features and it achieved a recall of 78.7%, precision of 83.2% and f-score of 80.9% for class exact span. We also found that using the dimensions of the vector representations as tokens decreases the accuracy of the system. This could be due to dimensionality curse. Our relationship extraction system achieved a 75.0% recall, precision of 65.1% and f-score of 69.7% for correctly matching relationships.

Conclusion

Overall, the results indicate that distributional semantic features aid clinical concept extraction. The next step would be to use hundreds of thousands of clinical records as unlabeled data from which we can only expect more increase in accuracy as there would be more words of the same type as problem, treatment and test.

Our future goal is to find the utility of distributional semantics for relationship extraction. A sentence or a pattern can be represented in word space as vector sum of the individual terms. Order can be encoded by using the permutational model of Sahlgren. For example: the vector for “hypertension was controlled on hydrochlorothiazide” would be $|\pi 0(\text{hypertension}) + \pi 1(\text{controlled}) + \pi 2(\text{hydrochlorothiazide})|$, where π is a random permutation and || is the L2-Norm. While this appears theoretically sound, empirically permutation model performed suboptimally for large windows since this increases the ratio of seed length and the number of dimensions, thus compromising the conditions of Johnson-LindenStrauss Lemma. Thus, we propose the use of concatenation as a means to encode order in word space. Thus, For n=2, the vector for “hypertension was controlled on hydrochlorothiazide" would be |C(^, hypertension) + C(hypertension, controlled) + C(controlled, hydrochlorothiazide) + C(controlled, hydrochlorothiazide) + C(hydrochlorothiazide, $)|, where ^ and $ are random vectors are assigned to the beginning and end of the sentence and C is the concatenation function.

References

1. Turian J, Opérationnelle R, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. In: *ACL*.Vol 51.; 2010:61801.
2. Sahlgren M, Holst A, Kanerva P. Permutations as a means to encode order in word space. In: *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*; 2008:23-26.
3. Saussure F, Bally C, Sechehaye A, et al. *Cours de linguistique générale*. Payot, Paris; 1922.
4. Minka T. *Discriminative models, not discriminative training*. Microsoft Research; 2005.
5. Sutton C, McCallum A. An Introduction to Conditional Random Fields for Relational Learning. In: *Introduction to statistical relational learning*. Cambridge, Massachusetts, USA: MIT Press; 2007.
6. Hirschman L, Grishman R, Sager N. From text to structured information. In: *Proceedings of the June 7-10, 1976, national computer conference and exposition on - AFIPS '76*. New York, New York; 1976:267.
7. Sager N. Sublanguage Grammers in Science Information Processing. *Journal of the American Society for Information Science*. 1975;26(1):10-16.
8. Friedman C. Towards a comprehensive medical language processing system: methods and issues. In: *AMIA*.; 1997.
9. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *AMIA*.; 2001.
10. Meystre S, Haug PJ. Evaluation of medical problem extraction from electronic clinical documents using MetaMap Transfer (MMTx). *Studies in health technology and informatics*. 2005;116:823.
11. Zeng QT, Goryachev S, Weiss S, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak*. 2006;6:30.
12. Savova GK, Kipper-Schuler KC, Buntrock JD, Chute CG. UIMA-based Clinical Information Extraction System. In: *LREC*.; 2008.
13. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128-144.
14. Uzuner O, Mailoa J, Ryan R, Sibanda T. Semantic relations for problem-oriented medical records. *Artificial Intelligence in Medicine*. 2010.
15. Roberts A, Gaizauskas R, Hepple M, Guo Y. Mining clinical relationships from patient narratives. *BMC Bioinformatics*. 2008;9(Suppl 11):S3.
16. Christensen LM, Harkema H, Haug PJ, Irwin JY, Chapman WW. ONYX: A system for the semantic analysis of clinical text. In: *Proceedings of the Workshop on BioNLP*.; 2009:19–27.
17. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*. 2008;15(1):87–98.
18. Wang X, Hripcsak G, Friedman C. Characterizing environmental and phenotypic associations using information theory and electronic health records. *BMC bioinformatics*. 2009;10(Suppl 9):S13.
19. Zhou X, Han H, Chankai I, Prestrud A, Brooks A. Approaches to text mining for clinical medical records. In: *Symposium on Applied Computing*. Dijon, France: ACM; 2006:235-239.
20. Jonnalagadda S, Leaman R, Cohen T, Gonzalez G. A Distributional Semantics Approach to Simultaneous Recognition of Multiple Classes of Named Entities. In: *Computational Linguistics and Intelligent Text Processing (CICLing)*.Vol 6008/2010. Lecture Notes in Computer Science.; 2010.
21. Ando RK. BioCreative II Gene Mention Tagging System at IBM Watson. In: *Proceedings of the Second BioCreative Challenge*.; 2007.
22. Liu C, Ng HT. Learning Predictive Structures for Semantic Role Labeling of NomBank. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics; 2007:208–215.
23. Cohen T, Widdows D. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*. 2009;42(2):390–405.
24. Widdows D, Ferraro K. Semantic vectors: a scalable open source package and online technology management application. In: *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.; 2008.
25. Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. In: *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.Vol 1036. Citeseer; 2000.
26. Johnson WB, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*. 1984;26(189-206):1-1.1.