Evaluating the Use of Empirically Constructed Lexical Resources for Named Entity Recognition

Siddhartha Jonnalagadda[1], Trevor Cohen[2], Stephen Wu[1], Hongfang Liu[1], Graciela Gonzalez[3]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

[2]School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA

[3]Department of Biomedical Informatics, Arizona State University, Phoenix, AZ, USA

Abstract – Because of privacy concerns and the expense involved in creating an annotated corpus, the existing small annotated corpora might not have sufficient number of examples for statistically learning to extract all the named-entities precisely. In this work, we evaluate what value may lie in automatically generated features based on distributional semantics when using machine-learning named entity recognition (NER). The features we generated and experimented with include n-nearest words, support vector machine (SVM)-regions, and term clustering, all of which are considered semantic (or distributional semantic) features. The addition of n-nearest words feature resulted in a greater increase in F-score than adding a manually constructed lexicon to a baseline system that extracts medical concepts from clinical notes. Although the need for relatively small annotated corpora for retraining is not obviated, lexicons empirically derived from unannotated text can not only supplement manually created lexicons, but replace them. This phenomenon is observed in extracting concepts both from biomedical literature and clinical notes.

Background

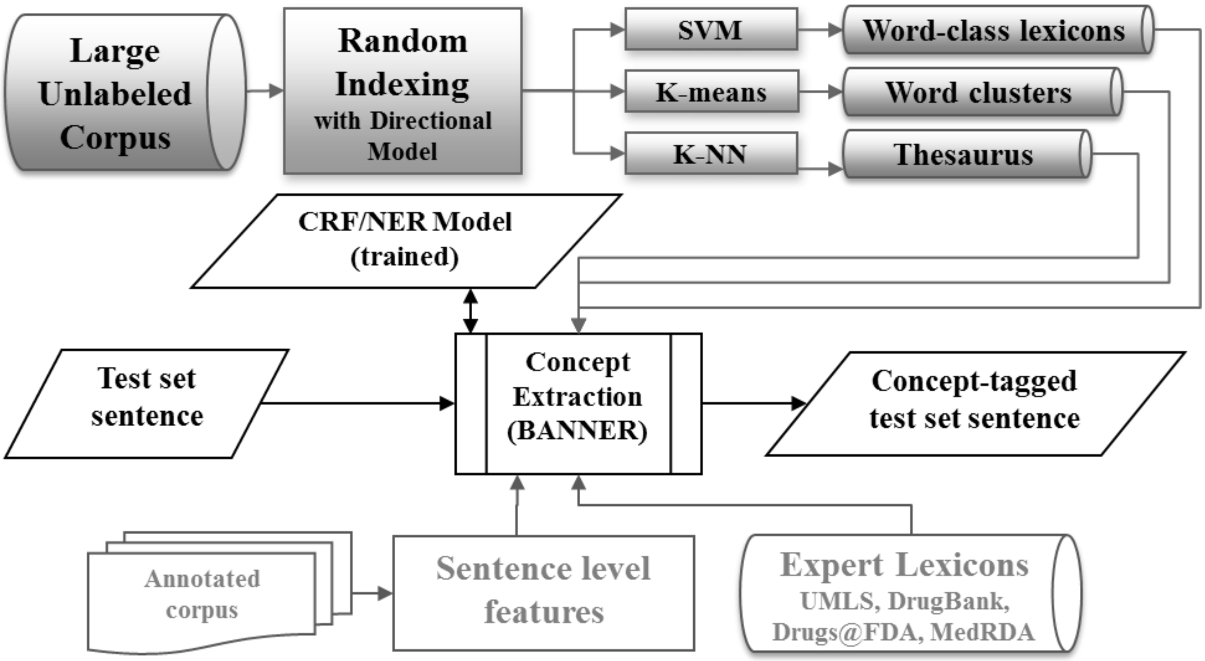
One of the most time-consuming tasks faced by a Natural Language Processing (NLP) researcher or practitioner trying to adapt a machine-learning–based NER system to a different domain is the creation, compilation, and customization of the needed lexicons. Lexical resources, such as lexicons of concept classes are considered necessary to improve the performance of NER. It is typical for medical informatics researchers to implement modularized systems that cannot be generalized (Stanfill et al. 2010). As the work of constructing or customizing lexical resources needed for these highly specific systems is human-intensive, automatic generation is a desirable alternative. It might be possible that empirically created lexical resources might incorporate domain knowledge into a machine-learning NER engine and increase its accuracy.

Although many machine learning–based NER techniques require annotated data, semi-supervised and unsupervised techniques for NER have been long been explored due to their value in domain robustness and minimizing labor costs. Some attempts at automatic knowledgebase construction included automatic thesaurus discovery efforts (Grefenstette 1994), which sought to build lists of similar words without human intervention to aid in query expansion or automatic dictionary construction (Riloff 1996). More recently, the use of empirically derived semantics for NER is used by Finkel and Manning (Finkel and Manning 2009a), Turian et al. (Turian et al. 2010), and Jonnalagadda et al. (Siddhartha Jonnalagadda et al. 2010). Finkel's NER tool uses clusters of terms built apriori from the British National corpus (Aston and Burnard 1998) and English gigaword corpus (Graff et al. 2003) for extracting concepts from newswire text and PubMed abstracts for extracting gene mentions from biomedical literature. Turian et al. (Turian et al. 2010) also showed that statistically created word clusters (P. F. Brown et al. 1992; Clark 2000) could be used to improve named entity recognition. However, only a single feature (cluster membership) can be derived from the clusters. Semantic vector representations of terms had not been previously used for NER or sequential tagging classification tasks before (Turian et al. 2010). Although Jonnalagadda et al. (Siddhartha Jonnalagadda et al. 2010) use empirically derived vector representation for extracting concepts defined in the GENIA (Kim, Ohta, and Tsujii 2008) ontology from biomedical literature using rule-based methods, it was not clear whether such methods could be ported to extract other concepts or incrementally improve the performance of an existing system . This work not only demonstrates how such vector representation could improve state-of-the-art NER, but also that they are more useful than statistical clustering in this context.

Methods

We designed NER systems to identify treatment, tests, and medical problem entities in clinical notes and proteins in biomedical literature. Our systems are trained using 1) sentence-level features using training corpus; 2) a small lexicon created, compiled, and curated by humans for each domain; and 3) distributional semantics features derived from a large unannotated corpus of domain-relevant text. Different models are generated through different combinations of these features. After training for each concept class, a Conditional Random Field (CRF)-based machine-learning model is created to process input sentences using the same set of NLP features. The output is the set of sentences with the concepts tagged. We evaluated the performance of the different models in order to assess the degree to which human-curated lexicons can be substituted by the automatically created list of concepts.

Figure 1: Overall Architecture of the System



The design of the system to identify concepts using machine learning and distributional semantics. The top three components are related to distributional semantics.

The architecture of the system is shown in Figure 1. We first use a state-of-the-art NER algorithm, CRF, as implemented by MALLET (McCallum 2002), that extracts concepts from both clinical notes and biomedical literature using several orthographic and linguistic features derived from respective training corpora. Then, we study the impact on the performance of the baseline after incorporating manual lexical resources and empirically generated lexical resources. The CRF algorithm classifies words according to IOB or IO -like notations (I=inside, O=outside, B=beginning) to determine whether they are part of a description of an entity of interest, such as a treatment or protein. We used four labels for clinical NER— "Iproblem," "Itest," and "Itreatment," respectively, for tokens that were inside a problem, test, or treatment, and "O" if they were outside any clinical concept. For protein tagging, we used the IOB notation, i.e., there are three labels— "Iprotein," "Bprotein," and "O."

Several sentence-level orthographic and linguistic features such as lower-case tokens, lemmas, prefixes, suffixes, n-grams, patterns such as "beginning with a capital letter" and parts of speech are used by the systems to build the NER model and tag the entities in input sentences. This configuration is referred to as MED_noDict for clinical NER and BANNER_noDict for protein tagging.

The UMLS (Humphreys and Lindberg 1993), DrugBank (Wishart et al. 2006), Drugs@FDA (Food 2009), and MedDRA (E. G. Brown, Wood, and Wood 1999) are used to create dictionaries for medical problems, treatments and tests. The guidelines of the i2b2/VA NLP entity extraction task (i2b2 2010) are followed to identify the corresponding UMLS semantic types for each of the three concepts. The other three resources are used to add more terms to our manual lexicon. In an exhaustive evaluation on the nature of the resources by Gurulingappa et al. (Gurulingappa et al. 2010), UMLS and MedDRA were found to be the best resources for extracting information about medical problems among several other resources. For protein tagging, BANNER, one of the best protein-tagging systems (Kabiljo, Clegg, and Shepherd 2009), uses the 344,000 single-word lexicon constructed using the BioCreative II gene normalization training set (Morgan et al. 2008). This configuration is referred to as MED_Dict for clinical NER and as BANNER_Dict for protein tagging.

Distributional Semantic Feature Generation

Here, we implemented automatically generated distributional semantic features based on a semantic vector space model trained from unannotated corpora. This model, referred to as the directional model, uses a sliding window that is moved through the text corpus to generate a reduced-dimensional approximation of a token-token matrix, such that two terms that occur in the context of similar sets of surrounding terms will have similar vector representations after training. As the name suggests, the directional model takes into account the direction in which a word occurs with respect to another by generating a reduced-dimensional approximation of a matrix with two columns for each word, with one column representing the number of occurrences to the left and the other column representing the number of occurrences to the right. The directional model is therefore a form of sliding-window based Random Indexing (Kanerva, Kristoferson, and Holst 2000), and is related to the Hyperspace Analog to Language (Lund and Burgess 1996). Sliding-window Random Indexing models achieve dimension reduction by assigning a reduced-dimensional index vector to each term in a corpus. Index vectors are high dimensional (e.g. dimensionality on the order of 1,000), and are generated by randomly distributing a small number (e.g. on the order of 10) of +1's and -1's across this dimensionality. As the rest of the elements of the index vectors are 0, there is a high probability of index vectors being orthogonal, or close-to-orthogonal to one another. These index vectors are combined to generate context vectors representing the terms within a sliding window that is moved through the corpus. The semantic vector for a token is obtained by adding the contextual vectors gained at each occurrence of the token, which are derived from the index vectors for the other terms it occurs with in the sliding window. The model was built using the open source Semantic Vectors package (Widdows and Cohen 2010).

The performance of distributional models depends on the availability of an appropriate corpus of domain-relevant text. For clinical NER, 447,000 Medline abstracts that are indexed as pertaining to clinical trials are used as the unlabeled corpus. In addition, we have also used clinical notes from Mayo Clinic and University of Texas Health Science Center to understand the impact of the source of unlabeled corpus. For protein NER, 8,955,530 Medline citations in the 2008 baseline release that include an abstract (2010) are used as the large unlabeled corpus. Previous experiments (Siddhartha Jonnalagadda et al. 2012) revealed that using 2000-dimensional vectors, five seeds (number of +1s and –1s in the vector), and a window radius of six is better suited for the task of NER. While a stop-word list is not employed, we have rejected tokens that appear only once in the unlabeled corpus or have more than three nonalphabetical characters.

Quasi-Lexicons of Concept Classes Using SVM

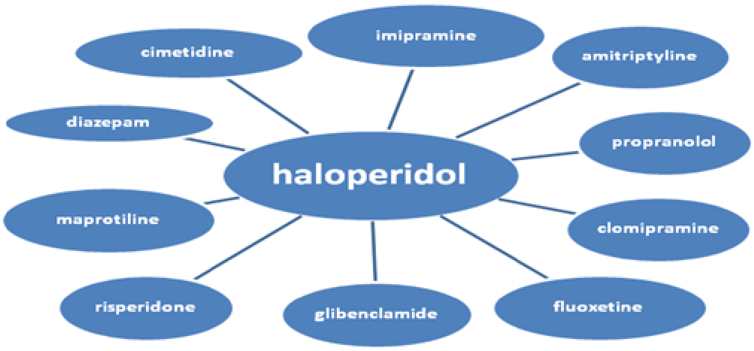
SVM (Cortes and Vapnik 1995) is designed to draw hyper-planes separating two class regions such that they have a maximum margin of separation. Creating the quasi-lexicons (automatically generated word lists) is equivalent to obtaining samples of regions in the distributional hyperspace that contain tokens from the desired (problem, treatment, test and none) semantic types. In clinical NER, each token in training set can belong to either one or more of the classes: problem, treatment, test, or none of these. Each token is labeled as "Iproblem," "Itest," "Itreatment" or "Inone." To remove ambiguity, tokens that belong to more than one category are discarded. Each token has a representation in the distributional hyperspace of 2,000 dimensions. Six (C[4, 2] = 4!/[2!*2!]) binary SVM classifiers are generated for predicting the class of any token among the four possible categories. During the execution of the training and testing phase of the CRF machine-learning algorithm, the class predicted by the SVM classifiers for each token is used as a feature for that token.

Clusters of Distributionally Similar Words Over K-Means

The K-means clustering algorithm (MacQueen 1967) is used to group the tokens in the training corpus into 200 clusters using distributional semantic vectors. The cluster identifier assigned to the target token is used as a feature for the CRF-based system for NER. This feature is similar to the Clark's automatically created clusters (Clark 2000), used by Finkel and Manning (Finkel and Manning 2009b), where the same number of clusters are used. We focused on using features generated from *semantic vectors* as they allow us to also create the other two types of features.

Quasi-Thesaurus of Distributionally Similar Words Using N-Nearest Neighbors

Figure 2: Nearest Tokens to Haloperidol



The closest tokens to haloperidol in the word space are psychiatric drugs. Using the nearest tokens to haloperidol as features, when haloperidol is not a manually compiled lexicon or when the context is unclear, would help to still infer (statistically) that haloperidol is a drug (medical treatment).

Cosine similarity of vectors is used to find the 20 nearest tokens for each token. These nearest tokens are used as features for the respective target token. Figure 2 shows the top few tokens closest in the word space to "haloperidol" to demonstrate how well the semantic vectors are computed. Each of these nearest tokens is used as an additional feature whenever the target token is encountered. Barring evidence from other features, the word "haloperidol" would be classified as belonging to the "medical treatment," "drug," or "psychiatric drug" semantic class based on other words belonging to that class sharing nearest neighbors with it.

Evaluation Strategy

The previous sub-sections detail how the manually created lexicons are compiled and how the empirical lexical resources are generated from semantic vectors (2000 dimensions). In the respective machine learning system for extracting concepts from literature and clinical notes, each manually created lexicon (three for the clinical notes task) contributes one binary feature whose value depends on whether a term surrounding the word is present in the lexicon. Each quasi-lexicon will also contribute one binary feature whose value depends on the output of the SVM classifier discussed before. The distributional semantic clusters together contribute a feature whole value is the id of the cluster the word belongs to. The quasi-thesaurus contributes 20 features that are the 20 distributionally similar words to the word for which features are being generated.

As a gold standard for clinical NER, the fourth i2b2/VA NLP shared-task corpus (i2b2 2010) for extracting concepts of the classes—problems, treatments, and tests—is used. The corpus contains 349 clinical notes as training data and 477 clinical notes as testing data. For protein tagging, the BioCreative II Gene Mention Task (Wilbur, Smith, and Tanabe 2007) corpus is used. The corpus contains 15,000 training set sentences and 5,000 testing set sentences.

Results

Comparison of Different Types of Lexical Resources on Extracting Clinical Concepts

Table 1: Clinical NER: Comparison of SVM-Based Features and Clustering-Based Features With N-Nearest Neighbors–Based Features

Setting	Exact F	Inexact F	Exact Increase	Inexact Increase
MED_Dict	80.3	89.7		
MED_Dict+SVM	80.6	90	0.3	0.3
MED_Dict+NN	81.7	90.9	1.4	1.2
MED_Dict+NN+SVM	81.9	91	1.6	1.3
MED_Dict+CL	80.8	90.1	0.5	0.4
MED_Dict+NN+SVM+CL	81.7	90.9	1.4	1.2

MED_Dict is the baseline, which is a machine-learning clinical NER system with several orthographic and syntactic features, along with features from lexicons such as UMLS, Drugs@FDA, and MedDRA. In MED_Dict+SVM, the quasi-lexicons are also used. In MED_Dict+NN, the quasi-thesaurus is used. In MED_Dict+CL, the clusters automatically generated are used in addition to other features in MED_Dict. Exact F is the F-score for exact match as calculated by the shared task software. Inexact F is the F-score for inexact match or matcing only a part of the other. Exact Increase is the increase in Exact F from previous row. Inexact Increase is the increase in Inexact F from previous row.

Table 1 shows that the F-score of the clinical NER system for exact match increases by 0.3% after adding quasi-lexicons, whereas it increases by 1.4% after adding quasi-thesaurus. The F-score slightly increases further with the use of both these features. The F-score for an inexact match follows a similar pattern. Table 1 also shows that the F-score for an exact match increases by 0.5% after adding clustering-based features, whereas it increases by 1.6% after adding quasi-thesaurus and quasi-lexicons. The F-score slightly decreases with the use of both the features. The F-score for an inexact match follows a similar pattern.

Overall Impact on Extracting Clinical Concepts

Table 2: Clinical NER: Impact of Distributional Semantic Features

Setting	Exact F	Inexact F	Exact Increase	Inexact Increase
MED_noDict	79.4	89.2		
MED_Dict	80.3	89.7	0.9	0.5
MED_noDict+NN+SVM	81.4	90.8	2.0	1.6
MED_Dict+NN+SVM	81.9	91.0	2.5	1.8

MED_noDict is the machine-learning clinical NER system with all the orthographic and syntactic features, but no features from lexicons such as UMLS, Drugs@FDA, and MedDRA. MED_noDict+NN+SVM also has the features generated using SVM and the nearest neighbors algorithm.

Table 2 shows how the F-score increased over the baseline (MED_noDict, which uses various orthographic and syntactic features). After manually constructed lexicon features are added (MED_Dict), it increased by 0.9%. On the other hand, if only distributional semantic features (quasi-thesaurus and quasi-lexicons) were added without using manually constructed lexicon features (MED_noDict+NN+SVM), it increased by 2.0% (*P*<0.001 using Bootstrap Resampling (Noreen 1989) with 1,000 repetitions). It increases only by 0.5% more if the manually constructed lexicon features were used along with distributional semantic features (MED_Dict+NN+SVM). The F-score for an inexact match follows a similar pattern.

Moreover, Table 3 shows that the improvement is consistent even across different concept classes, namely medical problems, tests, and treatments. Each time the distributional semantic features are added, the number of TPs increases, the number of FPs decreases, and the number of FNs decreases.

Impact of the Source of the Unlabeled Data

We utilized three sources for creating the distributional semantics models for NER from i2b2/VA clinical notes corpus. The first source is the set of Medline abstracts indexed as pertaining to clinical trials (447,000 in the 2010 baseline). The second source is the set of 0.8 million clinical notes (half of the total available) from the clinical data warehouse at the School of Biomedical Informatics, University of Texas Health Sciences Center, Houston, Texas (http://www.uthouston.edu/uth-big/clinical-data-warehouse.htm). The third source is the set of 0.8 million randomly chosen clinical notes written by clinicians at Mayo Clinic in Rochester. Table 3 shows the performance of the systems that use each of these sources for creating the distributional semantics features. Each of these systems has a significantly higher F-score than the system that does not use any distributional semantic feature (*P*<0.001 using Bootstrap Resampling (Noreen 1989) with 1,000 repetitions and a difference in F-score of 2.0%). The F-scores of these systems are almost the same (differing by <0.5%).

Table 3: Clinical NER: Impact of the Source of Unlabeled Corpus

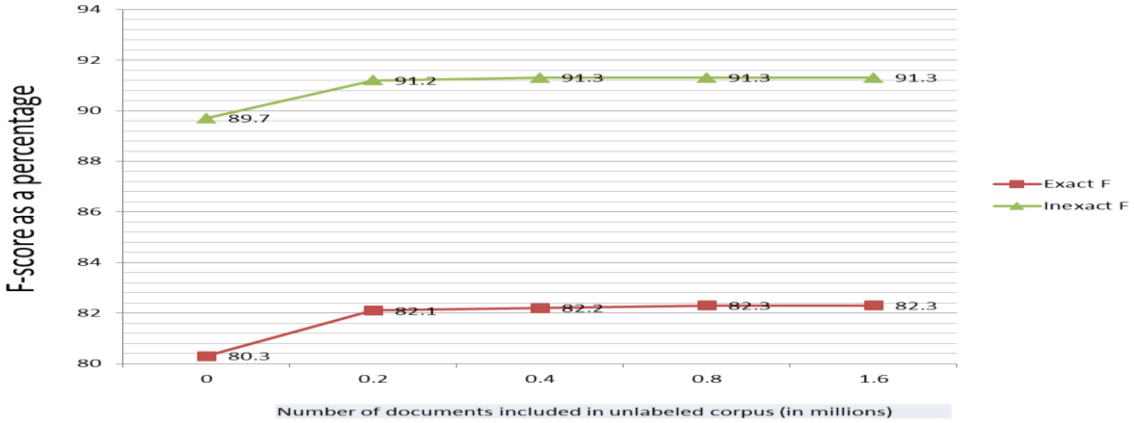
Unlabeled Corpus	Exact F	Inexact F
None	80.3	89.7
Medline	81.9	91.0
UT Houston	82.3	91.3
Mayo	82.0	91.3

None = The machine-learning clinical NER system that does not use any distributional semantic features. Medline = The machine-learning clinical NER system that uses distributional semantic features derived from the Medline abstracts indexed as pertaining to clinical trials. UT Houston = The machine-learning clinical NER system that uses distributional semantic features derived from the notes in the clinical data warehouse at University of Texas Health Sciences Center. Mayo = The machine-learning clinical NER system that uses distributional semantic features derived from the clinical notes of Mayo Clinic, Rochester, MN.

Impact of the Size of the Unlabeled Data

Using the set of 1.6 million clinical notes from the clinical data warehouse at the University of Texas Health Sciences Center as the baseline, we studied the relationship between the size of the unlabeled corpus used and the accuracy achieved. We randomly created subsets of size one-half, one-fourth, and one-eighth the original corpus and measured the respective F-scores. Figure 3 depicts the F-score for exact match and inexact match, suggesting a monotonic relationship with the number of documents used for creating the distributional semantic measures. While there is a leap from not using any unlabeled corpus to using 0.2 million clinical notes, the F-score is relatively constant from there. We might infer that by incrementally adding more documents to the unlabeled corpus, one would be able to determine what size of corpus is sufficient.

Figure 3: Impact of the Size of the Unlabeled Corpus



On the X-axis, N represents the system created using distributional semantic features from N-unlabeled documents. N=0 refers to the system that does not use any distributional semantic feature.

Impact on Extracting Protein Mentions

Table 4: Protein Tagging: Impact of Distributional Semantic Features on BANNER

Rank	Setting	Precision	Recall	F-score	Significance
1	Rank 1 system	88.48	85.97	87.21	6-11
2	Rank 2 system	89.30	84.49	86.83	8-11
3	BANNER_Dict+DistSem	88.25	85.12	86.66	8-11
4	Rank 3 system	84.93	88.28	86.57	8-11
5	BANNER_noDict+DistSem	87.95	85.06	86.48	10-11
6	Rank 4 system	87.27	85.41	86.33	10-11
7	Rank 5 system	85.77	86.80	86.28	10-11
8	Rank 6 system	82.71	89.32	85.89	10-11
9	BANNER_Dict	86.41	84.55	85.47	-
10	Rank 7 system	86.97	82.55	84.70	-
11	BANNER_noDict	85.63	83.10	84.35	-

The significance column indicates which systems are significantly less accurate than the system in the corresponding row. These values are based on the Bootstrap re-sampling calculations performed as part of the evaluation in the BioCreative II shared task (the latest gene or protein tagging task). BANNER_Dict+DistSem is the system that uses both manual and empirical lexical resources. BANNER_noDict+DistSem is the system that uses only empirical lexical resources. BANNER_Dict is the system that uses only manual lexical resources. This is the system available prior to this research, and the baseline for this study. BANNER_noDict is the system that uses neither manual nor empirical lexical resources. BANNER_Dict+DistSem is the system that is significantly more accurate than the baseline. It is equally important to the improvement that the accuracy of BANNER_noDict+DistSem is better than BANNER_noDict. The most significant contribution in terms of research is that an equivalent accuracy (BANNER_noDict+DistSem and BANNER_Dict) could be achieved even without using any manually compiled lexical resources apart from the annotated corpora.

In Table 4, the performance of BANNER with distributional semantic features (row 3) and without distributional semantic features (row 9) is compared with the top ranking systems in the most recent gene-mention task of the BioCreative shared tasks. Each system has an F-score that has a statistically significant comparison (P<0.05) with the teams indicated in the *Significance* column. The significance is estimated using Table 1 in the BioCreative II gene mention task (Wilbur, Smith, and Tanabe 2007). The performance of BANNER with distributional semantic features and no manually constructed lexicon features is better than BANNER with manually constructed lexicon features and no distributional semantic features. This demonstrates again that distributional semantic features (that are generated automatically) are more useful than manually constructed lexicon features (that are usually compiled and cleaned manually) as means to enhance supervised machine learning for NER.

Discussion

The evaluations for clinical NER reveal that the distributional semantic features are better than manually constructed lexicon features. The accuracy further increases when both manually created dictionaries and distributional semantic feature types are used, but the increase is not very significant (P=0.15 using Bootstrap Resampling (Noreen 1989) with 1,000 repetitions). This shows that distributional semantic features could supplement manually built lexicons, but the development of the lexicon, if it does not exist, might not be as critical as previously believed. Further, the N-nearest neighbor (quasi-thesaurus) features are better than SVM-based (quasi-lexicons) features and clustering-based (quasi-clusters) features for improving the accuracy of clinical NER (P<0.001 using Bootstrap Resampling (Noreen 1989) with 1,000 repetitions). For the protein extraction task, the improvement after adding the distributional semantic features to BANNER is also significant (P<0.001 using Bootstrap Resampling (Noreen 1989) with 1,000 repetitions). The absolute ranking of

BANNER with respect to other systems in the BioCreative II task improves from 8 to 3. The F-score of the best system is not significantly better than that of BANNER with distributional semantic features. We again notice that distributional semantic features are more useful than manually constructed lexicon features alone. The purpose of using protein mention extraction in addition to NER from clinical notes is to verify that the methods are generalizable. Hence, we only used the nearest neighbor or quasi-thesaurus features (as the other features contributed little) for protein mention extraction and have not studied the impact of the source or size of the unlabeled data separately. The advantages of our features are that they are independent of the machine-learning system used and can be used to further improve the performance of forthcoming algorithms.

The increment in F-scores after adding manually compiled dictionaries (without distributional semantic features) is only around 1%. However, many NER tools, both in the genomic domain (Leaman and Gonzalez 2008; Torii et al. 2009) and in the clinical domain (Friedman 1997; Savova et al. 2010) use dictionaries. This is partly because systems trained using supervised machine-learning algorithms are often sensitive to the distribution of data, and a model trained on one corpus may perform poorly on those trained from another. For example, Wagholikar (Wagholikar et al. 2012) recently showed that a machine-learning model for NER trained on the i2b2/VA corpus achieved a significantly lower F-score when tested on the Mayo Clinic corpus. Other researchers recently reported this phenomenon for part of speech tagging in clinical domain (Fan et al. 2011). A similar observation was made for the protein-named entity extraction using the GENIA, GENETAG, and AIMED corpora (Wang et al. 2009; Ohta et al. 2009), as well as for protein-protein interaction extraction using the GENIA and AIMED corpora (Siddhartha Jonnalagadda and Gonzalez 2010; S. Jonnalagadda and Gonzalez 2009). The domain knowledge gathered through these semantic features might make the system less sensitive. This work showed that empirically gained semantics are at least as useful for NER as the manually compiled dictionaries. It would be interesting to see if such a drastic decline in performance across different corpora could be countered using distributional semantic features.

Currently, very little difference is observed between using distributional semantic features derived from Medline and unlabeled clinical notes for the task of clinical NER. In the future, we would study the impact using clinical notes related to a specific specialty of medicine. We hypothesize that the distributional semantic features from clinical notes of a subspecialty might be more useful than the corresponding literature. Our current results lack qualitative evaluation. As we repeat the experiments in a subspecialty such as cardiology, we would be able to involve the domain experts in the qualitative analysis of the distributional semantic features and their role in the NER.

Conclusion

Our evaluations using clinical notes and biomedical literature validate that distributional semantic features are useful to obtain domain information automatically, irrespective of the domain, and can reduce the need to create, compile, and clean dictionaries, thereby facilitating the efficient adaptation of NER systems to new application domains. We showed this through analyzing results for NER of four different classes (genes, medical problems, tests, and treatments) of concepts in two domains (biomedical literature and clinical notes). Though the combination of manually constructed lexicon features and distributional semantic features has slightly better performance, suggesting that if a manually constructed lexicon is available, it should be used, the de-novo creation of a lexicon for purpose of NER is not needed.

The distributional semantics model for Medline and the quasi-thesaurus prepared from the i2b2/VA corpus and the clinical NER system's code is available at (http://diego.asu.edu/downloads/AZCCE/) and the updates to the BANNER system are incorporated at http://banner.sourceforge.net/.

Acknowledgments

This work was possible because of funding from possible sources: NLM HHSN276201000031C (PI: Gonzalez), NCRR 3UL1RR024148, NCRR 1RC1RR028254, NSF 0964613 and the Brown Foundation (PI: Bernstam), NSF ABI:0845523, NLM R01LM009959A1 (PI: Liu) and NLM 1K99LM011389 (PI: Jonnalagadda). We also thank the developers of BANNER (http://banner.sourceforge.net/), MALLET (http://mallet.cs.umass.edu/) and Semantic Vectors (http://code.google.com/p/semanticvectors/) for the software packages and the organizers of the i2b2/VA 2010 NLP challenge for sharing the corpus.

References

Aston, G., and L. Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh Univ Pr.

Brown, E G, L Wood, and S Wood. 1999. "The Medical Dictionary for Regulatory Activities (MedDRA)." *Drug Safety: An International Journal of Medical Toxicology and Drug Experience* 20 (2) (February): 109–117.

Brown, P. F, P. V Desouza, R. L Mercer, V. J.D Pietra, and J. C Lai. 1992. "Class-based N-gram Models of Natural Language." *Computational Linguistics* 18 (4): 467–479.

Clark, A. 2000. "Inducing Syntactic Categories by Context Distribution Clustering." In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language learning-Volume 7*, 91–94. Association for Computational Linguistics Morristown, NJ, USA.

Cortes, C., and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20: 273–297.

Fan, Jung-wei, Rashmi Prasad, Rommel M Yabut, Richard M Loomis, Daniel S Zisook, John E Mattison, and Yang Huang. 2011. "Part-of-speech Tagging for Clinical Text: Wall or Bridge Between Institutions?" *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2011: 382–391.

Finkel, J. R., and C. D. Manning. 2009a. "Joint Parsing and Named Entity Recognition." In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on ZZZ*, 326–334. Association for Computational Linguistics.

———. 2009b. "Nested Named Entity Recognition." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Food, U. S. 2009. "Drug Administration. Drugs@ FDA." *FDA Approved Drug Products*. http://www. accessdata.fda. gov/scripts/cder/drugsatfda.

Friedman, Carol. 1997. "Towards a Comprehensive Medical Language Processing System: Methods and Issues." In *AMIA*.

Graff, D., J. Kong, K. Chen, and K. Maeda. 2003. "English Gigaword." *Linguistic Data Consortium, Philadelphia*.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Gurulingappa, H., R. Klinger, M. Hofmann-Apitius, and J. Fluck. 2010. "An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature." In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, 15.

Humphreys, B L, and D A Lindberg. 1993. "The UMLS Project: Making the Conceptual Connection Between Users and the Information They Need." *Bulletin of the Medical Library Association* 81 (2) (April): 170–177.

i2b2. 2010. "Fourth i2b2/VA Shared-Task and Workshop." *Fourth i2b2/VA Shared-Task and Workshop*. https://www.i2b2.org/NLP/Relations.

Jonnalagadda, S., and G. Gonzalez. 2009. "Sentence Simplification Aids Protein-Protein Interaction Extraction." In *Languages in Biology and Medicine*.

Jonnalagadda, Siddhartha, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. "Enhancing Clinical Concept Extraction with Distributional Semantics." *Journal of Biomedical Informatics* 45 (1) (February): 129–140. doi:10.1016/j.jbi.2011.10.007.

Jonnalagadda, Siddhartha, and Graciela Gonzalez. 2010. "BioSimplify: An Open Source Sentence Simplification Engine to Improve Recall in Automatic Biomedical Information Extraction." In *AMIA Annual Symposium Proceedings*.

Jonnalagadda, Siddhartha, Robert Leaman, Trevor Cohen, and Graciela Gonzalez. 2010. "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.

Kabiljo, R., A. B Clegg, and A. J Shepherd. 2009. “A Realistic Assessment of Methods for Extracting Gene/protein Interactions from Free Text.” *BMC Bioinformatics* 10: 233.

Kanerva, P., J. Kristoferson, and A. Holst. 2000. “Random Indexing of Text Samples for Latent Semantic Analysis.” In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Vol. 1036.

Kim, J. D., T. Ohta, and J. Tsujii. 2008. “Corpus Annotation for Mining Biomedical Events from Literature.” *BMC Bioinformatics* 9 (January 8): 10.

Leaman, R., and G. Gonzalez. 2008. “BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition.” In *Pacific Symposium in Bioinformatics*.

Lund, K., and C. Burgess. 1996. “Hyperspace Analog to Language (HAL): A General Model of Semantic Representation.” *Language and Cognitive Processes*.

MacQueen, J. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*.

McCallum, AK. 2002. “MALLET: A Machine Learning for Language Toolkit.” http://mallet.cs.umass.edu.

Morgan, A., Z. Lu, X. Wang, A. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, and J. Hakenberg. 2008. “Overview of BioCreative II Gene Normalization.” *Genome Biology* 9 (Suppl 2): S3.

NLM. 2010. “MEDLINE®/PubMed® Baseline Statistics.” http://www.nlm.nih.gov/bsd/licensee/baselinestats.html.

Noreen, E. W. 1989. *Computer-intensive Methods for Testing Hypotheses: An Introduction*. New York: John Wiley & Sons, Inc.

Ohta, Tomoko, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun’ichi Tsujii. 2009. “Incorporating GENETAG-style Annotation to GENIA Corpus.” In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 106–107. BioNLP ’09. Stroudsburg, PA, USA: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1572364.1572379.

Riloff, Ellen. 1996. “An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains.” *Artif. Intell.* 85 (1-2) (August): 101–134. doi:10.1016/0004-3702(95)00123-9.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. “Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications.” *Journal of the American Medical Informatics Association* 17 (5) (September): 507–513. doi:10.1136/jamia.2009.001560.

Stanfill, Mary H, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. “A Systematic Literature Review of Automated Clinical Coding and Classification Systems.” *Journal of the American Medical Informatics Association: JAMIA* 17 (6) (November 1): 646–651. doi:10.1136/jamia.2009.001024.

Torii, M., Z. Hu, C. H. Wu, and H. Liu. 2009. “BioTagger-GM: A Gene/protein Name Recognition System.” *Journal of the American Medical Informatics Association* 16 (2): 247–255.

Turian, J., R. Opérationnelle, L. Ratinov, and Y. Bengio. 2010. “Word Representations: A Simple and General Method for Semi-supervised Learning.” In *ACL*, 51:61801.

Wagholikar, Kavishwar, Manabu Torii, Siddhartha Jonnala, and Hongfang Liu. 2012. “Feasibility of Pooling Annotated Corpora for Clinical Concept Extraction.” In Vol. Accepted for publication.

Wang, Yue, Jin-Dong Kim, Rune Sætre, Sampo Pyysalo, and Jun’ichi Tsujii. 2009. “Investigating Heterogeneous Protein Annotations Toward Cross-corpora Utilization.” *BMC Bioinformatics* 10 (1): 403. doi:10.1186/1471-2105-10-403.

Widdows, D., and T. Cohen. 2010. “The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics.” In *Fourth IEEE International Conference on Semantic Computing*, 1:43.

Wilbur, J., L. Smith, and T. Tanabe. 2007. “BioCreative 2 Gene Mention Task.” *Proceedings of the Second BioCreative Challenge Workshop* 7-16.

Wishart, David S, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. “DrugBank: a Comprehensive Resource for in Silico Drug Discovery and Exploration.” *Nucleic Acids Research* 34 (Database issue) (January 1): D668–672. doi:10.1093/nar/gkj067.