Size (and Domain) Matters: Evaluating Semantic Word Space Representations for Biomedical Text

Pontus Stenetorp[1] Hubert Soyer[2] Sampo Pyysalo[3]
Sophia Ananiadou[3] and Takashi Chikayama[2]
{[1]Graduate School of Information Science and Technology, [2]School of Engineering}
University of Tokyo, Tokyo, Japan
[3]National Centre for Text Mining and School of Computer Science,
University of Manchester, Manchester, United Kingdom
{pontus,soyerh,smp,chikayama}@logos.ic.i.u-tokyo.ac.jp
sophia.ananiadou@manchester.ac.uk

Abstract

Despite the availability of large corpora of unannotated biomedical scientific texts, domain machine learning-based systems tend to draw only on comparatively small manually annotated corpora. In this work, we explore opportunities to support supervised machine learning through the use of word representations induced from large unannotated corpora. We evaluate a number of established methods extrinsically, by studying the capacity of induced representations to support machine learning-based natural language processing tasks, specifically named entity recognition on three different corpora and semantic category disambiguation on a large automatically acquired corpus. Experiments demonstrate both a clear benefit of many semantic representations on both tasks and all corpora as well as a strong domain dependence, indicating that semantic representations should be induced on documents drawn from the domain relevant to the supervised learning tasks they aim to support. All of the code and resources introduced in this study are freely available from *http://wordreprs.nlplab.org/*

1 Introduction

In biomedical Natural Language Processing (NLP), supervised Machine Learning (ML) methods have been established to achieve state-of-the-art performance at a broad variety of tasks, ranging from Named Entity Recognition (NER) (Smith et al., 2008) to relation (Miwa et al., 2009) and event extraction (Kim et al., 2011). The success of these ML methods depends critically on the quantity and quality of manually annotated data used for training, and consequently annotated corpus resources such as GENIA (Ohta et al., 2002) and GENETAG (Tanabe et al., 2005) have played an important role in the advancement of domain NLP. However, despite the many person-years of expert effort invested into many manually annotated corpora, these annotations cover only a small fraction of the available literature: as of this writing, the PubMed[1] literature database contains over 20 million citations, while even the largest manually annotated corpora reach only into the thousands of annotated documents.

The unannotated texts available in large-scale databases, such as PubMed abstracts and PubMed Central open-access full texts[2], represent obvious opportunities for domain NLP: for example, information on word relatedness derived from unannotated data can help determine the correct treatment of unknown words (Lin et al., 2010). Yet, biomedical domain NLP methods typically forgo unannotated resources in favor of using only annotated corpora, despite the former being orders of magnitude larger. While there are a number of notable exceptions to this trend, these frequently involve dedicated methods and task-specific optimisation with unannotated data, such as the top-ranking submission at the BioCreative II Gene Mention task (Smith et al., 2008), based on the semisupervised Alternating Structure Optimisation (ASO) method (Ando and Zhang, 2005). These approaches are not readily integrated into existing tools and extraction pipelines.

[1] http://www.pubmed.com
[2] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

In this work, we study word representations induced from large unannotated corpora without reference to any single specific task or annotated corpus. The representations can be straightforwardly applied for various NLP tasks simply through the introduction of additional features for each word, and introduce few restrictions on the choice of ML method or general approach. Our study draws in part on the related recent work of Turian et al. (2010), who evaluated the contribution of various word representations to "general-domain" (newswire) NLP tasks. In this work, we focus in particular on the questions 1.) can word representation improve performance at biomedical NLP tasks? and 2.) are representations induced from general-domain corpora sufficient, or are in-domain representations required? We seek to answer these questions by evaluating the applicability of word representations to two supervised ML tasks, Named Entity Recognition and Semantic Category Disambiguation.

2 Approach

There are various possible approaches for acquiring information on the relatedness of words for supervised learning, including the use of manually curated resources such as WordNet (Miller, 1995). Here, we focus in particular on methods that require only unlabeled texts as input. Such approaches have many appealing aspects in terms of e.g. generality and cost of adaptation, in particular in specialised domains such as biomedicine where broad-coverage hand-crafted resources may not be available.

The methods considered in this work all build in some way on the observation formulated by Firth (1957) as "*You shall know a word by the company it keeps*". While they differ substantially in their specifics, the methods share the high-level approach of using the contexts in which words appear in a large corpus of unlabeled text to induce representations that in some way capture (at least) the relatedness of words, broadly based on the assumption that words sharing similar context have similar meaning (the distributional hypothesis). These representations can take various forms, such as word clusters or mappings to semantic spaces such as those induced by Latent Semantic Analysis (LSA) (Dumais et al., 1988).

We apply information on the relatedness of words through the use of a mapping from each word to a vector, the specific characteristics of which vary broadly depending on the method used to induce the mapping: for example, for mappings derived from hard-clustering, each vector will contain only a single non-zero feature identifying the cluster; for mappings derived from LSA, the vectors will be dense. Regardless of the characteristics of such mappings, they can be used uniformly in supervised machine learning tasks taking words as (part of) their input, simply by extending the feature representation with the vector to which each word is mapped.

Due to space constraints, we will not attempt a detailed description of specific approaches here. References to studies introducing the applied methods are given in the methods section; we refer the reader interested in the general category of approaches to the overviews presented by Lin and Wu (2009) and Turian et al. (2010).

3 Methods

3.1 Word representations

For our experiments we draw upon the following previously introduced word representations.

Brown clusters Brown clustering (Brown et al., 1992) is a hierarchical, bigram-based clustering algorithm. It introduces a binary tree on top of the vocabulary, refining information about the similarity of words with each branch. In the resulting representation, each word is associated with a string of binary decisions that lead from the root of that tree to the leaf that the word is assigned to.

The computational cost of the Brown clustering algorithm grows quadratically with respect to the number of clusters, which introduces limitations on the number of clusters that can be introduced.

Google N-gram clusters Lin et al. (2010) introduced a new N-gram corpus from web-scale data similar to that used to create the Google N-gram Corpus (Brants and Franz, 2009). This same work also introduced various derivatives of this data, including a phrase clustering created using higher-order N-grams to determine the contexts in which lower-order N-grams (including single tokens) appear. Further, their work presents clustering using

Name	Method	Domain	Unlabeled data size	Dim.	Size	Introduced by
Brown-news-100	Brown	newswire	63M words	100	4.6MB	
Brown-news-320	Brown	newswire	63M words	320	5.1MB	
Brown-news-1000	Brown	newswire	63M words	1,000	5.7MB	
Brown-news-3200	Brown	newswire	63M words	3,200	6.0MB	Turian et al. (2010)
HLBL-news	HLBL	newswire	63M words	100	395.7MB	
C&W-news-200d-0.1	C&W	newswire	63M words	200	848.1MB	
C&W-news-50d-0.3	C&W	newswire	63M words	50	210.2MB	
Google	K-means	web	1,000,000M words	1,000	327.0MB	Lin et al. (2010)
ClarkNE-bio	Clark-Ney-Essen	biomedical	31M words	45	8.4MB	McClosky et al. (2011)
Brown-bio-100	Brown	biomedical	13M words	100	6.5MB	
Brown-bio-320	Brown	biomedical	13M words	320	7.1MB	This study
Brown-bio-1000	Brown	biomedical	13M words	1,000	7.7MB	

Table 1: Applied word representations. The Dim. column gives the number of clusters for clustering-based representations and the dimensionality of the semantic space for others. (The size given for the Google clusters is for data restricted to single tokens only, size including phrases is 2.6GB.)

distributed K-means with distance defined by the dot product of vectors containing mutual information between the phrase and each of its context words (Lin and Wu, 2009).

Clark-Ney-Essen clusters Clark (2003) considered the task of unsupervised part-of-speech induction and introduced a bigram-based clustering approach incorporating morphological information in a Ney-Essen clustering model (Ney et al., 1994). The approach can be applied to produce a soft clustering, providing the strength of cluster membership for each word. Although initially proposed explicitly in the context of inducing parts-of-speech, both the method[3] and the produced clustering are generic and the clusters need not be exclusively interpreted as parts of speech.

HLBL and CW embeddings The Hierarchical Log-Bilinear embeddings (HLBL) (Mnih and Hinton, 2009) and the Collobert and Weston (2008) (CW) embeddings are distributed word representations. They are low dimensional, real valued vectors with mostly non zero components also referred to as *word embeddings*. The word embeddings are induced using neural network-like language models and while the CW embeddings are inferred directly from the model parameters, the HLBL embeddings are composed by condensing all model representations for all contexts of a given word.

3.2 NER methods

For NER experiments, we apply the publicly available regularized average perceptron-based NER tool of Ratinov and Roth (2009). This choice follows that of Turian et al. (2010), allowing comparison of general-domain and domain-specific results.

Since the tool requires its input to be split into sentences, we initially perform sentence splitting for NER corpora without existing sentence segmentation using the GENIA sentence splitter[4]. We then tokenise the data and add part-of-speech features using the GENIA tagger (Tsuruoka et al., 2005).

3.3 Semantic category disambiguation methods

For the semantic category disambiguation experiments, we use the seed words from McIntosh and Curran (2009) to generate word contexts for a wide array of semantic categories. We then classify these contexts, blinding the seed word, and assigning each a specific semantic category. The seed words were originally used to study semantic drift and the impact of seed choice on bootstrapping performance, but we use them simply to induce our contexts for training and evaluation, relying on them being semantically unambiguous. The main motivation for this choice of set-up is that it enables us to generate a very large amount of training data to study the potential benefits of word representations when the

[3] http://www.cs.rhul.ac.uk/home/alexc/pos2.tar.gz

[4] https://github.com/TsujiiLaboratory/geniass

size of the training data far surpasses that of manually curated corpora.

4 Resources

4.1 Word representations

We apply a broad selection of word representations introduced in previous work as well as a number of newly induced representations created using biomedical domain texts.

First, we consider all the word representations introduced by Turian et al. (2010). These were created on newswire texts (Reuters RCV1 corpus) and include Brown clusters (100, 320, 1,000 and 3,200 clusters), Collobert and Weston embeddings (50 and 200 dimensions) and HLBL embeddings (100 dimensions). These resources are applied without modification.

Second, we apply the clusters introduced by Lin et al. (2010) using very large web data (below, "Google clusters"). For comparability and consistency of processing with the other considered representations, we filter the set of word and phrase clusters introduced by Lin et al. to remove multi-word phrases.

Third, we use the Clark-Ney-Essen (ClarkNE) clusters introduced for feature extraction in the Stanford information extraction system for the BioNLP Shared Task 2011 (McClosky et al., 2011). Unlike the resources above, these clusters were generated on biomedical domain texts, namely PubMed abstracts.

Finally, we create a new set of Brown clusters (100, 320, and 1,000 clusters) using a randomly selected set of PubMed abstracts. This choice of method is motivated both by the availability of an implementation of Brown clustering[5] and the general-domain results of Turian et al. (2010), who showed Brown clusters to outperform the other considered word representations.

Table 1 summarises the word representations considered in this work.

[5]Namely, that of Percy Liang, available from http://www.cs.berkeley.edu/~pliang/software/brown-cluster-1.2.zip. Andriy Mnih generously provided us with the Mnih and Hinton (2009) implementation for creating HLBL embeddings, but due to time constraints we were regrettably not able to include domain-specific embeddings in the current experiments.

	Corpus		
	AnEM	BC2GM	NCBID
Words	91,420	450,991	174,062
Sentences	4,548	20,000	7,844
Entities	3,135	24,596	6,900

Table 2: Statistics of the NER corpora

4.2 NER corpora

The BioCreative II Gene Mention (BC2GM) corpus (Smith et al., 2008), an extension of the GENETAG corpus (Tanabe et al., 2005), is comprised of 20,000 sentences manually annotated for mentions of names of genes, proteins, and related entities such as protein complexes. The BC2GM corpus is a *de facto* standard for both the training and evaluation of ML-based NER methods targeting genes and proteins and has served as training material for various established tools such as BANNER (Leaman et al., 2008).

The Anatomical Entity Mention (AnEM) corpus (Ohta et al., 2012) is a recently introduced corpus consisting of 500 documents, half PubMed abstracts and half full-text extracts, annotated for mentions of anatomical entities (e.g. cells, tissues, and organs). The resource is distributed in two variants, a multi-class version including different entity types (e.g. CELL and TISSUE) and a single-class version. For consistency with the other considered corpora, we apply only the single-class version in this study.

The NCBI disease (NCBID) corpus (Islamaj Dogan and Lu, 2012) is an extension of the Arizona Disease Corpus (AZDC) (Leaman et al., 2009) that extends the sentence-level annotation of AZDC to mark all disease mentions in PubMed abstracts.

Table 2 shows statistics of the corpora used for the NER experiments.

4.3 Semantic category disambiguation data

To generate the dataset used for the semantic category disambiguation experiments, we used a subset of 1,200,000 of the over 20,000,000 citations contained in the PubMed 2012 baseline distribution and randomly separated them into training, development and test sets containing 3/6, 1/6 and 2/6 of the data, respectively. For each set we then extracted contexts for each of the seed words applied by McIn-

Category	Seed words
Antibodies	MAb IgG IgM rituximab infliximab
Cells	RBC HUVEC BAEC VSMC SMC
Cell lines	PC12 CHO HeLa Jurkat COS
Diseases	asthma hepatitis tuberculosis HIV malaria
Drugs	acetylcholine carbachol heparin penicillin tetracycline
Molecular functions	kinase ligase acetyltransferase helicase binding
Mutations and mutants	Leiden C677T C282Y 35delG null
Proteins and genes	p53 actin collagen albumin IL-6
Signs and symptoms	anemia hypertension hyperglycemia fever cough
Tumors	lymphoma sarcoma melanoma neuroblastoma osteosarcoma

Table 3: Semantic categories and seed words from McIntosh and Curran (2009).

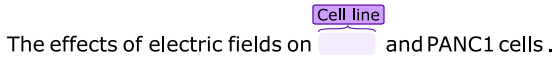

Figure 1: Example context induced by the "Jurkat" seed word for the "Cell lines" semantic category. Note that the seed word has been blinded.

tosh and Curran (2009), shown in Table 3. These seed words were selected by biomedical domain experts working with McIntosh and Curran to be "*as unambiguous as possible with respect to the other [semantic] categories*". We then assigned each context the semantic category label associated with the seed word that generated the given context. See Figure 1 for an example of a blinded induced context. This procedure yielded a total of 428,289 contexts, which dwarfs the size of the training data available for our NER experiments (Table 2).

However, a problem for data induced in this way is a risk of bias towards the semantic categories with the most frequent seed words. For example, in the development set there are 23,045 "Molecular functions" contexts and only 769 "Mutation and mutants" contexts. To remedy this when evaluating our models, we stratified the number of contexts that each seed word would generate, for the test set we took a random sub-sample of at most 150 contexts generated by each seed word which resulted in a more even distribution between the semantic categories.

4.4 Experimental Setup

For the NER experiments, we follow the standard train-test set splits provided with each of the applied corpora, running the NER tool of Ratinov and Roth (2009) with varying word representations but otherwise default parameters. Performance is evaluated in all cases as mention-level precision, recall and F_1-score (the harmonic mean between precision and recall), requiring exact matches between the boundaries of gold and predicted entities.[6]

For the semantic category disambiguation experiments we use the train, development and test sets introduced in this publication. As the task does not involve the concept of false negatives, we use accuracy as our primary performance measure. We use two baseline feature sets, Bag-of-Words (BoW) and Competitive (Comp). BoW is a weak baseline and uses only word features generated from the three word context of the seed word. Comp is a more involved model employing the same context size but using weighted positional word features along with trigrams generated from the context words. In accordance with standard evaluation procedures, the test sets are only used to generate the final results, with initial experiments and statistics generation only being carried out on the training and development sets.

5 Results and Discussion

In the NER experiments, we can see increases in performance over the baseline for most of the newswire-domain word representations (Table 4). The benefits of the word representations are particularly clear for the AnEM dataset, the most sparse out of the three, where the HLBL-news word representations increase performance by over 2 points of F_1-score over the baseline. However, the bio-domain Brown clusters boost performance even further, outperforming the newswire-domain representations for all three data sets. Somewhat surprisingly, the Clark-NE-bio representation shows mixed results when applied alone, falling below the baseline for two of the datasets. The Google clusters generated on web data show good general improvement, but do not attain the level achieved by Brown

[6] We note that this strict evaluation criterion implies our results are not comparable to (and in cases notably lower than) previous studies on these corpora using relaxed matching criteria, but are comparable in terms of evaluation to those of relevant previous studies of word representations.

Model	Dataset: AnEM Pre.	AnEM Rec.	AnEM F_1	BC2GM Pre.	BC2GM Rec.	BC2GM F_1	NCBID Pre.	NCBID Rec.	NCBID F_1	μ Pre.	μ Rec.	μ F_1
Baseline	73.33	45.54	56.19	80.76	75.55	78.07	73.72	63.14	68.02	75.94	61.41	67.43
Brown-news-100	72.79	45.14	55.73	81.68	76.07	78.77	73.13	65.85	69.30	75.86	62.35	67.93
Brown-news-320	71.10	44.27	54.56	81.03	75.55	78.19	73.08	65.52	69.10	75.07	61.78	67.29
Brown-news-1000	71.88	46.82	56.70	81.41	75.65	78.43	73.04	65.36	68.99	75.45	62.61	68.04
Brown-news-3200	70.22	45.62	55.31	82.28	75.86	78.94	72.84	65.69	69.08	75.11	62.39	67.78
Brown-bio-100	72.35	47.29	57.20	82.07	75.75	78.79	74.59	65.50	69.75	76.34	62.85	68.58
Brown-bio-150	70.90	42.28	52.97	81.39	76.90	79.08	74.31	65.66	69.72	75.53	61.61	67.26
Brown-bio-320	72.58	48.25	57.96	82.88	76.07	79.33	73.74	65.72	69.50	76.40	63.34	68.93
Brown-bio-500	73.27	53.90	62.11	**83.18**	76.69	79.81	73.60	66.53	69.88	76.68	65.71	70.60
Brown-bio-1000	72.04	**55.18**	**62.49**	82.38	77.84	80.04	74.77	**66.71**	**70.52**	76.40	**66.58**	71.02
ClarkNE-bio	71.22	41.96	52.81	82.00	75.86	78.81	72.62	63.63	67.83	75.28	60.48	66.48
HLBL-news	**74.36**	48.25	58.52	82.44	76.69	79.46	73.10	65.45	69.06	76.63	63.46	69.02
Google	70.80	54.62	61.66	80.75	**78.15**	79.43	74.58	65.39	69.68	75.38	66.05	70.26
HLBL-news+Brown-news-1000	73.92	48.97	58.91	82.24	76.59	79.31	73.09	65.74	69.22	76.42	63.76	69.15
HLBL-news+Brown-bio-1000	73.54	55.10	62.10	82.62	**78.15**	**80.32**	**74.80**	66.05	70.15	**76.99**	66.43	**71.16**

Table 4: Named Entity Recognition results.

Model	Accuracy
BoW	67.61
Comp	71.59
Comp-Brown-news-100	71.54
Comp-Brown-news-320	71.93
Comp-Brown-news-1000	71.45
Comp-Brown-news-3200	71.42
Comp-Google	**73.70**
Comp-ClarkNE-bio	72.05
Comp-Brown-bio-100	71.73
Comp-Brown-bio-320	72.03
Comp-Brown-bio-1000	72.31

Table 5: Semantic category disambiguation results.

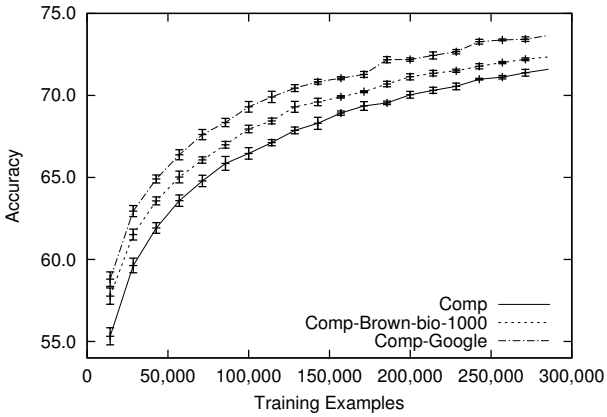

Figure 2: Learning curves for semantic category experiments

clusters generated on in-domain data.[7]

We further considered a small set of the many possible combinations of word representations. These experiments indicated that the HLBL-news+Brown-bio-1000 model performs the best out of the considered models, implying that combination of out-of-domain and in-domain representations can be beneficial. One reason for the success of this combination may be that the HLBL representations were induced from a larger set of unannotated data than any of the in-domain representations.

The semantic category disambiguation experiments show results similar to those for NER, with one major difference. The Google clusters substantially outperform even our in-domain representations (Table 5). One may speculate that this is due to the vastly greater amount of data used to generate these representations, but the question of why a similar effect is not seen for NER remains open. Unlike for our NER experiments, we failed to establish any clear benefit from combining different representations for this task.

From our learning curve (Figure 2) we find that although the amount of training data goes well beyond 100,000 examples we don't see any tendency for the baseline to converge with either of our models enhanced with word representations. This indicates that the benefits of these word representations

[7] We carried out also experiments using the CW representations, but failed to establish any consistent benefit over our baseline using them either alone or in combination with other word representations. For brevity, these results are not shown.

are not restricted to settings where only a comparatively small amount of annotated data is available, and speaks in favour of adopting word representations regardless of the amount of data available for supervised training.

Our evaluation using the newswire-domain representations introduced by Turian et al. (2010) broadly agree with their findings that Brown clusters are surprisingly effective compared to the other representations and that performance tends to improve with a larger number of Brown clusters.[8]

We found impressive performance gains for Google clusters in the semantic class disambiguation task, but a more limited advantage for NER, where the considerably smaller representation using in-domain Brown clusters provided competitive performance. This result is encouraging for the feasibility of achieving the best level of performance in a practical system, as the size of the Google data can make distribution and training challenging.[9]

6 Related Work

As described in the introduction, our work is in many ways closely related to the "general-domain" studies of Lin and Wu (2009) and Turian et al. (2010), who demonstrated significant benefits from their respective approaches for "general-domain" NLP tasks such as the CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003).

Recently, a number of studies have considered use of information derived from large unannotated corpora also in various biomedical domain NLP tasks. In the BioNLP Shared Task 2011, in addition to McClosky et al. (2011), word clusters were applied in support of information extraction systems also by the MSR-NLP (Quirk et al., 2011) and VIBGhent (Van Landeghem et al., 2011) teams. The clustering approach of Clark (2003) was also previously considered for biomedical NER by Finkel and Manning (2009). A new dedicated algorithm for NER using distributional semantics was recently proposed by Jonnalagadda et al. (2010).

[8] Although time constraints prevented us from completing a set of 3,200 Brown clusters, we will allow this computation to complete and release this set with the other resources introduced in this study.

[9] During NER system training with Google clusters, memory usage approached 100GB for some datasets.

While neither the comparative evaluation of word representations nor the application of models derived from unsupervised data in support of biomedical domain NER is novel by itself, this is to the best of our knowledge the first effort to systematically explore the benefits of multiple approaches for inducing word representations from unannotated data to biomedical domain NLP tasks as well as of their domain-dependence.

7 Conclusions and Future Work

We have presented an evaluation of the effectiveness of various word representations in support of biomedical domain NLP, finding that word representations can realize substantial benefits both for entity recognition and classification tasks and that representations induced on in-domain texts show greater and more consistent benefits than comparable representations induced on out-of-domain texts.

As an initial study, our work leaves open various opportunities for future study. In future work, we will aim to assess the contribution of in-domain HLBL and CW word embeddings and their combinations with other representations. We will also consider the effect of the amount of unlabeled data used to induce word representations on performance.

Code and resources introduced in this study are freely available from *http://wordreprs.nlplab.org/*

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable feedback.

This work was funded in part by UK Biotechnology and Biological Sciences Research Council (BBSRC) under project Automated Biological Event Extraction from the Literature for Drug Discovery (reference number: BB/G013160/1), by the Ministry of Education, Culture, Sports, Science and Technology of Japan under the Integrated Database Project and by the Swedish Royal Academy of Sciences.

References

R.K. Ando and T. Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853.

T. Brants and A. Franz. 2009. The Google Web 1T 5-gram Corpus version 1.1. LDC2006T13.

P.F. Brown, P.V. Desouza, R.L. Mercer, V.J. Della Pietra, and J.C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

A. Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL*, pages 59–66.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML 2008*, pages 160–167.

S.T. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

J.R. Finkel and C.D. Manning. 2009. Nested named entity recognition. In *Proceedings of EMNLP 2009*, pages 141–150.

J. Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*.

R. Islamaj Dogan and Z. Lu. 2012. An improved corpus of disease mentions in PubMed citations. In *Proceedings of BioNLP 2012*, pages 91–99.

S. Jonnalagadda, R. Leaman, T. Cohen, and G. Gonzalez. 2010. A distributional semantics approach to simultaneous recognition of multiple classes of named entities. *Computational Linguistics and Intelligent Text Processing*, pages 224–235.

J-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Shared Task*, pages 1–6.

R. Leaman, G. Gonzalez, et al. 2008. BANNER: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663.

R. Leaman, C. Miller, and G. Gonzalez. 2009. Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark. In *Proceedings of LBM 2009*.

D. Lin and X. Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of ACL-IJCNLP 2009*, pages 1030–1038.

D. Lin, K. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, et al. 2010. New tools for web-scale n-grams. In *Proceedings of LREC*.

D. McClosky, M. Surdeanu, and C. Manning. 2011. Event extraction as dependency parsing for BioNLP 2011. In *Proceedings of BioNLP Shared Task 2011*, pages 41–45.

T. McIntosh and J.R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of ACL 2009*, pages 396–404.

G.A. Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii. 2009. Protein–protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12):e39–e46.

A. Mnih and G.E. Hinton. 2009. A scalable hierarchical distributed language model. *NIPS*, 21:1081–1088.

H. Ney, U. Essen, and R. Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1):1–38.

T. Ohta, Y. Tateisi, and J-D. Kim. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT 2002*, pages 82–86.

T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of DSSD 2012*.

C. Quirk, P. Choudhury, M. Gamon, and L. Vanderwende. 2011. MSR-NLP entry in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task*, pages 155–163.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL 2009*, pages 147–155.

L. Smith, L.K. Tanabe, R.J. Ando, C.J. Kuo, I.F. Chung, C.N. Hsu, Y.S. Lin, R. Klinger, C.M. Friedrich, K. Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2):S2.

L. Tanabe, N. Xie, L. Thom, W. Matten, and W.J. Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(Suppl 1):S3.

E.F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent named entity recognition. In *Proceedings of HLT-NAACL 2003*, pages 142–147.

Y. Tsuruoka, Y. Tateishi, J-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

S. Van Landeghem, T. Abeel, B. De Baets, and Y. Van de Peer. 2011. Detecting entity relations as a supporting task for bio-molecular event extraction. In *Proceedings of BioNLP Shared Task*, pages 147–148.