Clinical Information Extraction Using Word Representations

Shervin Malmasi ♣ **Hamed Hassanzadeh** ♢ **Mark Dras** ♣

♣ Centre for Language Technology, Macquarie University, Sydney, NSW, Australia
♢ School of ITEE, The University of Queensland, Brisbane, QLD, Australia

shervin.malmasi@mq.edu.au, h.hassanzadeh@uq.edu.au
mark.dras@mq.edu.au

Abstract

A central task in clinical information extraction is the classification of sentences to identify key information in publications, such as intervention and outcomes. Surface tokens and part-of-speech tags have been the most commonly used feature types for this task. In this paper we evaluate the use of word representations, induced from approximately 100m tokens of unlabelled in-domain data, as a form of semi-supervised learning for this task. We take an approach based on unsupervised word clusters, using the Brown clustering algorithm, with results showing that this method outperforms the standard features. We inspect the induced word representations and the resulting discriminative model features to gain further insights about this approach.

1 Introduction

Evidence-based Medicine (EBM) is an approach to enhance clinical decision making by leveraging currently available evidence. The rationale behind EBM is that clinicians can make more judicious decisions with access to abundant clinical evidence about a particular medical case. This evidence is sourced from research outcomes which can be found in medical publications accessible via online repositories such as PubMed.[1] Although millions of publications are available, finding the most relevant ones is cumbersome using current search technology. Additionally, the rapid growth of research output makes manual analysis and synthesis of search results unfeasible. This has given rise to the need for methods to automatically extract relevant information from publications to support automatic summarization (Hassanzadeh et al., 2015). This is an emerging research area that has begun to attract increasing attention (Summerscales et al., 2011).

This information extraction is generally performed at the sentence level on the paper abstracts (Verbeke et al., 2012). Scholarly publications usually follow a common rhetorical structure that first defines the problem and research aims by introducing background information. They then describe the methodology and finally the outcomes of the research are presented. Abstracts, as the summary of the reported research, generally have the same structure. This information, which can be considered as *scientific artefacts*, can usually be found in the form of whole sentences within the abstracts. More specifically, the artefacts in the clinical research domain have been categorized as Intervention, Population or Problem, Comparison, and Outcome. This is known as the *PICO* scheme (Richardson et al., 1995). Another proposed approach to formalise the rhetorical structure of medical abstracts is the PIBOSO model (Kim et al., 2011), a refined version of the PICO criteria. It contains six classes, rather than four: (i) POPULATION: the group of individuals participating in a study; (ii) INTERVENTION: the act of interfering with a condition to modify it or with a process to change its course; (iii) BACKGROUND: material that places the current study in perspective, *e.g.* work that preceded the current study; information about disease prevalence; etc.; (iv) OUTCOME: a summarisation of the consequences of an intervention; (v) STUDY DESIGN: the type of study that is being described; and (vi) OTHER: other information in the publication.

By comparing these artefacts across publications clinicians can track the evolution of treatments and empirical evidence, allowing them to employ it in their decision making. However, finding and identifying these artefacts is a barrier. To facilitate this process, various approaches have

[1] http://www.ncbi.nlm.nih.gov/pubmed

Shervin Malmasi, Hamed Hassanzadeh and Mark Dras. 2015. Clinical Information Extraction Using Word Representations . In *Proceedings of Australasian Language Technology Association Workshop*, pages 66−74.

been devised to automatically recognise these scientific artefacts in publications (Hassanzadeh et al., 2014a). The most common approach, as discussed in §2, is the use of supervised learning to classify sentences into the various categories.

Separately, another recent trend in Natural Language Processing (NLP) has been the use of word representations to integrate large amounts of unlabelled data into such supervised tasks, a form of semi-supervised learning (Turian et al., 2010). This is something that has not been applied to scientific artefacts extraction.

Accordingly, the primary aim of the present work is to draw together the two areas, evaluating the utility of word representations for this task and comparing them against the most commonly used features to see if they can enhance accuracy. A secondary goal is to inspect the induced word representations and the resulting discriminative models to gain further insights about this approach.

The paper is structured as follows. We present related work on biomedical information extraction in §2. Word representations are introduced in §3 along with our unlabelled data and clustering method. The experimental setup is outlined in §4 followed by results in §5. In §6 we analyze the most discriminative features of our model and in §7 we present a brief error analysis. Finally, we conclude with a discussion in §8.

2 Related Work

The approaches for classifying scientific artefacts vary from having very coarse grained models of these artefacts, such as, publication zone/section identification (Teufel, 2000), to more fine grained ones, such as, sentence classification (Kim et al., 2011; Liakata et al., 2012). In this section, we review the literature that has a similar perspective as ours, that is, sentence-level classification.

Kim et al. (2011) perform classification in two steps using PIBOSO scheme. In the first step, a classifier identifies the sentences that contain PIBOSO concepts, while in the second step, a different classifier assigns PIBOSO classes to these sentences. The annotation is performed at the sentence level and one sentence may have more than one class (*i.e.* multi-label classification). They also employ a Conditional Random Field (CRF) as their classifier using features derived from the context, semantic relations, structure and the sequence of sentences in the text. Domain-specific information is obtained via Metamap. Their final feature vector includes a combination of: bag-of-words, bigrams, part-of-speech (POS) tags, semantic information, section headings, sentence position, and windowed features of the previous sentences.

Verbeke et al. (Verbeke et al., 2012), on the other hand, apply a statistical relational learning approach using a kernel-based learning (kLog) framework to perform classification using the NICTA-PIBOSO corpus. They exploit the relational and background knowledge in abstracts, but take into account only the sequential information at word level. More concretely, their feature set includes a sequence of class labels of the four previous sentences as well as of the two following ones, the lemma of the dependency root of the current sentence and the previous sentence, the position of the sentence, and the section information.

Finally, Sarker et al. (2013) use a set of binary Support Vector Machine (SVM) classifiers in conjunction with feature sets customised for each classification task to attain the same goal. Using the same NICTA-PIBOSO corpus, they use MetaMap to extract medical concepts, and in particular UMLS Concept Unique Identifiers (CUIs) and Semantic Types, to be then considered as domain-specific semantic features. The rest of the features they employ consist of n-grams, POS tags, section headings, relative and absolute sentence positions and sequential features adapted from Kim et al. (2011), as well as class-specific features for the POPULATION class. Similar to our approach, they use an SVM classifier.

A key commonality of previous research is that lexical features and POS tags constitute a set of core features that are almost always used for this task. Although some approaches have applied different external resources, from generic dictionaries such as WordNet to domain specific ontologies, no attempt has been made to leverage large-scale unlabelled data. The main aim of this work is to evaluate the feasibility of such an approach.

3 Word Representations

Word representations are mathematical objects associated with words. This representation is often, but not always, a vector where each dimension is a *word feature* (Turian et al., 2010). Various methods for inducing word representations have been proposed. These include *distributional* represen-

tations, such as LSA, LSI and LDA, as well as *distributed* representations, also known as *word embeddings*. Yet another type of representation is based on inducing a clustering over words, with Brown clustering (Brown et al., 1992) being the most well known method. This is the approach that we take in the present study.

Recent work has demonstrated that unsupervised word representations induced from large unlabelled data can be used to improve supervised tasks, a type of semi-supervised learning. Examples of tasks where this has been applied include dependency parsing (Koo et al., 2008), Named Entity Recognition (NER) (Miller et al., 2004), sentiment analysis (Maas et al., 2011) and chunking (Turian et al., 2010). Such an approach could also be applied to the clinical information extraction task where although we only have a very limited amount of labelled data, large-scale unlabelled data — hundreds of millions of tokens — is readily available to us.

Researchers have noted a number of advantages to using word representations in supervised learning tasks. They produce substantially more compact models compared to fully *lexicalized* approaches where feature vectors have the same length as the entire vocabulary and suffer from sparsity. They better estimate the values for words that are rare or unseen in the training data. During testing, they can handle words that do not appear in the labelled training data but are observed in the test data and unlabelled data used to induce word representations. Finally, once induced, word representations are model-agnostic and can be shared between researchers and easily incorporated into an existing supervised learning system.

3.1 Brown Clustering

We use the Brown clustering algorithm (Brown et al., 1992) to induce our word representations. This method partitions words into a set of c classes which are arranged hierarchically. This is done through greedy agglomerative merges which optimize the likelihood of a hidden Markov model which assigns each lexical type to a single class. Brown clusters have been successfully used in tasks such as POS tagging (Owoputi et al., 2013) and chunking (Turian et al., 2010). They have been successfully applied in supervised learning tasks (Miller et al., 2004) and thus we also adopt their use here.

3.2 Unlabelled Data

To obtain suitable unlabelled data, we followed two strategies to retrieve data from the PubMed repository: *(1)* based on user-defined clinical inquiries, and *(2)* using a generic query. In the first strategy we employed 456 clinical queries from the EBMSummariser corpus (Mollá and Santiago-martinez, 2011). The inquiries in this corpus are collected from the Clinical Inquiries section of the Journal of Family Practice.[2] This section of the journal contains a number of queries submitted by the users and their evidence-based answers by medical experts. We queried PubMed with these 456 inquiries and retrieved the results using their PM-IDs (*i.e.* PubMed's unique identifiers) via PubMed's eUtils API.[3] In total, 212,393 abstracts were retrieved, of which 22,873 abstracts did not contain valid text, leaving 189,520.

For the second retrieval strategy, we queried PubMed with the term *Randomised Controlled Trial*. This results in retrieving publications presenting medical cases and providing evidence (*i.e.* desirable for EBM practice). PubMed returned 491,357 results for this query. After removing duplicate results, *i.e.* those retrieved in the first strategy, we downloaded 200,000 abstracts. After removing empty abstracts, 171,662 remained.

The text of each abstract was extracted by parsing the PubMed XML file and it was then segmented into sentences; each sentence was then tokenized and lowercased. This resulted in a total of 96 million tokens across 3.7 million sentences, with 873k unique tokens.[4]

We next induced Brown clusters using this data. Five runs with clusters of size 100, 200, 300, 1000 and 3000 were performed for comparison purposes.

3.3 Clustering Results

We now turn to a brief analysis of the clustering results. Table 1 shows examples of both generic and domain-specific clusters taken from the run with 3,000 clusters. We observe that words were clustered according to both their semantic and grammatical properties, with some clusters containing highly domain-specific entries. These results show that the word clusters are very effective at captur-

[2] http://jfponline.com/articles/clinical-inquiries.html

[3] http://www.ncbi.nlm.nih.gov/books/NBK25497/

[4] We also note that this data has a much higher type-token ratio compared to other domains such as newswire text, indicating greater lexical variation in this domain.

Cluster Path	Top Words
00100111	article paper manuscript chapter commentary essay
001011011010	observations investigations evidences facts explorations
1000000011	evaluating investigating examining exploring
111010111011100	suggests indicates implies posits asserts contends
111010111011101	shows demonstrates reveals confirms concludes argues establishes assumes finds
1111011011001	mg/dl mmhg kg/m2 bpm beats/min u/ml mmol/mol
001111000101	antibiotics analgesics opioids antimicrobials placebos antihypertensives
11000100100	reconstruction dissection ligation instrumentation
010111101011110	oncology cardiology rheumatology psychiatry urology dermatology radiology
010111100011010	vaccination immunization inoculation immunisation immunizations revaccination

Table 1: Some example clusters and their top words (by frequency). Examples include both generic (top) and domain-specific (bottom) clusters.

ing lexical knowledge and organizing it by syntactic function. We will examine the cluster contents again in §6 as part of our feature analysis. We make these unsupervised clusters available for viewing or download from our website.[5]

4 Experimental Setup

We take a supervised classification approach, comparing previously used features against the unsupervised Brown cluster features.

As the primary focus of this work is the evaluation of word representation features, we limit the scope of our experiment in two ways: (1) we do not attempt multi-label classification, as explained in §4.1 and (2) we do not use sentence sequence information, as outlined in §4.2. These conditions allow us to focus on systematically comparing feature types in a controlled manner.

4.1 Data

We use the NICTA-PIBOSO corpus (Kim et al., 2011) in this experiment. Here each sentence is labelled with one or more classes, making it a multi-label classification task. Table 2 lists a breakdown of the per-class sentence statistics, showing that 9% of the sentences have more than one label. The multi-label characteristic of instances as well as imbalanced distribution of classes are two most common issues of many corpora in biomedical scientific artefacts classification task (Hassanzadeh et al., 2014b). As the scope of our work is limited to evaluating word representation features, we simplify our setup by excluding the multi-label instances, thus reducing the task to a multi-class classification one. This avoids the use of multi-label evaluation metrics, making it easier to draw direct comparisons between the performance of the standard features and the word representations. The sentences were tokenized in a preprocessing step.

	All	Multi-label
BACKGROUND	2,557	160 (6%)
INTERVENTION	690	350 (51%)
OUTCOME	4,523	71 (2%)
POPULATION	812	412 (51%)
STUDY DESIGN	228	114 (50%)
OTHER	3,396	0 (0%)
Total	12,206	1,107 (9%)

Table 2: Sentence counts in the NICTA-PIBOSO corpus. The multi-label column lists the number of sentences annotated with more than one label.

4.2 Classifier

We use a linear SVM to perform multi-class classification. In particular, we use the LIBLINEAR[6] package (Fan et al., 2008) which has been shown to be efficient for highly-dimensional text classification problems such as this (Malmasi and Dras, 2014; Malmasi and Dras, 2015b; Malmasi and Dras, 2015a).

Previous work (see §2) shows that CRF classifiers perform well for this task, exploiting the sequential structure of abstracts. As our aim is to evaluate the effectiveness of *intrinsic* word representation features we focus on the classification of individual sentences and do not use *extrinsic* features, *i.e.* the contents or predicted labels of preceding sentences in an abstract. In practice this means that the sentences are being classified independently.

[5] http://web.science.mq.edu.au/%7Esmalmasi/data/med3k/

[6] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

4.3 Features

We compare our proposed word representation features against the most commonly used features for this task, which we describe here.

Word n-grams Surface tokens are the most commonly employed feature type in this task using both bag-of-words (unigram) and n-grams. The length of the feature vector equals that of the vocabulary; n-gram vocabulary grows exponentially. We extracted word n-grams of order 1–3.

Part-of-Speech n-grams POS tags are another frequently used feature type and capture the syntactic differences between the different classes.[7] We tagged the sentences using the Stanford Tagger, which uses the Penn Treebank tagset containing 36 tags, and extracted n-grams of order 1–3.

Brown Cluster Features Brown clusters are arranged hierarchically in a binary tree where each cluster is identified by a bitstring of length ≤ 16 that represents its unique tree path. The bitstring associated with each word can be used as a feature in discriminative models, Additionally, previous work often also uses a p-length prefix of this bitstring as a feature. When p is smaller than the bitstring's length, the prefix represents an ancestor node in the binary tree and this superset includes all words below that node. We follow the same approach here, using all prefix lengths $p \in \{2, 4, 6, \ldots, 16\}$. Using the prefix features in this way enables the use of cluster supersets as features and has been found to be effective in other tasks (Owoputi et al., 2013). Each word in a sentence is assigned to a Brown cluster and the features are extracted from this cluster's bitstring.

4.4 Evaluation

We report our results as classification accuracy under k-fold cross-validation, with $k = 10$. These results are compared against a majority baseline and an oracle. The oracle considers the predictions by all the classifiers in Table 3 and will assign the correct class label for an instance if at least one of the the classifiers produces the correct label for that data point. This approach can help us quantify the *potential* upper limit of a classification system's performance on the given data and features (Malmasi et al., 2015).

[7] *e.g.* Our own analysis showed that OUTCOME sentences contained substantially more past tense verbs, comparative adverbs and comparative adjectives.

Feature	Accuracy (%)
Majority Baseline	40.1
Oracle	92.5
Part-of-Speech unigrams	64.6
Part-of-Speech bigrams	68.6
Part-of-Speech trigrams	67.4
Word unigrams	73.3
Word bigrams	66.0
Word trigrams	49.7
Brown (100 clusters)	70.4
Brown (200 clusters)	72.8
Brown (300 clusters)	74.3
Brown (1000 clusters)	74.8
Brown (1000 clusters) bigrams	73.9
Brown (3000 clusters)	**75.6**
Brown (3000 clusters) bigrams	74.9
Brown (3000 clusters) trigrams	70.7

Table 3: Sentence classification accuracy results for the features used in this study.

5 Results

The results for all of our experiments are listed in Table 3. All features performed substantially higher than the baseline. We first tested the POS n-gram features, with bigrams providing the best result of 68.6% accuracy and performance dropping with trigrams. Word n-grams were tested next, with unigrams achieving the best result of 73.3%. Unlike the POS features, word feature performance does not increase with bigrams.

Finally, the Brown cluster features were tested using clusters induced from the five runs of different cluster different sizes. Accuracy increases with the number of clusters; 200 clusters match the performance of the raw unigram features and the largest cluster of size 3000 yields the best result of 75.6%, coming within 17% of the oracle accuracy of 92.5%. Another variation tested was Brown cluster n-grams. Although they outperformed their word n-gram counterparts, they did not provide any improvement over the standard Brown features.

In sum, these results show that Brown clusters, using far fewer features, can outperform the widely used word features.

Class	Clusters of words
BACKGROUND	[have has had] — [describes presents examines discusses summarizes addresses] [objectives goal] — [emerged evolved attracted fallen arisen risen proliferated]
INTERVENTION	[received underwent undergoing taking] — [gel cream spray ointment] [orally intravenously subcutaneously intramuscularly topically intraperitoneally] [mg/kg mg/kg/day g/kg ml/kg μg/kg mg/kg/d microg/kg μg/kg]
POPULATION	[identified enrolled recruited contacted] — [aged] — [randomly] [twenty thirty forty sixty fifty eighty thirty-two twenty-eight . . .]
OUTCOME	[revealed showed suggests indicates implies] — [found observed noted noticed] [significantly] — [p n r r2] — [demonstrate indicate imply] [0.002 0.003 0.004 0.006 .02 0.008 0.007 .03 0.009 .04 . . .]
STUDY DESIGN	[cross-sectional case-control quasi-experimental sectional mixed-methods case-crossover case-controlled . . .] — [randomised randomized-controlled]
OTHER	[include] — [evaluate assess] — [obtained] — [measured] [articles papers publications literatures manuscripts]

Table 4: Some highly-weighted clusters associated with the NICTA-PIBOSO classes. Each cluster is a single feature in the model, but we have expanded them here to include their constituent words.

6 Feature Analysis

In this section we analyze some of the discriminative features in our model to gain better insight about the knowledge being captured by our models and the task in general. This was done by ranking the features according to the weights assigned by the SVM model. In this manner, SVMs have been successfully applied in data mining and knowledge discovery tasks such as identifying discriminant cancer genes (Guyon et al., 2002).

Table 4 lists several highly weighted Brown clusters for each of our classes. Although each cluster is a single feature in the model, we have expanded the clusters here to include their constituent words.

The BACKGROUND class is associated with words that are quite common in the introductory rhetoric of scientific publications. These are descriptive of the current and previous research, and are mostly in the present/past perfect tense.

The INTERVENTION class is mostly associated with clusters that include clinical vocabulary, including verbs such as *received*, *underwent* and *taking*; medication-related nouns like *gel* or *ointment*; dosage descriptors such as *mg/kg* and *mg/kg/day*; and adverbs describing the route of administration, for example *orally* and *intravenously*.

For POPULATION sentences, numerical quantities, likely relating to the number of participants,[8] as well as verbs that are related to participation, are very frequent.

Similarly, reporting verbs are more likely to occur in OUTCOME sentences. They are organized into different clusters according to their syntactic and semantic function. In addition, we also note that a cluster of decimal numbers is also common. These numbers are used in the sentences to report study results, including those from various statistical tests. This is accompanied by another cluster containing relevant tokens for reporting statistics, *e.g.* "p", "r", and "n" which could refer to "p-value", "Pearson correlation" and "number", respectively.

Overall, it can be seen that the clusters associated with the features are logical. Furthermore, these examples underline the clustering method's effectiveness, enabling us to encode a wide range of similar tokens (*e.g.* decimal values or dosage

[8]These are mostly spelled out as they appear at the start of a sentence.

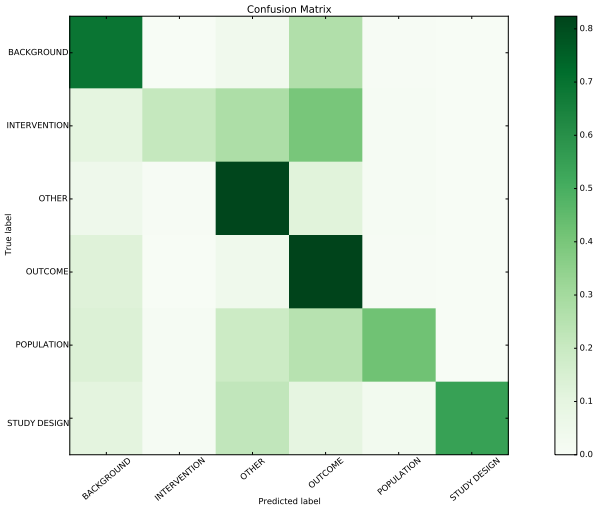

Figure 1: Normalized confusion matrix for results using Brown features (3000 clusters). The values are normalized due to the class size imbalance.

amounts) under a single cluster feature. This provides a substantial reduction in the feature space without the loss of information.

7 Error Analysis

We now turn to an analysis of the errors being committed by the classifier. The error distribution is illustrated by the confusion matrix in Figure 1. We note that the two largest classes, OUTCOME and OTHER, are the most correctly classified. Conversely, INTERVENTION sentences are highly misclassified and mostly confused for OUTCOME. To better understand these errors we segregated the subset of misclassified instances for analysis. Table 5 lists a number of these sentences from highly confused classes.

Our analysis suggests that the occurrences of similar domain-specific terminologies in both types of sentences, in INTERVENTION sentences as the explanation of the methodologies, and restating them in OUTCOME sentences in order to describe the effects of those methodologies, can be a reason for this confusion.

There is also some confusion between BACKGROUND and OUTCOME instances. Both of these classes commonly describe some challenges and findings of either previous studies (*i.e.* BACKGROUND sentences) or the current reporting study (*i.e.* OUTCOME). This narrative characteristic of these classes has similar rhetorical and linguistic attributes, *e.g.* they usually contain past tense verbs and similar structures. This is demonstrated by the two example OUTCOME sentences in Table 5 which are misclassified. Looking at the sentences, it can be challenging even for a human to correctly label them without knowing the context; they both describe the outcome of a study, but it is not clear if it is the reporting study or previous work. Only by reading it in the context of the abstract and the preceding sentence can we confidently determine that they are outcomes of the present study. This is the case for many of the misclassified instances.

However, this is not due to the feature types but rather the classification approach taken here and in many other studies for this task. The SVM does not model the sequential characteristics of sentences in an abstract, instead classifying them independently. It is mostly for these reasons that sequence labelling algorithms, *e.g.* Conditional Random Fields (CRF), have been found to be useful for this task, as we mentioned in §2. Hence, it has been noted that applying such methods with the most suitable features can considerably avoid such contextual errors and improve the overall accuracy (Jonnalagadda et al., 2015).

8 Discussion

We presented a semi-supervised classification approach for clinical information extraction based on unsupervised word representations, outperforming the most commonly used feature types. This is the first application of word representation features for this task; the promising results here inform current research by introducing a new feature class. We also made our word clusters available.

A positive byproduct of this approach is a substantial reduction in the feature space, and thus model sparsity. This has practical implications, resulting in more efficient models and enabling the use of simpler learning algorithms which are generally used with smaller feature sets. This would allow faster and more efficient processing of large amount of data which is an important practical facet of this task. For example, we conducted some preliminary experiments with multinomial Naïve Bayes and k-NN classifiers and our results showed that the Brown cluster features achieved faster and much more accurate results than a bag-of-words approach.

Actual	Predicted	Sentence
INTERVENTION	OUTCOME	Glucocorticoids were decreased and could be stopped as the neurologic deficits fully recovered.
INTERVENTION	OTHER	Subjects were examined before and 1 year after surgical treatment.
OUTCOME	BACKGROUND	Negative symptoms are associated with poor outcome, cognitive impairments, and incapacity in social and work domains.
OUTCOME	BACKGROUND	Patients suffering from mild TBI are characterized by subtle neurocognitive deficits in the weeks directly following the trauma.
POPULATION	OTHER	The aim of this study was to investigate this association in an Italian OCD study group.
POPULATION	OUTCOME	Five cases of biopsy- or Kveim test-proved sarcoidosis with MR findings consistent with MS are reported.

Table 5: Examples of misclassified sentences with their true and predicted labels.

One limitation here was the size of the unlabelled data we used for inducing the Brown clusters.[9] Future work could examine the effects of using more data on classification accuracy.

Having demonstrated the utility of the features, there are a number of directions for future work. We previously described that sequence labelling approaches have been found to be helpful for this task given the structured nature of the abstracts. At the same time, it has been shown that incorporating word representations can result in significant improvements for sequence labelling tasks (Huang and Yates, 2009; Turian et al., 2010; Miller et al., 2004). Therefore, the combination of these two approaches for this task seems like a natural extension.

The evaluation of these Brown cluster features on other datasets used for this task — such as the ART corpus (Liakata et al., 2012) — is another direction for research in order to assess if these results and patterns can be replicated.

Cross-corpus studies have been conducted for various data-driven NLP tasks, including parsing (Gildea, 2001), Word Sense Disambiguation (WSD) (Escudero et al., 2000) and NER (Nothman et al., 2009). While most such experiments show a drop in performance, the effect varies widely across tasks, making it hard to predict the expected drop. This is something that could be evaluated for this task by future work.

[9] *e.g.* Owoputi et al. (2013) used approx 850m tokens of unlabelled text compared to our 96m.

Finally, previous work has also found that combining different word representations can further improve accuracy, *e.g.* the results from Turian et al. (2010, §7.4). This is another avenue for further research in this area.

References

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 172–180.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Daniel Gildea. 2001. Corpus variation and parser performance. In *EMNLP*, pages 167–202.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014a. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159 – 170.

Hamed Hassanzadeh, Tudor Groza, Anthony Nguyen, and Jane Hunter. 2014b. Load balancing for imbalanced data sets: Classifying scientific artefacts for evidence based medicine. In *PRICAI 2014: Trends in Artificial Intelligence*, volume 8862 of *Lecture Notes in Computer Science*, pages 972–984.

Hamed Hassanzadeh, Diego Mollá, Tudor Groza, Anthony Nguyen, and Jane Hunter. 2015. Similarity Metrics for Clustering PubMed Abstracts for Evidence Based Medicine. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*.

Fei Huang and Alexander Yates. 2009. Distributional representations for handling sparsity in supervised sequence-labeling. In *ACL*, pages 495–503. Association for Computational Linguistics.

Siddhartha R. Jonnalagadda, Pawan Goyal, and Mark D. Huffman. 2015. New directions in biomedical text annotations: definitions, guidelines and corpus construction. *Systematic Reviews*, 4(78):16.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support Evidence Based Medicine. *BMC Bioinformatics*, 12(Suppl 2):S5.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *ACL*, pages 595–603.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014. Chinese Native Language Identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14)*, pages 95–99, Gothenburg, Sweden, April. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2015a. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *NAACL*, pages 1403–1409, Denver, CO, USA, June.

Shervin Malmasi and Mark Dras. 2015b. Multilingual Native Language Identification. In *Natural Language Engineering*.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, Colorado, June.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *HLT-NAACL*, pages 337–342.

Diego Mollá and Maria Elena Santiago-martinez. 2011. Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop*.

Joel Nothman, Tara Murphy, and James R Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 612–620. Association for Computational Linguistics.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*. Association for Computational Linguistics.

W.S. Richardson, M.C. Wilson, J. Nishikawa, and R.S. Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13.

Abeed Sarker, Diego Molla, and Cecile Paris. 2013. An Approach for Automatic Multi-label Classification of Medical Sentences. In *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis*, Sydney, NSW, Australia.

Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Huperff, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 372–377. IEEE.

Simone Teufel. 2000. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Mathias Verbeke, Vincent Van Asch, Roser Morante, Paolo Frasconi, Walter Daelemans, and Luc De Raedt. 2012. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 579–589, Jeju Island, Korea.