Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP

Kirk Roberts
School of Biomedical Informatics
University of Texas Health Science Center at Houston
Houston, TX, USA
kirk.roberts@uth.tmc.edu

Abstract

The proliferation of deep learning methods in natural language processing (NLP) and the large amounts of data they often require stands in stark contrast to the relatively data-poor clinical NLP domain. In particular, large text corpora are necessary to build high-quality word embeddings, yet often large corpora that are suitably representative of the target clinical data are unavailable. This forces a choice between building embeddings from small clinical corpora and less representative, larger corpora. This paper explores this trade-off, as well as intermediate compromise solutions. Two standard clinical NLP tasks (the i2b2 2010 concept and assertion tasks) are evaluated with commonly used deep learning models (recurrent neural networks and convolutional neural networks) using a set of six corpora ranging from the target i2b2 data to large open-domain datasets. While combinations of corpora are generally found to work best, the single-best corpus is generally task-dependent.

1 Introduction

The use of vector representations in natural language processing (NLP) has a solid foundation (Turian et al., 2010; Collobert et al., 2011). These enable dense representations that often encode semantic properties and are particularly useful for machine learning tasks as an alternative to extremely sparse, "one-hot" vocabulary-length vector representations. Many ways of building these vectors exist, including random indexing (Sahlgren, 2006), clustering (Brown et al., 1992), regression (Pennington et al., 2014), and neural (Mikolov et al., 2013) methods. This paper focuses on the last such type of vector representation, often referred to as *embeddings*, and exemplified by the popular method `word2vec` (Mikolov et al., 2013).

Embeddings are particularly useful in neural network architectures, which due to their heavy use of matrix multiplication typically favor low-dimensional, dense representation. In particular, neural network models that utilize multiple layers of operations to find abstractions in the data (collectively referred to as deep learning models) are a natural fit for these dense semantic representations.

In what is typically a semi-supervised process, word embeddings are generated from a large, representative sample of data. Then, a smaller manually annotated sample is used to train the deep learning models. However, this results in a common problem for clinical NLP: large representative corpora (at least comparable to those used in much open-domain NLP research) are not often available for building these embeddings. This is due to the significant restrictions on the use of electronic health record (EHR) data, especially narrative notes, for research purposes. Clinical NLP researchers and practitioners are often then left with a trade-off: using a small-but-representative corpus versus a large-but-unrepresentative corpus. The former may not be large enough to properly capture the necessary semantics, while the latter might not be representative enough to capture the semantics of some of the most important words in the corpus. For instance, a large open-domain corpus might associate the abbreviation *ms* with *millisecond* (or *Mississippi*) rather than *multiple sclerosis* (or *mitral stenosis*).

In theory, one could simply experiment with multiple corpora to see what works best for a given task. But in practice this may be overly burdensome, especially in the context of deep learning models that have many, many other important parameters and architectural choices to consider, in addition to their

Proceedings of the Clinical Natural Language Processing Workshop,
pages 54–63, Osaka, Japan, December 11-17 2016.

long training times. What would be useful, then, is some intuitive notion or rule-of-thumb on what corpora to use for building word embeddings for clinical NLP. From a practical point-of-view, one can see two ideal scenarios:

1. A small target corpus (several hundred or a few thousand documents) that is highly representative of the annotated notes in the clinical NLP task (possibly including the annotated notes themselves).
2. A large corpus (millions of documents) that is completely general-purpose (likely not containing clinical note text at all).

If the first scenario were to result in optimal system performance, this would be quite easy for the clinical NLP practitioner: for each NLP task, generate a set of embeddings specific to the corpus. The second scenario is even easier: simply use an "off-the-shelf" set of word embeddings. However, there are many possible compromise solutions between these two extremes. For example, a medium-size corpus of clinical notes from a different corpus, or a large corpus of scientific articles, or even a combination of two or more of these. The goal of this paper is to explore this size vs. similarity trade-off, specifically for clinical NLP purposes. A handful of corpora ranging from a small target corpus to a large general-purpose corpus are used to build embeddings. Experiments using two common deep learning models in combination with two standard clinical NLP datasets are used to evaluate this trade-off.

The remainder of this paper is organized as follows. Section 2 describes related work with word embeddings, including its use in clinical NLP. Section 3 describes the tasks used to evaluate the embeddings. Section 4 describes the datasets used to generate the embeddings. Section 5 describes the experimental setup, including the parameters for generating the word embeddings as well as the parameters for the deep learning models. Section 6 shows the results of the experiments. Section 7 discusses the implications, with some practical considerations.

2 Related Work

As mentioned above, there are various types of word vector representations for use in NLP (Brown et al., 1992; Sahlgren, 2006; Mikolov et al., 2013; Pennington et al., 2014). By themselves, these are well-known to be easily integrateable into common NLP tasks (Turian et al., 2010; Collobert et al., 2011). Generally, the best types of representations have semantic properties, notably that synonyms are nearby in vector space, and certain types of vector operations (addition and subtraction) roughly correspond to semantic operations. This largely holds for neural word embeddings, which allow for the induction of additional semantic properties, such as hypernymy relations (Fu et al., 2014). As embeddings become more and more important in NLP, work continues on analyzing their usefulness, such as how to interpret specific vector dimensions (Luo et al., 2015), but most work focuses on applying embeddings to well-defined NLP tasks.

Further, the increased importance of deep learning methods in NLP has resulted in a significant number of uses of embeddings to represent words. Wang and Manning (2013) help clarify the relationship between embeddings and deep learning models: these models excel with low-dimensional, continuous representations, but offer no benefit over more traditional models like conditional random fields (CRF) (Lafferty et al., 2001) when used with high-dimensional, discrete representations. Embeddings for NLP are commonly used in sequence classification tasks such as part-of-speech tagging and chunking (Huang et al., 2015), named entity recognition (Chiu and Nichols, 2016; Lample et al., 2016), and semantic role labeling (Zhou and Xu, 2015). Typically, these sequence models are based on recurrent neural networks (RNN). Classification models, on the other hand, are often based on convolutional neural networks (CNN). These models are more adept at picking out a relevant piece of information in a relatively long span of text, and so are often used for sentence classification (Kim, 2014; Zhang and Wallace, 2016), or sentiment and topic prediction (Zhang et al., 2015). Note that many other deep learning methods are possible with embeddings, such as sentiment classification with recursive autoencoders (Socher et al., 2011), but this paper focuses on the use of RNNs and CNNs specifically for clinical NLP.

While less explored than the open domain, research exists on the uses of word embeddings for clinical NLP (though less so in the context of a deep learning model). Several non-neural vector representations

3. **Echocardiogram** on **DATE[Nov 6 2007] , showed **ejection fraction** of 55% , **mild mitral insufficiency** , and **1+ tricuspid insufficiency** with **mild pulmonary hypertension** .
DERMOPLAST TOPICAL TP Q12H PRN **Pain** **DOCUSATE SODIUM** 100 MG PO BID PRN **Constipation** **IBUPROFEN** 400-600 MG PO Q6H PRN **Pain**
The patient is struggling to breathe at this time , and she is **tachypneic** , and she might have to be **intubated** right now but ; however , the patient 's family did not wish the patient to be **intubated** even after I explained to them that she could potentially die if she was not on a **breathing machine** ; but however , the patient 's family stressed to me again and wished that they do not want her mother to be on a **breathing machine** .
The patient had **headache** that was relieved only with **oxycodone** . A **CT scan of the head** showed **microvascular ischemic changes** . A **followup MRI** which also showed **similar changes** . This was most likely due to **her multiple myeloma** with **hyperviscosity** .

Table 1: Examples of concepts (**Problem**, **Treatment**, and **Test**) from the i2b2 2010 corpus.

have been used for named entity recognition style tasks, notably random indexing (Jonnalagadda et al., 2012; Henriksson et al., 2014). Most uses of neural embeddings have likewise been through non-deep learning models. Wu et al. (2015) explored different feature representations for embeddings, showing that for CRFs both binarized and distributed prototype embeddings (Guo et al., 2014) out-performed the raw embeddings. Related, but outside of clinical NLP, Tang et al. (2014) study the use of word representations, including embeddings, for gene/protein NER, also within the context of CRF features.

Finally, there has been some study on the use of multiple word embeddings in the context of deep learning models. Luo et al. (2014) learn new task-specific embeddings from multiple pre-trained embeddings for the purpose of search ranking and text similarity. Yin and Schütze (2015) treat multiple word embeddings as different channels in a CNN. This achieves great performance, but requires all the embeddings be of the same dimension. In contrast, the method in this paper uses simple concatenation, which does not require equal dimensions, but Yin and Schütze (2015) may still have some desirable semantic properties. Finally, Zhang et al. (2016) proposes a multi-group norm constraint CNN (MGNC-CNN) that separates the convolutional layers for different sets of embeddings. This model also has a lot of promise, but it beyond the scope of this work. Additionally, all of these multi-embedding models have focused on CNNs, while it is not clear whether Yin and Schütze (2015) or Zhang et al. (2016) could be successfully applied to RNNs. However, the focus in this paper is on devising an intuition behind choosing the right sets of embeddings (or ideally, only one set of embeddings).

3 Tasks

Two common clinical NLP tasks are considered: sequence classification and multi-class text classification. While sequence classification is often a type of multi-class text classification (if there is more than one type of phrase to be recognized), it nonetheless is often treated differently in regards to the "default" machine learning algorithm (i.e., SVM vs. CRF). For each type of task, a specific task from the i2b2 2010 Shared Task (Uzuner et al., 2011b) is selected for the experiments. While the deep learning-based models used for each task are mentioned here, Section 5 contains more details on the actual implementations.

3.1 Sequence Classification

Word embeddings for sequence classification are evaluated using the i2b2 2010 concept recognition task. A medical concept in this task is a problem (e.g., disease or symptom), treatment (e.g., drug or therapeutic procedure), or test (e.g., diagnostic procedure). This is an especially difficult problem in clinical NLP due to the compact nature of text in EHR notes. Table 1 shows examples of different concept types from the i2b2 2010 corpus, while Table 2 shows their distributions in the train and test sets.

To model concept recognition, a bi-directional recurrent neural network (RNN) using long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) is used. LSTM-RNNs are heavily used in named entity recognition and other sequence-based NLP tasks (Hammerton, 2003; Huang et al., 2015; Zhou and Xu, 2015; Chiu and Nichols, 2016; Lample et al., 2016).

	Train Set	Test Set	Total
Documents	349	477	826
Concepts	27,831	45,009	72,840
Problems	11,967	18,550	30,517
Treatments	8,496	13,560	22,056
Tests	7,368	12,899	20,267

Table 2: Frequencies of concept types in the i2b2 2010 corpus.

Present	... a short - term temporary measure , and after **her pneumonia** gets better ... ... demonstrating a **low fibrinogen** , **positive D-dimer** , and ... ... admitted with **vomiting** and **fever** and found to have **urinary tract infection** ...
Absent	... patient 's left **back pain** was evaluated and ruled out for MI and the back pain ... His neck was supple with no **jugular venous distention** or **thyromegaly** . He denied any **fevers** , **chills** , or **night sweats** .
Hypothetical	The patient was instructed to report any new or **increased shortness of breath** ... The patient is to expect **some blood in his urine** for the first couple of days . ... her steroid inhalers and PO prednisone for **COPD exacerbation** .
Possible	... who came to the hospital with what appears to be **acute coronary syndrome** . ... multiple bilateral pulmonary nodules compatible with **inflammatory disease** . The patient did not have any EKG changes consistent with **hyperkalemia** .
Conditional	... chest tightness (pressure) approximately every three months with **stress** . He reports **severe dyspnea** on exertion . ... pt **slightly lightheaded** and with **increased HR** when getting up out of bed .
Associated with Someone Else	She has no family history of **gallbladder** or **pancreatic disease** . His mother and father both died secondary to **myocardial infarction** . The patient's sister has a history of **cervical cancer** .

Table 3: Examples of assertions types for **Problems** from the i2b2 2010 corpus.

3.2 Multi-class Text Classification

Word embeddings for multi-class text classification are evaluated using the i2b2 2010 assertion task. An assertion is a belief state about a medical problem (present, absent, hypothetical, etc.). This is especially important in clinical NLP as diagnoses are often ruled out or speculated about during the diagnostic process. Table 3 shows examples of different assertion types from the i2b2 2010 corpus, while Table 4 shows their distributions in the train and test sets.

To model assertion classification, a 2-layer convolutional neural network (CNN) with a max-pooling layer and softmax classifier is used. While more noteworthy for imaging tasks, CNNs have been heavily utilized in text classification as well (Collobert et al., 2011; Kim, 2014; Zhang et al., 2015; Zhang and Wallace, 2016).

4 Data

Six datasets are utilized for generating word vectors (see Table 5):

i2b2 is the "target" dataset. This is a combination of multiple years worth of i2b2 shared tasks: 2010 (Uzuner et al., 2011b), 2011 (Uzuner et al., 2011a), and 2012 (Uzuner et al., 2013) tasks. However, the vast majority come from the same data pull(s) used to build the training and testing data (87%) for the 2010 tasks described in Section 3. This dataset corresponds to the first ideal scenario described above, since it would be practical if this data alone would be sufficient to generate optimal word embeddings as additional corpora would never be needed. However, as is often the case in practice, far less data is available in this dataset compared to what is typically used to generate word embeddings.

	Train Set	Test Set	Total
Documents	349	477	826
Problems	11,967	18,550	30,517
Present	8,052	13,025	21,077
Absent	2,535	3,609	6,144
Hypothetical	651	883	1,534
Possible	535	717	1,252
Conditional	103	171	274
Associated with Someone Else	92	145	237

Table 4: Frequencies of assertion types in the i2b2 2010 corpus.

MIMIC3 (Johnson et al., 2016) is a freely-accessible database of intensive care unit (ICU) encounters from a large hospital. It is significantly larger than the i2b2 dataset, and some of the i2b2 data was even drawn from MIMIC-II. MIMIC-III represents the next-best case scenario to having a large clinical target dataset: it is both large and fairly similar to the i2b2 data. MIMIC is commonly used to generate word embeddings for clinical NLP, but its exact utility in comparison to the target dataset is rarely, if ever, measured.

MEDLINE is a collection of scientific article abstracts maintained by the National Library of Medicine. While a large dataset, these are not clinical notes and lack many of the peculiarities of clinical notes (e.g., abbreviations, telegraphic text). Further, while clinical notes are written by clinicians largely to communicate with other clinicians, MEDLINE abstracts are written by researchers largely to communicate with other researchers. However, MEDLINE does discuss almost all the diseases, conditions, treatments, and techniques that are described in clinical notes.

WebMD Forum is a collection of forum posts on the WebMD Community[1]. The forum posts are written largely by health consumers, who are known to write health-related text quite differently than clinicians (Roberts and Demner-Fushman, 2016). This dataset is intended to represent a small-to-medium-size medically-related corpus that is nonetheless quite different from clinical notes.

Wikipedia is a large, online encyclopedia. Wikipedia has extensive coverage of medical topics, but also many other topics as well. Wikipedia represents the other best-case scenario for generating word embeddings: if near-optimal performance could be obtained using such a general corpus, it could be used in all experiments without the need to generate new word embeddings for each task.

Gigaword is a large newswire corpus (Parker et al., 2009). It has extensive coverage of topics that typically dominate the news media, including politics and sports, but its coverage of medicine is largely limited to newsworthy studies and announcements. Gigaword represents a control corpus: it should be less useful than Wikipedia, but if it were to be beneficial then one could argue that using several arbitrary corpora simultaneously (like an ensemble) is useful simply to provide multiple views of each word, or even just more free parameters for the neural network to work with.

Instead of creating word embeddings for each combination of corpora, the embeddings are built for each individual corpus independently. This has several advantages. First, it prevents the smaller, more similar corpora from being "drowned out" by the larger, more distant corpora. Second, it dramatically reduces the time needed to produce the embeddings since only N embeddings are needed. Third, providing the neural networks with multiple sets of embeddings allows for a kind of domain adaptation to take place: the networks can learn to take different information from different corpora, which it would not be able to do with a single, unified embedding vector built from all the data. As mentioned above, multi-embedding models have been utilized for neural networks before. The implementation here is intentionally one of the simplest forms of embedding combinations: simple concatenation of the em-

[1] http://exchanges.webmd.com/

Corpus	# Documents	# Sentences	# Tokens	*% diabetes*	*% myocardial*	*% tumor*
i2b2	3k	158k	1.7m	2.9e-4%	2.4e-4%	1.3e-4%
MIMIC	876k	17m	366m	1.0e-4%	1.2e-4%	9.1e-5%
MEDLINE	24m	138m	3.7b	2.2e-4%	1.5e-5%	7.9e-4%
WebMD	232k	1.5m	24m	1.3e-4%	4.5e-7%	3.4e-5%
Wikipedia	4.8m	96m	2.1b	7.0e-6%	1.0e-6%	1.1e-5%
Gigaword	8.5m	169m	4.1b	9.3e-6%	N/A	7.5e-6%

Table 5: Basic corpus statistics, including the proportion of three important clinical terms (*diabetes*, *myocardial*, *tumor*) to illustrate how representative each corpus is of clinical text. Note that this excludes common clinical abbreviations (e.g., *dm* or *dm2* for *diabetes*). "N/A" indicates the word was not in the top 100k terms and thus not included in the embeddings.

bedding vectors. Other methods are possible (Zhang et al., 2016), but it is unclear whether these more specialized methods would produce results as generalizable as simple vector concatenation.

5 Experimental Setup

Both word embeddings and deep learning models have very many possible parameters that can impact downstream tasks. The following experimental description is by no means likely to be optimal for the tasks, but was made based on a combination of default parameters, conventional wisdom, and practical necessity. In some cases experiments were conducted to test parameter impact on the downstream tasks (mostly with the more crucial deep learning model parameters). See Section 7.1 for a discussion of the limitations of these experiments.

5.1 Word Vectors

Each corpus was pre-processed with tokenization and sentence segmentation. Case was removed. Numbers were altered to just the most significant digit (e.g., 929 becomes 900). Word occurring less than 5 times were changed to `UNK`. Finally, a maximum vocabulary of 100k word types was applied, keeping only the most frequent words. The numbers in Table 5 reflect these transformations. The gensim (Řehůřek and Sojka, 2010) version of `word2vec` was then applied to create 100-dimensional embeddings largely using default parameters (CBOW, α=0.025, 5-word window, 50 epochs).

5.2 Recurrent Neural Network

The RNN uses a bi-directional, 3-layer LSTM implemented in TensorFlow (Abadi et al., 2015). Each LSTM cell uses 256 hidden units. Dropout is set to 0.5. A maximum sequence length of 50 tokens per sentence is used, which includes 98.4% of the concepts in the test set. Of the 30k sentences in the training set, 5k are used as a validation set for early stopping, evaluated up to 100 training epochs. The i2b2 concepts are represented in IOB format.

5.3 Convolutional Neural Network

The CNN uses 2 convolutional layers with a ReLU activation followed by a max-pooling layer and a softmax classifier, again implemented in TensorFlow. Optimization is performed with the Adam algorithm. Filters of sizes 1, 2, 3, and 4 are used, each replicated 400 times. No dropout is used. A context window of 3 tokens around the problem's first token is used for a total input width of 7 tokens. Of the 12k problems in the training set, 1k are used as a validation set for early stopping, evaluated up to 300 training epochs.

6 Results

The results of the experiments are shown in Table 6 and Table 7.

Concept recognition is measured in precision, recall, and micro-averaged F_1-measure. The single best corpus for this task was the MIMIC data, which out-performed the target i2b2 corpus in F_1 by 2.7 points. It also outperformed the more general-purpose Wikipedia and Gigaword corpora in F_1 by 4.7 and 7.5

Corpus	P	R	F_1
i2b2	74.47	80.12	77.19
MIMIC	77.99	81.97	**79.93**
MEDLINE	76.64	82.83	79.61
WebMD	71.95	77.72	74.72
Wikipedia	72.40	78.25	75.21
Gigaword	71.64	76.98	74.22
Corpus combination, starting with i2b2			
+ MIMIC	78.30	82.86	80.52
+ MEDLINE	79.65	83.71	**81.63**
+ WebMD	79.10	83.99	81.47
+ Wikipedia	79.64	83.62	81.58
+ Gigaword	78.78	83.89	81.25

Table 6: Results for RNN-based concept recognition on the i2b2 2010 corpus, measured with precision (P), recall (R), and F_1-measure.

Corpus	Accuracy	P	A	H	B	C	O
i2b2	**91.29**	96.24	95.45	87.14	86.27	81.58	91.84
MIMIC	91.16	96.10	96.15	85.60	85.58	81.63	92.73
MEDLINE	90.98	95.70	95.92	86.59	90.51	85.25	91.59
WebMD	90.22	95.40	95.14	86.65	88.50	83.87	92.59
Wikipedia	90.36	95.71	94.84	86.05	86.37	82.35	96.06
Gigaword	90.33	95.59	95.39	86.11	85.60	78.38	94.44
Corpus combination, starting with i2b2							
+ MIMIC	91.26	96.29	95.36	86.44	86.58	86.96	85.48
+ MEDLINE	91.56	96.24	96.29	85.11	88.25	83.78	95.16
+ WebMD	91.39	96.46	95.43	85.76	85.79	86.75	89.33
+ Wikipedia	**91.58**	96.35	96.79	82.61	88.06	80.00	92.65
+ Gigaword	91.57	96.42	95.85	87.60	86.17	80.52	83.78

Table 7: Results for CNN-based assertion classification on the i2b2 2010 corpus, measured with accuracy, along with the F_1-measure for present (P), absent (A), hypothetical (H), possible (B), conditional (C), and associated with someone else (O).

points, respectively. MEDLINE did almost as well as MIMIC, while WebMD did poorly, only slightly better than Gigaword. Results improve when the corpora are combined. The best overall results are achieved by combining i2b2, MIMIC, and MEDLINE. Adding in the other corpora hurt performance slightly, by at most 0.4.

Assertion classification is measured in accuracy, with F_1-measures for the individual assertion type provided in Table 7. Unlike concepts, the single best corpus is the target i2b2 data. All other corpora performed close, with the worst performance being WebMD with a 1.1 point drop in accuracy. Only slight gains are seen by adding in other corpora, the best being all corpora except Gigaword for a 0.3 point improvement, but no substantial losses are seen either.

7 Discussion

It would first be useful to compare the results obtained above with the state-of-the-art methods for the concept and assertion tasks (Uzuner et al., 2011b). In both cases, the results are less than the best performing scores on the tasks, but they are quite close. The best concept RNN would have placed 6th overall (out of 22) and well above the median (77.78). The best assertion CNN would have done a bit worse, performing near the median. However, these models were built using not particularly well-optimized parameters and furthermore they only had access to word information. The many features used

by participants in the i2b2 tasks (e.g., UMLS (Lindberg et al., 1993), NegEx (Chapman et al., 2001), and task-specific patterns) could be incorporated into these models for superior performance. The fact that near state-of-the-art performance is achieved without any medical knowledge or custom features speaks to the power of these models.

Regarding the ideal scenarios for embeddings discussed in the Introduction (target data only and general-purpose only embeddings), these turned out to unfortunately not be the best performing conditions. i2b2 was the single best corpus for assertions, but not for concepts. Rather, MIMIC and MEDLINE greatly outperformed i2b2 for concepts, and were only slightly behind for assertions. This difference is likely due to the small number of relevant phrases that indicate assertion types compared to the vast vocabulary of medical concepts. The second ideal scenario, using a general-purpose corpus only, performs quite poor as a single corpus for both tasks. If only one set of embeddings can be used, then, it seems a compromise corpus such as MIMIC might be best.

The multi-embedding experiments reveal an important point, however. Combining multiple sets of embeddings can help quite a bit (e.g., i2b2 + MIMIC + MEDLINE did 1.7 points better than MIMIC alone for concepts), while adding "bad" corpora only will only hurt slightly (adding a single corpus never brought the score down more than 0.3 points). Therefore, if it is not possible to perform many experiments with embeddings on the task data (a common case in many applied clinical NLP settings), using several corpora at once seems relatively safe.

7.1 Limitations

This paper seeks to identify best practices experimentally, so its limitations revolve around reasons why the results may not be generalizable. In this sense, the possible limitations are vast, including:

- Only two clinical NLP datasets were evaluated, so the results obtained here may vary greatly with other tasks.
- Only a handful of experiments (just less than a week of computing time) were conducted to optimize the parameters of the various models: every choice made in Section 5 may be suboptimal. This may have reduced performance inconsistently, changing the relative performance of the various corpora.
- As an explicit example of the above point, the use of 100-dimension embeddings is less than what is typically used (often 300). Since embedding combination was an intentional goal of this paper, the embedding dimensionality was kept small to reduce training time (e.g., 600 vs. 1800 dimensions for the final experiment).
- It would have been useful to evaluate on more corpora–clinical, medical, and general-purpose–to measure intra-domain variance.
- Multi-embedding methods (Yin and Schütze, 2015; Zhang et al., 2016) could have improved results over simple vector concatenation.

Despite the extent to which these limitations may reduce the ability to generalize the experiments, the results largely do match the intuitions gained elsewhere in NLP. For ensembles, for example, adding additional weak classifiers is more likely to have a strong positive effect than a strong negative effect, which is consistent with the above results.

8 Conclusion

This paper presented a series of experiments to evaluate the trade-off between small-but-representative corpora versus large-but-unrepresentative corpora for building word embeddings for clinical NLP tasks. Two standard clinical NLP tasks (i2b2 2010 concepts and assertions) were used in combination with two appropriate deep learning methods (RNNs and CNNs) to evaluate six text corpora of varying size and similarity to the target corpus. While using only the small target corpus or a large general-purpose corpus would have been ideal from a practical standpoint, empirically it was found that combining multiple corpora, especially a corpus like MIMIC, is the safest option for choosing embeddings.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

PF Brown, PV deSouza, RL Mercer, VJ Della Pietra, and JC Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, pages 467–479.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310, October.

Jason P.C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning Semantic Hierarchies via Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1199–1209.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting Embedding Features for Simple Semi-supervised Learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 110–120.

James Hammerton. 2003. Named Entity Recognition with Long Short-Term Memory. In *Proceedings of the Seventh Conference on Natural Language Learning*.

Aron Henriksson, Hercules Dalianis, and Stewart Kowalski. 2014. Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. In *Proceedings of the IEEE Conference on Bioinformatics and Biomedicine*, pages 450–457.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, , and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 45(1):129–140.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Donald A.B. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

Yong Luo, Jian Tang, Jun Yan, Chao Xu, and Zhang Chen. 2014. Pre-trained multi-view word embedding using two-side neural network. In *Proceedings on the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1982–1988.

Hongyin Luo, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2015. Online Learning of Interpretable Word Embeddings. pages 1687–1692.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword Fourth Edition. *The LDC Corpus Catalog.*, LDC2009T13.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: A comparison of consumer and professional questions. *Journal of the American Medical Informatics Association*.

Magnus Sahlgren. 2006. *Vector-Based Semantic Analysis: Representing Word Meanings Based On Random Labels*. Ph.D. thesis, Stockholm University.

Richard Socher, Jeffrey Pennington, Eric Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. volume 2014, page 240403.

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.

Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett South. 2011a. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 18:552–556.

Özlem Uzuner, Brett South, Shuying Shen, and Scott L. DuVall. 2011b. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18:552–556.

Özlem Uzuner, Anna Rumshisky, and Weiyi Sun. 2013. 2012 i2b2 challenge on temporal relations. *Journal of the American Medical Informatics Association*, (in submission).

Mengqiu Wang and Christopher D. Manning. 2013. Effect of Non-linear Deep Architecture in Sequence Labeling. In *Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. 2015. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. In *Proceedings of the AMIA Annual Symposium*, pages 1326–1333.

Wenpeng Yin and Henrich Schütze. 2015. Multichannel Variable-Size Convolution for Sentence Classification. In *Proceedings of the Nineteenth Conference on Natural Language Learning*, pages 204–214.

Ye Zhang and Byron Wallace. 2016. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv:1510.03820.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, pages 649–657.

Ye Zhang, Stephen Roller, and Byron C. Wallace. 2016. MGNC-CNN: A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1522–1527.

Jie Zhou and Wei Xu. 2015. End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.