Medical Semantic Similarity with a Neural Language Model

Lance De Vine[1], Guido Zuccon[2], Bevan Koopman[3,2], Laurianne Sitbon[1], Peter Bruza[2]
[1]Electrical Engineering & Computer Science, Queensland University of Technology, Brisbane, Australia
[2]Information Systems, Queensland University of Technology, Brisbane, Australia
[3]Australian e-Health Research Centre, CSIRO, Brisbane, Australia
l.devine@student.qut.edu.au, g.zuccon@qut.edu.au, bevan.koopman@csiro.au,
laurianne.sitbon@qut.edu.au, p.bruza@qut.edu.au

ABSTRACT

Advances in neural network language models have demonstrated that these models can effectively learn representations of words meaning. In this paper, we explore a variation of neural language models that can learn on concepts taken from structured ontologies and extracted from free-text, rather than directly from terms in free-text.

This model is employed for the task of measuring semantic similarity between medical concepts, a task that is central to a number of techniques in medical informatics and information retrieval. The model is built with two medical corpora (journal abstracts and patient records) and empirically validated on two ground-truth datasets of human-judged concept pairs assessed by medical professionals. Empirically, our approach correlates closely with expert human assessors (≈ 0.9) and outperforms a number of state-of-the-art benchmarks for medical semantic similarity.

The demonstrated superiority of this model for providing an effective semantic similarity measure is promising in that this may translate into effectiveness gains for techniques in medical information retrieval and medical informatics (e.g., query expansion and literature-based discovery).

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]

General Terms: Theory, Experimentation, Measurement

Keywords: Neural Language Model; Skip-gram; Distributed Representations; Word2Vec; Semantic Similarity; Medical Information Retrieval.

1. INTRODUCTION

A variety of neural network-based methods have emerged as effective approaches for generating representations of words [4, 16, 12]; these are referred to as neural language models. These methods learn word embeddings based on the optimisation of an objective function. The term "word embeddings" generally refers to representations for words occupying a real valued vector space where the similarity between words is measured by cosine similarity. An objective function that is often used for training word embeddings is to learn a vector for a target word which predicts the vectors for words occurring near to it (Skip-gram).

Recent research has demonstrated that neural language models (NLM) based on the continuous Skip-gram model proposed by Mikolov et al. [11] are highly effective in determining semantic relationships between words [13]. It is still not clear, however, whether these neurally inspired models are better than traditional distributional semantic methods. For example, Lebret et. al. [10] report results that suggest that computing a Hellinger PCA of a word co-occurence matrix provides similar results to neural network models on natural language processing tasks. On the other hand, Baroni et. al [3] report on comparisons between standard distributional semantic models and neural network models and conclude that neural network models do indeed provide superior word representations. They note however that not all neural network word models are equal.

Semantic similarity measures are central to several techniques used in health informatics and medical information retrieval, e.g., query expansion [6] and literature-based discovery [1]. A number of previous corpus-based approaches have been employed for semantic similarity measurements and have been evaluated by how well they correlate with human-judged similarity [15, 9]. These approaches were applied to medical concepts taken from the UMLS medical thesaurus and extracted from medical free-text. The results from these studies show that although corpus-based measures of similarity do correlate with human judgments, there is considerable room for improvement. Motivated by this and the recent findings in neural language models, we explore a variation to the original continuous Skip-gram NLM of Mikolov et al. [11], where instead of learning a distributed vector representation over sequences of terms, we train the model over sequences of UMLS medical concepts. This approach is evaluated over two human-judged semantic similarity datasets and is trained using two corpora: a collection of clinical records and a large set of MEDLINE medical journal abstracts. The empirical results of this study demonstrate that the proposed neural language models outperform a number of benchmark corpus-based approaches, strongly correlating with semantic similarity judgements provided by medical, expert judges.

2. SKIP-GRAM NEURAL LANGUAGE MODEL

The effectiveness of corpus-driven approaches relies on the distributional hypothesis [8, 14], which states that the degree of semantic similarity between two terms (or some other

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661974.

linguistic units) can be modelled as a function of the degree of overlap of their linguistic contexts. In practice, the counts of contextual features are generally accumulated into a term-context matrix and a transformation is then applied which re-weights the accumulated counts.

Neural language models also construct representations for terms based on linguistic contexts; however, they do so by optimising an objective function involving the target term and its linguistic context. The representations produced are often called "word embeddings". Word embeddings were first developed in the context of language modelling to overcome some of the well known problems relating to data sparsity that existed with n-gram based language models [4]. While NLMs were originally developed to model sequential term dependencies departing from the n-gram approach, a by-product of these models is that the constructed word representations were found to have useful semantic properties [11]. NLMs have more recently been employed for a large variety of natural language processing tasks, such as semantic role labelling, part-of-speech tagging, chunking, sentiment analysis and named entity recognition [7, 13]; they were found to be as good as, or better than, other state-of-the-art methods.

A particular instance of a NLM is the continuous Skip-gram model of Mikolov et al. [11]. The Skip-gram model constructs term representations by optimising their ability to predict the representations of surrounding terms. In this paper, we evaluate the continuous Skip-gram model on the task of predicting the semantic similarity of concept pairs. We employ the Skip-gram model in a way not previously seen in the literature; specifically, we use it to learn embedding vectors for concepts taken from structured ontologies rather than for terms. While previous work has considered the use of compound terms (e.g., named entities) in NLMs [13], these compound terms are not actually used as features; in addition, ontology concepts have not been used (to our knowledge).

Given a sequence $\mathcal{W} = \{w_1, \ldots, w_t, \ldots, w_n\}$ of training words, the objective of the Skip-gram model is to maximise the following average log probability

$$\frac{1}{2r}\sum_{i=1}^{2r}\sum_{-r\leq j\leq r, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

where r is the context window radius. The context window radius determines which words surrounding the target term w_t are considered for the computation of the log probability; the window is centred around the target term. The probability of an output word is computed according to

$$p(w_O|w_I) = \frac{\exp(v_{w_O}^{\top} v_{w_I})}{\sum_{w=1}^{W}\exp(v_w{}^{\top}, v_{w_I})} \tag{2}$$

where the v_{w_I} and v_{w_O} are the vector representations of the input and output vectors, respectively, and $\sum_{w=1}^{W}\exp(v_w^{\top}, v_{w_I})$ is the normalisation factor, whose role is to normalise the inner product results across all vocabulary words (W is the vocabulary size). In practice, a hierarchical approximation to this probability is used to reduce computational complexity [11]. At initialisation, the vector representations of the words are assigned random values; these vector representations are then optimised using gradient descent with decaying learning rate by iterating through sentences observed in the training corpus.

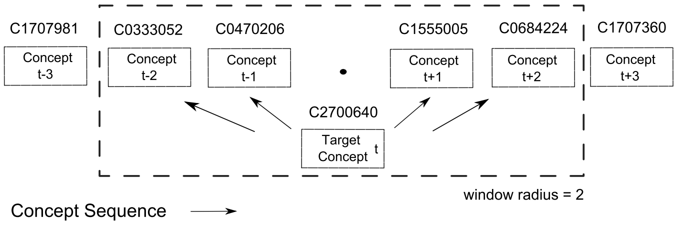

Figure 1: Skip-gram Neural Language Model applied to sequences of UMLS concept identifiers. In this example, the context radius r is set to 2.

In this paper, we explore a variation of the described Skip-gram NLM, where sequences of terms are substituted with sequences of UMLS concept identifiers. Thus, in practice, training is performed by iterating through sequences of concepts as shown in Figure 1. This method builds representations of concepts that are predictive of nearby concepts. It is this feature that, we hypothesise, would enhance semantic similarity measurements between medical concepts.

3. EXPERIMENT SETTINGS

3.1 Corpora and Human-judged Datasets

In this paper we adopted the evaluation framework setup by Koopman et al. [9], who empirically evaluated a number of different corpus-driven measures of semantic similarities for medical concepts. We refer to that work for details about the evaluation framework that are not reported in this paper. The evaluation framework comprised of two datasets:

- Ped: 29 UMLS medical concept pairs developed by Pedersen et al. [15]. Semantic similarity judgements were provided by 3 physician and 9 clinical terminologists, with an inter-coder correlation of 0.85.
- Cav: 45 MeSH/UMLS concept pairs developed by Cavides and Cimino [5]. Similarity between concept pairs was judged by 3 physicians, with no exact consensus value reported by Cavides and Cimino.

In addition, two corpora were used in the evaluation framework for learning concept representations:

- MedTrack: a collection of 17,198 clinical patient records used in the TREC 2011 and 2012 Medical Records Track [17]. Average document length was 932 tokens (words).
- OHSUMED: a collection of 348,566 MEDLINE medical journal abstracts used in TREC 2000 Filtering Track. Average document length was 100 tokens (words).

We specifically focus on two corpora because previous work has found that the effectiveness of corpus-based measures is influenced by corpus characteristics [9]. In particular, previous work has found that measures suited to the characteristics of one corpus are often not suited to those of another corpus, yielding significant differences in performance across corpora.

In accordance with [9], documents in both corpora where pre-processed using MetaMap v11.2, a state-of-the-art biomedical concept identification system [2], which converted the free-text into sequences of UMLS concept identifiers. Converting the test corpora to concepts allowed for direct comparison of the concept pairs contained in both the Ped and

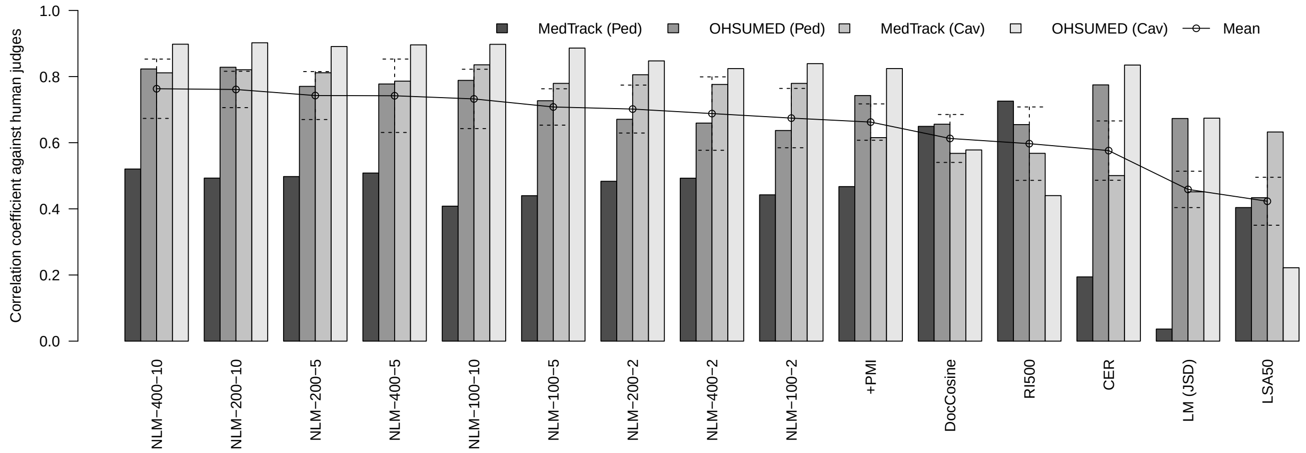

Figure 2: Pearson correlation coefficient against expert judged semantic similarity for the NLM and benchmark comparison methods. Correlations are computed for two gold standard datasets (Ped & Cav) using two corpora (MedTrack & OHSUMED). Methods are ordered from left to right by decreasing correlation averaged across all datasets/corpora, which is summarised by the trendline. Error bars for points in the trendline signify confidence intervals at 95% for the mean correlation value.

Cav datasets. Three concepts appearing in the Ped dataset were not found in the translated corpora (two in Medtrack and one in OHSUMED) and were, therefore, removed.[1]

3.2 Benchmark Comparison Methods

A number of other corpus-based measures of semantic similarity were included as benchmarks for comparison against the neural language model approach:

1. Random Indexing (RI)
2. Latent Semantic Analysis (LSA)
3. Document Vector Cosine Similarity (DocCosine)
4. Positive Pointwise Mutual Information (+PMI)
5. Cross Entropy Reduction (CER)
6. Language Model + Jensen-Shannon divergence (LM JSD)

A previous evaluation of the above models on the same task found that these were the most effective in terms of correlations with human judges [9]. We refer the reader to [9] for a description of each method.

3.3 Parameters Settings

For the benchmark comparison methods (e.g., RI and LSA) we selected the parameter settings, e.g., latent space dimensionality, that produced the highest correlations with human experts as reported in previous work [9][2].

For the Skip-gram NLM, we adopted the `word2vec` implementation provided by Mikolov et al. [11][3]. We used the hierarchical soft-max classification layer and set the "min-count" parameter to 1, thus effectively not excluding any concept occurrence from the computation of statistics. Each corpora was processed using only one thread so that processing was purely sequential. We studied the effect of window radius and embedding dimensionality (i.e. the dimensionality of the reduced space) on semantic similarity by considering 2, 5 and 10 as window radius and 100, 200 and 400 as latent dimensions; these are values typical of the range generally reported in the NLM literature [7, 11, 13].

4. RESULTS & DISCUSSION

Results showing the Pearson correlation coefficient against human judges for each semantic similarity method are reported in Figure 2. The methods on the x-axis are ordered from left to right in decreasing correlation averaged across all datasets/corpora: the leftmost method exhibited the highest overall correlation with human experts. Significance intervals are also reported for the mean correlation values.

According to the empirical results reported in Figure 2, the mean correlation between different settings of the Skip-gram neural language model provides overall higher correlations with human assessed semantic similarity than the other benchmark methods. In particular, the NLM approach is found to consistently outperform the benchmarks for all but one datasets-corpora combinations, with DocCosine and RI providing stronger correlations with human experts in the Ped dataset when trained with the Medtrack corpus.

When MedTrack is used to train the methods, the correlation between NLM semantic similarity estimations and the expert assessments for the Ped dataset is less strong than that obtained by the DocCosine and RI benchmarks. This may suggest that NLM does not appropriately use evidence encoded in the Medtrack corpus to construct effective concept representations. However, this is not confirmed when analysing the results on the Cav dataset: in the latter case NLM is found to strongly correlate with expert assessments when using MedTrack. Previous work has found that there is no single method that does consistently outperform any other method across all datasets-corpora combinations considered in this evaluation framework: it was the choice of corpora used to prime the measures that affected their performance [9]. While NLM does not provide strong correlations on Ped when using Medtrack, the use of this corpus

[1] Removed concepts were `C0702166, C0224701, C0029456`.

[2] Tested dimensionalities: 50, 150, 300 and 500.

[3] `http://word2vec.googlecode.com/`.

does not seem to detriment NLM's performance when considering the Cav dataset.

We now consider how window radius and embedding dimensionality affect performance of the studied NLM. We found that the best performing model was the Skip-gram model with the largest dimensionality and window radius. Overall we found that increasing both the embedding dimensionality and the window radius helped to improve performance, with larger window radius contributing more than larger dimensionalities. While not true in every case, the overall trend suggests as a guideline for building NLM models for this tasks, that vectors with larger window radius and larger embedding dimensionality should be used.

The empirical results highlight that the investigated Skip-gram NLM constructs representations for concepts that, when used as a measure of semantic relations, strongly correlate with semantic similarity judgements provided by medical experts. We conjecture that the predictive nature of the objective function used by the considered Skip-gram NLM is the core feature that produces such strong performance. The validation of this intriguing conjecture would require further investigation; this is left for future work.

5. CONCLUSIONS

Neural network language models (NLM) have recently attracted attention because of promising results obtained in a number of natural language processing tasks, e.g., semantic role labelling and sentiment analysis, among others. The intuition behind these models is that effective representations that synthesise word meaning can be learnt by iteratively observing word occurrences in the close surroundings of target words along with the optimisation of a task-specific function.

In this paper, we have explored a variation of a specific NLM approach, the Skip-gram model, applied to the task of measuring the semantic similarity between medical concepts. While the traditional Skip-gram model creates distributed vector representations of words, the model in this study leverages distributed representations of UMLS concepts extracted from medical corpora, including clinical records and medical journal abstracts.

Empirical findings demonstrate that the concept-based Skip-gram NLM correlates more strongly to expert judgement of semantic similarity than established benchmark approaches. Window radius *in primis*, along with embedding dimensionality, are factors that influence performance, with representations learnt with larger radius and dimensionalities more strongly correlating with expert judgements.

This work opens up a number of avenues for future research. One important research question is *why* the predictive nature of the objective function used by the Skip-gram NLM is conducive of such strong performance. We also conjecture that the use of "mixed" features, e.g., learning representations from both term and concept corpora, may result in further improvements. Another factor that may influence performance is the ordering of the training data, considering that the importance of data samples varies according to the learning rate parameter included in the gradient descent procedure.

6. REFERENCES

[1] P. Agarwal and D. B. Searls. Can literature analysis identify innovation drivers in drug discovery? *Nature reviews. Drug discovery*, 8(11):865–78, Nov. 2009.

[2] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 17(3):229–236, 2010.

[3] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[5] J. E. Caviedes and J. J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *Journal of biomedical informatics*, 37(2):77–85, Apr. 2004.

[6] T. Cohen and D. Widdows. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009.

[7] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[8] Z. S. Harris. Distributional structure. *Word*, 1954.

[9] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2439–2442. ACM, 2012.

[10] R. Lebret, J. Legrand, and R. Collobert. Is deep learning really necessary for word embeddings? In *NIPS Workshop on Deep Learning*, 2013.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[14] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

[15] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

[16] H. Schwenk and J.-L. Gauvain. Training neural network language models on very large corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 201–208. Association for Computational Linguistics, 2005.

[17] E. Voorhees and R. Tong. Overview of the TREC Medical Records Track. In *Twentieth Text REtrieval Conference (TREC 2011)*, MD, USA, 2011.