Predicting sense convergence with distributional semantics: an application to the CogALex-IV 2014 shared task

Laurianne Sitbon
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Australia
laurianne.sitbon@qut.edu.au

Lance De Vine
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Australia
l.devine@student.qut.edu.au

Abstract

This paper presents our system to address the CogALex-IV 2014 shared task of identifying a single word most semantically related to a group of 5 words (queries). Our system uses an implementation of a neural language model and identifies the answer word by finding the most semantically similar word representation to the sum of the query representations. It is a fully unsupervised system which learns on around 20% of the UkWaC corpus. It correctly identifies 85 exact correct targets out of 2,000 queries, 285 approximate targets in lists of 5 suggestions.

1 Introduction

How humans draw associations between words or concepts has been the object of many studies by psychologists, and for many years computer scientists have attempted to model this human mental lexicon by means of symbolic methods (Enguix et al., 2014) or statistical models (Baroni and Lenci, 2013). These models and methods have in turn been used to improve natural language processing systems (Lewis and Steedman, 2013), search technologies (Deerwester et al., 1990) and have since been evaluated in the view of supporting such systems more than helping users directly. The Shared Task CogALex-IV 2014 aims to evaluate how these models can support a user with deficiencies in their lexical access. The task is set as one of retrieving one target word when being presented with 5 cue (associated) words. After submissions of all systems, the organisers revealed that the cue words were the 5 words most often associated with the target words. They have been collected from a large number of users who were presented with the target word and invited to produce one associate. In this paper we present our preliminary investigations to address the task with a neural net language model learning representations for words on the UkWaC corpus (M. Baroni and Zanchetta, 2009). We propose a strict evaluation (accuracy of finding the target word) as well as a retrieval based evaluation that we believe is closer to the aim of helping user find their words.

2 Approach and methodology

2.1 Neural Net Language Model

In 2003 Bengio et. al. (Bengio et al., 2003) introduced a neural net based method for language modelling that learns simultaneously 1) a distributed representation for each word and 2) the probability function for word sequences, expressed in terms of the distributed representations. Generalization to unseen word sequences is obtained because such sequences receive a high probability if they are composed of words that are similar to words from an already seen sequence. An outcome of this approach is the learning of "word embeddings", which are vectors representing the meanings of words relative to other words (via a mechanism akin to word distribution). For this task, we used our own implementation of the continuous Skip-Gram neural language model introduced by (Mikolov et al., 2013). We refer to this model hereafter as skip-gram. The implementation is similar to the word2vec software package. Neither

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 64–67,
Dublin, Ireland, August 23, 2014.

sub-sampling nor negative sampling were used. A small context term window radius of size two and a vector dimensionality of 128 were used. We use the cosine between the word embedding representations (vectors) to estimate the similarity between the words in the evaluation task. The parameters were not tuned for the task and so it is probable that further improvements can be made.

2.2 Combined similarity

Once semantic vectors are created with skip-gram it allows us to measure the distances between words and retrieve the words most similar to another word, or those with a vector most similar to any vector, such as the sum of several word vectors.

In the CogALex-IV shared task, we are provided with 5 words (cues) associated to a target word to be found. If we consider that these words are effectively a unique semantic context for the word to be found, then it makes sense to add their vectors and find the unique word most similar. This approach is of course inspired by vectorial models of information retrieval and adopted widely when testing distributional models for more than single words (see for example (Deerwester et al., 1990)).

However we found that this strategy has limitations, because in the case of some polysemous words, some of the cues were from radically different contexts, and therefore summing up the vector did not necessarily make sense. For such situations, it makes more sense to find the lists of words most related to each of the cues, and then combine these lists. To do this we first selected 10 candidate targets for each cue, which are the 10 words with a representation most similar to the cue, according to the cosine between their words embeddings and that of the cue. We then ranked the words according to their number of occurrences in the 5 lists. We did not consider the distance as measured by cosine similarity (the actual value) because while cosine is a good measure to rank terms by similarity, we do not believe that this leads to an absolute estimate for actual semantic similarity. Additionally, we chose not to assign weights to the terms depending on which cue they were associated to as there was no reason to believe that the cues were ordered in any way (that is, by manual inspection, we did not find that cues early in the lists were most likely to lead to the target than cues later in the list were). The results were not as good then as those with the summed vectors. We then adopted a third strategy, which was to consider the sum of the cues as a 6th cue when generating the lists of candidates, but also to decide on priority when selecting a unique target in case there are several candidates ranked first with an equal number of occurrences in the 6 lists. On the training set, this allowed us to find 92 correct answers for the 2,000 cases.

court	law	judge	judges	courts	SUM
courts	laws	judges	appellants	court	court
sheriff	legislation	pettiti	judge	rackets	courts
tribunal	jurisprudence	court	defendants	**magistrates**	judge
prosecution	statutes	**sheriff**	respondents	badminton	judges
judge	statute	prosecutor	panellists	sharia	**sheriff**
justiciary	litigation	dredd	jury	squash	law
judicature	antiunion	jury	organizers	tribunals	**magistrates**
consistory	sharia	coroner	complainants	prosecutors	prosecution
leet	criminology	appellant	winners	proceedings	prosecutors
magistrates	arbitration	defendant	plaintiffs	parliaments	prosecutor
prosecutor	llm	magistrate	**magistrates**	rulings	tribunal
contactfulhamba	regulation	complainant	appellant	law	consistory
appeal	courts	**magistrates**	senatus	prosecution	rulings
palace	penal	appeal	chairmen	leagues	judicature

Figure 1: Example of lists of 14 most similar words for the 5 cues “court law judge judges courts” and their sum vector

Figure 1 shows an example where the cues are “court law judge judges courts” and the target was

magistrates. We present for each cue as well as for the sum of all cues the lists of 14 most similar terms. In gray the words that were cues or plural of a cue were ignored. In bold we show how "sheriff" would have been picked if we considered the sum only, while when considering the individual sets of similar terms in addition to the sum we could find that magistrates was a more likely target.

2.3 Training corpus

The corpus we used for learning the word representations is the UKWaC corpus (M. Baroni and Zanchetta, 2009). This is the corpus suggested by the organisers of the CogALex-IV 2014 Shared task, and contains web pages from the UK domain. We pre-processed the corpus by a) lower-casing all terms, b) replacing contractions with their expansion, eg. "it's" becomes "it is", c) removing all punctuation and d) replacing all digits with the single character '7'. The Skip-gram model that we used is able to scale to the whole UKWaC corpus (approximately 2 billion terms) but because of time constraints we selected only the first 20% of the corpus, and then processed the remainder of the corpus, adding to our corpus subset all sentences containing words that were present in the training set but not in the initial 20% subset. This was to ensure that representations for all the words in the training set could be learned.

3 Results

3.1 Shared task evaluation

The evaluation proposed in the shared task is the exact accuracy of a single proposed target for each query composed of 5 words. There were 2,000 queries in each of the training and test set, and the results are expressed in total number. All our results according to this metric are situated between 4% and 5%. We have included in table 1 the results according to the task metric, on the training and on the test set, for both the sum vectors and the combination of results from a sum and the individual words. The latter is the one that was submitted to the shared task.

Method	Train	Test
Sum	81	75
Combination	84	85

Table 1: Accuracy of the methods on the training and on the test corpus

3.2 Retrieval evaluation

We now consider a task where a system would support a user in finding a word that they describe using the 5 associations. In such a tip-of-the-tongue context, users would immediately recognise the word they are looking for when presented in a list and also if it is presented with a different inflection (ie. "run" instead of "running"). Therefore, if presented a list of words containing the target word or variation of the target word, the outcome of such a system would be considered successful. While it would be impractical to consider very long lists in a usability context, we have measured the accuracy for lists of 2, 3, 4 and 5 words, with a measure of 1 where the word (or at least one of its inflections) is in the list and 0 otherwise. In other words, the accuracy is the number of target words that appear as is or as an inflection in suggestion lists of varying sizes.

The results presented on Figure 2 show that taking inflections into account leads to only marginal improvements, but more importantly considering additional targets (as a list) can really improve outcomes for the users, with almost 13% of targets being retrieved in lists of 5 suggestions with the combined approach.

4 Discussion and conclusion

The accuracy of our approach, even when considering lists of 5 suggestions and inflections of words, show that results are still very low if one would consider a usable assistance system for users with lexical access issues. This is consistent with previous findings on a similar task in French (Sitbon et al., 2008).

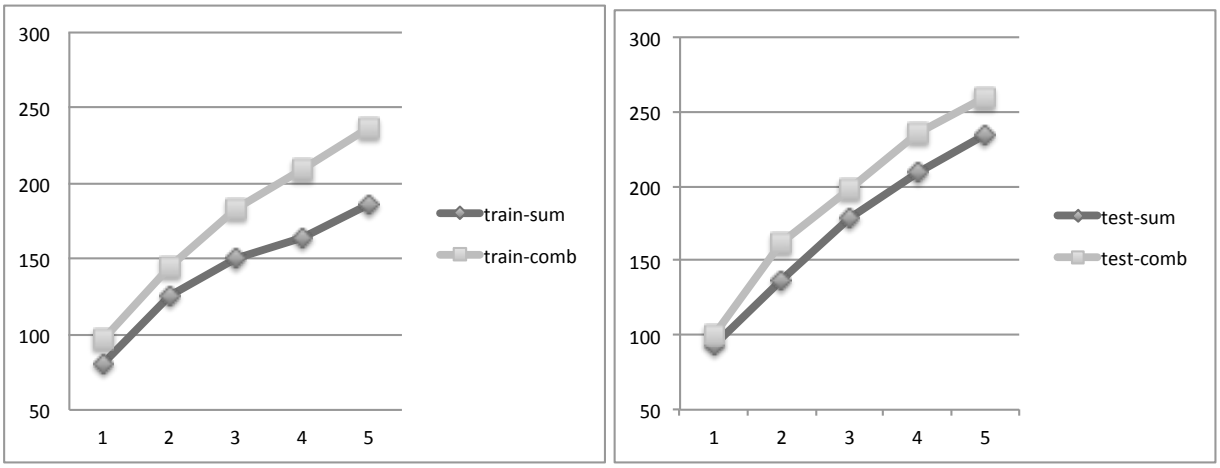

Figure 2: Total number of targets on the left, or inflections of target on the right, out of 2,000, found in lists of 1 to 5 results, in the training set and in the test set

This work suggested that a combination of resources encoding various types of semantic relations would be best, along with user models. CogALex-IV task was not based on associations drawn by a single user, but rather by majority associations drawn by many users, so this would not apply to the task specifically. However we believe that including definitional associations such as that drawn from an ESA model on the Wikipedia would be a way to dramatically improve the accuracy, at least when considering lists of results. Additionally it would be interesting to inspect a number of variables to weigh the contribution of each cue (depending on their specificity for example). In this paper we found that adding the vectors representing each word let to better results than only considering the words individually. This mode of combination is one of many proposed by (Mitchell and Lapata, 2010) and in future work we will experiment with alternative combination models. Finally, an area for future work would be to consider cleaning up the dataset so as to avoid effects such as several cues being inflections of one another (i.e.. "courts" and "court") or even the target being an inflection of one of the cues, as we have observed in the CogALex-IV dataset.

References

Marco Baroni and Alessandro Lenci. 2013. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Gemma Bel Enguix, Reinhard Rapp, and Michael Zock. 2014. A graph-based approach for computing free word associations. In *Proceedings of LREC 2014*.

Mike Lewis and Mark Steedman. 2013. Combined distributional and logical semantics. *TACL*, 1:179–192.

A. Ferraresi M. Baroni, S. Bernardini and E. Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

J. Mitchell and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, (34):1388–1429.

L. Sitbon, P. Bellot, and P. Blache. 2008. Evaluation of lexical resources and semantic networks on a corpus of mental associations. In *Proceedings of the Language Ressources and Evaluation Conference*.