Causality in vectors space language models

Márton Makrai
Hungarian Academy of Sciences, Research Institute for Linguistics

Abstract

Vector space language models (VSM) are tools for various human language understanding tasks in which decomposition of word meaning is also possible. We explore the geometric relation between a vector representing a cause like *hurt* and the vector representing the corresponding effect (*ache*). In a VSM called SENNA, we find that lines fitting to various causal pairs taken from WordNet run close to a common center point. This observation offers the interpretation that the meaning of an effect word is a combination of the meaning of its cause and a causal element independent of the actual pair.

1 Introduction

Computational linguistics targets problems, where the computer has to understand human texts, written or spoken. We say that the computer understood the text if the user has the impression of being understood. Understanding is needed for machine translation, automatic question answering, and intelligent web search among others.

All these tasks require a so-called language model that, in the simplest case, computes the probability (naturalness) of a word sequence. Probabilities are estimated using (relative) frequencies. As there are infinitely many possible sentences but the model is trained on a finite sample, the main point is in generalization. A simple and effective approach to language modeling is the family of n-gram models that make the simplifying assumption that the probability of a word in a context depends only on preceding words of some fixed number (four in most applications). Thus the probability of the Hungarian word string *minden madár társat választ* (every bird is choosing a partner)[1] is computed as a product of conditional probabilities:

[1] This sentence is from the song that gave the title of the Spring Wind Conference.

$$P(\texttt{\^{} minden madár társat választ \$}) =$$

$$P(\ \texttt{minden} \mid \texttt{\^{}}\ ) \cdot P(\ \texttt{madár} \mid \texttt{minden}\ ) \cdot P(\ \texttt{társat} \mid \texttt{madár}\ ) \cdot$$
$$\cdot P(\ \texttt{választ} \mid \texttt{társat}\ ) \cdot P(\ \texttt{\$} \mid \texttt{választ}\ )$$

$P(\ \texttt{madár} \mid \texttt{minden}\ )$ denotes the probability of the word *madár* given that the preceding word was *minden*. `^` and `$` denote the beginning and the end of the string respectively. While n-gram models are easy to understand and useful in application, they have the disadvantage of not capturing morphological and semantic relations between words.

2 Vector space language models

In recent years, vectors space language models (VSM) have gained popularity. These resources model each word with a vector in multidimensional real vector space (say 25 to 200 dimensional), see Figure 1. The motivation for this is that different dimensions of the models can capture different properties of word forms. Vector space language models are computed by machine learning instead of human work what makes interpretation of individual dimensions difficult.

There are different ways to compute VSMs. It is possible to take a matrix in which rows correspond to words and columns correspond to contexts. Context means the words preceding and following the focus word to some fixed distance. Elements of the matrix are frequencies of the corresponding word in the corresponding context, and model vectors are computed from the matrix by rank reduction.

A more novel and very successful architecture is when word vectors are machine-learned in a neural net simultaneously with the parameters of the net. There is a third method as well: in Makrai et al. (2013) we computed a VSM from the graph representation of a computational lexical resource `4lang`.

As the VSMs created from the `4lang` concept lexicon were used in the present experiment, we characterize it briefly. `4lang` is a lexicon for general purpose human language understanding. The resource contains a 1000-word basic vocabulary in 50 languages, defined in a formal language. The basic vocabulary is appropriate for defining all the words in everyday language and can be extended for specific tasks. Word forms in 50 languages include translations in four languages created by human labor (English, Hungarian, Polish and Latin), while others were collected automatically (Ács, 2014). The definitions are logical formulas that can be translated to graphs as well,

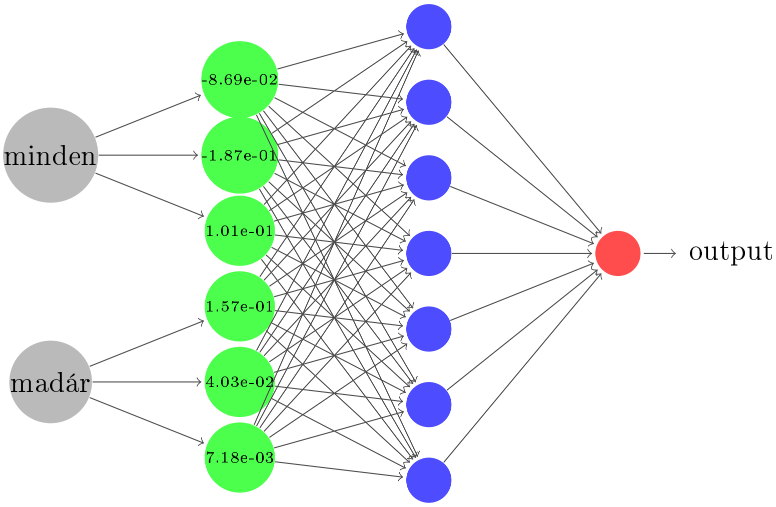

Figure 1: VSMs represent words by vectors (3-dimensional in the picture, and 500 to 200 in reality. The output is computed from the concatenation of vectors for words in a window of length n (*minden madár* in the Hungarian example. In this picture, computation is illustrated by a neural net, and architecture which is very successfully applied today.

see Figure 3. The definitions were created by human labor, taking typologically highly different languages into account. The resource is called a concept dictionary because items are more abstract than those in dictionaries made either for human or for machine use: in lexical meaning, a monosemic approach is followed and part of speech differences are also factored out as attributed to syntax.

Mikolov et al. (2013) shows that many semantic elements like 'gender', which is the systematic difference between words pairs like $(king, queen)$, $(uncle, aunt)$, $(Mrs., Mr.)$, etc. are mirrored in vector space languages models (see Figure 2).

In the present work we take a semantic relation with rich literature in philosophy and application in knowledge representation, causality (see Figure 3). We are interested in the geometric function mapping a vectors representing some cause (e.g. *hurt*) to the vector representing its effect (*ache*).

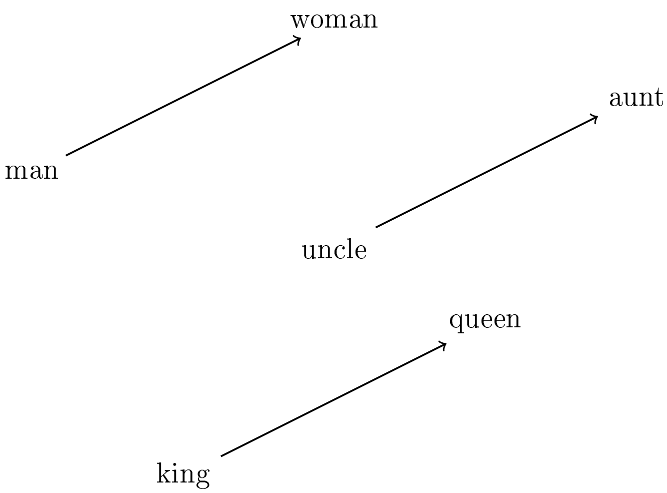

Figure 2: The vector offset method (Mikolov et al., 2013) is based on the fact that vectors representing words with a systematic difference in meaning differ in approximately the same vector, making lexical decomposition like $\mathbf{v}\,(\text{king}) + \mathbf{v}\,(\text{female}) = \mathbf{v}\,(\text{queen})$ possible.

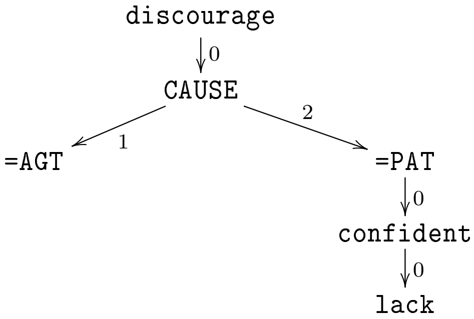

Figure 3: The definition of *discourage* in the 4langconcept lexicon exemplifies the use of ‘cause’ in associative network representations of linguistics knowledge. The graph expresses that *discourage* means, that the agent (=AGT) causes the participant that is called patient in linguistics (=PAT) to lack confidence.

3 Results

In this section we describe resources, methods, and results. As the results are preliminary, we outline directions of further research as well.

We took causal word pairs from a natural language processing resource containing lexical information of various kinds, WordNet (Miller, 1995). The pairs are listed in Table 1. We took several VSMs: SENNA (Collobert et al., 2011), Turian et al. (2010); Huang et al. (2012), HLBL (Mnih and Hinton, 2009), the English Polyglot (Al-Rfou' et al., 2013), and 24, variants of the model created from `4lang`. Casual pairs were projected to a 2-dimensional plane by principal component analysis, a machine learning technique often used for visualizing high-dimensional data. The visualization suggested that that there is a center in the vector space representing the words, that approximately fits the lines containing each causal pair, see Figure 4.

For testing the centrality property in the original, unreduced space, we took random word pairs of the same number as we have causal pairs. The point closest to all the lines fitting each pair was computed for both the real and the random sample of word pairs using a formula by Han and Bancroft (2010). Distances of the lines to the corresponding center was also computed. Centrality implies that the expected value of the distances if lower in the real case than in the random case. An unpaired t-test showed that this condition holds in the case of SENNA ($p < 0.001$).

Some of the models created from `4lang` also show significant ($p < 0.05$) difference, but this statistical result has to be taken with caution, because of the phenomenon known as *multiple testing* (Domingos, 2012).

> Standard statistical tests assume that only one hypothesis is being tested, but modern learners can easily test millions before they are done. As a result what looks significant may in fact not be. [...] This problem can be combatted by correcting the significance tests to take the number of hypotheses into account [...]

Multiplying the p values by 24 significance is lost, so we should motivate the choice of some specific model among all `4lang` models on some independent grounds to make results significant. This remains a problem for further research.

give	have
show	see
encourage	hope
feed	eat
kill	die
raise	rise
⋮	⋮

Table 1: Word causes and effects in WordNet. WordNet contains semantic relations like is-a (a chair is a furniture), instance-of (Mozart is an instance of 'composer'), antonym (cold and hot), part-of (Monday is a part of 'week') as well.

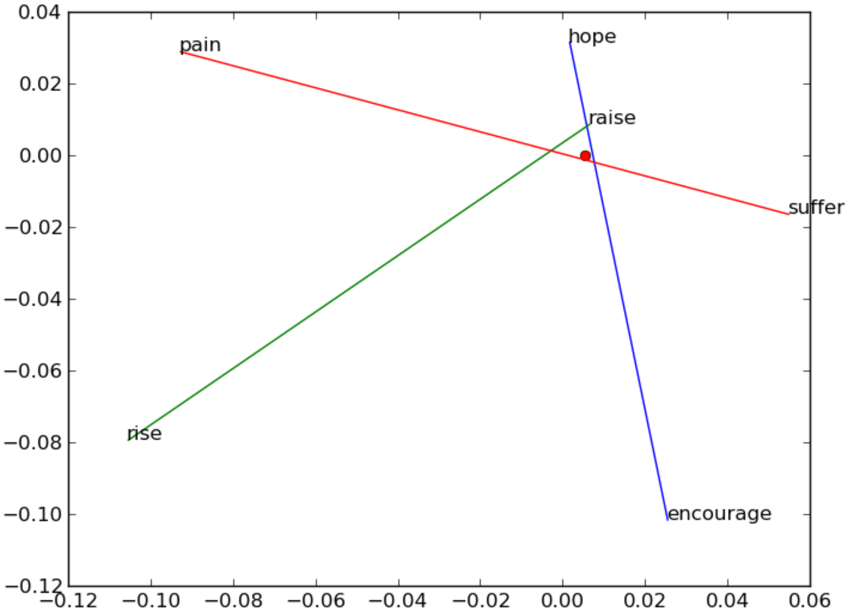

Figure 4: A 2-d visualization of causal pairs in the VSMs suggest that lines connecting causal pairs run close to a common center point.

4 Conclusion

Looking for an insightful interpretation of causality in VSMs, we have found a center point $\mathbf{c}$ in the VSM SENNA with the property that the lines connecting the two members of causal word pairs run close to $\mathbf{c}$. In algebraic terms this means that

$$\mathbf{v}\,(\text{effect}) \approx \lambda \mathbf{v}\,(\text{cause}) + (1-\lambda)\mathbf{c} \quad \lambda \in \mathbb{R},$$

reflecting the linguistic intuition that the meaning of the effect is a combination of the meaning of the cause and a causal element.

Further research should be made to discover more sophisticated connections between cause and effect vectors that apply to more models, possibly all models obtained by one or more of the three mentioned methods (co-occurrence matrices, neural nets, and lexicon graphs).

5 Acknowledgment

I would like to thank my advisor András Kornai for his helpful ideas and encouragement, and Balázs Szalkai for reminding me to the problem of multiple testing.

References

Judit Ács. Pivot-based multilingual dictionary building using Wiktionary. In *The 9th edition of the Language Resources and Evaluation Conference*, May 2014.

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-3520`.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 2011.

Pedro Domingos. A few useful things to know about machine learning. In *Communications of the ACM*, volume 55, pages 78–87. ACM New York, NY, USA, October 2012.

Lejia Han and John C. Bancroft. Nearest approaches to multiple lines in n-dimensional space. In *CREWES Research Report*, volume 22. University of Calgary, 2010.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2390524.2390645`.

Márton Makrai, Dávid Márk Nemeskey, and András Kornai. Applicative structure in vector space models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 59–63, Sofia, Bulgaria, August 2013. ACL. URL `http://www.aclweb.org/anthology/W13-3207`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *Proc. ICLR 2013*, 2013.

George A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21: 1081–1088, 2009.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 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, 2010.