

# Filtering Wiktionary triangles by linear mapping between distributed models

Márton Makrai

Research Institute for Linguistics  
Hungarian Academy of Sciences  
Benczúr u. 33, H-1068 Budapest  
makrai.marton@nytud.mta.hu

*Abstract content*

Word translations arise in dictionary-like organization as well as via machine learning from corpora. The former is exemplified by Wiktionary, a crowd-sourced dictionary with editions in many languages. Ács et al. (2013) obtain word translations from Wiktionary with the pivot-based method, also called triangulation, that infers word translations in a pair of languages based on translations to other, typically better resourced ones called pivots. Triangulation may introduce noise if words in the pivot are polysemous. The reliability of each triangulated translation basically is measured by the number of pivot languages (Tanaka and Umemura, 1994).

Mikolov et al. (2013b) introduce a method for generating or scoring word translations. Translation is formalized as a linear mapping between distributed vector space models (VSM) of the two languages. VSMS are trained on monolingual data, while the mapping is learned in a supervised fashion, using a seed dictionary of some thousand word pairs. The mapping can be used to associate existing translations with a real-valued similarity score.

This paper exploits human labor in Wiktionary combined with distributional information in VSMS. We train VSMS on gigaword corpora, and the linear translation mapping on direct (non-triangulated) Wiktionary pairs. This mapping is used to filter triangulated translations based on scores. The motivation is that scores by the mapping may be a smoother measure of merit than considering only the number of pivot for the triangle. We evaluate the scores against dictionaries extracted from parallel corpora (Tiedemann, 2012). We show that linear translation really provides a more reliable method for triangle scoring than pivot count.

The methods we use are language-independent, and the training data is easy to obtain for many languages. We chose the German-Hungarian pair for evaluation, in which the filtered triangles resulting from our experiments are the greatest freely available list of word translations we are aware of.

## 1. Triangulation

A method for creating dictionaries is triangulation through better-resourced ones called the *pivot* (Tanaka and Umemura, 1994). The idea is that if the English translation of the Hungarian word *letartóztat* is *arrest*, and the German translation of *arrest* is *verhaften*, then the German translation of *letartóztat* is *verhaften*, see Figure 1.

Triangles are corrupted by ambiguity in the pivot word (the one in the middle): German *Dose* can be translated as *can*

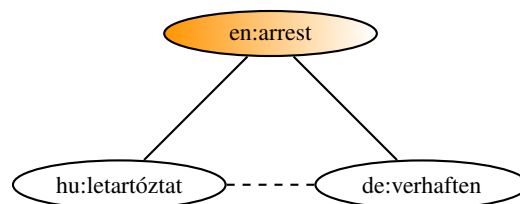


Figure 1: Triangulation

to English (as a synonym of *tin*), which, as a verb, translates to *tud* in Hungarian, which is unrelated to *Dose*. Saralegi et al. (2011) analyze two methods for pruning wrong triangles: one based on exploiting the structure of the source dictionaries, and the other based on distributional similarity computed from comparable corpora. This paper is more similar to the later in that it uses distributional information applying a method connected to neural language modeling.

## 2. Vector space language models

In this section we introduce vector space language models as two interrelated families of word representations. The traditional method takes the co-occurrence matrix as a starting point, while more recent representations are learned as weights (*word embeddings*) in neural networks (*deep learning*). Interestingly, the probably most popular neural language model, skip-gram with negative sampling (Mikolov et al., 2013a) is not so deep in architecture, and has been shown (Levy and Goldberg, 2014) to be equivalent with a method based on the co-occurrence matrix, *shifted point-wise mutual information*.

The primary source of information about the meaning of a word is how often it is used in different contexts, an idea called the *distributional hypothesis* by linguists going back to Harris (1951), and often quoted in the form that “You shall know a word by the company it keeps” (Firth, 1957). One simple formalization of word distribution in a corpus is the *co-occurrence* matrix whose rows correspond to words in the vocabulary, columns to contexts, and cells contain the occurrence count of the word corresponding to the row appearing in the context corresponding to the column. What is meant by context depends on the application. In Latent Semantic Analysis (Deerwester et al., 1990), columns of the original (unreduced) matrix correspond to documents. In matrix-based vector space language models (Turney and Pantel, 2010) on the other hand, columns originally correspond to words, and counts express how often the words

corresponding to the row and the column collocate in a window of some fixed length (say 5). Both in LSA and co-occurrence based VSMs, the number of contexts is at least in the thousands and gets reduced to some hundred dimensions for computation efficiency.

Neural language models (Bengio et al., 2003), on the other hand, are neural nets, trained on gigaword corpora by iterating over words in their contexts and updating some weights of the model at each word. The resulting VSMs represent similar words with similar vectors, and VSMs also reflect relational similarities between words like *king – queen*  $\approx$  *man – woman* (Mikolov et al., 2013c).

### 3. Linear translation

Mikolov et al. (2013b) discovered that VSMs of different languages have such similarities that a linear mapping can map representations of source language words to the representation of their translations. The method belongs to the paradigm of supervised machine learning: specifically it makes use of a great amount of monolingual data i.e. gigaword corpora for training, needing to be supervised by a seed dictionary of some thousand words. Mikolov et al. formalize translation as linear mapping  $W \in \mathbb{R}^{d_2 \times d_1}$  from the source (monolingual) VSM  $\mathbb{R}^{d_1}$  to the target one  $\mathbb{R}^{d_2}$ : the translation  $z_i \in \mathbb{R}^{d_2}$  of a source word  $x_i \in \mathbb{R}^{d_1}$  is approximately its image  $Wx_i$  by the mapping. The translation model is trained with linear regression on the seed dictionary

$$\min_W \sum_i \|Wx_i - z_i\|^2$$

and can be used to collect translations for the whole vocabulary (by choosing  $z_i$  to be the nearest neighbor of  $Wx_i$ ) or to score a translation  $z$  coming from some other source (with the score being the distance between  $Wx_i$  and  $z_i$ ).<sup>1</sup> In the original setting of the collection mode, evaluation is done on another thousand seed pairs.

A common error in linear translation is when there are target words that are returned as the translation of many words, which is wrong in most of the cases. Dinu et al. (2015) propose a method for downplaying the importance of such target words they call *global correction*. Our experiments use this method.

### 4. Data

Direct and triangulated Wiktionary translations have been extracted with wikt2dict<sup>2</sup> (Ács et al., 2013) that handles 43 editions of Wiktionary.

The German VSMs have been trained on SdeWaC (Baroni et al., 2009) and the Hungarian on the concatenation of the Hungarian Webcorpus (Halácsy et al., 2004) and

<sup>1</sup>Mikolov et al. use a surprising combination of vector distances, Euclidean distance in training and cosine similarity (and distance) in collection (and, respectively, scoring) of translations. This choice is theoretically unmotivated, but we (Makrai, 2015) also found it to work better than more consistent combinations of metrics. However, see Xing et al. (2015) for opposing results.

<sup>2</sup><https://github.com/juditacs/wikt2dict>

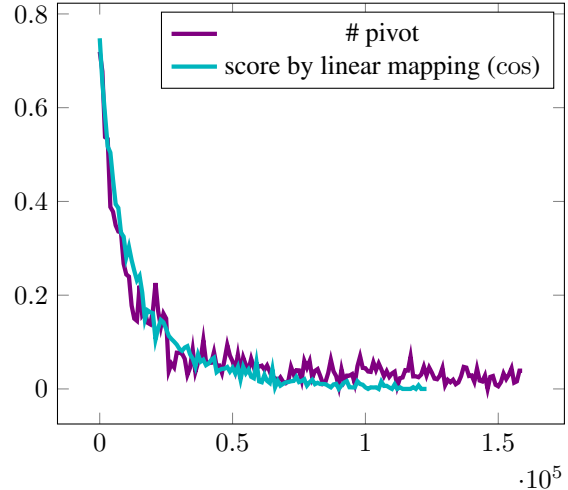


Figure 2: Accuracy of triangles sorted by pivot count as baseline, or cos. The later is more monotonic.

the Hungarian National Corpus (Oravecz et al., 2014) with word2vec<sup>3</sup> (Mikolov et al., 2013a).<sup>4</sup>

For training and using the linear mapping, we forked<sup>5</sup> the implementation by Dinu et al. (2015). The German to Hungarian mapping was trained on the 5K direct word pairs that are supported by the most pivots in Wiktionary. All the triangles were scored. The scoring has been evaluated against a dictionary extracted from the OpenSubtitles2013 parallel corpus in the OPUS project<sup>6</sup> (Tiedemann, 2012). Glue code we wrote for this can be found in the repo <https://github.com/makrai/efnilex-vect>.

### 5. Evaluation

We evaluated the vector-based scoring of triangulated translational word pairs (*triangles*) in a comparison with a dictionary created from the parallel corpus OpenSubtitles2013. For each (German) word, we consider as gold translations all the (Hungarian) words that are listed in the OpenSubtitles2013 dictionary as its translation.

For evaluation, we sort the triangles in two orders: as baseline, by the number of pivots for the triangle, and more importantly, by the score in the linear mapping (cos). Then in each order, we compute accuracy on each 1000-word slice of the list taking OpenSubtitles2013 translations as gold. Figure 2 shows that in sort by cos accuracy descends in a more monotonic fashion than in sort by pivot count. The last 22.73% of the nearly 160 K triangles is out of the vocabulary of one or both of the VSMs, so cos cannot be computed.

### 6. Acknowledgement

I would like to thank Vladimír Benko for information on corpora, Judit Ács and Sergey Bartunov for help with their

<sup>3</sup><https://code.google.com/p/word2vec/>

<sup>4</sup>The German VSM has been a continuous bag of words model in 300 (infrequent words have been cut off at 100), the Hungarian a 600 dimensional one (with a cut-off of 10). The choice of meta-parameters was not fully systematic.

<sup>5</sup><https://github.com/makrai/dinu15/>

<sup>6</sup><http://opus.lingfil.uu.se/>

tools, and Márton Miháltz, Csaba Oravecz, Bálint Sass, Attila Zséder, and András Kornai for useful discussions. Work supported by the EFNILEX project of the European Federation of National Institutions for Language.

## References

- Judit Ács, Katalin Pajkossy, and András Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. ACL.
- M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *LREC 2009*, volume 3, pages 209–226, 2009.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.
- Scott C. Deerwester, Susan T Dumais, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
- Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR 2015, Workshop Track*, 2015.
- John R. Firth. A synopsis of linguistic theory. In *Studies in linguistic analysis*, pages 1–32. Blackwell, 1957.
- Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*, pages 203–210, 2004.
- Zellig Harris. *Methods in Structural Linguistics*. University of Chicago Press, 1951.
- Omer Levy and Yoav Goldberg. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, chapter Linguistic Regularities in Sparse and Explicit Word Representations, pages 171–180. Association for Computational Linguistics, 2014. URL <http://aclweb.org/anthology/W14-1618>.
- Márton Makrai. Comparison of distributed language models on medium-resourced languages. In Attila Tanács, Viktor Varga, and Veronika Vincze, editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, 2015. ISBN 978-963-306-359-0.
- 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*, 2013a.
- Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168, 2013b.
- Tomas Mikolov, Wen-tau Yih, and Zweig Geoffrey. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT-2013*, pages 746–751, 2013c.
- Csaba Oravecz, Tamás Váradi, and Bálint Sass. The Hungarian Gigaword Corpus. In *Proceedings of LREC 2014*, 2014.
- Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011.
- Kumiko Tanaka and Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics, 1994.
- Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.
- Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.
- Chao Xing, Chao Liu, RIIT CSLT, Dong Wang, China TNList, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*, 2015.