# Filtering Wiktionary triangles by linear mapping between distributed word models

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

**Abstract**

Triangulation infers word translations in a pair of languages based on translations to other, typically better resourced ones called pivots. This method may introduce noise if words in the pivot are polysemous. The reliability of each triangulated translation is basically estimated by the number of pivot languages (Tanaka and Umemura, 1994).

Mikolov et al. (2013b) introduce a method for 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 supervised fashion, using a seed dictionary of some thousand word pairs.

We apply linear mapping to filter triangulated translations, and show that scores by the mapping are smoother measure of merit than the number of pivots. 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.

**Keywords:** word triangulation, word embedding, Wiktionary

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 is basically estimated 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.

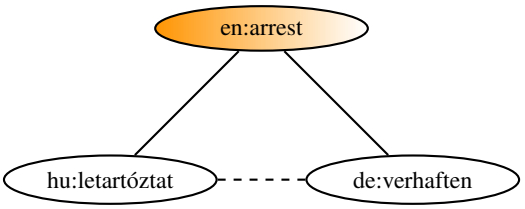


Figure 1: Triangulation

## 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 fig. 1.

Triangles are corrupted by ambiguity in the pivot word (the one in the middle): German *Dose* can be translated as *can* 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 pointwise 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 $\mathbf{king} - \mathbf{queen} \approx \mathbf{man} - \mathbf{woman}$ (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 $z_i$).[1]

[1]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.

| | |
|---|---|
| documents | 3208 |
| sentences | 3.2 M |
| German tokens | 23.3 M |
| Hungarian tokens | 19.7 M |
| extracted word pairs | 29.1 K |

Table 1: The German Hungarian section of the OpenSubtitles2013 parallel corpus (Tiedemann 2012)

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[2] (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 the Hungarian National Corpus (Oravecz et al., 2014) with word2vec[3] (Mikolov et al., 2013a).[4]

For training and using the linear mapping, we forked[5] 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. Glue code we wrote for this project is freely available[6].

The scoring has been evaluated against a dictionary in the OPUS project[7] that has been extracted by Tiedemann (2012) from the OpenSubtitles2013 parallel corpus, a collection of translated movie subtitles[8]. OpenSubtitles2013 contains 59 languages. Some sizes of the German Hungarian section are shown in table 1.

Most of our training data are general in their *domain*: web corpora (SdeWaC, the Hungarian Webcorpus), a curated corpus (the Hungarian National Corpus, as far as a corpus of 754 million words may be curated), and a crowd-sourced but otherwise causal dictionary (Wiktionary). One may ask whether the domain of the reference dictionary extracted

[2]https://github.com/juditacs/wikt2dict

[3]https://code.google.com/p/word2vec/

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

[5]https://github.com/makrai/dinu15/

[6]https://github.com/makrai/efnilex-vect

[7]http://opus.lingfil.uu.se/

[8]http://www.opensubtitles.org/

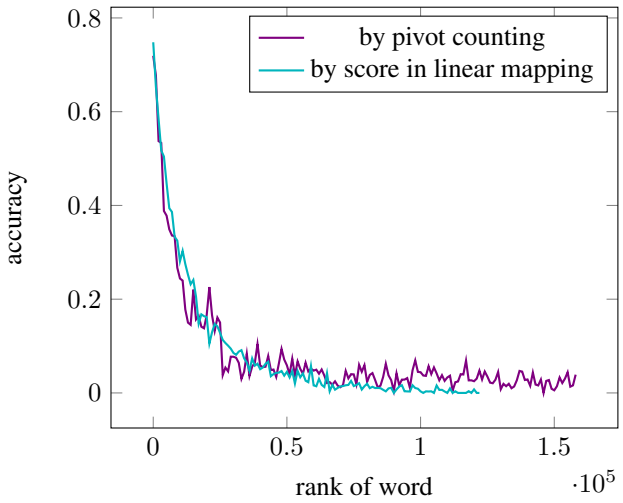


Figure 2: Accuracy curve of triangles sorted by their pivot count as baseline, or score in linear translations (cos). The later is smoother.

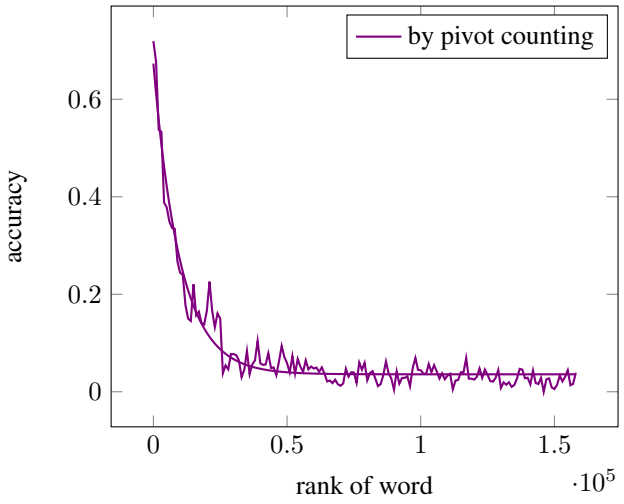


Figure 3: The accuracy curve of pivot counting approximated by an exponential function.

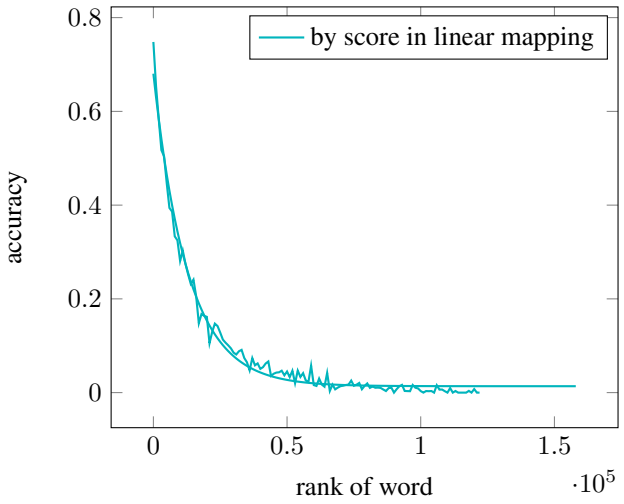


Figure 4: The accuracy curve of scores by the linear mapping approximated by an exponential function.

from movie subtitles is general to an appropriate extent, or how far a problem of domain mismatch between train and test may arise. We hypothesize that the mismatch is negligible and defer a more subtle analysis to further research.

## 5. Evaluation

We evaluated the vector-based scoring of triangulated translational word pairs (*triangles*) in 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 (e.g. triangles 1–1000, then 1001–2000, etc.) taking OpenSubtitles2013 translations as gold.

While overall accuracy of the linear scoring (8.58%) is slightly worse than that of pivot counting (9.32%), fig. 2 suggests that in sort by cos, accuracy descends more smoothly 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.) Now we turn to a more quantitative support of this visual analysis.

### 5.1. Quantitative analysis of smoothness

We measure the smoothness of the accuracy curves by how well they can be approximated by a function in some parametric family, see figs. 3 to 6. We tried two families with similar results. The fist family is exponential functions of the form

$$a \cdot \exp(-bx) + c,$$

where $x$ is the index of the vocabulary slice (0 for words 0–1000, 1 for 1001–2000, etc), and $a, b$, and $c$ are parameters to fit. The second family is that of power law functions

$$a \cdot (bx + c)^k,$$

where $k$ is another parameter to fit, and the remaining variables play similar roles as in the exponential case. The error of the fit (i. e. the lack of smoothness) is quantified as the mean squared error (MSE) between the two curves. The MSE of the two accuracy curves (scoring translations by pivot counting or cosine score) approximated by the two families (exponential or power law functions) are shown in table 2. The MSE of the accuracy curve in pivot counting is 2.51 (resp. 4.42) times more than that in scoring by the linear mapping, when both curves are modeled as exponential (resp. power law) functions. It is probably also worth mentioning that the accuracy is slightly better for 20–30 000 higher-ranked words in the proposed method than in the baseline, see figs. 7 and 8.

## 6. Acknowledgement

I would like to thank Vladimír Benko for information on corpora, the reviewers for useful and thorough comments, Judit Ács and Sergey Bartunov for help with their tools, and Bálint Daróczy, Márton Miháltz, Csaba Oravecz, Bálint

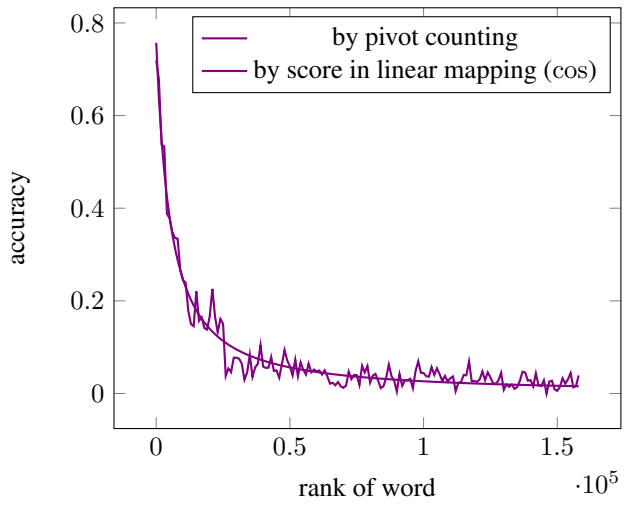


Figure 5: Accuracy curves of scores by pivot count approximated by power law functions.

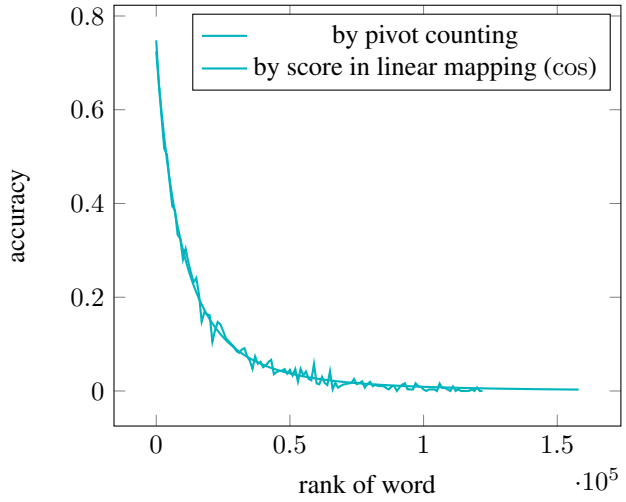


Figure 6: Accuracy curves of scores by the linear mapping approximated by power law functions.

Sass, Attila Zséder, and András Kornai for useful discussions. Work supported by the EFNILEX project of the European Federation of National Institutions for Language.

## 7. Bibliographical References

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

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Deerwester, S. C., Dumais, S. T., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dinu, G., Lazaridou, A., and Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. ICLR 2015, Workshop Track.

Firth, J. R. (1957). A synopsis of linguistic theory. In *Studies in linguistic analysis*, pages 1–32. Blackwell.

Harris, Z. (1951). *Methods in Structural Linguistics*. University of Chicago Press.

Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180. Associa-

| scoring method | exp | power law |
|---|---|---|
| pivot counting | 6.1859e-04 | 5.2182e-04 |
| linear mapping | **2.4574e-04** | **1.1789e-04** |
| ratio | 2.51 | 4.42 |

Table 2: The mean squared error of fitting parametric curves to the accuracy values obtained by translation scoring methods. Linear mapping produces a smoother accuracy decay than pivot counting.

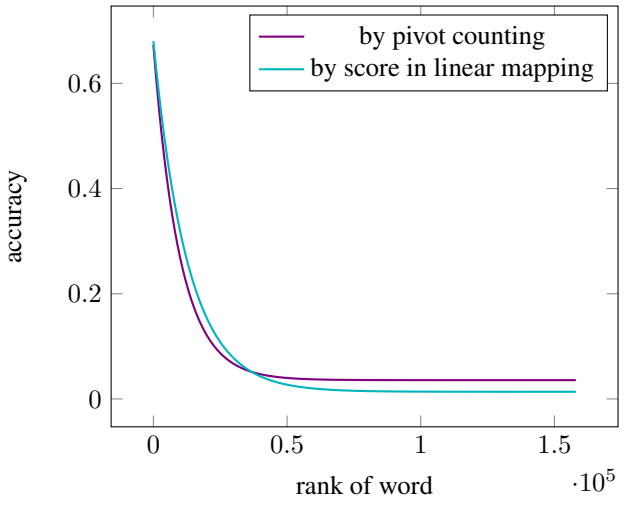


Figure 7: The exponential approximations of the accuracy curves.

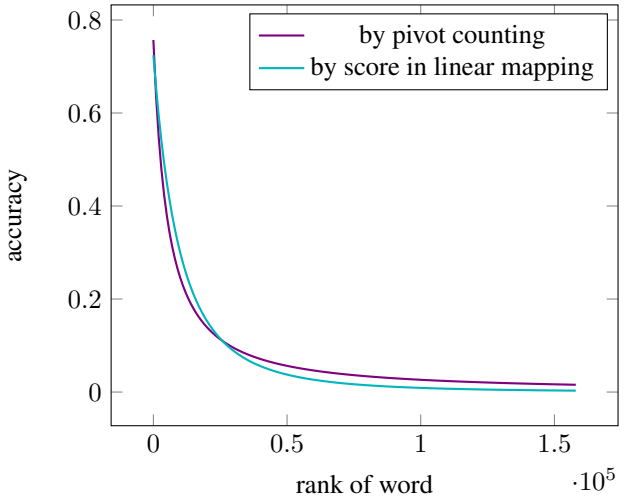


Figure 8: The power law approximations of the accuracy curves.

tion for Computational Linguistics.

Makrai, M. (2015). Comparison of distributed language models on medium-resourced languages. In Attila Tanács, et al., editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, pages 22–33. Szegedi Tudományegyetem Informatikai Tanszékcsoport.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Y. Bengio et al., editors, *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, 05.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Saralegi, X., Manterola, I., and Vicente, I. S. (2011). 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.

Tanaka, K. and Umemura, K. (1994). 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.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Xing, C., Liu, C., Wang, D., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*, pages 1005–1010.

## 8. Language Resource References

Ács, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., and Trón, V. (2004). Creating open language resources for Hungarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 203–210. ELRA.

Oravecz, C., Váradi, T., and Sass, B. (2014). The Hungarian Gigaword Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, editor, *LREC*, Istanbul, Turkey, 05. European Language Resources Association (ELRA).