# Disambiguated linear word translation in medium European languages

Márton Makrai

Research Institute for Linguistics of the Hungarian Academy of Sciences
1068 Budapest VI., Benczúr u. 33., Hungary
E-mail: makrai.marton@nytud.mta.hu

**Abstract for demonstration**

***Abstract*—An earlier paper used triangulated word translations as seed in linear translation between medium European languages. The present work improves upon it by handling word ambiguity both in the main (i.e. source and target) languages and in the pivot.**

Many successful computation formalisms have been motivated by modeling the brain, going back to finite state automata and artificial neural networks, the latter of which has recently gave state-of-the-art performance in speech technology (Seide et al., 2011), object recognition (Krizhevsky and Sutskever, 2012), and human language understanding, which is the topic of the present work. Neural language models (Bengio et al., 2003) predict probabilities of word sequences based on *word embeddings* that represent words in vector spaces of some hundred dimensions (*vector space models*, VSMs). Experiments in predicting brain activities associated with the meanings of words via vector space representations (Mitchell et al., 2008) place neural language models in the perspective of brain-computer interfacing.

Neural language models, being special cases of neural nets, are trained on gigaword corpora by iterating over words in their contexts and updating some parameters of the model at each word. This can be extended to the problem of modeling *multiple languages* in three alternative ways. In the *mapping* approach, models of the different languages are trained separately and a mapping is learned either from the source model to the target (Mikolov et al., 2013) or from more languages to a smaller common model (Faruqui and Dyer, 2014). *Adaptation* exploits a well-trained model for some resource-fortunate language attached with some inter-lingual constraints (Zou et al., 2013). Other architectures, including the seminal (Klementiev et al., 2012) and the more recent Luong et al. (2015) train aligned models of more languages *simultaneously*.

One of the greatest difficulties in translation, done either by a human or a machine, is *ambiguity* of word forms: words have to be translated differently according to context. VSMs with multiple prototypes for each meaning of a word have been proposed (Reisinger and Mooney, 2010), specifically in neural ones by Huang et al. (2012) who cluster word occurrences to a uniform number of meanings prior to training the models. (The term *prototype* has been taken from psychological concept modeling.) Neelakantan et al. (2014) improve upon the former work in three aspects: they do word sense discrimination simultaneously with the training of word embeddings, determine the number of meanings based on the similarities of exemplars to global vectors, and make the system more efficient.

The present work is part of the EFNILEX project that pilots the use of machine translation tools for lexicography in medium European languages. The first phase of the project used parallel corpora (Héja and Takács, 2012). More recently Makrai (2015) has utilized Mikolov et al. (2013)'s *linear method*, which mainly just needs monolingual gigaword corpora to be trained, supervised by a seed dictionary of some thousand words. They 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 translations coming from some other source (with the score being the distance between $Wx_i$ and $z_i$). [1] In collection mode, evaluation is done on another thousand seed pairs.

The seed dictionary in Mikolov et al. was the Google translation of the most frequent 5 K words, while we tried different seed dictionaries including ones populated by *triangulation* of Wiktionary data (Ács et al., 2013). The idea in triangulation is that if the Hungarian translation of the English word *guild* is *céh*, and the Romanian translation of the later is *breaslă*, then the Romanian translation of *guild* is *breaslă*. 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*.

[1] Mikolov et al. use Euclidean distance in training and cosine similarity (and distance) in collection (and, respectively, scoring) of translations, a theoretically unmotivated choice, which we also found to work better than more consistent combinations of metrics, but see Xing et al. (2015) for opposing results).

| | | # words |
|---|---|---|
| Czech | CNK-SYN (Hnátková et al., 2014) | 2.2 B |
| Croatian | hrWaC2.0 (Ljubešić and K. 2014) | 2.0 B |
| Slovenian | slWaC (Ljubešić and Erjavec, 2011) | 1.6 B |
| Serbian | srWaC (Ljubešić and Klubička, 2014) | 1.0 B |
| Hungarian | HNC (Oravecz et al., 2014) | 0.8 B |
| Hungarian | webcorpus (Halácsy et al., 2004) | 0.7 B |

Fig. 1. Gigaword corpora in some medium European languages

In this demo we improve upon our previous work by handling word ambiguity both in the main (i.e. source and target) languages and in the pivot. The former is done by training multi-prototype vector space models, while triangles created by ambiguity in the pivot are filtered based on scores computed by a linear model trained with direct (non-triangulated) translations. We use the implementation by Dinu et al. (2015). Languages in focus are medium European ones, see corpora in Figure 1. The tools and data we used and also the ones we created are ope-source and can be found on the project page http://corpus.nytud.hu/efnilex-vect/ along with details of our approach to disambiguated translation and some technical details.

## 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.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR 2015, Workshop Track*, 2015.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *EACL*. Association for Computational Linguistics, 2014.

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.

M. Hnátková, M. Křen, P. Procházka, and H. Skoumalová. The syn-series corpora of written czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164. ELRA, 2014. ISBN 978-2-9517408-8-4.

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.

Enikő Héja and Dávid Takács. An online dictionary browser for automatically generated bilingual dictionaries. In *Proceedings of EURALEX2012*, pages 468–477, 2012.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. *COLING*, 2012.

A. Krizhevsky and G. Sutskever, I.and Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'2012*, 2012.

Nikola Ljubešić and Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer, 2011.

Nikola Ljubešić and Filip Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of NAACL-HLT*, pages 151–159, 2015.

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, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168, 2013.

T. M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2014.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. The Hungarian Gigaword Corpus. In *Proceedings of LREC 2014*, 2014.

Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.

F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*, pages 437–440, 2011.

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.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D

Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.