# Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation

**Gábor Borbély**
Department of Algebra
Budapest University of Technology
Egry József u. 1
1111 Budapest, Hungary
borbely@math.bme.hu

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

**Dávid Nemeskey**
Institute for Computer Science
Hungarian Academy of Sciences
Kende u. 13-17
1111 Budapest, Hungary
nemeskeyd@sztaki.mta.hu

**András Kornai**
Institute for Computer Science
Hungarian Academy of Sciences
Kende u. 13-17
1111 Budapest, Hungary
andras@kornai.com

## Abstract

Multi-sense word embeddings (MSEs) model different meanings of word forms with different vectors. We propose two new methods for evaluating MSEs, one based on monolingual dictionaries, and the other exploiting the principle that words may be ambiguous as far as the postulated senses translate to different words in some other language.

## 1 Introduction

Gladkova and Drozd (2016) calls polysemy "the elephant in the room" as far as evaluating embeddings are concerned. Here we attack this problem head on, by proposing two methods for evaluating multi-sense word embeddings (MSEs) where polysemous words have multiple vectors, ideally one per sense. Section 2 discusses the first method, based on sense distinctions made in traditional monolingual dictionaries. We investigate the correlation between the number of senses of each word-form in the embedding and in the manually created inventory as a proxy measure of how well embedding vectors correspond to concepts in speakers' (or at least, the lexicographers') mind.

The other evaluation method, discussed in Section 3, is bilingual, based on the method of Mikolov et al. (2013b), who formulate word translation as a linear mapping from the source language embedding to the target one, trained on a seed of a few thousand word pairs. Our proposal is to perform such translations from MSEs, with the idea that what are different senses in the source language will very often translate to different words in the target language. This way, we can use single-sense embeddings on the target side and thereby reduce the noise of MSEs.

Altogether we present a preliminary evaluation of four MSE implementations by these two methods on two languages, English and Hungarian: the released result of the spherical context clustering method `huang` (Huang et al., 2012); the learning process of Neelakantan et al. (2014) with adaptive sense numbers (we report results using their release MSEs and their tool itself, calling both `neela`); the parametrized Bayesian learner of Bartunov et al. (2015) where the number of senses is controlled by a parameter $\alpha$ for semantic resolution, here referred to as AdaGram; and `jiweil` (Li and Jurafsky, 2015). MSEs with multiple instances are suffixed with their most important parameters, i.e. the learning rate for AdaGram ($a = 0.5$); the number of multi-prototype words and whether the model is adaptive (NP) for release `neela`; and the number of induced word senses ($s = 4$) for our non-adaptive `neela` runs.

Some very preliminary conclusions are offered in Section 4, more in regards to the feasibility of the two evaluation methods we propose than about the merits of the systems we evaluated.

## 2 Comparing lexical headwords to multiple sense vectors

Work on the evaluation of MSEs (for lexical relatedness) goes back to the seminal Reisinger and Mooney (2010), who note that usage splits words

more finely (with synonyms and near-synonyms ending up in distant clusters) than semantics. The differentiation of word senses is fraught with difficulties, especially when we wish to distinguish homophony, using the same written or spoken form to express different concepts, such as Russian *mir* 'world' and *mir* 'peace' from polysemy, where speakers feel that the two senses are very strongly connected, such as in Hungarian *nap* 'day' and *nap* 'sun'. To quote Zgusta (1971) "Of course it is a pity that we have to rely on the subjective interpretations of the speakers, but we have hardly anything else on hand". Etymology makes clear that different languages make different lump/split decisions in the conceptual space, so much so that translational relatedness can, to a remarkable extent, be used to recover the universal clustering (Youna et al., 2016).

Another confounding factor is part of speech (POS). Very often, the entire distinction is lodged in the POS, as in *divorce* (Noun) and *divorce* (Verb), while at other times this is less clear, compare the verbal *to bank* 'rely on a financial institution' and *to bank* 'tilt'. Clearly the former is strongly related to the nominal *bank* 'financial institution' while the semantic relation 'sloping sideways' that connects the tilting of the airplane to the side of the river is somewhat less direct, and not always perceived by the speakers. This problem affects our sources as well: the Collins-COBUILD (CED, Sinclair (1987)) dictionary starts with the semantic distinctions and subordinates POS distinctions to these, while the Longman dictionary (LDOCE, Boguraev and Briscoe (1989)) starts with a POS-level split and puts the semantic split below. Of the Hungarian lexicographic sources, the Comprehensive Dictionary of Hungarian (NSZ, Ittzés (2011)) is closer to CED, while the Explanatory Dictionary of Hungarian (EKSZ, Pusztai (2003)), is closer to LDOCE in this regard. The corpora we rely on are UMBC Webbase (Han et al., 2013) for English and Webkorpusz (Halácsy et al., 2004) for Hungarian. For the Hungarian dictionaries, we relied on the versions created in Miháltz (2010); Recski et al. (2016). We simulate the case of languages without a machine-readable monolingual dictionary with `OSub`, a dictionary extracted from the OpenSubtitles parallel corpus (Tiedemann, 2012) automatically: the number of the senses of a word in a source language is the number of words it translates to, averaged among many languages. More precisely, we use the unigram perplexity of the translations instead of their count to reduce the considerable noise present in automatically created dictionaries.

| Resource | 1 | 2 | 3 | 4 | 5 | 6+ | Size | Mean | Std |
|---|---|---|---|---|---|---|---|---|---|
| CED | 80,003 | 1,695 | 242 | 69 | 13 | 2 | 82,024 | 1.030 | 0.206 |
| LDOCE | 26,585 | 3,289 | 323 | 56 | 11 | 1 | 30,265 | 1.137 | 0.394 |
| `OSub` | 58,043 | 14,849 | 2,259 | 431 | 111 | 25 | 75,718 | 1.354 | 0.492 |
| AdaGram | 122,594 | 330,218 | 11,341 | 5,048 | 7,626 | 0 | 476,827 | 1.836 | 0.663 |
| `huang` | 94,070 | 0 | 0 | 0 | 0 | 6,162 | 100,232 | 1.553 | 2.161 |
| `neela`.30k | 69,156 | 0 | 30,000 | 0 | 0 | 0 | 99,156 | 1.605 | 0.919 |
| `neela`.NP.6k | 94,165 | 2,967 | 1,012 | 383 | 202 | 427 | 99,156 | 1.101 | 0.601 |
| `neela`.NP.30k | 71,833 | 20,175 | 4,844 | 1,031 | 439 | 834 | 99,156 | 1.411 | 0.924 |
| `neela`.s4 | 574,405 | 0 | 0 | 4,000 | 0 | 0 | 578,405 | 1.021 | 0.249 |
| EKSZ | 66,849 | 628 | 57 | 11 | 1 | 0 | 121,578 | 1.012 | 0.119 |
| NSZ (b) | 5,225 | 122 | 13 | 3 | 0 | 0 | 5,594 | 1.029 | 0.191 |
| `OSub` | 159,843 | 9,169 | 229 | 3 | 0 | 0 | 169,244 | 1.144 | 0.199 |
| AdaGram | 135,052 | 76,096 | 15,353 | 5,448 | 6,513 | 0 | 238,462 | 1.626 | 0.910 |
| `jiweil` | 57,109 | 92,263 | 75,710 | 39,624 | 15,153 | 5,997 | 285,856 | 2.483 | 1.181 |
| `neela`.s2 | 767,870 | 4,000 | 0 | 0 | 0 | 0 | 99,156 | 1.005 | 0.072 |
| `neela`.s4 | 767,870 | 0 | 0 | 4,000 | 0 | 0 | 99,156 | 1.016 | 0.215 |

Table 1: Sense distribution, size (in words), mean, and standard deviation of the English and Hungarian lexicographic and automatically generated resources

Table 1 summarizes the distribution of word senses (how many words with 1,. . . ,6+ senses) and the major statistics (size, mean, and variance) both for our lexicographic sources and for the automatically generated MSEs.

While the lexicographic sources all show roughly exponential decay of the number of senses, only some of the automatically generated MSEs replicate this pattern, and only at well-chosen hyperparameter settings. `huang` has a hard switch between single-sense (94% of the words) and 10 senses (for the remaining 6%), and the same behavior is shown by the released Neela.300D.30k (70% one sense, 30% three senses). The English AdaGram and the Hungarian `jiweil` have the mode shifted to two senses, which makes no sense in light of the dictionary data. Altogether, we are left with only two English candidates, the adaptive (NP) `neelas`; and one Hungarian, AdaGram, that replicate the basic exponential decay.

The figure of merit we propose is the correlation between the number of senses obtained by the automatic method and by the manual (lexicographic) method. We experimented both with Spearman $\rho$ and Pearson $r$ values, the entropy-based measures Jensen-Shannon and KL divergence, and cosine similarity and Cohen's $\kappa$. The entropy-based measures failed to meaningfully distinguish between the various resource pairs. The cosine similarities and $\kappa$ values would also have to be taken with a grain of salt: the former does not take the exact number of senses into account, while the latter penalizes all disagreements the same, regardless of how far the guesses are. On the other hand, the Spearman and Pearson values are so highly correlated that Table 2 shows only $\rho$ of sense numbers attributed to each word by different resources, comparing lexicographic resources to one another (top panel); automated to lexicographic (mid panel); and different forms of automated English (bottom panel). The top two values in each column are highlighted in the last two panels, $n$ is the number of headwords shared between the two resources.

| Resources compared | $n$ | $\rho$ |
|---|---|---|
| LDOCE vs CED | 23702 | 0.266 |
| EKSZ vs NSZ (b) | 3484 | 0.648 |
| `neela`.30k vs CED | 23508 | 0.089 |
| `neela`.NP.6k vs CED | 23508 | 0.084 |
| `neela`.NP.30k vs CED | 23508 | 0.112 |
| `neela`.30k vs LDOCE | 21715 | 0.226 |
| `neela`.NP.6k vs LDOCE | 21715 | **0.292** |
| `neela`.NP.30k vs LDOCE | 21715 | 0.278 |
| `huang` vs CED | 23706 | 0.078 |
| `huang` vs LDOCE | 21763 | **0.280** |
| `neela`.s4 vs EKSZ | 45401 | 0.067 |
| `jiweil` vs EKSZ | 32007 | 0.023 |
| AdaGram vs EKSZ | 26739 | 0.086 |
| AdaGram.a05 vs EKSZ | 26739 | 0.088 |
| `neela`.30k vs `huang` | 99156 | 0.349 |
| `neela`.NP.6k vs `huang` | 99156 | **0.901** |
| `neela`.NP.30k vs `huang` | 99156 | 0.413 |
| `neela`.s4 vs `jiweil` | 283083 | 0.123 |
| AdaGram vs `neela`.s4 | 199370 | **0.389** |
| AdaGram vs `jiweil` | 201291 | 0.140 |

Table 2: Word sense distribution similarity between various resources

The dictionaries themselves are quite well correlated with each other. The Hungarian values are considerably larger both because we only used a subsample of NSZ (the letter *b*) so there are only 5,363 words to compare, and because NSZ and EKSZ come from the same Hungarian lexicographic tradition, while CED and LDOCE never shared personnel or editorial outlook. Two English systems, `neela` and `huang`, show perceptible correlation with a lexical resource, LDOCE, and only two systems, AdaGram and `neela`, correlate well with each other (ignoring different parametrizations of the same system, which of course are often well correlated to one another).

### 2.1 Parts of speech and word frequency

Since no gold dataset exists, against which the results could be evaluated and the errors analyzed, we had to consider if there exist factors that might have affected the results. In particular, the better correlation of the adaptive methods with LDOCE than with CED raises suspicions. The former groups entries by part of speech, the latter by meaning, implying that the methods in question might be counting POS tags instead of meanings.

Another possible bias that might have influenced the results is word frequency (Manin, 2008). This is quite apparent in the release version of the non-adaptive methods `huang` and `neela`: the former expressly states in the README that the 6,162 words with multiple meanings "roughly cor-

| Resources compared | $n$ | $\rho$ |
|---|---|---|
| CED vs POS | 42532 | 0.052 |
| LDOCE vs POS | 28549 | **0.206** |
| OSubvs POS | 48587 | **0.141** |
| EKSZ vs POS | 52158 | 0.080 |
| NSZ vs POS | 3532 | 0.046 |
| huang vs POS | 98405 | 0.026 |
| AdaGram vs freq | 399985 | 0.343 |
| huang vs freq | 94770 | 0.376 |
| CED vs freq | 36709 | 0.124 |
| LDOCE vs freq | 27859 | 0.317 |
| neela.s4 vs freq | 94044 | **0.649** |
| neela.NP.30k vs freq | 94044 | 0.368 |
| neela.NP.6k vs freq | 94044 | **0.635** |
| UMBC POS vs freq | 136040 | -0.054 |

Table 3: Word sense distribution similarity with POS tag perplexity (top panel) and word frequency (bottom panel)

respond to the most frequent words".

To examine the effect of these factors, we measured their correlation with the number of meanings reported by the methods above. For each word, the frequency and the POS perplexity was taken from the same corpora we ran the MSEs on: UMBC for English and Webkorpusz for Hungarian. Table 3 shows the results for both English and Hungarian. The correlation of automatically generated resources with POS tags is negligible: all other embeddings correlate even weaker than huang, the only one shown. From the English dictionaries, LDOCE produces the highest correlation, followed by OSub; the correlation with CED, as expected, is very low. The Hungarian dictionaries are around the level of CED.

In comparison, the correlation between sense numbers and word frequency is much more evident. Almost all English resources correlate with the word frequency by at least 0.3 (the notable exception being CED which is the closest to a gold standard we have); furthermore, the highest correlation we measured are between two versions of neela and the word frequency. Adding to this the low correlation of the gold CED against the other resources (see Table 2), it appears the multi-prototype embeddings included in the study were trained to assign more vectors to frequent words instead of trying this for truly polysemous ones.

To disentangle these factors further, we performed partial correlation analysis with the effect of frequency (or its log) or POS perplexity removed. Recall that LDOCE and CED originally correlated only to $\rho = 0.266$. After removing POS, we obtain 0.545, removing frequency yields 0.546, and removing log frequency brings this up to 0.599. Full discussion would stretch the bounds of this paper, but on select embeddings such as neela.NP.6k correlations with CED improve from a negligible 0.093 to a respectable 0.397 if POS, and an impressive 0.696 if log frequency is factored out.

## 3 Cross-linguistic treatment of concepts

Since monolingual dictionaries are an expensive resource, we also propose an automatic evaluation of MSEs based on the discovery of Mikolov et al. (2013b) that embeddings of different languages are so similar that a linear transformation can map vectors of the source language words to the vectors of their translations.

The method uses a seed dictionary of a few thousand words to learn translation as a linear mapping $W : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ from the source (monolingual) embedding to the target: 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 $Wx_i$.

We follow Mikolov et al. (2013b) in using different metrics, Euclidean distance in training and cosine similarity in collection of translations. Though this choice is theoretically unmotivated, it

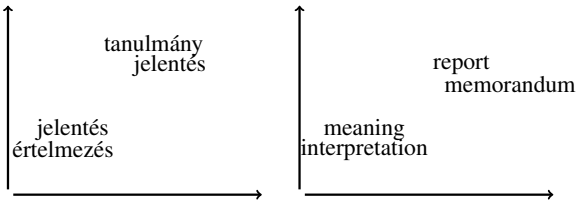


Figure 1: Linear translation of word senses. The Hungarian word *jelentés* is ambiguous between 'meaning' and 'report'. The two senses are identified by the "neighboring" words *értelmezés* 'interpretation' and *tanulmány* 'memorandum'.

seems to work better than more consistent use of metrics; but see (Xing et al., 2015) for opposing results.

In a multi-sense embedding scenario, we take a multi-sense embedding as source model, and a single-sense embedding as target model. We evaluate a specific source MSE model in two ways referred as *single*, and *multiple*.

The tools that generate MSEs all provide fallbacks to singe-sense embeddings in the form of so called global vectors. The method *single* can be considered as a baseline; a traditional, single-sense translation between the global vectors and the target vectors. Note that the seed dictionary may contain overlapping translation pairs: one word can have multiple translations in the gold data, and more than one word can have the same translation. In the *multiple* method we used the same translation matrix, trained on the global vectors, and inspected the translations of the different senses of the same source word. Exploiting the multiple sense vectors one word can have more than one translation.

Two evaluation metrics were considered, *lax* and *strict*. In lax evaluation a translation is taken to be correct if any of the source word's senses are translated into any of its gold translations. In strict evaluation the translations of the source word are expected to cover all of its gold translations. For example if *jelentés* has two gold translations, *report* and *meaning*, and its actual translations are 'report' and some word other than 'meaning', then it has a lax score of 2, but a strict score of 1.

The quality of the translation was measured by training on the most frequent 5k word pairs and evaluating on another 1k seed pairs. We used `OSub` as our seed dictionary. Table 4 shows the percentage of correctly translated words for **s**ingle-sense and **m**ulti-sense translation.

| embedding | | lax | strict |
|---|---|---|---|
| AdaGram 800 a.05 m100 | s | 26.0% | 21.7% |
| | m | 30.5% | 25.1% |
| AdaGram 800 a.01 m100 | s | 12.8% | 10.8% |
| | m | 24.4% | 21.0% |
| jiweil | s | 39.1% | 32.2% |
| | m | 9.7% | 8.3% |

Table 4: Hungarian to English translation. Target embedding from Mikolov et al. (2013a)

## 4 Conclusions

To summarize, we have proposed evaluating word embeddings in terms of their semantic resolution (ability to distinguish multiple senses) both monolingually and bilingually. Our monolingual task, match with the sense-distribution of a dictionary, yields an intrinsic measure in the sense of Chiu et al. (2016), while the bilingual evaluation is extrinsic, as it measures an aspect of performance on a downstream task, MT. For now, the two are not particularly well correlated, though the low/negative result of `jiweil` in Table 1 could be taken as advance warning for the low performance in Table 4. The reason, we feel, is that both kinds of performance are very far from expected levels, so little correlation can be expected between them: only if the MSE distribution of senses replicates the exponential decay seen in dictionaries (both professional lexicographic and crowdsourced products) is there any hope for further progress.

The central linguistic/semantic/psychological property we wish to capture is that of a *concept*, the underlying word sense unit. To the extent standard lexicographic practice offers a reasonably robust notion (this is of course debatable, but we consider a straight correlation of 0.27 and and a frequency-effect-removed correlation of 0.60 over a large vocabulary a strong indication of consistency), this is something that MSEs should aim at capturing. We leave the matter of aligning word senses in different dictionaries for future work, but we expect that by (manual or automated) alignment the inter-dictionary (inter-annotator) agreement can be improved considerably, to provide a more robust gold standard.

At this point everything we do is done in software, so other researchers can accurately reproduce these kinds of evaluations. Some glue code for this project can be found at `https://github.com/hlt-bme-hu/multiwsi`. Whether a 'gold' sense-disambiguated dictionary should be produced beyond the publicly available CED is not entirely clear, and we hope workshop participants will weigh in on this matter.

## Acknowledgments

Research partially supported by National Research, Development, and Innovation Office NKFIH grant #115288.

## References


Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. Breaking sticks and ambiguities with adaptive skip-gram. *ArXiv preprint* .

Branimir K. Boguraev and Edward J. Briscoe. 1989. *Computational Lexicography for Natural Language Processing*. Longman.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In Omer Levy, editor, *Proc. RepEval (this volume)*. ACL.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In Omer Levy, editor, *Proc. RepEval (this volume)*. ACL.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In *Proc. LREC2004*. pages 203–210.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*. pages 44–52.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 873–882.

Nóra Ittzés, editor. 2011. *A magyar nyelv nagyszótára III-IV*. Akadémiai Kiadó.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *EMNLP*.

Dmitrii Y. Manin. 2008. Zipf's law and avoidance of excessive synonymy. *Cognitive Science* .

Márton Miháltz. 2010. *Semantic resources and their applications in Hungarian natural language processing*. Ph.D. thesis, Pázmány Péter Catholic University.

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

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

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

Ferenc Pusztai, editor. 2003. *Magyar értelmező kéziszótár*. Akadémiai Kiadó.

Gábor Recski, Gábor Borbély, and Attila Bolevácz. 2016. Building definition graphs using monolingual dictionaries of Hungarian. In Attila Tanács, Viktor Varga, and Veronika Vincze, editors, *XI. Magyar Számitógépes Nyelvészeti Konferencia [11th Hungarian Conference on Computational Linguistics]*.

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

John M. Sinclair. 1987. *Looking up: an account of the COBUILD project in lexical computing*. Collins ELT.

Jörg Tiedemann. 2012. 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)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Chao Xing, Chao Liu, RIIT CSLT, Dong Wang, China TNList, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*.

Hyejin Youna, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya.

2016. On the universal structure of human lexical semantics. *PNAS* 113(7):1766–1771.

Ladislav Zgusta. 1971. *Manual of lexicography*. Academia, Prague.