Comparison of distributed language models on medium-resourced languages

Márton Makrai

Research Institute for Linguistics of the Hungarian Academy of Sciences
e-mail: makrai.marton@nytud.mta.hu

Abstract. word2vec and GloVe are the two most successful open-source tools that compute distributed language models from gigaword corpora. word2vec implements the neural network style architectures skip-gram and cbow, learning parameters using each word as a training sample, while GloVe factorizes the cooccurrence-matrix (or more precisely a matrix of conditional probabilities) as a whole. In the present work, we compare the two systems on two tasks: a Hungarian equivalent of a popular word analogy task and word translation between European languages including medium-resourced ones e.g. Hungarian, Lithuanian and Slovenian.

Keywords: distributed language modeling, relational similarity, machine translation, medium-resourced languages

1 Introduction

The empirical support for both the syntactic properties and the meaning of a word form consists in the probabilities with that the word appears in different contexts. Contexts can be documents as in latent semantic analysis (LSA) or other words appearing within a limited distance (window) from the word in focus. In these approaches, the corpus is represented by a matrix with rows corresponding to words and columns to contexts, with each cell containing the conditional probability of the given word in the given context. The matrix has to undergo some regularization to avoid overfitting. In LSA this is achieved by approximating the matrix as the product of special matrices.

Neural nets are taking over in many filed of artificial intelligence. In natural language processing applications, training items are the word tokens in a text. Vectors representing word forms on the so called embedding layer have their own meaning: Collobert and Weston [1] trained a system providing state of the art results in several tasks (part of speech tagging, chunking, named entity recognition, and semantic role labeling) with the same embedding vectors. Mikolov et al. [2] trained an embedding with the skip-gram (sgram) architecture, that not only encode similar word with similar vectors but reflects *relational similarities* (similarities of relations between words) as well. The system answers analogical questions. For more details see Section 2.

The two approaches, one based on cooccurrence matrices and the other on neural learning are represented by the two leading open-source tools for computing distributed language models (or simply vector space language models, VSM) from gigaword corpora, GloVe and word2vec respectively. Here we compare them on a task related to statistical machine translation. The goal of the EFNILEX project has been to generate protodictionaries for European languages with fewer speakers. We have collected translational word pairs between English, Hungarian, Slovenian, and Lithuanian

We took the method of Mikolov et al. [3] who train VSMs for the source and the target language from monolingual corpora, and collect word translation by learning a mapping between these supervised by a seed dictionary of a few thousand items.

Before collecting word translations, we test the models in an independent and simpler task, the popular analogy task. For this, we created the Hungarian equivalent of a the test question set by Mikolov et al. [2, 4].[1]

The only related work evaluating vector models of a language other than English on word analogy tasks we know is Sen and Erdogan [5] that compares different strategies to deal with the a morphologically rich Turkish language. Application of GloVe to word translations seems to be a novelty of the present work.

2 Monolingual analogical questions

Measuring the quality of VSMs in a task-independent way is motivated by the idea of representation sharing. VSMs that capture something of language itself are better that ones tailored for the task. We compare results in the monolingual and the main task in Section 5.4.

Analogical questions (also called relational similarities [6] or linguistic regularities [2]) are such a measure of merit for vector models. This test has gained popularity in the VSM community in the recent year. Mikolov et al. observe that analogical questions like *good* is to *better* as *rough* is to ... or *man* is to *woman* as *king* is to ... can be answered by basic linear algebra in neural VSMs:

$$\text{good} - \text{better} \approx \text{rough} - \mathbf{x} \tag{1}$$

$$\mathbf{x} \approx \text{rough} - \text{good} + \text{better} \tag{2}$$

So the vector nearest to the right side of (2) is supposed to be *queen*, which is really the case.

We created a Hungarian equivalent of the analogical questions made publicly available by Mikolov et al. [2, 4][2].

[1] For data and else visit the project page `http://corpus.nytud.hu/efnilex-vect`.

[2] More precisely, we follow the main ideas reported in Mikolov et al. [2] and target the sizes of the data-set accompanying Mikolov et al. [4].

Analogical pairs are divided to morphological ("grammatical") and semantic ones. The morphological pairs in Mikolov et al. [2] were created in the following way:

> [We test] base/comparative/superlative forms of adjectives; singular/plural forms of common nouns; possessive/non-possessive forms of common nouns; and base, past and 3rd person present tense forms of verbs. More precisely, we tagged 267M words of newspaper text with Penn Treebank POS tags [7]. We then selected 100 of the most frequent comparative adjectives (words labeled JJR); 100 of the most frequent plural nouns (NNS); 100 of the most frequent possessive nouns (NN POS); and 100 of the most frequent base form verbs (VB).

English		Hungarian	
plural	singular	plural	singular
decrease	decreases	lesznek	lesz
describe	describes	állnak	áll
eat	eats	tudnak	tud
enhance	enhances	kapnak	kap
estimate	estimates	lehetnek	lehet
find	finds	nincsenek	nincs
generate	generates	kerülnek	kerül

Table 1. Morphological word pairs

The Hungarian morphological pairs were created in the following way: For each grammatical relationship, we took the most frequent inflected forms from the Hungarian Webcorpus [8]. The suffix in question was restricted to be the last one. See sizes in Table 2. In the case of opposite, we restricted ourselves to forms with the derivational suffix *-tlan* (and its other allomorphs) to make the task morphological rather then semantic. plural-noun includes pronouns as well.

For the semantic task, data were taken from Wikipedia. For the capital-common-countries task, we choose the one-word capitals appearing in the Hungarian Webcorpus most frequently. The English task city-in-state contains USA cities with the states they are located in. The equivalent tasks `county-center` contains counties (*megye*) with their centers (*Bács-Kiskun – Kecskemét*) currency contains the currencies of the most frequent countries in the Webcorpus. The `family` task targets gender distinction. We filtered the pairs where the gender distinction is sustained in Hungarian (but dropping e.g. *he – she*). We put some relational nouns in the possessive case (*bátyja – nővére*). We note that this category contains the royal "family" as well, e.g. the famous *king – queen*, and even *policeman – policewoman*.

Both morphological and semantic questions were created by matching every pair with every other pair resulting in e.g. $\binom{20}{2}$ questions for family.

	English		Hungarian
	# questions	# pairs	# questions
gram1-adjective-to-adverb	32	992	40
gram2-opposite	812	29	30
gram3-comparative	37	1332	40
gram4-superlative	34	1122	40
gram5-present-participle	33	1056	40
gram6-nationality-adjective	41	1599	41
gram7-past-tense	40	1560	40
gram8-plural-noun	37	1332	40
gram9-plural-verb	30	870	40
capital-common-countries	23	506	20
capital-world	116	4524	166
city-in-state	2467	68	
county-center			19
county-district-center			175
currency	30	866	30
family	23	506	20

Table 2. Sizes of the question sets

English		Hungarian	
Athens	Greece	Budapest	Magyarország
Baghdad	Iraq	Moszkva	Oroszország
Bangkok	Thailand	London	Nagy-Britannia
Beijing	China	Berlin	Németország
Berlin	Germany	Pozsony	Szlovákia
Bern	Switzerland	Helsinki	Finnország
Cairo	Egypt	Bukarest	Románia

Table 3. Semantic word pairs

English				Hungarian			
Athens	Greece	Baghdad	Iraq	Budapest	Magyarország	Moszkva	Oroszország
Athens	Greece	Bangkok	Thailand	Budapest	Magyarország	London	Nagy-Britannia
Athens	Greece	Beijing	China	Budapest	Magyarország	Berlin	Németország
Athens	Greece	Berlin	Germany	Budapest	Magyarország	Pozsony	Szlovákia
Athens	Greece	Bern	Switzerland	Budapest	Magyarország	Helsinki	Finnország
Athens	Greece	Cairo	Egypt	Budapest	Magyarország	Bukarest	Románia

Table 4. Analogical questions

3 Word translations with vector models

For collection of word translations, we take the method of Mikolov et al. [3] that starts with creating a VSM for the source and the target language from monolingual corpora in the magnitude of billion(s) of words. VSMs represent words in vector spaces of some hundred dimensions. The key point of the method is learning a linear mapping from the source vector space to the target space supervised by a seed dictionary of 5 000 words. Training word pairs are taken from among the most frequent ones skipping pairs with a source of target word unknown to the language model. The learned mapping is used to find a translation for each word in the source model. The computed translation is the target word with a vector closest to the image of the source word vector by the mapping. The closeness (cosine similarity) between the image of the source vector and the closest target vector measures the goodness of the translation, the similarity of the source and the computed target word. Best results are reported when the dimension of the source model is 2–4 times the dimension of the target model, e.g. 800 → 300.

We generate word translations between the following language pairs: Hungarian-Lithuanian, Hungarian-Slovenian, and Hungarian-English.

The method provides a measure of confidence for each translational pair, namely the distance of the vector computed by mapping the source word vector, and the nearest target word vector. This measure makes a tuning between precision and recall possible (Table 10). With a higher cosine similarity cut-off (column cos >), we get word translations for a smaller vocabulary (vocab) with a higher precision, while lower cosine similarities produce a greater vocabulary with translations of a lower precision. prec@1 is the ratio of words, for which the first candidate translation coincides with that provided in the seed dictionary, prec@5 is the ratio of words with the seed translation in the first 5 candidates. These are strict metrics, as synonyms of the gold translation count as incorrect. gold is the number of words with a gold translation in the corresponding part of the test data.

We follow Mikolov et al. [2] in using least squares of the Euclidean distance for training, and, surprisingly, cosine similarity for translation generation, which is the only combination of the two distances that works.

4 Data

4.1 Corpora and vectors

For English, we use vector models downloaded from the home pages of the tools, while for the medium-resourced languages, we train new models on the corpora in Table 5, using the tokenization provided by the authors of the corpora.

4.2 Seed dictionaries

Mikolov et al. [3] use Google translate as a seed dictionary. We have been experimenting with three seed dictionaries: (1) efnilex12, the protodictionaries collected

language	corpus	# words
Lithuanian	webcorpus [9]	1.4 B
Slovenian	slWaC [10]	1.6 B
Hungarian	webcorpus [8]	0.7 B
Hungarian	HNC [11]	0.8 B

Table 5. Corpora for medium-resourced languages. Word counts are given in billions.

within the EFNILEX project [12], (2) word pairs collected using wikt2dict with and without triangulation (See Ács et al. [13], and, for sizes, Table 6), and (3) dictionaries from the opus collection (Europarl, OpenSubtitles2012 and OpenSubtitles2013) [**?**][3]. efnilex12 contains directed dictionaries (ranked by the conditional probability of the (cooccurrence of the) target word conditioned on the source word).

	efnilex12	wikt	wikt triang	OSub12	OSub13	Europarl
en-hu	83 K	47 K	+134 K	97 K	19 K	21 K
hu-lt	152 K	6 K	+21 K	11 K	9 K	27 K
hu-sl	235 K	2 K	+26 K	63 K	45 K	29 K

Table 6. Number of translational word pairs in the seed dictionaries

5 Results

Throughout the following two sections, the these abbreviations will be used: d for dimension, w for window radius ($w = 15$ means that (a maximum of) 15 words are considered on both sides of the word in focus), i for number of training iterations over the corpus (epochs), m for minimum word count in the vocabulary cutoff, and n for number of negative samples (in the case of word2vec).

5.1 Analogical questions

For comparing the Hungarian analogical questions to the English ones, we trained sgram models on the concatenation of HNC and the Hungarian Webcorpus with $d = 300, m = 5$ comparing negative sampling to hierarchical softmax (two techniques to avoid computing a the denominator of softmax that is a sum with as many terms as there are words in the embedding) and the effect of subsampling of frequent words, see [14] for details. In Table 7, it can be seen that we (bellow the line) get similar results in the Hungarian equivalent of the original tasks

[3] `http://opus.lingfil.uu.se/`

(Mikolov et al. [14] are above the line) in the morphological questions, while Hungarian results in the semantic questions are worse, This suggest that the semantic questions are too hard. This problem has to be investigated further.

		morph		semant		total	
en [14]	$n = 5$	61		58		60	
	$n = 15$	61		61		61	
	HS	52		59		55	
hu	$n = 5$	63.0	3419/5430	38.5	269/699	60.2	3688/6129
	$n = 15$	61.9	3359/5430	39.2	274/699	59.3	3633/6129
	HS	48.9	2653/5430	22.5	157/699	45.8	2810/6129

Table 7. Comparison of results in word translations to those of Mikolov et al [3]

5.2 Protodictionary generation

In this section we report our results in Slovenian/Hungarian/Lithuanian to English protodictionary generation. We take four source embeddings: two Slovenian ones trained on slWaC, one trained on the Hungarian Webcorpus, and one on the Lithuanian webcorpus by Zséder et al. [9], all in $d = 600$. One of the Slovenian models is a GloVe one, the other models are cbow models with $n = 15$ and $w = 10$. The target model is always glove.840B.300d from the GloVe site, the seed dictionary is OpenSubtitles2012. The source (rs), the target (rt) embedding, or both (rst) was restricted to words accepted by Hunspell. In Table 8 we compare our results (bellow the line) to those of Mikolov et al. [3] (above the line) with slightly different metaparameters. The vocabulary cutoff m of the source embedding is specified for each word2vec model we trained.

	prec@1	prec@5
en → sp	33	51
sp → en	35	52
en → cz	27	47
cz → en	23	42
en → vn	10	30
vn → en	24	40
glove-sl → en rs	44.80	63.40
word2vec-sl → en $m = 100$ rs	41.70	60.40
word2vec-hu → en $m = 50$ rst	32.80	54.70
word2vec-lt → en $m = 100$ rt	21.20	36.50

Table 8. Results in protodictionary collection

source word	cos	translations			
öt	0.9101	five	six	eight	three
jó	0.8961	good	really	too	very
de	0.8957	but	though	even	just
bár	0.8955	though	but	even	because
hit	0.8904	faith	belief	salvation	truth
ugyan	0.888	though	but	even	because
vöröshagymát	0.8878	onion	garlic	onions	tomato

Table 9. Example word translations. cos is the cosine similarity of the image of the source word vector by the learned mapping and the nearest target vector. Words in the target language are listed in the (descending) order of their similarity to the image vector.

cos >	vocab	gold	prec@1	prec@5
0.7	3803	301	68.4%	84.4%
0.6	9967	711	54.7%	74.1%
0.5	12949	958	46.6%	65.6%
0.4	13451	988	45.3%	64.0%

Table 10. Trade-off between precision and recall in Hungarian to English word translation.

5.3 word2vec, LBL4word2vec and GloVe

We compared word2vec, its modification LBL4word2vec[4], and GloVe with two parameter settings in the two tasks. The two parameter settings were needed because the default (recommended) values of d, w, i and m are different in the two architectures, see Table 11 with the more computation-intensive setting in bold. We trained two models with each architecture on HNC: a small one with

	word2vec	GloVe
d	**100**	50
w	5	**15**
i	5	**25**
m	**5**	10

Table 11. Default values of parameters shared by word2vec and GloVe

the less computation-intensive one of the two default values and a big one with the lesser one (except for using $d = 52$ in small for historical reasons). For the number of negative samples, which is specific for word2vec, we use the default

[4] `https://github.com/qunluo/LBL4word2vec`

$n = 5$. See results in Table 12. Note that GloVe results could be further improved by taking the average other the two vectors learned by the model for each word.

		morph		sem		total	
small	word2vec sgram	49.0%	2703	20.3%	156	45.5%	2859
	LBL4word2vec sgram	46.6%	2567	19.4%	149	43.2%	2716
	word2vec cbow	49.9%	2751	15.7%	121	45.7%	2872
	glove	41.3%	2277	11.1%	85	37.6%	2362
big	word2vec sgram	57.8%	3186	42.0%	323	55.8%	3509
	LBL4word2vec sgram	55.5%	3058	36.3%	279	53.1%	3337
	glove	58.1%	3206	31.3%	241	54.9%	3447
	word2vec cbow	57.8%	3187	30.7%	236	54.5%	3423

Table 12. Comparison of models trained in different architectures Rows within each model "size" are sorted by precision in semantic task that we consider more relevant to lexicography than morphology. The total number of questions that do not contain out-of-vocabulary words is 5514 in morphological questions and 6283 in semantic ones.

5.4 Comparison of results in the two tasks

In Figure 1 we show the results of some Hungarian VSMs in the analogical and the word translation task plotted against each other. The horizontal axis shows precision in the semantic analogical questions, while the vertical axis shows precision (@5) in protodictionary generation to the Google News model[5] restricted to words accepted by Hunspell and using seed pairs collected with wikt2dict. It can be seen that result in the two tasks are unfortunately uncorrelated.

6 Parameter analysis

6.1 Corpus

Quality In Table 13, we compare on analogical questions models trained on the Hungarian National Corpus (September 12 snapshot) [11] that is a curated corpus of Hungarian, and on the Hungarian Webcorpus [8] that is a similarly sized webcorpus. The numbers suggest that a curated corpus is more suitable for the analogical task.

Size Table 14 shows how the performance depends on the size of the corpus. It is clear that a much larger corpus is needed to answer semantic questions.

[5] `https://code.google.com/p/word2vec/#Pre-trained_word_and_phrase_vectors`

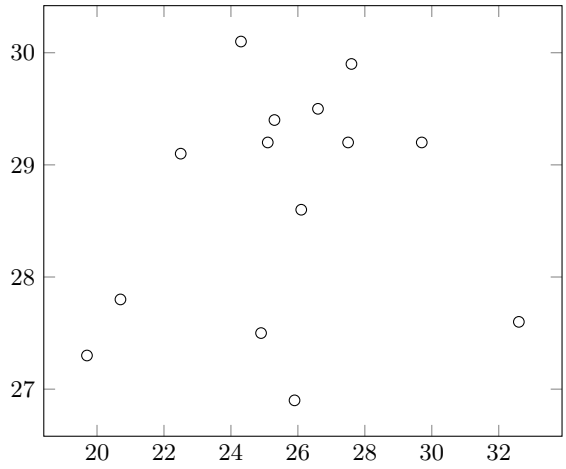

Fig. 1. Precision in monolingual (horizontal axis) vs bilingual (vertical axis) task

model	question type	Webcorpus		HNC	
word2vec	morphological	**54.9**	**2924**/5326	51.8	2856/5514
	semantic	8.3	40/482	**16.0**	**123**/769
	total	**51.0**	2964/5808	47.4	**2979**/6283
glove	morphological	47.4	2525/5326	**48.2**	**2658**/5514
	semantic	9.3	45/482	**14.4**	**111**/769
	total	**44.2**	2570/5808	44.1	**2769**/6283

Table 13. Comparison of results on two different corpora. The denominator of each fraction is the number of questions with all three words known to the vector model, while the numerator is the number of correct answers for these questions. Parameters: $d = 152$, $m = 10$, $i = 5$ in both models. For word2vec, $w = 5$ and $n = 5$ while for glove, $w = 3$. The different window sizes mean that these results are not suitable for comparing the models just the corpora.

	morph		sem		total	
1M	1.8	58/3256	0.0	0/84	1.7	58/3340
2M	6.1	191/3130	0.0	0/60	6.0	191/3190
10M	24.9	986/3954	7.4	8/108	24.5	994/4062
100M	55.1	2530/4594	31.4	37/118	54.5	2567/4712
754M	63.2	3486/5514	49.8	383/769	61.6	3869/6283

Table 14. The effect of corpus size.

6.2 word2vec

Hierarchical softmax and negative samples We also tried whether hierarchical softmax (HS) and negative sampling can be combined to get better result with either of the techniques. A negative answer can be seen in Table 15 (HNC, $d = 100, w = 5, i = 5, m = 5$).

	morph		semant		total	
cbow $hs = 0, n = 5$	59.4%	**3276**/5514	24.1%	185/769	55.1%	3461/6283
cbow $hs = 1, n = 0$	49.0%	2702/5514	13.9%	107/769	44.7%	2809/6283
cbow $hs = 1, n = 5$	49.5%	2730/5514	14.3%	110/769	45.2%	2840/6283
sgram $hs = 0, n = 5$	59.1%	3261/5514	33.6%	**258**/769	56.0%	3519/6283
sgram $hs = 1, n = 0$	49.8%	2744/5514	23.1%	178/769	46.5%	2922/6283
sgram $hs = 1, n = 5$	50.4%	2781/5514	23.1%	178/769	47.1%	2959/6283

Table 15. Hierarchical softmax (HS) and negative sampling.

6.3 Protodictionaries: Seed dictionary

We compare result obtained in the protodictionary generation task with different English-Hungarian seed dictionaries in Table 16. The source language model is always glove.840B.300d[6], the target model is also a GloVe model trained on HNC ($d = 300, m = 1, w = 15, i = 25$). For details of the seed dictionaries see Section 4.2.

seed dictionary	prec@1	prec@5
Europarl	17.70%	34.10%
wikt triang	13.10%	25.30%
wikt	12.50%	25.40%
OpenSubtitles2012	10.30%	23.40%
efnilex12 en→hu	10.10%	23.80%

Table 16. Accuracy of protodictionary generation with different seed dictionaries

7 Acknowledgements

I would like to thank Tamás Váradi for supervision, András Kornai, Csaba Oravecz and Attila Zséder for ideas and advices, and Mehmet Umut Sen for translating the essence of [5] to English. Work was supported by the EFNILEX project.

[6] `http://nlp.stanford.edu/projects/glove/`

References

1. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08, New York, NY, USA, ACM (2008) 160–167
2. Mikolov, T., Yih, W.t., Geoffrey, Z.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT-2013. (2013)
3. Mikolov, T., Le, Q.V., Sutskeve, I.: Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168 (2013)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In Bengio, Y., LeCun, Y., eds.: Proc. ICLR 2013. (2013)
5. Sen, M., Erdogan, H.: Learning word representations for turkish. In: Signal Processing and Communications Applications Conference (SIU), 2014 22nd. (2014) 1742–1745
6. Turney, P.D.: Similarity of semantic relations. Computational Linguistics **32** (2006) 379–416
7. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn treebank. Computational Linguistics **19** (1993) 313–330
8. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004). (2004) 203–210
9. Zséder, A., Recski, G., Varga, D., Kornai, A.: Rapid creation of large-scale corpora and frequency dictionaries. In: Proceedings to LREC 2012. (2012) 1462–1465
10. Ljubešić, N., Erjavec, T.: hrwac and slwac: Compiling web corpora for croatian and slovene. In Habernal, I., Matousek, V., eds.: Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings. Lecture Notes in Computer Science, Springer (2011) 395–402
11. Oravecz, Cs., Váradi, T., Sass, B.: The hungarian gigaword corpus. In: Proceedings of LREC 2014. (2014)
12. Héja, E., Takács, D.: An online dictionary browser for automatically generated bilingual dictionaries. In: Proceedings of EURALEX2012. (2012) 468–477
13. Ács, J., Pajkossy, K., Kornai, A.: Building basic vocabulary across 40 languages. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, Association for Computational Linguistics (2013) 52–58
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., eds.: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 3111–3119