Building a global dictionary for semantic technologies

Master's Thesis

Author: Eszter Iklódi
Supervisor: Gábor Recski

Budapest University of Technology and Economics
Department of Automation and Applied Informatics

2018

M Ű E G Y E T E M 1 7 8 2

Thesis objectives

- Study state-of-the-art multilingual embeddings
- Propose a new method
- Run experiments on benchmark data
- Run experiments on data extracted from PanLex

Table of Contents

Natural Language Processing (NLP)

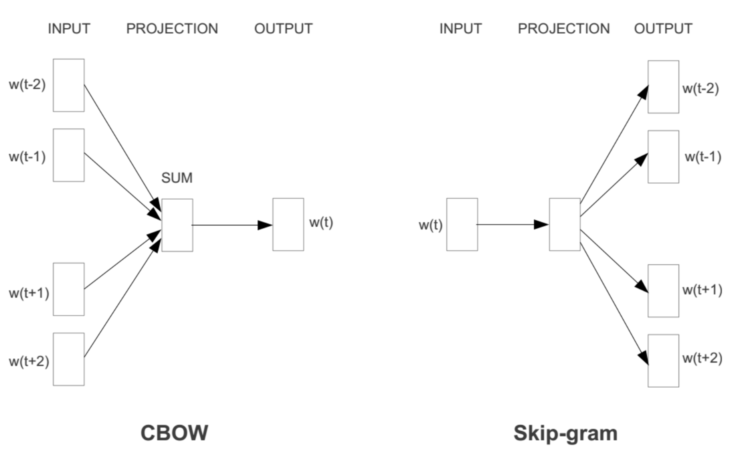
- Aim: use natural languages for human-machine communication
- Common tasks
- Vibrant research field: Amazon, Apple, Facebook, Google

Spam detection	Sentiment analysis	Question answering
Part-of-speech (POS) tagging	Word sense disambiguation	Dialogues
Named Entity Recognition (NER)	Syntactic parsing	Summarization
	Machine translation	Paraphrases

Word embeddings

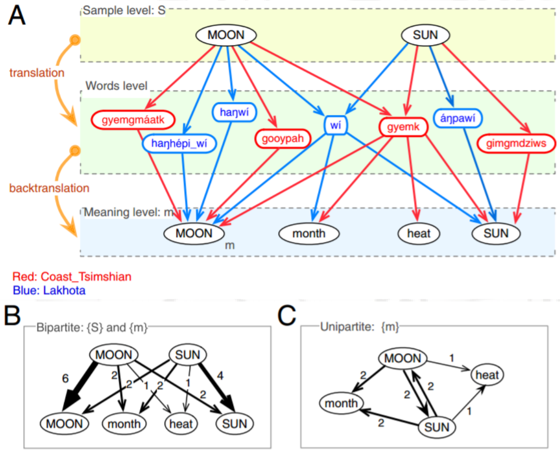
- Vector representation of words
- Mikolov et al. (2013a) *word2vec*

vec(king) - vec(man) + vec(woman) = vec(queen)



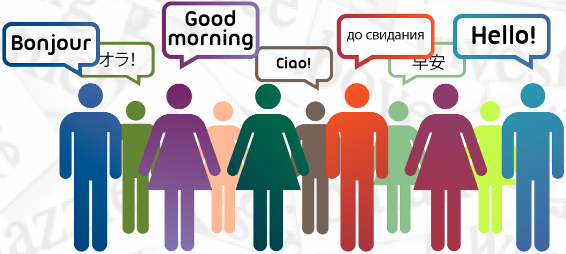
Multilingual word embeddings - Motivation

Theoretical background: Youn et al. (2016): meaning is independent of geography, environment, literary traditions



Multilingual word embeddings - Tasks

- Cross-language part-of-speech tagging
- Cross-language super sense tagging
- Machine translation
- Under-resourced languages
- Already existing applications e.g.: Facebook Recommendations, M Suggestions



Dinu's data

- Dinu et al. (2014)
- English-Italian gold dictionary
- Benchmark data for word translation tasks
- Built from Europarl en-it
- Test: 1869 word pairs
  - $5 \cdot 300 = 1500$ English words
  - frequency bins: [1-5K], [5K-20K], [20K-50K], [50K-100K], [100K-200K]
- Train: 5000 word pairs
  - top 5000 translation pairs

Set	Language	No. words
train	eng	3442
	ita	4549
test	eng	1500
	ita	1849

State-of-the-art multilingual word embeddings

- Mikolov et al. (2013b)

$$\min_W \sum_{i=1}^{n} ||Wx_i - z_i||^2 \quad (1)$$

- Faruqui and Dyer (2014)
  - Canonical Correlation Analysis (CCA)
- Dinu et al. (2014)
  - hub problem
- Smith et al. (2017)
  - orthogonal, SVD
  - inverted softmax
- Conneau et al. (2017)
  - unsupervised method
  - `fastText` embedding

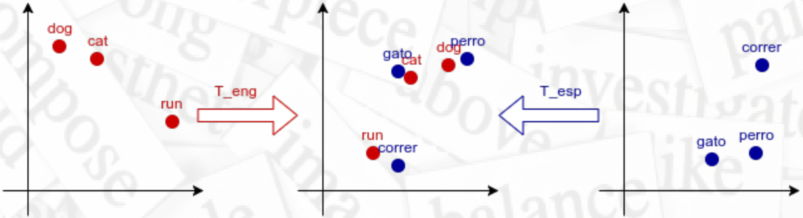
Precision	@1	@5	@10
Mikolov et al. (2013b)	0.338	0.483	0.539
Faruqui and Dyer (2014)	0.361	0.527	0.581
Dinu et al. (2014)	0.385	0.564	0.639
Smith et al. (2017)	0.431	0.607	0.664
Conneau et al. (2017)	**0.662**	**0.804**	**0.834**

Precision	@1	@5	@10
Mikolov et al. (2013b)	0.249	0.410	0.474
Faruqui and Dyer (2014)	0.310	0.499	0.570
Dinu et al. (2014)	0.246	0.454	0.541
Smith et al. (2017)	0.380	0.585	0.636
Conneau et al. (2017)	**0.587**	**0.765**	**0.809**

Proposed method

$$cos\_sim = \cos\theta = \frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\cdot||\vec{b}||} \tag{2}$$

$$\frac{1}{|TP|}\cdot\sum_{\substack{L_1,L_2\\ \in L}}\sum_{\substack{(w_1,w_2)\\ \in TP}} cos\_sim(w_1\cdot T_1, w_2\cdot T_2) \tag{3}$$



Parameter adjustment - Learning rate

Dinu's train's split: train ($\sim 90\%$), valid ($\sim 10\%$)
Batch size $= 64$

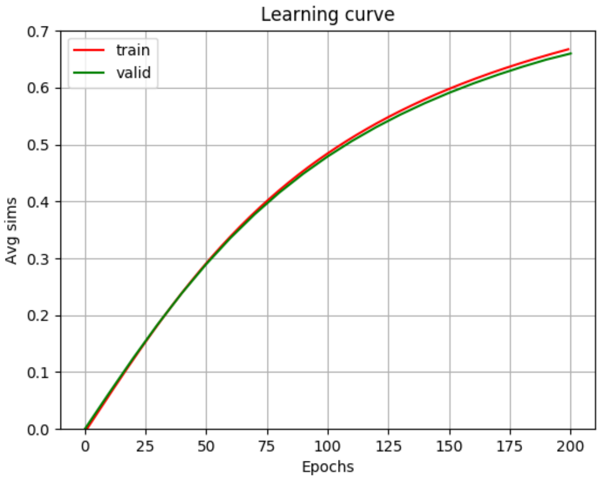
LR	cos_sim	English - Italian Precision			Italian - English Precision		
		@1	@5	@10	@1	@5	@10
0.001	0.988743	0.1831	0.1831	0.3721	0.1667	0.2851	0.3494
0.003	0.995905	0.3401	0.5058	0.5669	0.3032	0.4799	0.5462
0.01	0.998957	0.4651	0.6366	0.6802	0.4036	0.6185	0.6586
0.03	0.999824	0.5262	0.7006	0.7645	0.4438	0.6506	0.6988
0.1	0.999994	**0.5407**	**0.7297**	**0.7645**	**0.4618**	**0.6546**	0.6948
0.3	1.000000	0.5407	0.7151	0.7645	0.4478	0.6526	**0.7028**
1	1.000000	0.4535	0.6483	0.6977	0.3554	0.5542	0.6265
3	1.000000	0.0698	0.1599	0.1890	0.0462	0.0462	0.1586

Parameter adjustment - Batch size

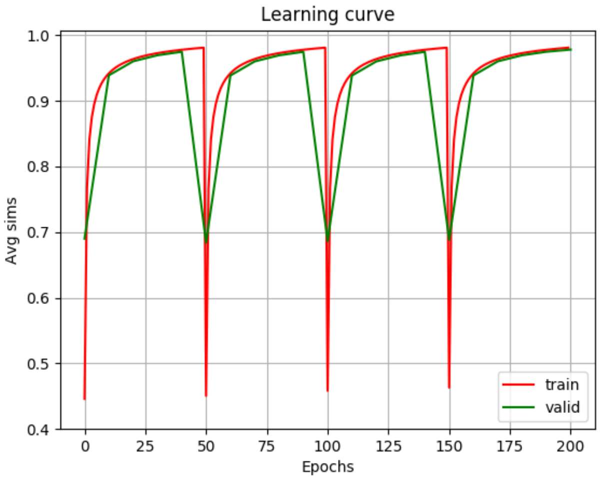
Dinu's train's split: train ($\sim 90\%$), valid ($\sim 10\%$)
Learning rate = 0.1

BS	cos_sim	English - Italian Precision			Italian - English Precision		
		@1	@5	@10	@1	@5	@10
16	1.000000	0.5320	0.7209	0.7616	0.4418	0.6446	0.7008
32	1.000000	0.5203	0.7064	0.7558	0.4398	0.6446	0.6948
64	0.999994	**0.5465**	0.7209	**0.7878**	**0.4578**	**0.6627**	0.7068
128	0.999946	0.5407	**0.7267**	0.7645	0.4458	0.6586	**0.7129**
256	0.999949	0.5320	0.7093	0.7645	0.4398	0.6627	0.7088

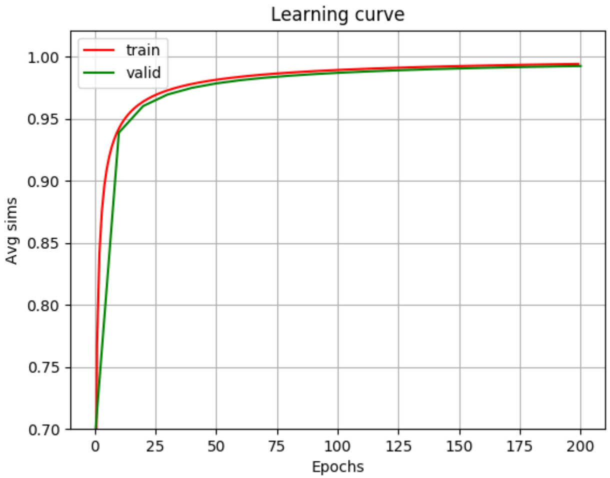
Parameter adjustment - Without SVD



Parameter adjustment - SVD on every 50th epoch

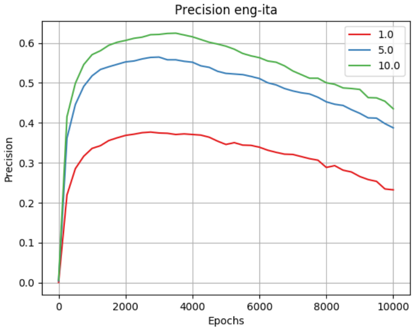


Parameter adjustment - SVD only at the beginning



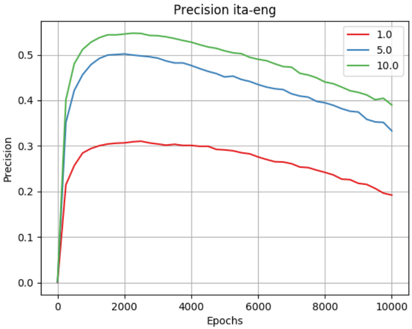
Best system on Dinu's data: English-Italian scores

Eng-Ita	@1	@5	@10
Mikolov et al. (2013b)	0.338	0.483	0.539
Faruqui and Dyer (2014)	0.361	0.527	0.581
Dinu et al. (2014)	0.385	0.564	0.639
Smith et al. (2017)	0.431	0.607	0.651
Conneau et al. (2017)	**0.662**	**0.804**	**0.834**
Proposed method (fastText)	0.377	0.565	0.625



Best system on Dinu's data: Italian-English scores

Ita-Eng	@1	@5	@10
Mikolov et al. (2013b)	0.249	0.410	0.474
Faruqui and Dyer (2014)	0.310	0.499	0.570
Dinu et al. (2014)	0.246	0.454	0.541
Smith et al. (2017)	0.380	0.585	0.636
Conneau et al. (2017)	**0.587**	**0.765**	**0.809**
Proposed method (fastText)	0.310	0.502	0.547



The PanLex database

Aim: to build a multilingual lexical database, in all languages
Confidence values: [1, 9]

English	Italian	Confidence values
Sarajevo	Sarajevo	9
euro	euro	9
simple	semplice	8
difficult	difficile	8
college	università	7
plausible	verisimile	7
sea	mare	6
sky	cielo	6
better	meglio	5
inform	informare	5
combustible	combustibile	4
office	ufficio	4
sorcerer	conscitore	3
it	ella	3
Great Wall of China	Grande muraglia cinese	2
factory workers	lavoratori dell'industria	2
stay	restare	1
sometimes	qualche volta	1

English-Italian dataset creation from PanLex data

- Applying the method of Dinu et al. (2014)
- Used as English-Italian gold dictionary
- Confidence value $>= 7$
- Test:
  - $5 \cdot 300 = 1500$ English words
  - frequency bins: [1-5K], [5K-20K], [20K-50K], [50K-100K], [100K-200K]
- Train:
  - top 5000 translation pairs, sorted according to English frequency

conf.val.	number of word pairs $>=$ conf.val.
9	66
8	580
7	69623
6	69690
5	96168
4	99004
3	146481
2	182663
1	187601

PanLex experiments

Learning rate = 0.1, Batch size = 64, SVD at the beginning

	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
first 5k	0.0093	0.0253	0.0367	0.0000	0.0007	0.0007
Eng. words retaining one translation	0.1120	0.2073	0.2427	0.1114	0.2052	0.2440
Eng. words only with one translation	**0.1960**	**0.3087**	**0.3440**	**0.1838**	**0.3059**	**0.3443**

	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
1k	0.1500	0.2847	0.3340	0.1391	0.2761	0.3256
3k	**0.2127**	**0.3473**	**0.3933**	**0.2232**	**0.3650**	**0.4152**
5k	0.1980	0.3193	0.3620	0.2212	0.3555	0.4030
10k	0.1613	0.2807	0.3227	0.1879	0.3012	0.3372

Comparing Dinu and PanLex experiments

Test on Dinu	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
train:PanLex - test:old	0.3770	0.5647	0.6245	0.3103	0.5018	0.5474
train:PanLex - test:new	**0.3560**	**0.5407**	**0.5978**	**0.2917**	**0.4792**	**0.5215**
train:Dinu - test:new	0.1360	0.2309	0.2594	0.1361	0.2556	0.2965
train:Dinu+PanLex - test:new	0.2930	0.4349	0.4861	0.2910	0.4556	0.5090

Test on PanLex	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
train:PanLex - test:old	0.1960	0.3087	0.3440	0.1838	0.3059	0.3443
train:PanLex - test:new	0.1812	0.2858	0.3196	0.1668	0.2835	0.3213
train:Dinu - test:new	**0.2295**	**0.4171**	**0.4839**	0.2227	0.3763	0.4199
train:Dinu+PanLex - test:new	0.2295	0.3712	0.4275	**0.2498**	**0.4026**	**0.4495**

Continuing the baseline system with PanLex data

	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
original	**0.3770**	**0.5647**	**0.6245**	0.3103	**0.5018**	0.5474
cont from 2000	0.3426	0.5256	0.5802	**0.3229**	0.4882	**0.5535**
cont from 3000	0.3535	0.5416	0.5970	0.3229	0.4840	0.5465
cont from 4000	0.3510	0.5273	0.5911	0.3118	0.4701	0.5243

English-Italian-Spanish parallel training

Pairwise	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
eng-ita	0.2080	0.3280	0.3687	0.2082	0.3386	0.3904
eng-spa	0.2840	0.4320	0.4800	0.2883	0.4331	0.4836
spa-ita	0.3920	0.5340	0.5813	0.3655	0.5291	0.5750

Parallel	eng-ita			ita-eng		
Precision	@1	@5	@10	@1	@5	@10
eng-ita	0.1573	0.2667	0.3127	0.1638	0.2942	0.3386
eng-spa	0.1947	0.2973	0.3447	0.2350	0.3538	0.4064
spa-ita	0.2520	0.3640	0.4160	0.2568	0.3723	0.4162

Conclusion and future work

- A novel method was proposed for finding linear mappings between word embeddings
- Parameter adjustment:
  - best learning rate: 0.1, best batch size: 64
  - Applying SVD on the transformation matrices
    - Makes the learning process faster
    - Best way: doing it only once, at the beginning
- The best system:
  - Outperforms Mikolov et al.'s baseline system
  - Comparable with more sophisticated systems: Faruqui and Dyer, Dinu et al.
  - Significantly worse than Conneau et al.'s state-of-the-art system
- Dinu's data provides better results than the PanLex dataset
- Slight improvement on Italian-English scores when continuing the baseline system with the PanLex data
- Multilingual experiments
  - Possible parallel training with many languages
  - But pairwise results are always better

References I

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

G. Dinu, A. Lazaridou, and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, 2014.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

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

S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

H. Youn, L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771, 2016.

Thank you for your attention!