Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Simon Šuster*, Ivan Titov◊, and Gertjan van Noord*

*CLCG, University of Groningen
◊ILLC, University of Amsterdam

Crosslingual supervision

Multilingual learning: more accurate monolingual models

- Ambiguity in L1 often different than ambiguity in L2
  [Snyder and Barzilay, 2010]
- Resolving polysemy in L1 by also looking at *translations*
  - well known in WSD (”translation as sense”)
    [Diab and Resnik, 2002]
  - but little explored in representation learning
    [Guo et al., 2014, Ettinger et al., 2016]
- Availability of parallel corpora

En-Es intuition

track: a course of study; a piece of music, a rough path...

sent$_{L1}$: *Choose a track that interests you*

En-Es intuition

<u>track</u>: a course of study; a piece of music, a rough path...

sent$_{L1}$: *Choose a track that interests you*

sent$_{L2}$: *Pon una canción que te gusta*

Related embedding research

- Multi-sense [Neelakantan et al., 2014, Li and Jurafsky, 2015]
  - deal with polysemy explicitly
  - monolingual
- Multilingual
  - embeddings in the same semantic space [Gouws et al., 2014, Klementiev et al., 2012]
  - use target-language signal for better source-language embeddings [Hill et al., 2014, Faruqui and Dyer, 2014]

- Better L1 **multi-sense** embeddings **with L2 signal**?

Inspiration from autoencoders

At output, reconstruct input by relying on a latent representation

- Latent sense representation is a categorical variable
- Reconstruct some part of the input (i.e. a word) based on another word and its sense
- Cf. discrete autoencoders
  [Marcheggiani and Titov, 2016, Ammar et al., 2014]

Model structure

- **Encoding**: $p(s|x_i, C_i, C_i', \theta)$
  - learn a sense mapping with a log-linear model
  - choose the sense of the pivot x_i using the combination of L1 (C_i) and L2 (C_i') contexts
- **Reconstruction**: $p(x_j|x_i, s, \theta)$
  - learn sense-specific word embeddings with Skip-gram
  - predict a context word x_j based on the pivot and its sense

Model structure

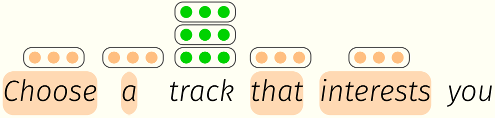
- **Encoding**: $p(s|x_i, C_i, C'_i, \theta)$
  - learn a sense mapping with a log-linear model
  - choose the sense of the pivot x_i using the combination of L1 (C_i) and L2 (C'_i) contexts
- **Reconstruction**: $p(x_j|x_i, s, \theta)$
  - learn sense-specific word embeddings with Skip-gram
  - predict a context word x_j based on the pivot and its sense

- Both components jointly optimized
- Induce a sense mapping that facilitates inferring context words

En-Es illustration

Encoding:

sense selection $p(s|x_i, C_i, C_i', \theta)$



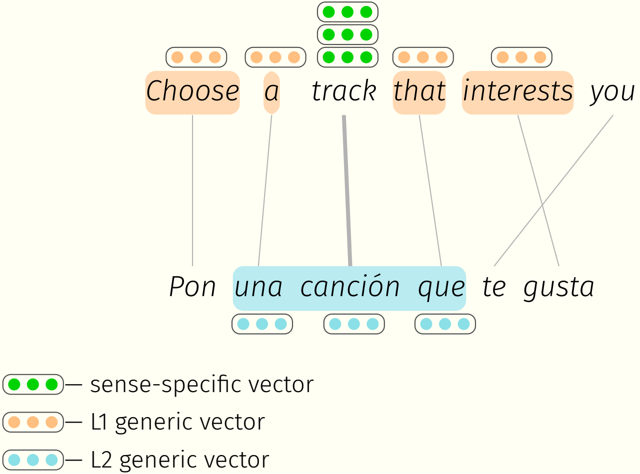
— sense-specific vector

— L1 generic vector

En-Es illustration

Encoding:

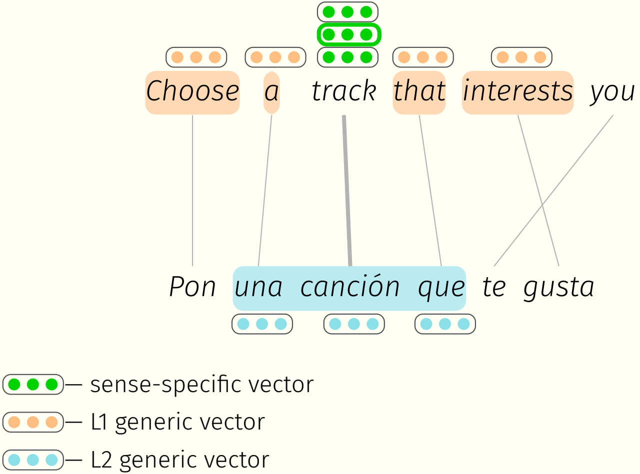
sense selection $p(s|x_i, C_i, C'_i, \theta)$



En-Es illustration

Encoding:

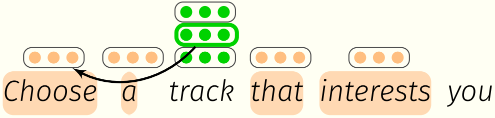
sense selection $p(s|x_i, C_i, C_i', \theta)$



En-Es illustration

Reconstruction:

context-word prediction $p(x_j | x_i, s, \theta)$



Model set-up

- **BIMU**: Multi-sense (n=3) trained with bilingual signal
- MU: Multi-sense trained monolingually
- SG: Skipgram

In training of the multi-sense models, use

- entropy regularization to sharpen the encoder posteriors
- or hard updates

Experimental set-up

Data

- Use word-aligned parallel corpora in training
- L1: English, paired with
  - French (GigaFrEn)
  - Czech (CzEng)
  - Russian (Yandex)
  - Es, De, Cs, Fr, Ru (NewsCommentary)
- At test time, use only L1 since L2 not available

Experimental set-up

Data

- Use word-aligned parallel corpora in training
- L1: English, paired with
  - French (GigaFrEn)
  - Czech (CzEng)
  - Russian (Yandex)
  - Es, De, Cs, Fr, Ru (NewsCommentary)
- At test time, use only L1 since L2 not available

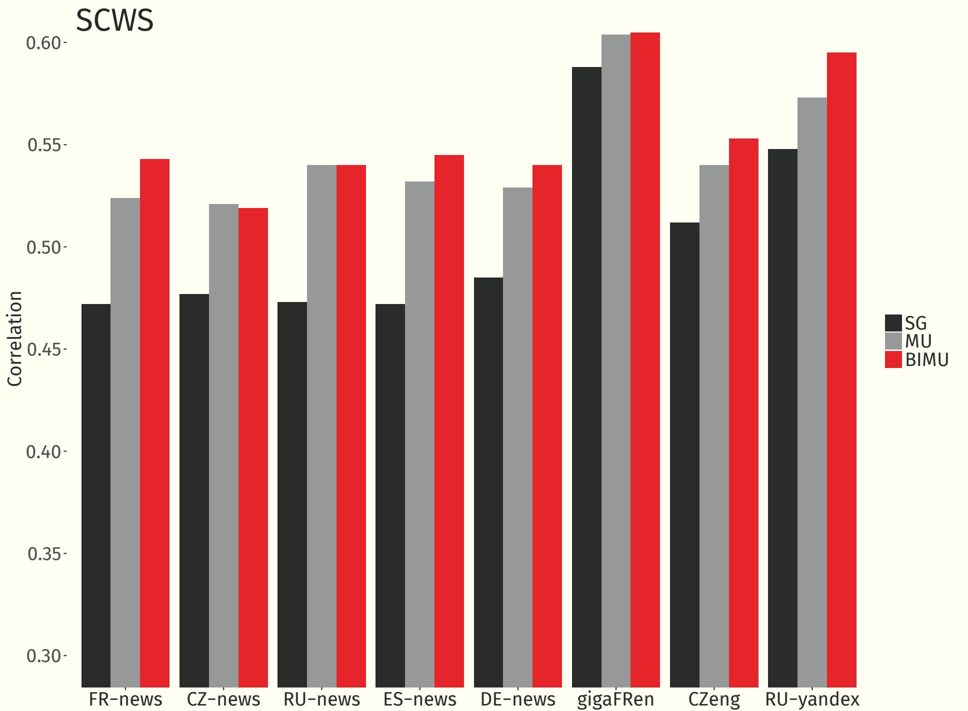
Tasks

	Context?	Representation
Sem. similarity: SCWS	✓	weighted avg.
Sem. similarity: 12 benchmarks	✗	uniform avg.
Qvec	✗	uniform avg.
Neural POS tagger	✓	weighted avg.

Nearest neighbors*

$\overbrace{\text{track}_1}^{\approx\textit{‘to follow’}}$	$\overbrace{\text{track}_2}^{\approx\textit{sports}}$	$\overbrace{\text{track}_3}^{\approx\textit{‘railroad line’}}$
monitor	jumping	railroad
keep	race	deck
analyze	scramble	mainline
validate	flight	wye
check	sledge	fence
manage	cycling	gate
evaluate	rowing	rail
assess	pitches	sidings

*BIMU, GigaFrEn


SCWS
Correlation
0.60
0.55
0.50
0.45
0.40
0.35
0.30
FR-news
CZ-news
RU-news
ES-news
DE-news
gigaFRen
CZeng
RU-yandex
SG
MU
BIMU

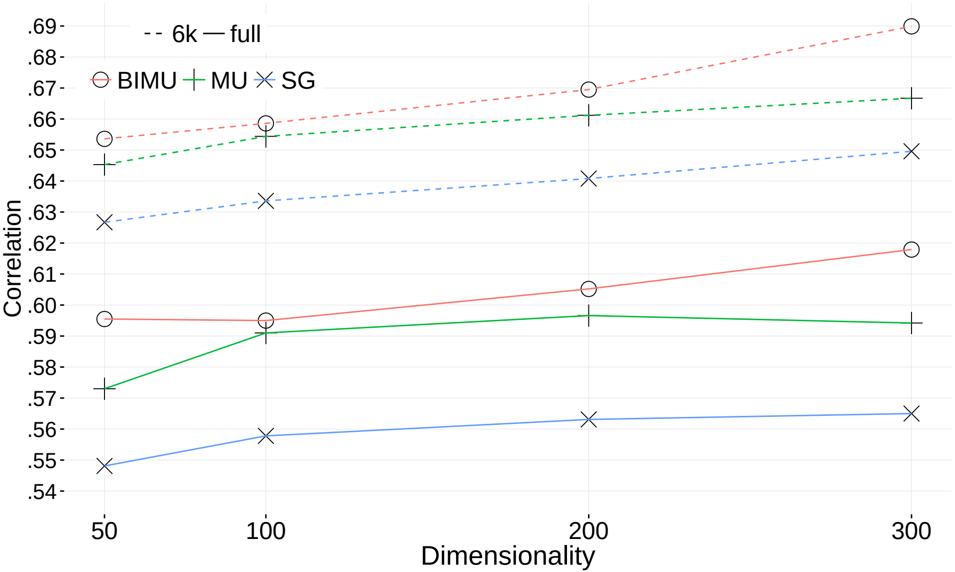
Results on other evaluation tasks

- Semantic similarity & Qvec:
  - despite the lack of context, large improvements over MU and SG with Russian
  - uniformly averaged sense embeddings might represent rare senses better than SG
- POS tagging:
  - bilingual signal somewhat beneficial
  - overall multi-sense models less robust compared to SG
  - NN might be disentangling the senses in SG embeddings, cf. [Li and Jurafsky, 2015]

Other findings

- Are word alignments necessary?
  - very large L2 context windows (=entire sentence) work well too
- Corpus domain matters
  - e.g. embeddings trained on Yandex (23M) almost as good those trained on GigaFrEn (670M)
- Robustness to increased dimensionality

Effect of embedding dimensionality (SCWS)



Recap

- Bilingual learning affects monolingual quality of English embeddings positively
- Bilingual signal not used at test time
- Some benefits even without word alignments

Recap

- Bilingual learning affects monolingual quality of English embeddings positively
- Bilingual signal not used at test time
- Some benefits even without word alignments

Future directions

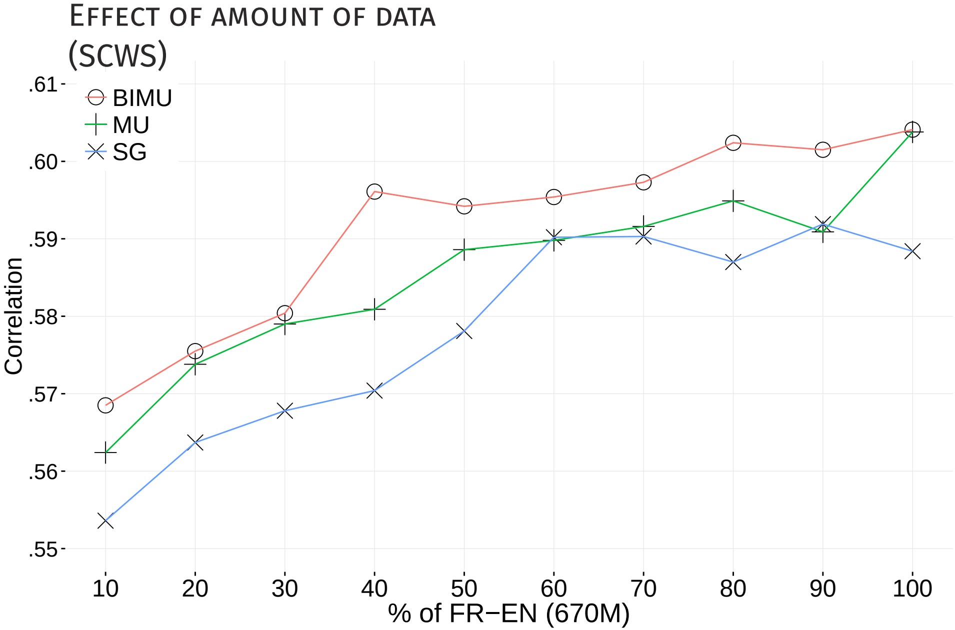
- Effect of language choice and language distance
- Bilingual $\rightarrow$ multilingual signal

Thank you!

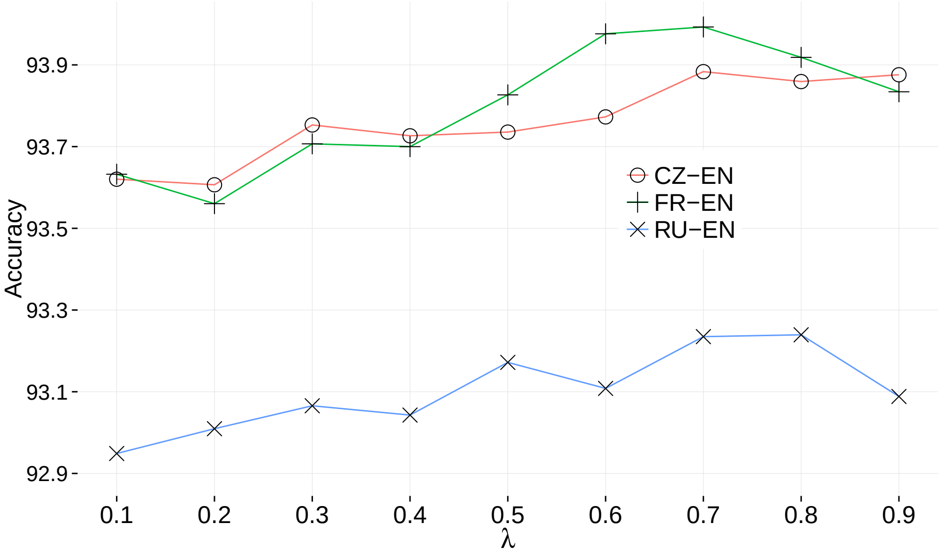
Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders

Simon Šuster, Ivan Titov and Gertjan van Noord

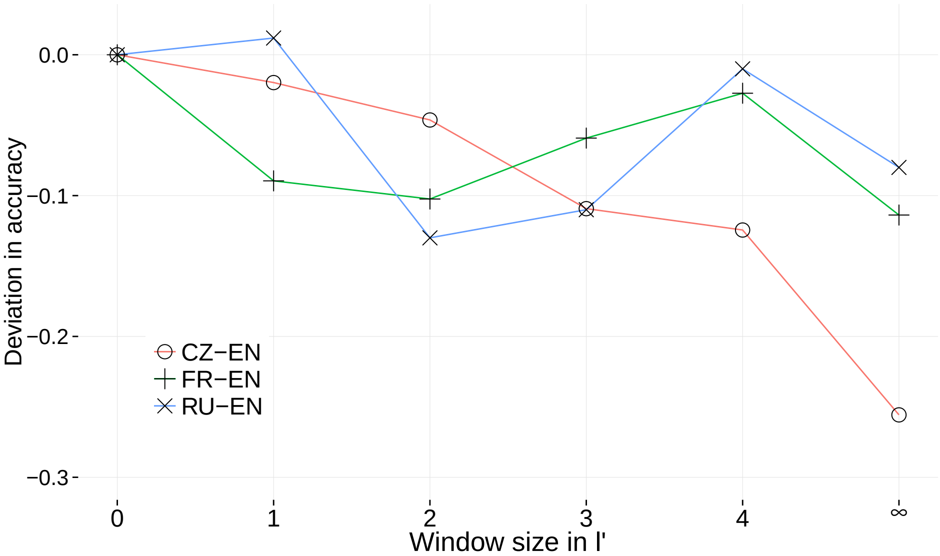
Code: `github.com/rug-compling/bimu`


Effect of amount of data
(SCWS)
BIMU
MU
SG
Correlation
.61
.60
.59
.58
.57
.56
.55
10
20
30
40
50
60
70
80
90
100
% of FR−EN (670M)

Weighting L2 (POS tagging)



Effect of L2 window size (POS tagging)



Model	RU-EN	CZ-EN	FR-EN
MU	.63	.59	.64
BIMU, $m = \infty$	**.66**	**.62**	**.64**

Table : Comparison of SCWS correlation scores of BIMU trained with infinite l' window to the MU baseline (vocabulary of top-6000 words).

Model (300-dim.)	SCWS
Sg	.65
Mu	.66
BiMu	.69
[Chen et al., 2014]	.68
[Neelakantan et al., 2014]	.69
[Li and Jurafsky, 2015]	.70

Table : Comparison to other works (reprinted), for the vocabulary of top-6000 words. Our models are trained on RU-EN, a much smaller corpus than those used in previous work.

Task	Corpus	Sg	Mu	BiMu
Similarity	RU-EN	37.8	41.2	**46.3**
	CZ-EN	39.5	36.9	**41.9**
	FR-EN	**46.3**	42.0	43.5
	FR-EN (NC)	17.9	26.0	**27.6**
	RU-EN (NC)	19.3	27.3	**28.4**
	CZ-EN (NC)	15.8	**26.6**	25.4
	DE-EN (NC)	20.7	28.4	**30.8**
	ES-EN (NC)	19.9	27.2	**31.2**
Qvec	RU-EN	55.8	56.0	**56.5**
	CZ-EN	**56.6**	56.5	55.9
	FR-EN	57.5	57.1	**57.6**
POS	RU-EN	**93.5**	93.2	93.3
	CZ-EN	**94.0**	93.7	**94.0**
	FR-EN	**94.1**	93.8	94.0

Table : Results, per-row best in bold. Sg and Mu are trained on the English part of the parallel corpora. In BiMu-Sg, we report the difference between BiMu and Sg, together with the 95% CI of that difference. The Similarity scores are averaged over 12 benchmarks. For POS tagging, we report the accuracy.

Ammar, W., Dyer, C., and Smith, N. A. (2014).
Conditional random field autoencoders for unsupervised structured prediction.
In *NIPS*.

Chen, X., Liu, Z., and Sun, M. (2014).
A unified model for word sense representation and disambiguation.
In *EMNLP*.

Diab, M. and Resnik, P. (2002).
An unsupervised method for word sense tagging using parallel corpora.
In *ACL*.

Ettinger, A., Resnik, P., and Carpuat, M. (2016).
Retrofitting sense-specific word vectors using parallel text.
In *NAACL-HLT*.

Faruqui, M. and Dyer, C. (2014).
Improving vector space word representations using multilingual correlation.

In *EACL*.

Gouws, S., Bengio, Y., and Corrado, G. (2014).
BilBOWA: Fast Bilingual Distributed Representations without Word Alignments.
arXiv preprint arXiv:1410.2455.

Guo, J., Che, W., Wang, H., and Liu, T. (2014).
Learning sense-specific word embeddings by exploiting bilingual resources.
In *COLING*.

Hill, F., Cho, K., Jean, S., Devin, C., and Bengio, Y. (2014).
Embedding word similarity with neural machine translation.
arXiv preprint arXiv:1412.6448.

Klementiev, A., Titov, I., and Bhattarai, B. (2012).
Inducing crosslingual distributed representations of words.
In *COLING*.

Li, J. and Jurafsky, D. (2015).
Do multi-sense embeddings improve natural language understanding?

In *EMNLP*.

Marcheggiani, D. and Titov, I. (2016).
Discrete-state variational autoencoders for joint discovery and factorization of relations.
Transactions of the Association for Computational Linguistics, 4.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014).
Efficient non-parametric estimation of multiple embeddings per word in vector space.
In *EMNLP*.

Snyder, B. and Barzilay, R. (2010).
Climbing the Tower of Babel: Unsupervised Multilingual Learning.
In *ICML*.