Universität Stuttgart

Moritz Stiefel
& Ngoc Thang Vu

Code-switching Language Modeling with Bilingual Word Embeddings

Introduction

Code-switching (CS)

Code-switch examples from the SEAME (Lyu et al., 2015) corpus:

- 二十五 号 november
- sidetrack 一下
- content 是 讲 什么
- 不 interesting 我 不想 讲
- then 就 爱 上 each other

Set	train	dev	eval
# Speakers	139	8	8
# Utterances	39,240	1,961	1,235
# Tokens	466,741	23,773	13,547

Table: Statistics of the SEAME corpus

Corpus	BOLT	OS16
# Utterances	15,764	2,484,640
# Tokens	430,279	21,613,900

Table: Statistics of BOLT and OS16

Code-switching occurs with syntactic sensibility (e.g. Poplack, 1981)!

Language modeling (LM)

Task Predict the next word in a sentence given preceding words

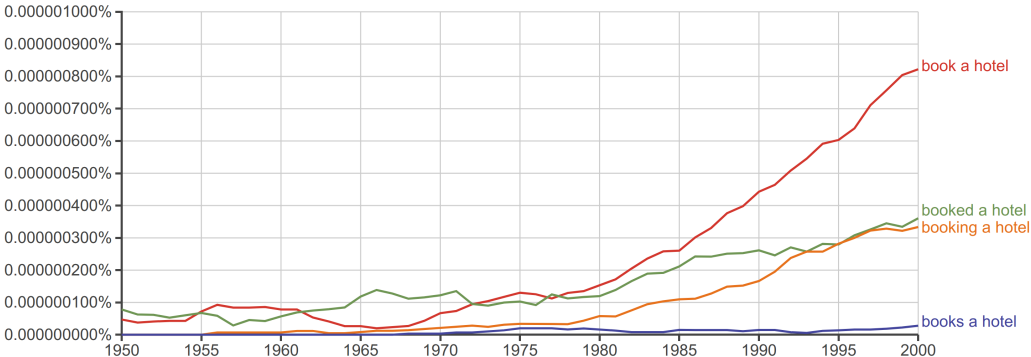


Figure from `https://books.google.com/ngrams/info`, obtained June 22nd 2017.

Language modeling (LM)

Task Predict the next word in a sentence given preceding words

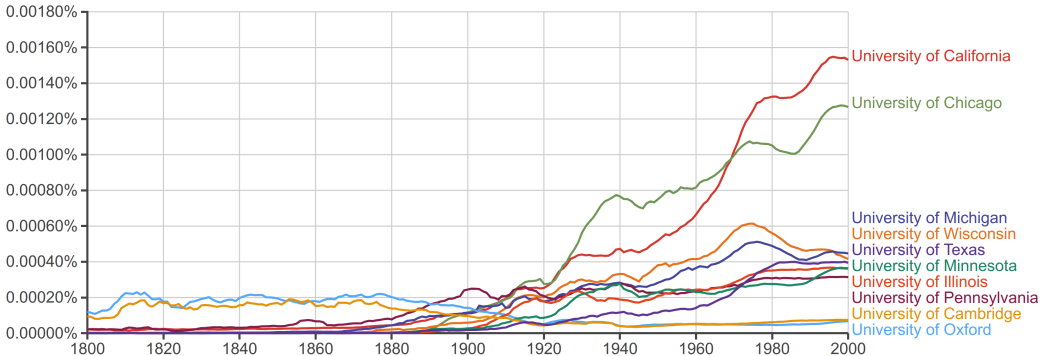


Figure from `https://books.google.com/ngrams/info`, obtained June 22nd 2017.

Word embeddings (WE)

- are dense vector representations of words
- consider context
- do not need supervision in training
- perform well across tasks

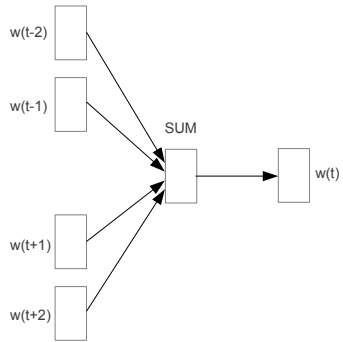
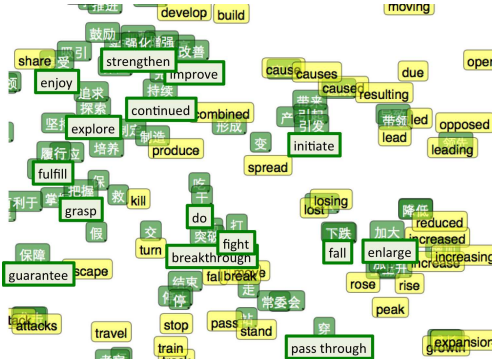


Figure from Mikolov et al. (2013).

Motivation to use bilingual WE (BWE)

- BWE been shown to improve even monolingual quality of embeddings (Luong et al., 2015)
- BWE can map similar words across languages close to each other in a shared vector space

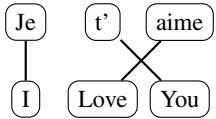


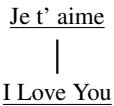
⇒ We need a shared representation with predictive quality for two languages!

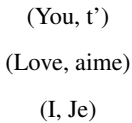
Figure from Zou et al. (2013).

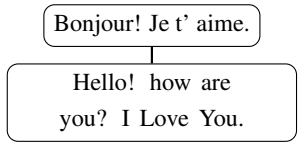
Approaches

Other BWE approaches









BiSkip
(Luong et al., 2015)

word-level alignments

BiCVM
(Hermann and Blunsom, 2014)

sentence-level alignments

BiCCA
(Faruqui and Dyer, 2014)

translation lexicon

BiVCD
(Vulic and Moens, 2015)

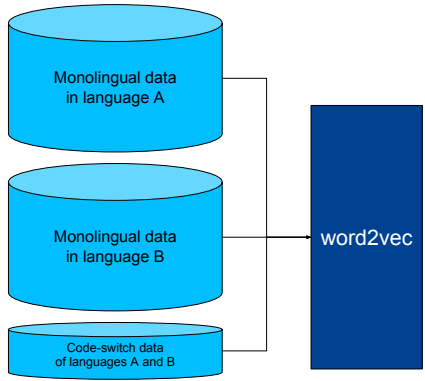
comparable documents

Figure from Upadhyay et al. (2016).

Our approach

Model word2vec skip-gram model with hierarchical softmax and max. skip distance of 5 no frequency cutoff

What's new? Use CS data in the training data, alongside monolingual data



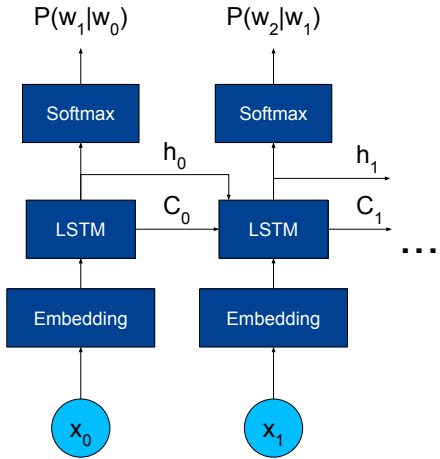
Results

Baseline LM results and evaluation model

We use TheanoLM* for task-based evaluation:

Models	dev	test
SRILM 3-gram	378.9	337.6
SRILM 4-gram	381.7	340.6
Random init, $d = 100$	334.8	336.5
Random init, $d = 200$	321.8	325.9
CBOW-w2v, $d = 100$	343.7	347.3
CBOW-w2v, $d = 200$	330.1	333.2
Skip-gram w2v, $d = 100$	277.4	283.1
Skip-gram w2v, $d = 200$	**274.6**	**280.9**

Table: Baseline results (SEAME only)



LM results with BWE

test

Models	dev	test
Skip-gram (BOLT)	332.5	331.8
Bi-CS (SEAME+BOLT)	276.5	280.1
Skip-gram (OS16)	328.9	328.2
Bi-CS (SEAME+OS16)	266.4	**265.7**
Skip-gram (BOLT+OS16)	331.9	331.0
Bi-CS (SEAME+BOLT+OS16)	265.6	267.6

Table: w2v results with and *without* SEAME data.

Models	dev	test
Bi-CS (SEAME+BOLT+OS16)	265.6	267.6
BOLT+OS16		
BiCCA-skip	291.1	292.3
BiCVM-add	281.1	282.9
BiCVM-bi	282.9	284.2
BiSkip	271.0	271.2
SEAME+BOLT+OS16		
BiCCA-skip	286.2	286.8
BiCVM-add	270.1	271.7
BiCVM-bi	275.1	277.3
BiSkip	260.4	**258.1**

Table: Comparison between Bi-CS, BiCCA, BiCVM and BiSkip trained with and *without* SEAME data.

Conclusion

Conclusion

- CS data is a valuable resource, even in small quantities
- Strong models and algorithms deliver good results given the right data

Conclusion

- CS data is a valuable resource, even in small quantities
- Strong models and algorithms deliver good results given the right data

Generated samples:

- so 有 年 让 人家 的 那种 分数 给 你 知道 你 可以 哥哥 没有 摔 gossip out 吗

Conclusion

- CS data is a valuable resource, even in small quantities
- Strong models and algorithms deliver good results given the right data

Generated samples:

- so 有 年 让 人家 的 那种 分数 给 你 知道 你 可以 哥哥 没有 摔 gossip out 吗
- especially that day 跟 我们 去 rock

Universität Stuttgart

Moritz Stiefel & Ngoc Thang Vu
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart

eMail moritz.stiefel@ims.uni-stuttgart.de
Telefon +49-711-685 813 60
Fax +49-711-685 813 66

References

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 462–471. The Association for Computer Linguistics, 2014. URL http://aclweb.org/anthology/E/E14/E14-1049.pdf.

Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 58–68. The Association for Computer Linguistics, 2014. URL http://aclweb.org/anthology/P/P14/P14-1006.pdf.

Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In Phil Blunsom, Shay B. Cohen, Paramveer S. Dhillon, and Percy Liang, editors, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, pages 151–159. The Association for Computational Linguistics, 2015. URL http://aclweb.org/anthology/W/W15/W15-1521.pdf.

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. Mandarin-english code-switching speech corpus in south-east asia: SEAME. *Language Resources and Evaluation*, 49 (3):581–600, 2015. doi: 10.1007/s10579-015-9303-x. URL http://dx.doi.org/10.1007/s10579-015-9303-x.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL http://arxiv.org/abs/1301.3781.

Shana Poplack. Syntactic structure and social function of code-switching. In Richard P. Durán, editor, *Latino Language and Communicative Behavior*, chapter 10, pages 169–184. ABLEX Publishing Corp., New Jersey, 1981. ISBN 0893910384.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. URL http://aclweb.org/anthology/P/P16/P16-1157.pdf.

Ivan Vulic and Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 719–725. The Association for Computer Linguistics, 2015. URL http://aclweb.org/anthology/P/P15/P15-2118.pdf.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1393–1398. ACL, 2013. URL http://aclweb.org/anthology/D/D13/D13-1141.pdf.