Bilingual Embeddings for Phrase-Based Machine Translation

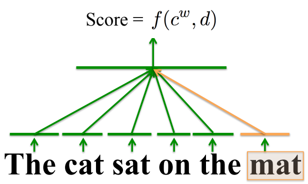
Will Y. Zou, Richard Socher, Daniel Cer and Christopher D. Manning

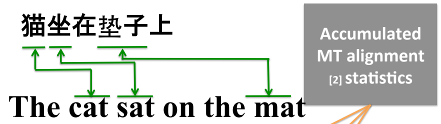
1 Abstract

We introduce bilingual word embeddings: semantic embeddings associated across two languages in the context of neural language models. We propose a method to learn bilingual embeddings from a large unlabeled corpus, while utilizing MT word alignments to constrain translational equivalence. The new embeddings significantly out-perform baselines in word semantic similarity. A single semantic similarity feature induced with bilingual embeddings adds near half a BLEU point to the results of NIST08 Chinese-English machine translation task.

3 Neural language models and bilingual semantics

Neural language models [1][4] learn distributed representations of words and offer a framework to incorporate cross-language constraints.



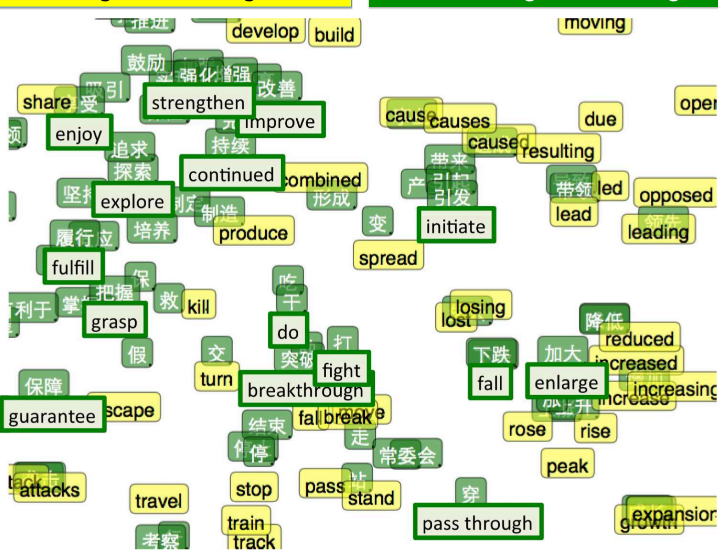


Initialization: $W_{t\text{-}init} = \sum_{s=1}^{S} \frac{C_{ts}+1}{C_t+S} W_s$

Optimization objective [4]:

$$J_{CO}^{(c,d)} = \sum_{w^r \in V_R} \max(0, 1 - f(c^w, d) + f(c^{w^r}, d)) + \quad J_{TEO\text{-}en \to zh} = \|V_{zh} - A_{en \to zh} V_{en}\|^2$$

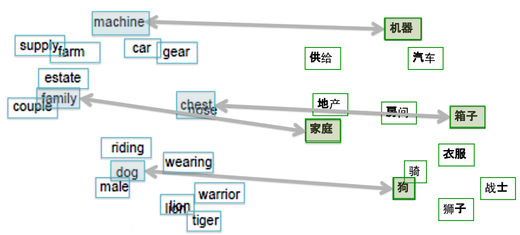
Monolingual Embeddings | **Bilingual Embeddings**



2 Motivation

A. un cas de force majeure ←?→ *case of absolute necessity*
(an event of) (unavoidable accident)

B. 依然故我 ←?→ persist in a stubborn manner
(as before)(old)(self)



Difficult for classical Statistical Machine Translation systems: require enough co-occurrences to identify semantic equivalence.

Word Embeddings using Neural Language Models map words into low-dimensional semantic space

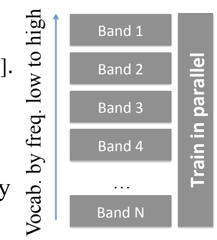
Large-scale embeddings of words across languages

Better semantically-informed MT systems

4 Optimization and training

The 100k Mandarin Chinese embeddings contain **5 million parameters**. We train these embeddings on the **Chinese Gigaword corpus** using **mini-batch LBFGS** across 19 days.

We perform **Band Curriculum Training** [3]. The vocabulary is sorted by frequency to band-sizes {5k, 10k, 25k, 50k, 100k}. All bands are trained in parallel for 100k iters per curriculum. Finally the entire vocabulary is trained for 500k iters.



5 Application pipeline for phrase-based MT

Phrase in Lan A: *依然故我* → Average embeddings → **Phrase representation**

Phrase in Lan B: persist stubbornly → Average embeddings → **Phrase representation**

Phrase representation, Phrase representation → Cosine Similarity → Phrase-pair Feature → MT decoder

6 Results

Word semantic similarity

Method	*Sp. Corr.* (×100)	*K. Tau* (×100)
Prior work (Jin and Wu, 2012)		5.0
Tf-idf		
Naive tf-idf	41.5	28.7
Pruned tf-idf	46.7	32.3
Word Embeddings		
Align-Init	52.9	37.6
Mono-trained	59.3	42.1
Biling-trained	**60.8**	**43.3**

Named Entity Recognition

Embeddings	*Prec.*	*Rec.*	*F1*	*Improve*
Align-Init	0.34	0.52	0.41	
Mono-trained	0.54	0.62	**0.58**	**0.17**
Biling-trained	0.48	0.55	0.52	0.11

Vector matching alignment

Embeddings	*Prec.*	*Rec.*	*AER*
Mono-trained	0.27	0.32	0.71
Biling-trained	0.37	0.45	**0.59**

BLEU score on NIST08 Chinese-English translation task

Method	*BLEU*
Our baseline	30.01
Embeddings	
Random-Init Mono-trained	30.09
Align-Init	30.31
Mono-trained	30.40
Biling-trained	**30.49**

7 References

[1] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin. A Neural Probabilistic Language Model. JMLR 2003
[2] P. Liang, B. Taskar and D. Klein. Alignment by Agreement. NAACL 2006
[3] Y. Bengio, J. Louradour and J. Weston. Curriculum Learning. ICML 2009
[4] E. H. Huang, R. Socher, C. D. Manning and A. Y. Ng. Improving Word Representations via Global Context and Multiple Word Prototypes. ACL 2012

http://ai.stanford.edu/~wzou/mt/