OVERVIEW
Semantically similar words are close across languages in a crosslingual vector space
Leverage monolingual + bilingual resources
- Sentence-parallel corpus
- Monolingual corpora in both languages
Composition
- Training at phrase level
- Flexible w.r.t. composition function
Training-speed scales to large corpora
- Vocabulary > 150,000 tokens
- Corpora > 100 million tokens

MODEL
To obtain a vector representation of a phrase
1.) Look up word vectors 2.) Apply composition function
Bilingual Objective (Sentence-parallel data, EuroParl)
Minimize distance between bilingualy aligned sentences

Hinge Loss. Sub-phrase closer to mother-phrase than to random other phrase
\[
\max(0, m + \|a_{c_{outer}} - a_{c_{inner}}\|^2 - \|a_{c_{outer}} - b_{noise}\|^2) + \|a_{c_{outer}} - a_{c_{inner}}\|^2
\]

Monolingual Objective (Monolingual data, RCV1)

Mini-batch AdaGrad, BackProp into word vectors

EVALUATION
Crosslingual document classification of news text
- Reuters news text - German - English
- 4 categories - ~5000 test docs - 1000 training docs
Train classifier on text in language 1 (word reprs: l1)
Apply classifier to text in language 2 (word reprs: l2)

Examples (Nearest Neighbors)
("hard cases" - tokens appear only in monolingual data)

1Inducing crosslingual distributed representations of words
Klementiev, Alexandre, Titov, Ivan, and Bhattarai, Binod