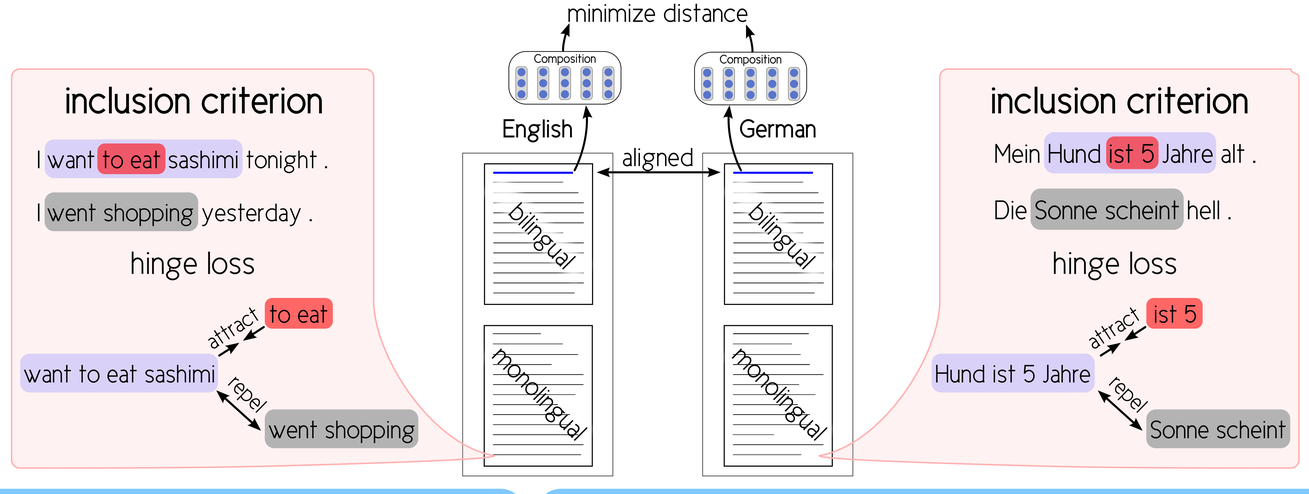
LEVERAGING MONOLINGUAL DATA FOR CROSSLINGUAL COMPOSITIONAL WORD REPRESENTATIONS

Hubert Soyer* *hubert.soyer@gmail.com*

Pontus Stenetorp† *pontus@stenetorp.se*

Akiko Aizawa* *aizawa@nii.ac.jp*

* National Institute of Informatics, Tokyo, Japan
† University College London, London, UK



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

Compositional
- 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

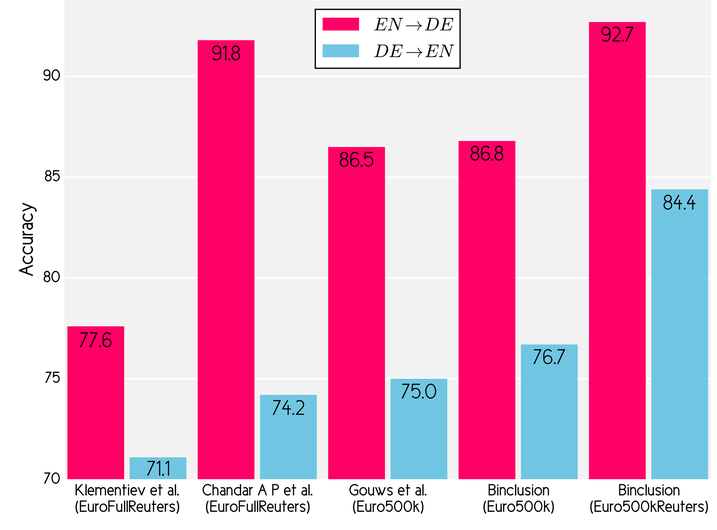
Bilingual Objective (Sentence-parallel data, EuroParl)
Minimize distance between bilingually aligned sentences

Monolingual Objective (Monolingual data, RCV1)
Hinge Loss. Sub-phrase closer to mother-phrase than to random other phrase

$$\underbrace{[max(0, m + \|a_c^{outer} - a_c^{inner}\|^2 - \|a_c^{outer} - b_c^{noise}\|^2)}_{\text{hinge loss}} + \|a_c^{outer} - a_c^{inner}\|^2] \cdot \frac{len(a^{inner})}{len(a^{outer})}$$

Mini-batch AdaGrad, BackProp into word vectors

EVALUATION



Crosslingual document classification of news text[1]
- 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)

English	soybeans	s&p	stockholders
German	mais	ratings	aktionärsschutz
	alkoholherstellung	ratingindustrie	minderheitenaktionäre
	silomais	ratingbranche	aktionärsrechte
	genmais	ratingstiftung	aktionäre
	gluten	kreditratingagenturen	minderheitenaktionären

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