Word embeddings for 1250 languages through multi-source projection

Robert Östling and Murathan Kurfalı

Department of Linguistics, Stockholm University

2018-09-28

The problem

- Multilingual word embeddings are important for NLP
- Good methods for high/medium-resource languages (100+)
- What about low-resource languages (1000+)?

What we have

- Good multilingual embeddings for high-resource languages
- Short parallel texts (mainly Bible translations)
- **No** extra monolingual data for low-resource languages

Method

- Multi-source projection through parallel text
- 27 source languages (best ones from Smith et al. 2017)
- 1250 target languages

Conclusions

- Naive projection works, but...
  - a single source language $\rightarrow$ bad (unless very closely related)
  - several unrelated languages $\rightarrow$ acceptable and realistic
  - several related languages $\rightarrow$ better but unrealistic

Out-of-domain effects

Source	**police** say that the **truck** driver was not drunk at the time .
Translation	**vakterna** påstå att den **vagnen** förare hade inte drucken vid den tiden .
Glossing	the-**guards** claim that the **wagon** driver had not drunken by that time .

Hidden research question

How to translate modern news text into something a time traveler from 2000 years ago could relate to?

The embeddings

http://mumin.ling.su.se/fotran2018

Results

	Eng to Swe		Swe to Eng	
	p@1	p@5	p@1	p@5
ind	0.137	0.344	0.173	0.335
ind+fin	0.231	0.462	0.223	0.394
ind+fin+hun	0.255	0.493	0.234	0.399
ind+fin+hun+tur	**0.269**	0.501	0.235	**0.400**
ind+fin+hun+tur+est	0.267	**0.504**	**0.236**	0.395
Smith et al. (2017)	0.501	0.686	0.525	0.722

- Word translation results using nearest-neighbor search in multilingual embeddings
- Only non-IE languages used to simulate low-resource scenario
- Difficult but realistic constraint