Orthographic features for bilingual lexicon induction

Parker Riley and Daniel Gildea

University of Rochester

July 17, 2018

Outline

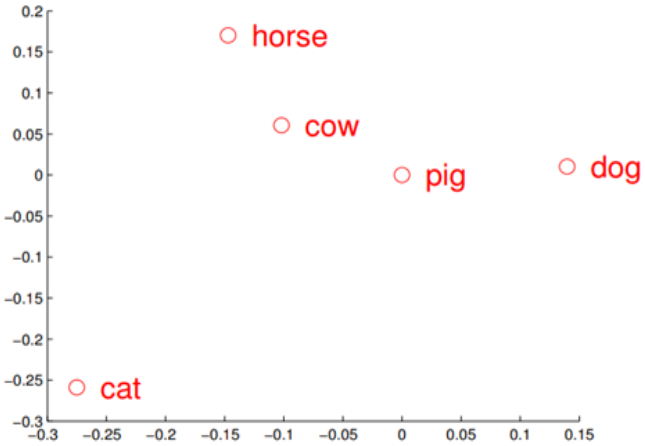
- Overview
  - Research question
  - Task and general approach
- Baseline system
- Proposed modifications
- Results
- Conclusion

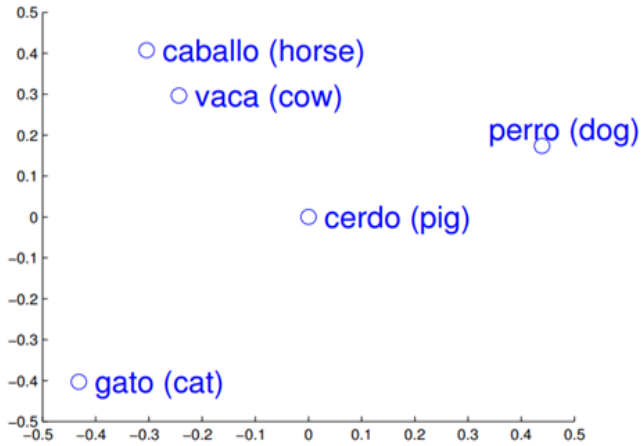
Overview - Research question

- Can orthographic (spelling) information enable better word translations in low-resource contexts?
  - Languages with common ancestors and/or borrowing exhibit increased lexical similarity
  - Spelling of words can carry signal for translation
  - Low-resource pairs are most in need of additional signal

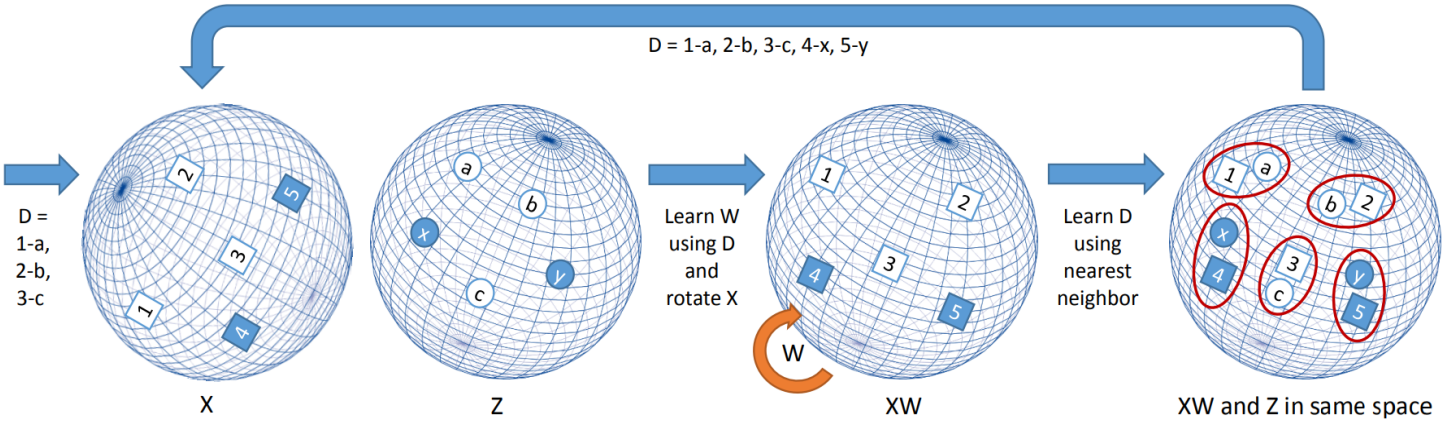
Overview - Task and general approach

- Bilingual lexicon induction: single-word translations (*modern-moderno*)
- Operate on word embeddings
  - Haghigi et al. (2008): orthographic features
  - Mikolov et al. (2013): word2vec, linear mapping
- Minimal supervision





Baseline: Artetxe et al. (2017)



- Start with dictionary D (inferred from numerals)
- Learn matrix W minimizing Euclidean distance between target (Z) and mapped source (XW) embeddings of pairs in D
- Use nearest neighbors as entries in new dictionary
- Repeat until convergence

Baseline: Artetxe et al. (2017) - Problems

Language	English Word	Baseline's Prediction	Reference
German	*unevenly*	*gleichmäßig* (*evenly*)	*ungleichmäßig*
German	*Ethiopians*	*Afrikaner* (*Africans*)	*Äthiopier*
Italian	*autumn*	*primavera* (*spring*)	*autunno*
Finnish	*Latvians*	*ukrainalaiset* (*Ukrainians*)	*latvialaiset*

- Suffers from clustering problems present in word2vec
  - Similar distributions $\rightarrow$ similar embeddings
- Hints of correct translation present in spelling

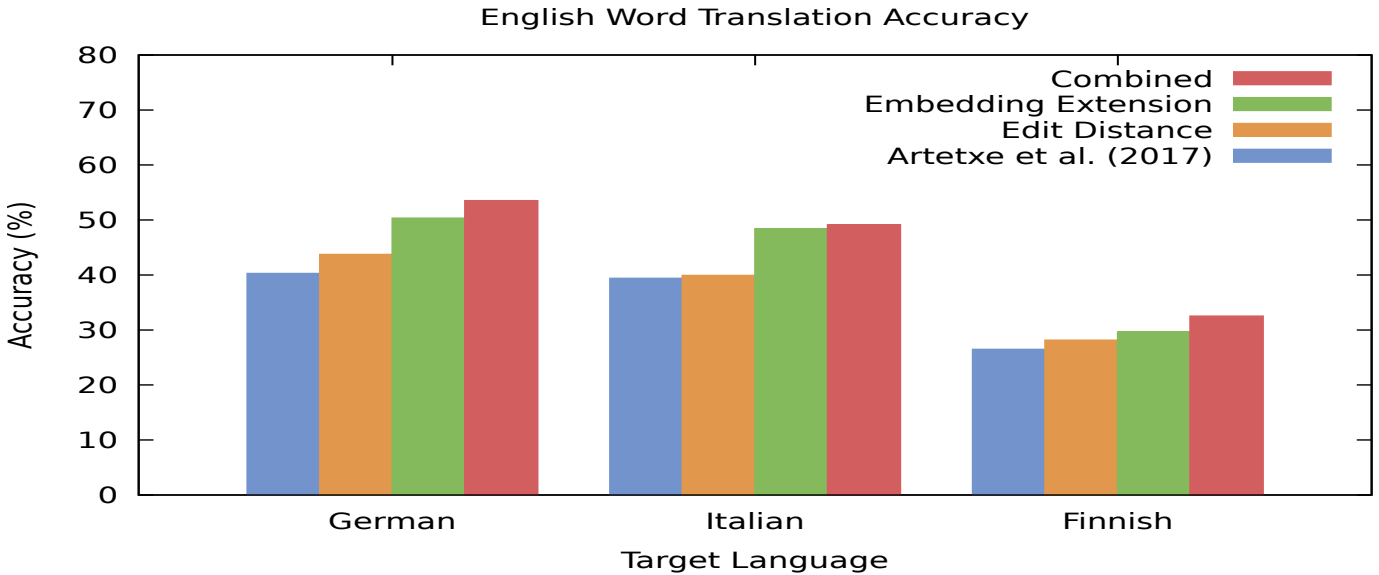
Proposed modifications

1. Use normalized edit distance in nearest-neighbor calculation
   - During dictionary induction, distances between similarly-spelled words are reduced
2. Extend embedding vectors with character counts
   - Extend vectors with scaled counts of letters in both language's alphabets (scale constant $k \leq 1$)

Word	d_1	d_2
aba	0.123	0.456

Word	d_1	d_2	a	b
aba	0.123	0.456	$2k$	$1k$

Quantitative results



- Universally outperform baseline
- Best when combined; largest contribution from embedding extension
- Improvement less pronounced for English-Finnish (linguistic dissimilarity)

Qualitative results

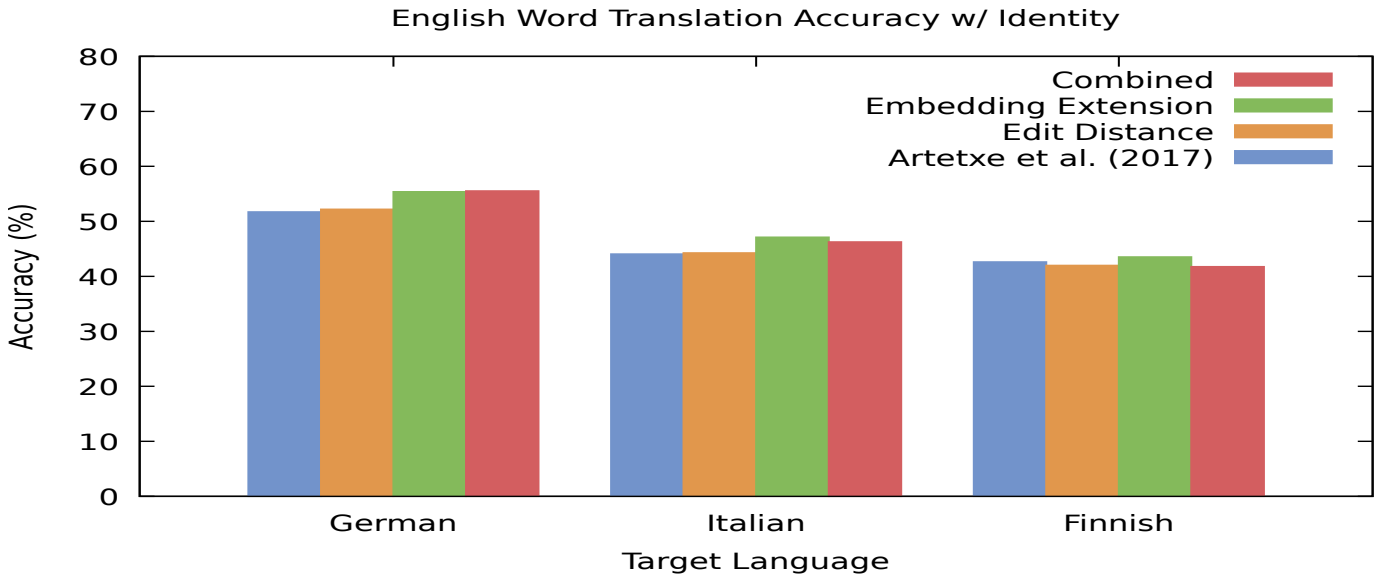
Language	English Word	Baseline's Prediction	Our Prediction
German	*unevenly*	*gleichmäßig* (*evenly*)	*ungleichmäßig*
German	*Ethiopians*	*Afrikaner* (*Africans*)	*Äthiopier*
Italian	*autumn*	*primavera* (*spring*)	*autunno*
Finnish	*Latvians*	*ukrainalaiset* (*Ukrainians*)	*latvialaiset*

- Use orthographic information to disambiguate semantic clusters
- Significant gains in adequacy

Conclusion

- Orthographic information can improve unsupervised bilingual lexicon induction, especially for language pairs with high lexical similarity.
- These techniques can be incorporated into other embedding-based frameworks.

Results with Identity



Proof of optimal W

$$
\begin{aligned}
W^* &= \arg\min_W \sum_{i=1}^{|V^X|} \sum_{j=1}^{|V^Z|} D_{ij} \|X_{i*}W - Z_{j*}\|^2 \\
&= \arg\min_W \sum_{i=1}^{|V^X|} \|X_{i*}W - (DZ)_{i*}\|^2 \\
&= \arg\min_W \sum_{i=1}^{|V^X|} \|X_{i*}W\|^2 + \|(DZ)_{i*}\|^2 - 2X_{i*}W((DZ)_{i*})^\intercal \\
&= \arg\min_W \sum_{i=1}^{|V^X|} -2X_{i*}W((DZ)_{i*})^\intercal = \arg\max_W \sum_{i=1}^{|V^X|} X_{i*}W((DZ)_{i*})^\intercal \\
&= \arg\max_W \mathrm{Tr}(XWZ^\intercal D^\intercal)
\end{aligned}
$$

Proof of optimal W, continued

$$
\begin{aligned}
W^* &= \arg\max_W \mathrm{Tr}(XWZ^\intercal D^\intercal) \\
&= \arg\max_W \mathrm{Tr}(Z^\intercal D^\intercal XW) \\
&= \arg\max_W \mathrm{Tr}(U\Sigma V^\intercal W) \quad [U\Sigma V^\intercal = \mathrm{SVD}(Z^\intercal D^\intercal X)] \\
&= \arg\max_W \mathrm{Tr}(\Sigma V^\intercal WU) \\
&= VU^\intercal
\end{aligned}
$$

Method	English-German	English-Italian	English-Finnish
Artetxe et al. (2017)	40.27	39.40	26.47
Artetxe et al. (2017)+id	51.73	44.07	42.63
Embedding extension	50.33	48.40	29.63
Embedding extension+id	55.40	47.13	**43.54**
Edit distance	43.73	39.93	28.16
Edit distance+id	52.20	44.27	41.99
Combined	53.53	**49.13**	32.51
Combined+id	**55.53**	46.27	41.78