Obtaining SMT dictionaries for related languages

Miguel Rios, Serge Sharoff
University of Leeds

Centre for Translation Studies
University of Leeds

30 July 2015

Outline

Motivation

- Extracting **cognates** for related languages in Romance and Slavonic language families
- Reducing the number of **unknown** words on SMT training data
- Learning regular differences in words **roots/endings** shared across related languages

Outline

Method

- Produce n-best lists of cognates using a family of distance measures from **comparable** corpora
- Prune the n-best lists by **ranking** Machine Learning (ML) algorithm trained over **parallel** corpora
- **Motivation** n-best list allows surface variation on possible cognate translations

Similarity metrics

- Compare words between frequency lists over comparable corpora
  Produce n-best lists
- **L** matching between the languages using Levenshtein distance:
  maladie → *malattia*
- **L-R** Levenshtein distance computed separately for the roots and for the endings:
  aceito (pt) vs **acept**o (es)
  rejeit**o** (pt) vs rechaz**o** (es)
- **L-C** Levenshtein distance over words with similar number of starting characters (i.e. prefix):
  introdução (pt) vs **introdu**cción (es)
  introduziu (pt) vs **introdu**jo (es)

Search space constraints

- **Motivation** Exhaustive method compares all the combinations of source and target words
- Order the target side frequency list into **bins** of similar frequency
  Compare each source word with target bins of similar frequency around a **window**
- **L-C** metric only compares words that share a given n prefix (characters)

Ranking

- **Motivation** Prune n-best lists by ranking ML algorithm
- Training data come from aligned parallel corpora where the **rank** is given by the **alignment** probability from GIZA++
- Simulate cognate **training** data by pruning pairs of words below a Levenshtein threshold

Features

- Similarity metric L
- Number of times of each edit operation, the model assigns a different weight to each operation
- Cosine between the distributional vectors of the source and target words
  vectors from word2vec
  mapped to same space via a learned transformation matrix
- SVM ranking default configuration (RBF kernel)
- Easy-adapt features given different domains (Wikipedia, subtitles)

Outline

Data description

- **n-best** lists from Wikipedia **dumps** (frequency lists)
- **ML training** Wiki-titles, parallel data from inter language links from the tittles of the Wikipedia articles 500K aligned links (i.e. 'sentences')
  Opensubs, 90K training instances
  Zoo proprietary corpus of subtitles produced by professional translators, 20K training instances
- **Ranking test** Heldout data from training
- **Manual cognate test** Wikipedia most frequent words
- **SMT test** Zoo data

Language pairs

- **Romance** Source: Portuguese, French, Italian Target: Spanish
- **Slavonic** Source: Ukrainian, Bulgarian Target: Russian

Results on heldout data

- Error score on heldout data
- **E** Edit distance features
- **EC** Edit distance plus distributed vectors features

	Zoo error%		Opensubs error%		Wiki-titles error%	
Lang pairs	Model E	Model EC	Model E	Model EC	Model E	Model EC
Romance						
pt-es	53.31	53.72	54.81	48.31	12.22	9.87
it-es	56.00	42.86	63.95	63.03	8.44	11.23
fr-es	59.05	53.00	43.00	41.19	10.75	10.09
Slavonic						
uk-ru	47.90	40.84	37.06	30.19	10.71	10.72
bg-ru	54.17	43.98	49.12	57.89	18.72	17.13

Manual evaluation

- Results on sample of 100 words
  Accuracy at 1, 10
- n-best lists **L**, **L-R**, **L-C**
- ranking model **E**

	List L		List L-R		List L-C	
Lang Pairs	acc@1	acc@10	acc@1	acc@10	acc@1	acc@10
pt-es	20	60	22	59	32	70
it-es	16	53	18	45	44	66
fr-es	10	48	12	51	29	59

Addition of lists SMT

- Moses phrase-based SMT
- 1-best lists with **L-C** and **E** ranking
- pt-es: 80K training sentences, 100K cognate pairs
  BLEU score baseline: 20.68 and augmented:20.86, +0.18 not significant
- uk-ru: 140K training sentences, 100K cognate pairs
  BLEU score baseline: 28.72 and augmented: 29.56, +0.93 not significant

Out-of-vocabulary reduction

- pt-es (OOV): 623 types (**21.1%**) to 337 types (**11.4%**)
- uk-ru (OOV): 756 types (**21.6%**) to 545 types (**15.6%**)

Outline

Conclusions

- MT dictionaries extracted from comparable resources for related languages
- Positive results on the n-bes lists with **L-C**
- Frequency **window** heuristic shows poor results
- ML models are able to rank similar words on the top of the list
- Preliminary results on an SMT system show modest improvements compare to the baseline
- The OOV rate shows improvements around **10%** reduction on word types

Future work

- Morphology features for the n-best list (Unsupervised)
  Instead of prefix heuristic (**L-C**) and stemmer (**L-R**)
- Contribution for all the produced cognate lists on SMT
  Using char-based transliteration model trained on Zoo plus n-best lists
  Motivation alignment learns useful transformations: e.g. introdu**ção** (pt) vs introdu**cción** (es)