Word’s Vector Representations meet Machine Translation

Eva Martínez Garcia
Cristina España-Bonet
Jörg Tiedemann
Lluís Màrquez

SSST-8 Workshop
Doha, Qatar
October 25, 2014
Summary

- We learn distributed vector representations of **bilingual** word pairs

  - *Motivation*: better characterize ambiguous words for MT
    - desk|mesa vs. desk|mostrador vs. desk|escritorio

- Bilingual representations capture information from source and target language contexts simultaneously

- We present two preliminary evaluations
Summary

- We learn distributed vector representations of **bilingual** word pairs
- **Motivation**: better characterize ambiguous words for MT
  - \(\text{desk} | \text{mesa} \) vs. \(\text{desk} | \text{mostrador} \) vs. \(\text{desk} | \text{escritorio} \)
- Bilingual representations capture information from source and target language contexts simultaneously
- We present two preliminary evaluations
- Future plan: use bilingual models in MT for improving lexical selection and document-level semantic coherence
We use the word2vec software (Mikolov et al. 2013) with parallel corpora and automatic word-alignments.

Parallel corpora: Opus (Europarl, UN, OpenSubtitles, etc.)

Word alignments: GIZA++ (one to one)

We train bilingual and monolingual vector models
  - Size: $\sim 700$ Mw (EN) – $1,100$ Mw (ES)

Parameters
  - Vector dimensionality
  - Context window
Eval I: Ability to Capture Relational Similarities

- Solving semantic analogies with vector models:
  - **Athens** is to **Greece** as **Paris** is to ?
  - Paris – Athens + Greece = France
Eval I: Ability to Capture Relational Similarities

- Solving semantic analogies with vector models:
  - **Athens** is to **Greece** as **Paris** is to ?
  - Paris – Athens + Greece = France

- Bilingual version of the same task

- Test set of 19,520 questions in 11 categories
  - EN: available in the *work2vec* data distribution
  - EN|ES (and also ES): translated and manually built by a Spanish native speaker
Eval II: Cross-Lingual Lexical Substitution

- Find the best translation of a given ambiguous word in context (source and target)
  - Same setting as SemEval-2010 task 2
Find the best translation of a given ambiguous word in context (source and target)
  ▶ Same setting as SemEval-2010 task 2

Test set from News Commentary 2010
  ▶ Ambiguous words (lemma level) automatically detected
  ▶ Stop word list to filter out non content words
Eval II: Cross-Lingual Lexical Substitution

- Find the best translation of a given ambiguous word in context (source and target)
  - Same setting as SemEval-2010 task 2

- Test set from News Commentary 2010
  - Ambiguous words (lemma level) automatically detected
  - Stop word list to filter out non content words

- Method:
  - Context Vector: $v = \sum_{i=1}^{n} \tilde{w}(t \pm i)$
  - Best translation: word pair that minimizes distance to $v$ (i.e. best fit to the bilingual context)
Conclusions

- Results in both evaluation are modest, but suggest that the bilingual vector models:
  - capture information useful to uncover semantic relations
  - can help MT lexical selection

- Limitations:
  - quality of translations, alignments, coverage, etc..
Conclusions

- Results in both evaluation are modest, but suggest that the bilingual vector models:
  - capture information useful to uncover semantic relations
  - can help MT lexical selection

- Limitations:
  - quality of translations, alignments, coverage, etc..

- More experimentation and extensions to come soon

Visit out poster for more details. Thanks!