Machine Translation Projects

Fabienne Braune

2017-05-23

Bilingual Word Embeddings

TOPIC: Evaluating BWEs

Induction with pseudo-parallel data (task 1)

- Read paper: Gouws et al, Simple task-specific bilingual word embeddings, Proc. ACL 2015.
- Create monolingual word embeddings using word2vec with large monolingual corpora.
- Given a seed dictionary, reimplement the method in the paper to create bilingual word embeddings.
- Evaluate the obtained (bilingual) embeddings using bilingual lexicon induction.

Bilingual Word Embeddings

TOPIC: Evaluating BWEs

Bilingual lexicon induction (task 2)

- Using parallel data automatically create bilingual word-pairs.
- How good are the created word pairs? Manually correct errors.
- Use the word pairs to evaluate the induced word embeddings: for each source word in the list compute a translation using the nearest neighbor in the bilingual vector space.
- Compute top-1 and top-10 accuracies.
- Give a qualitative analysis of the errors. Are there any systematic errors?

Bilingual Word Embeddings

TOPIC: Bilingual phrases

Representation of bilingual phrases (task 1)

- Read papers:
  - Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality, Proc. NIPS 2013.
  - Mikolov et al., Exploiting Similarities among Languages for Machine Translation, arXiv preprint arXiv:1309.4168, 2013.

- Using the method described in the first paper (Section 4) to create phrases up to 4-grams in your data.

- Create monolingual embeddings for words and phrases using word2vec with large monolingual corpora.

Bilingual Word Embeddings

TOPIC: Bilingual phrases

Bilingual phrase extraction (task 2)

- Reimplement the mapping method presented in the second paper to create bilingual word and phrase embeddings.
    - For this task, you need a seed phrase-table.
    - Using parallel data, create a phrase-table with the Moses toolkit. Select the 5000 most frequent 1 to 4-grams.
    - How good is your seed table? Manually correct errors.
- Download a small comparable corpus. Create 1-to 4-gram phrases in one language. Use embeddings to mine translations.
- Provide a quantitative analysis of the found phrase-pairs.