Supplementary Materials to "Embedding Learning Through Multilingual Concept Induction"

Philipp Dufter[1], Mengjie Zhao[2], Martin Schmitt[1], Alexander Fraser[1], Hinrich Schütze[1]
[1] Center for Information and Language Processing (CIS) LMU Munich, Germany
[2] École Polytechnique Fédérale de Lausanne, Switzerland
{philipp,martin,fraser}@cis.lmu.de, mengjie.zhao@epfl.ch

1 List of used editions

Files: `editions.txt`

List of all 1664 bible translations (editions) which are used in the paper. The selection of the editions is based on verse coverage in the New Testament.

2 Mapping of edition names to 4-character representation

Files: `name_map.txt`

A mapping which maps edition names as provided in the original corpus to a 4-character representation. The new representation consists of the 3-character ISO 639-3 codes followed by an alphanumeric character (e.g. "eng1" or "enge").

3 List of pivot languages

Files: `pivot_languages.txt`

List of the 10 pivot languages.

4 List of used verses

Files: `verses.txt`

List of all 7958 verse-ids which are used in the paper. The first 2 digits of the verse-id represent the book of the bible, the following 3 digits indicate the chapter and the last 3 digit indicate the verse number. For details see (Mayer and Cysouw, 2014). The selection of verse-ids aims at ensuring high coverage across all selected languages.

5 Split of verses into training, development and test set

Files: `verses_split.txt`

Split of the 7958 verses into 6458 training (trn), 500 development (dev) and 1000 testing (tst) verses. Note that throughout this paper we merge the development and testing verses to a single test set (comprising 1500 verses). We do not perform any model or hyperparameter selection for the roundtrip evaluation and use 5-fold cross-validation on the training set for the tuning of the sentiment classifiers. Note, that throughout the paper and the supplementary the test set always refers to the merged development and testing verses.

6 Sentiment silver standard

Files: `sent_silver.txt`

The silver standard is generated in a 3-step procedure with intensive quality checks in each step to ensure high quality of the standard.

1. First a small seed dataset of 1000 verses is created. Across all 32 English editions we concatenate each verse. For each verse the sentiment is predicted using the VADER classifier (Gilbert, 2014) within the NLTK framework (Bird et al., 2009). To ensure a clear signal, we pick 500 verses with the strongest positive or negative signal (non-neutral) and 500 most neutral verses. These verses were manually inspected.

2. This seed dataset is then expanded as follows. We train 2 linear SVMs on the seed dataset (for positive vs. neutral and negative vs. neutral classification, respectively) using a bag-of-word approach (based on one-hot vectors). We then apply the trained SVMs to all verses to produce 816 non-neutral and 500 neutral verses which are added to the seed dataset. Again the new verses are quality checked.

3. Last, we train again 2 SVMs on the extended seed dataset on a subset of 5 English editions (basic, common, goodnews, kingjames, standard). Again, we apply this classifier to all verses. To accommodate for the fact, that verses tend to be long and multiple sentiments can occur within one verse, we classify the verses using a rolling window of size 10 tokens. Given we receive multiple predictions for each verse, we only select verses which are 80% consistent in its predictions (both across the rolling windows and across the 5 editions). This results in 632 non-neutral and 193 neutral verses which are added to the silver standard.

Note that it is possible in the expansion steps to find verses which are already contained in the extended seed dataset. Obviously we can only train and test the SVMs on verses which are in the silver standard. We follow our training and test verse split and use 5-fold cross validation for the model selection. Overall we observe the following verse counts.

	positive	negative	neutral	total
training	839	487	821	2147
test	162	135	179	476
total	1001	622	1000	2623

The verse ids can be found in the attached files.

7 Roundtrip-translation ground truth

Files: `rt_word.txt`, `rt_char.txt`, `swadesh.txt`

The ground truth contains 70 queries selected from Swadesh (1946)'s list of universal concepts. There are two ground truth sets, one for WORD evaluation and one for CHAR evaluation.

Note, that in the attached files "@" replaces the space character.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *8th International Conference on Weblogs and Social Media.*

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*.

Morris Swadesh. 1946. South Greenlandic (Eskimo). In Cornelius Osgood, editor, *Linguistic Structures of Native America*. Viking Fund Inc. (Johnson Reprint Corp.), New York.