Research Target: Comparative Study of Word Embedding, BERT and LASER for binary Text classification(informative and non-informative)

(This work looks exemplarily at three types of data representations: context-less word embeddings (word2vec), contextualised word embeddings (BERT) and sentence embeddings (LASER).)

Problem for Multilingual: Not all disaster-related information is broadcasted in the local language. since sufficient train data is not available for all languages

Solution for Multilingual: models that are able to apply their learned classification skills from one language with already existing large corpora

will compare state-of-the-art language-agnostic and multilingual word and sentence embeddings on the task of cross-lingual text classification in the case of disaster tweets, which are classified according to their informativeness.

Dataset: Multilingual crisis dataset required. Need uniform label for evaluation

Choose: Binary classification of Non informative and informative( Tweets relevant to crisis but contains no useful information will be regarded as non informative)

Choose tree languages: English Spanish and Italian: reason: whether the amount of training resources or the linguistic closeness of source and target language is more important for successful transfer learning.

Datasets: SOSItalyT4 (T4) ChileEarthquakeT1 (T1)(Spanish)

CrisisLexT26 (T26)(Large dataset with multiple languages)（Problem: Some advertisement make use of the trending keywords, so the context of the whole tweets needs to be understood）(Language Identification: Three out of four agreement.

Problem: Code Switching, in order to test the language-agnostic sentence representations, keep the tweets

Translation:

Amazon translation

maintain the newly added UL, UQ and RT tokens and not accidentally translate them, they are added to custom terminology.

Baseline:

Word2Vec with SVM

BERT: self-supervised learning on unlabeled data. Task1: Mask 15% words and predict them. Task2 : Next sentence prediction

mBERT(Multilingual): drawbacks ( during the fine tuning, parameters changed, so that the same meaning of different language get different representation)

LASER: This means they encode sentences of similar semantic content with similar representations in one shared vector space, regardless of their language.

LASER encodings are fixed, not tuned towards the language of the task-specific labelled data, the embeddings remain truly multilingual which is optimal for a zero-shot application. Therefore, it is even possible to train the classifier on different source languages at the same time.

Experiment:

Monolingual: Only consider Spanish and Italian: cause they are the target for the cross-lingual study Method: Compare SVM using Word2vec to BERT model. Dataset split 10-1, 10 Fold cross validation and 10% of Test data are changed to validation set.

Cross lingual: Study the transfer ability of text classification in different language:

English – Spanish, English – Italian, Spanish- Italian

Method: 1)zero shot: No translation in train and test set 2) train-translation: translate train set and conduct evaluation in the language of target language 3) test-translation: translate the test set and conduct evaluation in source language. Problem: Word2vec can not be performed in zero shot LASER: Do not need translation, can be directly applied.

BootStrap Test: ?

Implementation:

Python3

SVM: scikit learn

BERT:( TODO read implement details)

LASER: Evaluate with feedforward neural network

Result:

Monolingual: F1 score fluctuate between train and test set. only in BERT

SVM is surprisingly good.

TODO: BERT IN THIS WORK AND BERT IN Alam et al. [2020]

Cross lingual: mBERT(TODO) can be improved by translating, except English-Spanish(train translation)

Improved in experiment: training in Spanish, test in Italian

But Italian-Spanish did not improved by translating train data. And translating test data the model performed even worse.

Bis Seite 38