Tweet embedding production
29 Haziran - 16 Temmuz 18

Data

Only tweets in tr.

Sources

1) somemto: some commercial tweets.
- 1M tweets

2) 2013-3months: various tweets from 2013 Jan-March, on various subject.
- 250K tweets

3) 20M tweets: various tweets, date unknown. (Collected by [1])
- 20M tweets.

Preprocessing

1) Duplicate and less-than-3-word tweets removed
2) Noisy characters, symbols removed. Emojis removed.
3) Twitter usernames, URLs and numbers replaced by *@USER*, *URL*, *NUM*, respectively

After preprocessing, the number of tweets per dataset
1) somemto: 214K tweets

2) 2013-3month: 165K tweets
3) 20Mtweets: 3.5M tweets

Vectors

Two sets produced:
1) Set1: vector_size: 50; vocabulary_size: 500K
2) Set2: vector_size: 100; vıcabulary_size: 1M

For the production of both sets we used:

- Model: word2vec [2]
- Algorithm: skip-gram
- Hyperparameters: min_count=3, window=5, alpha=0.025, max_vocab_size=None

We decided on this model and its hyperparameters after having many tests on small datasets and by the guidance we gained from the relevant works.

Evaluation

1) Extrinsic

Tweet sentiment classification

- Features: avg of word vectors in tweets
    - No preprocessing, no stopword removal; only tokenization
- Dataset: 2K tweets from various topics in tr. labels: POS, NEG.
- Classifier: SGD classifier (in scikit-learn)

Results

- Two different experiments, one with vectors of size 50 and the other with those of size 100.
  1) Set1_size-50
      - Performance is fluctuating between 0.75-0.77 for accuracy and 0.71-0.74 for f1-score, throughout more than 10 runs.
  2) Set2_size-100
      - Same as the other set but the values are larger and are less fluctuating.

2) Intrinsic

Analogy test.
Given a set of pairs of words like ((king, man), (queen, woman)), we guess one of the words using the angles between the words.

Test1 - Analogy by derivational or inflectional suffixes
Pairs are collected by [3].

Test2 - Analogy by semantic values
Semantic criteria: akrabalık, zıtanlamlılık, başkentler
Pairs are collected by [4].

Accuracy values are generally very low for these tests and such a method of evaluation is still being discussed [5, 6]; yet no remarkable progress has been done.

Vector set	Accuracy on Test1	Accuracy on Test2
Set1	0.016	0.031
Set2	0.023	0.053
Sabanci_embeddings [3]	0.020	0.28
Boun_embeddings [7]	0.013	0.1

- Test2 appears to be more reliable.
- Our Set2 is more qualified compared to Set1 whose vector size is 50, lower.
- Existing vector sets for turkish, produced from wikipedia articles, news and tweets, [3, 7], also give similar results except a remarkable performance on Test2 (accuracy: 0.28) by [7] has been recorded.

Resources
[1] http://www.kemik.yildiz.edu.tr/data/File/20milyontweet.rar
[2] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.
[3] Sen, Mehmet Umut, and Hakan Erdogan. "Learning word representations for Turkish." Signal Processing and Communications Applications Conference (SIU), 2014 22nd . IEEE, 2014.

[4] Onur Gungor, Eray Yildiz, "Linguistic Features in Turkish Word Representations - Türkçe Sözcük Temsillerinde Dilbilimsel Özellikler", 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017.
[5] Faruqui, Manaal, et al. "Problems with evaluation of word embeddings using word similarity tasks." *arXiv preprint arXiv:1605.02276* (2016).
[6] Drozd, Aleksandr, Anna Gladkova, and Satoshi Matsuoka. "Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen." *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016.
[7] Okur, Eda, Hakan Demir and Arzucan Özgür, "Named Entity Recognition on Twitter for Turkish using Semi-supervised Learning with Word Embeddings", LREC. 2016.