Dear Ansgar and Aline,

Sorry for the lack of communication from my side, but I've been working non-stop on the project and I believe that the results I've gathered prove to be satisfactory to a degree, but I need to corroborate with you.

This is the outline and results of the 3 major tests that I've performed in the past weeks.

**Data Pre-processing**:

* Used VNCs: VNC Examples found in the VNC-Tokens Dataset that. Examples were removed if they presented any of the following characteristics:
  + Target sentence in the BNC XML dataset was not present.
  + Window Size in which the VNC appeared in the target sentence exceeded 7 tokens.
    - E.g. See Stars -> Bob **saw** the best , most beautiful , and amazing **stars** .
  + Conjugation of the Verb or Noun present in the VNC was problematic for lemmatization:
    - E.g. See Stars -> **SAW** (which is past form) is problematic since Lemmatizer considers it as the ‘saw’ object.
* Set tokens to lower-case for training and generation of word and sentence embeddings.

Word Embedding’s Description:

* Word2Vec:
  + Vector Size: 300
  + Trained using the available BNC XML Corpus
  + Notes: Sentence Embeddings are obtained by averaging the token embeddings in the target sentence.
* Skip-Thoughts:
  + Vector Size: 4800 (2400 Uni-Skip Model + 2400 Bi-Skip Model)
  + Using pre-trained model available, which was trained using BookCorpus corpora.
* Siamese CBOW:
  + Vector Size: 300
  + Using pre-trained model available, trained with Wikipedia (INEX) Corpus [Snapshot of Wikipedia in 2012].
* ELMo:
  + Vector Size: 1024
  + Using pre-trained model available in Tensor-Hub, trained with “1 Billion Word Language Model Benchmark”.

**Test 1 [Supervised Approach] – Linear SVM**:

* Target: Reproduce results by King and Cook (2018) + Use ELMo Embeddings.
* Model Used:
  + SVM
    - *C* parameter: [0.01, 0.1, 1, 10, 100]
    - Kernel: Linear
* Performed 10-Fold Cross-Validation to find best *C*.
* Data Used:
  + Sentence Embeddings of the Word2Vec, Siamese CBOW, Skip-Thoughts and ELMo Embeddings.
  + Appended CForm as performed by King and Cook (2018)
* Results:
  + Word2Vec:
    - Average Test Accuracy: 88.61%
    - Average Test F-1: 92.85%
  + Word2Vec + CForm:
    - Average Test Accuracy: 89.48%
    - Average Test F-1: 93.36%
  + Skip-Thoughts:
    - Average Test Accuracy: 84.81%
    - Average Test F-1: 91.02%
  + Skip-Thoughts + CForm:
    - Average Test Accuracy: 85.58%
    - Average Test F-1: 91.43%
  + Siamese CBOW:
    - Average Test Accuracy: 78.76%
    - Average Test F-1: 88.10%
  + Siamese CBOW + CForm:
    - Average Test Accuracy: 78.23%
    - Average Test F-1: 87.66%
  + ELMo:
    - Average Test Accuracy: 87.31%
    - Average Test F-1: 92.15%
  + ELMo + CForm:
    - Average Test Accuracy: 88.06%
    - Average Test F-1: 92.57%

**Test 2 [Unsupervised Approach] – Cosine Similarity:**

* Target: Use Cosine Similarity metric with a threshold under the assumption that values closer to 0 will have more idiomatic use. The Cosine Similarity is
* Settings Used:
  + Cosine Similarity Threshold: **0.6**
* Results:
  + Word2Vec:
    - Accuracy:
    - F-1:
  + Skip-Thoughts:
    - Accuracy:
    - F-1:
  + Siamese CBOW:
    - Accuracy:
    - F-1:
  + ELMo:
    - Accuracy:
    - F-1:

**Test 2 [Unsupervised Approach] – Cosine Similarity + Overall Fixedness:**

* Target: Use a weighted average of Cosine Similarity and Overall Fixedness (Fazly et. Al 2009) of target VNC.   
    
  Cosine Similarity is obtained as follows:  
    
  Calculated between the embedded forms of the Target VNC and the Sentence it appears in:
  + E.g.: Lose Head -> They said he 'd **lost his head** and started shooting —

The weighted average would be calculated as follows:

* Settings Used:
  + Beta: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
  + Threshold: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
* Data Used:
  + Sentence Embeddings of the Word2Vec, Siamese CBOW, Skip-Thoughts and ELMo Embeddings for target VNC as used and the sentence which contains it.
  + Overall Fixedness of the target VNC.
* Results: (Best results were obtained for and )
  + Word2Vec ( and ):
    - Accuracy: 70.65%
    - F-1: 80.20%
  + Skip-Thoughts ( and ):
    - Accuracy: 73.99%
    - F-1: 84.87%
  + Siamese CBOW ( and ):
    - Accuracy: 77.58%
    - F-1: 87.35%
  + ELMo ( and ):
    - Accuracy: 65.39%
    - F-1: 75.57%

Please feel free to make any comments regarding my approaches. All the code is available in my public Github repo: <https://github.com/itsStrobe/IdiomDetection>

Best regards,

Jose Zavala