NormAFE: Unsupervised Creation of Normalisation Dictionaries for Micro-Blogs in Arabic, French and English

Amal Htait, Sébastien Fournier and Patrice Bellot

CICLing 2018 – Hanoi, Vietnam

Goal : Creating text normalisation dictionaries using distributed representations of words, known also as "word embedding", applied on Arabic, French and English Languages.

Step 1 : Create Models

- **Arabic :**
  238M tweets → Model of 9M word
- **French :**
  48M tweets → Model of 683k word
- **English :**
  1B tweets → Model of 5M word

Step 2 : Extract Similar words

The measure of cosine distance is used between the vector of standard-form word and the vectors of every other word in the word embedding model.

The antonym of the standard-form word is used to refine the results. For example, the antonyms exclusion eliminates the possibility of extracting the word *inactive as a* similar word to the word *active*.

Step 3 : Filter

The method is to find the longest contiguous matching subsequence that contains no different elements.

The purpose is to give matches that "look right" to people.

- **Evaluation of Dictionaries' Content :**

  For Arabic langage, an average of :
  89.5% in Normalisation success,
  83.7% in Correction success.

  For French langage, an average of :
  85% in Normalisation success,
  73.6% in Correction success.

  For English langage, an average of :
  96% in Normalisation success,
  86% in Correction success.

Dictionaries Annotation Example

Misspelled	Standard-word	Correction	Normalisation
aiiiiiime (looooooove)	aime (love)	✓	✓
decevra (will disappoint)	decevoir (disappoint)	✗	✓
deballer (unpack)	deprimer (depress)	✗	✗
ممتاذ (misspelled excellent)	ممتاز (excellent)	✓	✓
اكرهه (hate him)	اكره (hate)	✗	✓
اهبل (dump)	غبي (stupid)	✗	✗

- **Evaluation of Dictionaries with Sentiment Analysis Tool :** *

  Echo*, an open source software for sentiment analysis based on supervised machine learning algorithm, is used as a test tool. Text Normalisation is applied on both training and testing datasets, of tweets in English language.

  Echo f-mesure results without Normalisation : 55.64
  Echo f-mesure results with Normalisation : **56.22**

Future Work :

This work can be a resource for many domains in Natural Language Processing, like Sentiment Analysis and Arabic Dialects Normalisation.

Standard Word	Dialect Source	Dialect Word
غبي (Stupid)	Egypt Arabic	عبيط
ضايقك (bothered you)	Gulf Arabic	ضايقج

*** This work is available as open source on : **https://github.com/amalhtait/NormAFE & https://github.com/OpenEdition/NormAFE**
http://amalhtait.com/tools.html