IHLT-MAI
Final Project

Jordi Armengol - Joan Llop

Universitat Politècnica de Catalunya
Master in Artificial Intelligence

Preliminary experiments

- General class/interface: The Model class.
- Cross validation
- Investigate preprocessing, features,...

Jaccard distance of some basic features

- Preprocessing: Stop-words filter, alphanumeric.
- Word tokenization.
- Lemmas.
- Part-of-Speech.
- Most common synsets.
- Cross validation: 0.5744

Using words

- Preprocessing: lowercase, alphanumeric characters and stop-words filtering.
- Word tokenize
- Counting frequencies (local approach)
- Cosine similarity
- Cross validation: 0.5968

Using of lemmas

- Preprocessing: lowercase, alphanumeric characters and stop-words filtering.
- Lemmatizer (Wordnet)
- Counting frequencies (local approach)
- Cosine similarity
- Cross validation: 0.60764

Using stems

- Preprocessing: lowercase, alphanumeric characters and stop-words filtering.
- Stems (PorterStemmer)
- Counting frequencies (local approach)
- Cosine similarity
- Cross validation: 0.6488

Bigrams vector representation

- Preprocessing: $[0-9] \cup [a-z] \cup [' ']$
- Bigrams vector
- Cosine similarity
- Cross validation: 0.6637

Synsets

- Preprocessing: lowercase, alphanumeric characters and stop-words filtering.
- Union of Synsets
- Cosine similarity
- Cross validation: 0.4602

Noun phrases and verbs comparison

- Preprocessing: lowercase, alphanumeric characters and stop-words filtering.
- Noun phrases (spacy)
- Verbs (spacy)
- Named entities (spacy)
- Counting substrings
- Cross validation: 0.4953

Preprocessing

- Stems
- Chars $\rightarrow$ Bigrams, Trigrams
- We have also tried: pos tags, named entities, lemmas and synsets

Feature extraction

- Bag of stems (binary and integers)
- tf–idf (weighting factor)
- Bigrams
- Trigrams
- Edit distance (chars)
- Set kernel distance
- Cosine similarity
- Jaccard distance
- Overlap distance
- Scaling.

Model selection

Hyperparameter optimization + cross-validation.

- Linear, quadratic... SVMs.
- MLP.
- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **RBF Kernel SVM**
- We have selected the best 3 with cross-validation and taken their mean (ensemble).
- **Once selected, Test: 0.7527**.

Conclusions

- Test: 0.7527 $\rightarrow$ 11th
- Without embeddings/transfer learning.

Addendum: Transfer learning

- On a parallel front, we investigated the use of transfer learning approaches (SOTA).
- A linear regression with only one feature, cosine similarity of pre-trained FastText word embeddings, obtains relatively similar results and avoids feature engineering.
- Hybrid model?
- Contextual embeddings: BERT and derived models usually output token-level embeddings:
  - Aggregate them with some sort of pooling.
  - Add new layer(s) and fine-tune.
  - Best alternative: **Sentence-BERT**.