Lexical Text Simplification

Paula Gombar, Ivan Katanić

FER

June 13, 2016

Contents

SemEval-2012 Task 1

- finding less complex semantically-appropriate words or phrases and replacing those that are difficult to comprehend
- common pipeline in such a system:
  1. **complexity analysis:** finding out which words or phrases are considered complex
  2. **substitute lookup:** retrieving adequate replacements, simpler than the original word or phrase
  3. **context-based ranking:** ranking of substitutes to produce the final replacement

Example

Sentence: *The incisions will feel constricted for the first 24-48 hours.*
We identify the word *constricted* as complex, retrieve the possible substitutes:
{uncomfortable} {tight} {stretched} {compressed} {constricted}.
We score each candidate on simplicity and context-adequacy, rank them and determine the simplest one, e.g. *tight*.

Features extracted

- **inverse word length**
- **number of synsets in WordNet**
  - i.e. the word *fundamental* has the following synsets:
    1. (n) fundamental (any factor that could be considered important to the understanding of a particular business)
    2. (n) fundamental, fundamental frequency, first harmonic (the lowest tone of a harmonic series)
- **frequency in Simple Wikipedia**
- **frequency in Wikipedia**
- **corpus complexity**

$$C_w = \frac{f_{w,English}}{f_{w,Simple}}$$

where $f_{w,c}$ is the frequency of candidate w in corpus c.

Features extracted

- **context similarity**

$$csim(w, c) = \sum_{w' \in C(w)} cos(\mathbf{v_c}, \mathbf{v_{w'}})$$

where $C(w)$ is the set of context words of the original word w and $\mathbf{v_c}$ is the GloVe vector of the replacement candidate c.

- **semantic similarity**

$$ssim(w, c) = cos(\mathbf{v_w}, \mathbf{v_c})$$

where $\mathbf{v_w}$ is the GloVe vector of the original word w.

Methods used

Ranking SVM with RBF kernel.

Table : Optimal hyperparameters for Ranking SVM with RBF kernel.

Hyperparameter	Optimal value	Possible values
Scaler	Standard	Standard, MinMax, None
PolyFeatures degree	1	1, 2
C	2^5	$[2^{-15}, ..., 2^8]$
γ	0.00098	$[2^{-15}, ..., 2^8]$

Ranking SVM with linear kernel.

Table : Optimal hyperparameters for Ranking SVM with linear kernel.

Hyperparameter	Optimal value	Possible values
Scaler	Standard	Standard, MinMax, None
PolyFeatures degree	1	1, 2
C	2^4	$[2^{-15}, ..., 2^8]$

Methods used

Linear combination of features.

Table : Optimal hyperparameters for linear combination of features.

Feature	Weight
Inverse word length	1
WordNet synsets	0
Simple Wikipedia frequency	10
English Wikipedia frequency	0
Corpus complexity	0
Context similarity	9
Semantic similarity	0

Unsupervised approach. Scale the data using MinMax scaler, declare all coefficients as 1.

Baselines

L-Sub Gold. This baseline uses the gold-standard annotations from the Lexical Substitution corpus of SemEval-2007 as is.
Random Randomizes the order, allowing ties.
Simple Freq. Uses the frequency of the substitutes as extracted from the Google Web 1T Corpus.

Table : Baseline kappa scores on Trial and Test datasets.

Baseline	Trial	Test
L-Sub Gold	0.050	0.106
Random	0.016	0.012
Simple Freq.	**0.397**	**0.471**

Results

Table : Implemented methods kappa scores on the Test dataset.

Method name	Test score
Ranking SVM with RBF kernel	**0.461**
Ranking SVM with linear kernel	0.443
Linear combination of features	0.459
Unsupervised approach	0.313

Conclusion

- four different methods, using both context-dependent and context-independent features, as well as external resources such as state-of-the-art word vector representations and simplified corpora.
- the performance of supervised approaches is likely to improve with larger training sets
- very strong relation between distributional frequency of words and their perceived simplicity

The end

Thank you! Any questions?