A Semantic Similarity Measure Based on Lexico-Syntactic Patterns

Alexander Panchenko, Olga Morozova and Hubert Naets
Centre for Natural Language Processing (CENTAL) – Université catholique de Louvain – Belgium
E-mail: {Firstname.Lastname}@uclouvain.be

Introduction

We present a novel semantic similarity measure based on lexico-syntactic patterns such as those proposed by Hearst (1992). The measure achieves a correlation with human judgements up to 0.739. Additionally, we evaluate it on the tasks of semantic relation ranking and extraction. Our results show that the measure provides results comparable to the baselines without the need for any fine-grained semantic resource such as WordNet.

Lexico-Syntactic Patterns

- 18 patterns which aim at extracting hypernymic and synonymous relations:
  - Such NP as NP, NP[,] and/or NP:
  - NP, for example, NP, NP[,] and/or NP:
  - NP, especially NP, NP[,] and/or NP:
  - NP, also known as NP:
  - NP, also called NP:
  - NP alias NP:
  - NP a.k.a NP:

- Patterns are encoded in the form of a cascade of FST with the Unitex tool:

- Patterns are applied to corpora:

  - WacCorda + wiCorda: 2.694.815, 2.607.109, 3.368.147, 5.88 Gb

Patterns exact concordances:
- (traditional[food]), such as {[sandwich]}, {[burger]}, and {[fly]}
- such (non-alcoholic [sodas]) as {[root beer]}, {[cream soda]}
- such NP as NP, NP[,] and/or NP:
- NP, such as NP, NP[,] and/or NP:
- NP, NP[,] or other NP:
- NP, NP[,] and other NP:
- NP, including NP, NP[,] and/or NP:
- NP, in other words:
- NP, also known as NP:
- NP, also called NP:
- NP alias NP:
- NP a.k.a NP:

Evaluation and Results

Evaluation and Results

- 18 patterns which aim at extracting hypernymic and synonymous relations:
  - Such NP as NP, NP[,] and/or NP:
  - NP, for example, NP, NP[,] and/or NP:
  - NP, especially NP, NP[,] and/or NP:
  - NP, also known as NP:
  - NP, also called NP:
  - NP alias NP:
  - NP a.k.a NP:

Patterns are encoded in the form of a cascade of FST with the Unitex tool:

Patterns are applied to corpora:
- WacCorda + wiCorda: 2.694.815, 2.607.109, 3.368.147, 5.88 Gb

Patterns exact concordances:
- (traditional[food]), such as {[sandwich]}, {[burger]}, and {[fly]}
- such (non-alcoholic [sodas]) as {[root beer]}, {[cream soda]}
- such NP as NP, NP[,] and/or NP:
- NP, such as NP, NP[,] and/or NP:
- NP, NP[,] or other NP:
- NP, NP[,] and other NP:
- NP, including NP, NP[,] and/or NP:
- NP, in other words:
- NP, also known as NP:
- NP, also called NP:
- NP alias NP:
- NP a.k.a NP:

Semantic Similarity Measures

Algorithm 1: Similarity measure PatternSim.

Input: Terms C, Corpus D
Output: Similarity matrix, S [C x C]
1. $\text{K} \leftarrow \text{extract}\text{.}\text{concord}(D)$
2. $\text{K}_{\text{lem}} \leftarrow \text{lemmatize}\text{.}\text{concord}(K)$
3. $\text{K} \leftarrow \text{filter}\text{.}\text{concord}(K_{\text{lem}}, C)$
4. $S \leftarrow \text{get}\text{.}\text{extraction}\text{.}\text{freq}(C, K)$
5. $S \leftarrow \text{rename}(S, C, D)$
6. $S \leftarrow \text{normalize}(S)$
7. return $S$

We present a similarity measure based on manually-crafted lexico-syntactic patterns.

- The measure was evaluated on the five ground-truth datasets and the semantic relation extraction task.

The measure provides results comparable to the baseline WordNet-, dictionary-, and corpus-based measures and does not require semantic resources.

Future work -- using a supervised model to:
- combine different factors;
- tune the meta-parameters.

Conclusion

We presented a similarity measure based on manually-crafted lexico-syntactic patterns.

- The measure was evaluated on the five ground-truth datasets and the semantic relation extraction task.

- The measure provides results comparable to the baseline WordNet-, dictionary-, and corpus-based measures and does not require semantic resources.

Data: http://cental.fltr.ucl.ac.be/team/~panchenko/sim-eval/
Code: http://github.com/cental/patternsim/
Demo: http://serelex.cental.be/