A Semantic Similarity Measure Based on Lexico-Syntactic Patterns

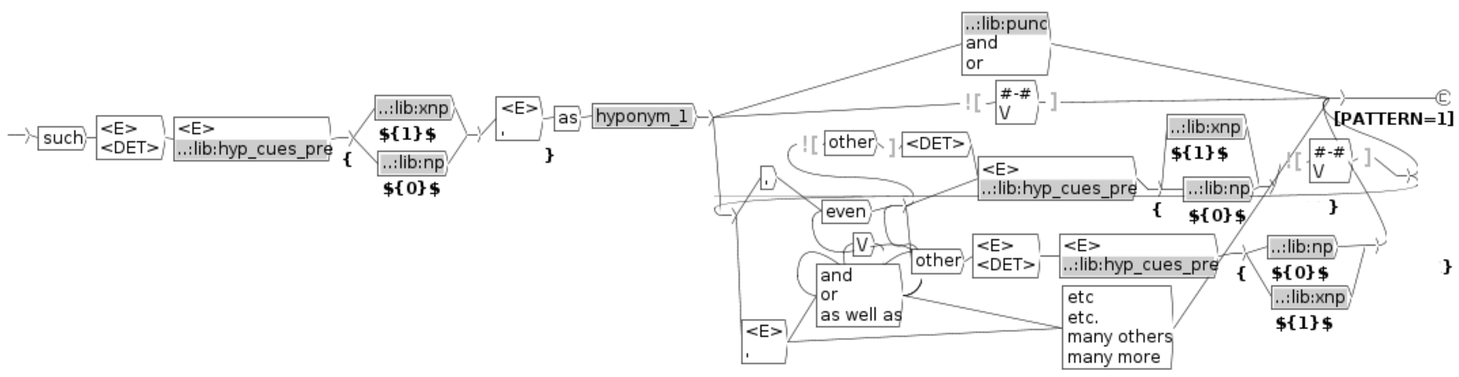
Alexander Panchenko, Olga Morozova and Hubert Naets
Centre for Natural Language Processing (CENTAL) – Université catholique de Louvain – Belgium
E-mail: {Firstname.Lastname}@uclouvain.be

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 synonymic relations:
  - *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, especially NP, NP [,] and/or NP;*
  - *NP: NP, [NP,] and/or NP;*
  - *NP is DET ADJ.Superl NP;*
  - *NP, e. g., NP, NP[,] and/or NP;*
  - *NP, for example, NP, NP[,] and/or NP;*
  - *NP, i. e.[,] NP;*
  - *NP (or NP);*
  - *NP means the same as NP;*
  - *NP, in other words[,] NP;*
  - *NP, also known as NP;*
  - *NP, also called NP;*
  - *NP alias NP;*
  - *NP aka NP.*
- Patterns are encoded in the form of a cascade of FST with the Unitex tool:



- Patterns are applied to corpora:

Name	# Documents	# Tokens	# Lemmas	Size
WaCypedia	2.694.815	$2.026 \cdot 10^9$	3.368.147	5.88 Gb
ukWaC	2.694.643	$0.889 \cdot 10^9$	5.469.313	11.76 Gb
WaCypedia + ukWaC	5.387.431	$2.915 \cdot 10^9$	7.585.989	17.64 Gb

- Patterns extact concordances:
  - *{traditional[food]}, such as {[sandwich]},{[burger]}, and {[fry]}[PATTERN=2]*
  - *such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]} [PATTERN=1]*

Semantic Similarity Measures

Algorithm 1: Similarity measure *PatternSim*.

Input: Terms C, Corpus D
Output: Similarity matrix, $\mathbf{S}$ $[C \times C]$

$K \leftarrow extract\_concord(D)$;
$K_{lem} \leftarrow lemmatize\_concord(K)$;
$K_C \leftarrow filter\_concord(K_{lem}, C)$;
$\mathbf{S} \leftarrow get\_extraction\_freq(C, K)$;
$\mathbf{S} \leftarrow rerank(\mathbf{S}, C, D)$;
$\mathbf{S} \leftarrow normalize(\mathbf{S})$;
return $\mathbf{S}$;

- **Efreq**

$$s_{ij} = e_{ij}$$

s_{ij} – semantic similarity between c_i and c_j
e_{ij} – frequency of extractions between the terms $c_i, c_j \in C$

- **Efreq-Rfreq**

$$s_{ij} = \frac{2 \cdot \alpha \cdot e_{ij}}{e_{i*} + e_{*j}},$$

$e_{i*} = \sum_{j=1}^{|C|} e_{ij}$ – a number of concordances containing word c_i
α – an expected number of semantically related words per term

- **Efreq-Rnum**

$$s_{ij} = \frac{2 \cdot \mu_b \cdot e_{ij}}{b_{i*} + b_{*j}},$$

$b_{i*} = \sum_{j:e_{ij} \geq \beta} 1$ – number of extractions with a frequency $\geq \beta$
$\mu_b = \frac{1}{|C|} \sum_{i=1}^{|C|} b_{i*}$ – an average number of related words per term

- **Efreq-Cfreq**

$$s_{ij} = \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$$

$P(c_i, c_j) = \frac{e_{ij}}{\sum_{ij} e_{ij}}$ – extraction probability of the pair $\langle c_i, c_j \rangle$
$P(c_i) = \frac{f_i}{\sum_i f_i}$ – probability of the word c_i
f_i – frequency of c_i in the corpus

- **Efreq-Rnum-Cfreq** $s_{ij} = \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}.$

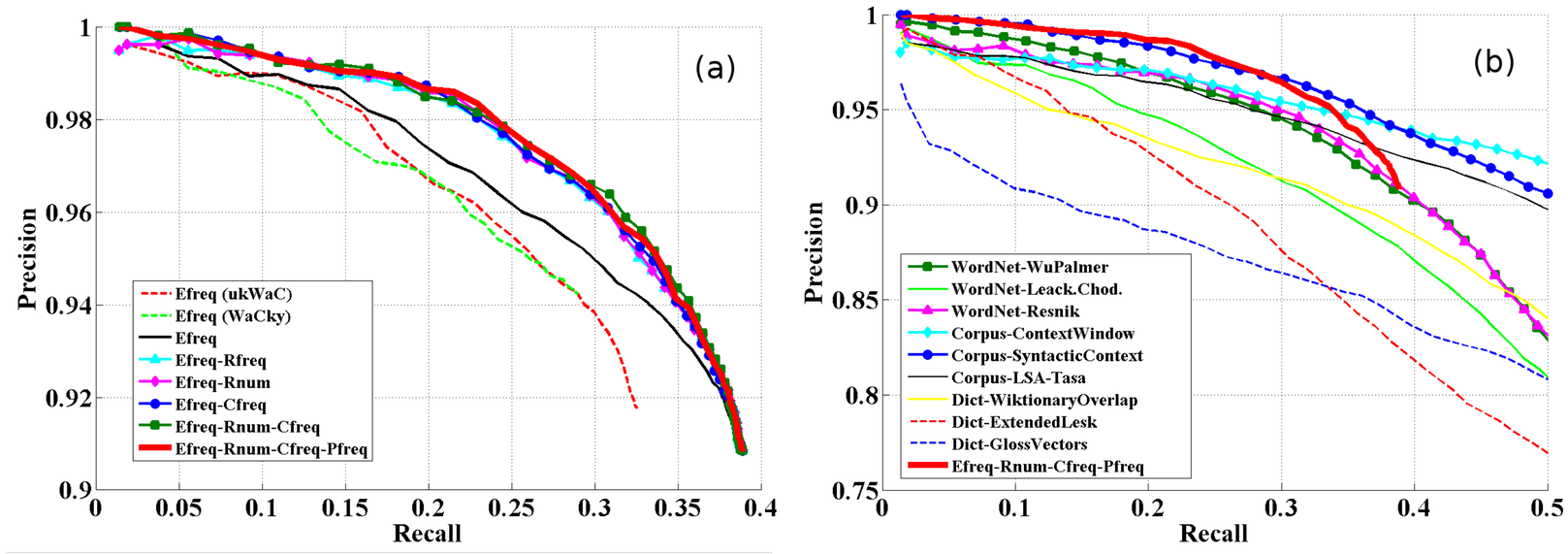
- **Efreq-Rnum-Cfreq-Pnum**

$$s_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}.$$

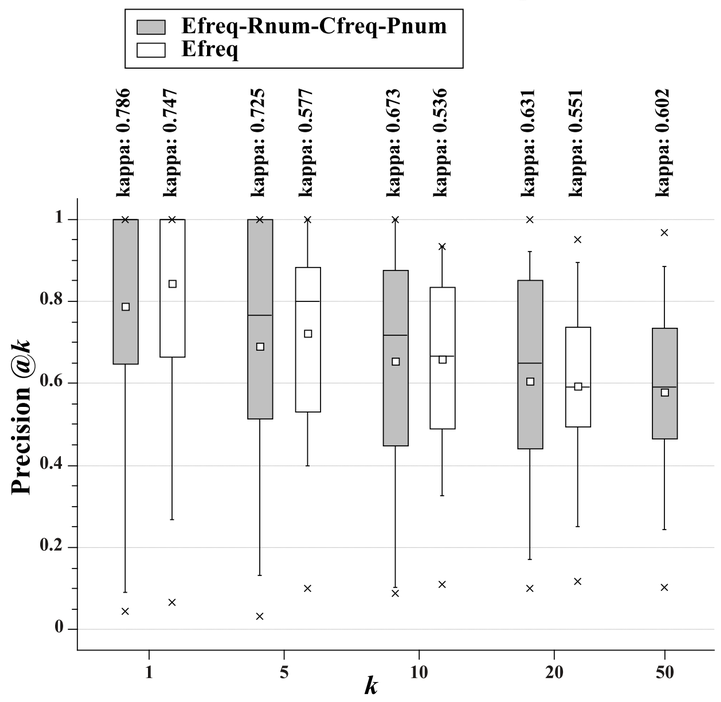
$p_{ij} = \overline{1, 18}$ – number of patterns extracted given pair of terms $\langle c_i, c_j \rangle$

Evaluation and Results

Precision-Recall graphs -- BLESS dataset



Semantic relation extraction: precision at *k*



Performance on human judgement (MC, RG, WS) and semantic relation (BLESS and SN) datasets

Similarity Measure	MC	RG	WS	BLESS (hypo,cohypo,mero,attri,event)				SN (syn, hypo, cohypo)				BLESS (hypo, cohypo)			
	ρ	ρ	ρ	*P(10)*	*P (20)*	*P(50)*	*R(50)*	*P(10)*	*P(20)*	*P(50)*	*R(50)*	*P(10)*	*P(20)*	*P(50)*	*R(50)*
Random	*0.056*	*-0.047*	*-0.122*	0.546	0.542	0.544	0.522	0.504	0.502	0.499	0.498	0.271	0.279	0.286	0.502
WordNet-WuPalmer	0.742	0.775	0.331	0.974	0.929	0.702	0.674	0.982	0.959	0.766	**0.763**	0.977	0.932	0.547	**0.968**
WordNet-Leack.Chod.	0.724	**0.789**	0.295	0.953	0.901	0.702	0.648	**0.984**	0.953	0.757	0.755	0.951	0.897	0.542	0.957
WordNet-Resnik	0.784	0.757	0.331	0.970	0.933	0.700	0.647	0.948	0.908	0.724	0.722	**0.968**	0.938	0.542	0.956
Corpus-ContextWindow	0.693	0.782	0.466	0.971	0.947	0.836	**0.772**	0.974	0.932	0.742	0.740	0.908	0.828	0.502	0.886
Corpus-SynContext	**0.790**	0.786	0.491	0.985	0.953	0.811	0.749	0.978	0.945	0.751	0.743	0.979	0.921	0.536	0.947
Corpus-LSA-Tasa	0.694	0.605	0.566	0.968	0.937	0.802	0.740	0.903	0.846	0.641	0.609	0.877	0.775	0.467	0.824
Dict-WiktionaryOverlap	0.759	0.754	0.521	0.943	0.905	0.750	0.679	0.922	0.887	0.725	0.656	0.837	0.769	0.518	0.739
Dict-GlossVectors	0.653	0.738	0.322	0.894	0.860	0.742	0.686	0.932	0.899	0.722	0.709	0.777	0.702	0.449	0.793
Dict-ExtenedLesk	0.792	0.718	0.409	0.937	0.866	0.711	0.657	0.952	0.873	0.655	0.654	0.873	0.751	0.464	0.820
WikiRelate-Gloss	0.460	0.460	0.200	–	–	–	–	–	–	–	–	–	–	–	–
WikiRelate-Leack.Chod.	0.410	0.500	0.480	–	–	–	–	–	–	–	–	–	–	–	–
WikiRelate-SVM	–	–	**0.590**	–	–	–	–	–	–	–	–	–	–	–	–
Efreq (WaCky)	0.522	0.574	0.405	0.971	0.950	0.942	0.289	0.930	0.912	0.897	0.306	0.976	0.937	0.923	0.626
Efreq (ukWaC)	0.384	0.562	0.411	0.974	0.944	0.918	0.325	0.922	0.905	0.869	0.329	0.971	0.926	0.884	0.653
Efreq	0.486	0.632	0.429	0.980	0.945	0.909	**0.389**	0.938	0.915	0.866	**0.400**	0.976	0.929	0.865	**0.739**
Efreq-Rfreq	**0.666**	**0.739**	0.508	0.987	0.955	0.909	**0.389**	0.951	0.922	0.867	**0.400**	0.983	0.940	0.865	**0.739**
Efreq-Rnum	0.647	0.720	0.499	0.989	0.955	0.909	**0.389**	0.951	0.922	0.867	**0.400**	0.983	0.940	0.865	**0.739**
Efreq-Cfreq	0.600	0.709	0.493	0.989	0.956	0.909	**0.389**	0.949	0.920	0.867	**0.400**	0.986	0.948	0.865	**0.739**
Efreq-Cfreq (concord.)	**0.666**	**0.739**	0.508	0.986	0.954	0.909	**0.389**	0.952	0.921	0.867	**0.400**	0.984	0.944	0.865	**0.739**
Efreq-Rnum-Cfreq	0.647	0.737	0.513	0.988	0.959	0.909	**0.389**	**0.953**	0.924	0.867	**0.400**	**0.987**	0.947	0.865	**0.739**
Efreq-Rnum-Cfreq-Pnum	0.647	0.737	**0.520**	**0.989**	0.957	0.909	**0.389**	0.952	0.924	0.867	**0.400**	0.985	0.947	0.865	**0.739**

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

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

Wallonie-Bruxelles International