TAXI at SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-a Patterns, Substrings and Focused Crawling

Alexander Panchenko[1], Stefano Faralli[2], Eugen Ruppert[1], Steffen Remus[1], Hubert Naets[3], Cedrick Fairon[3], Simone Paolo Ponzetto[2] and Chris Biemann[1]

[1] TU Darmstadt, Germany [2]University of Mannheim, Germany [3]UCLouvain, Belgium

Introduction

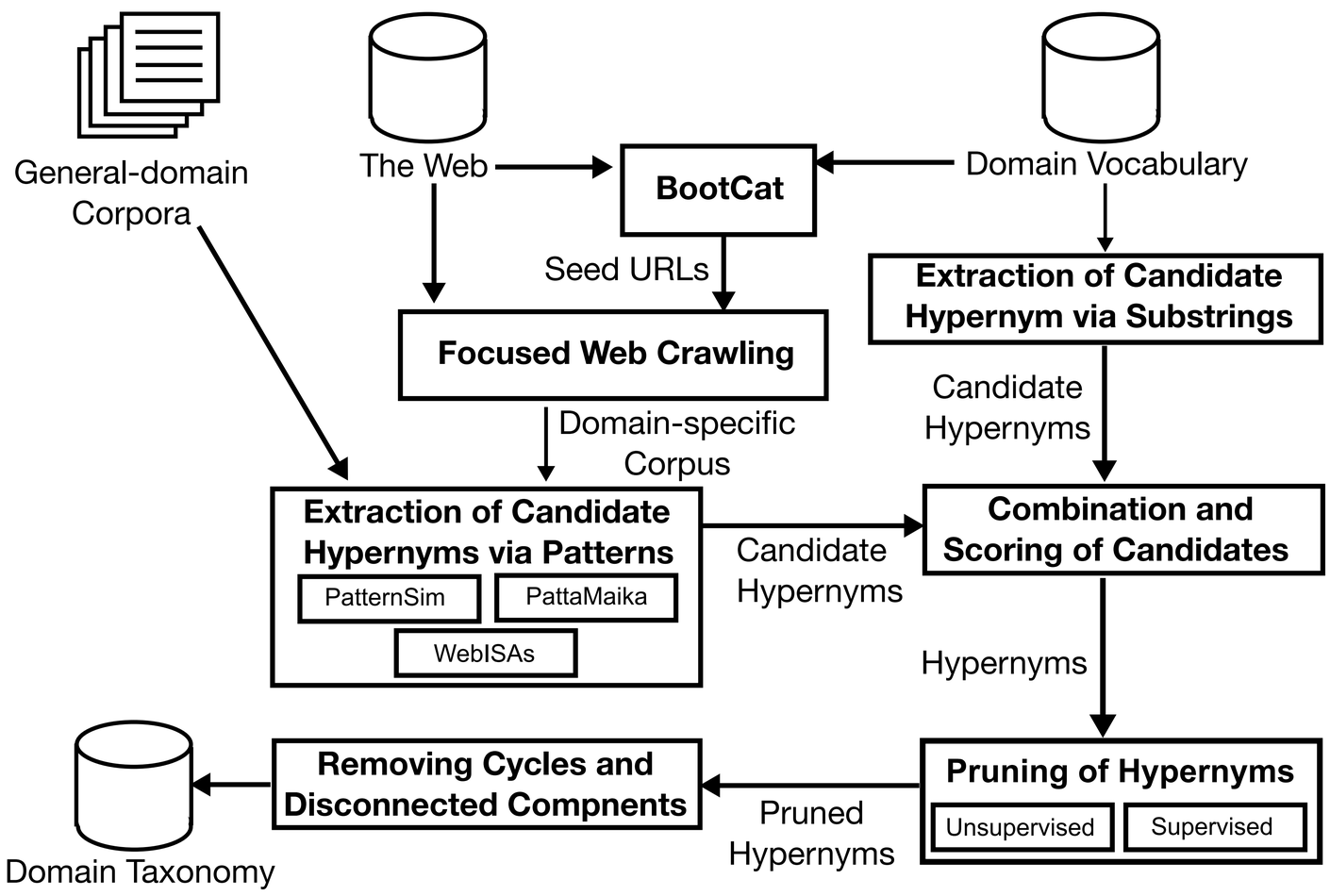
Task:

- Given a domain vocabulary construct a taxonomy
- 24 domain-specific vocabularies
- **Languages:** English, French, Dutch, Italian
- **Domains**: Science, Food, Environment
- **Golden Standard**: WordNet, EuroVoc
- 150 – 1500 terms per language-domain pair

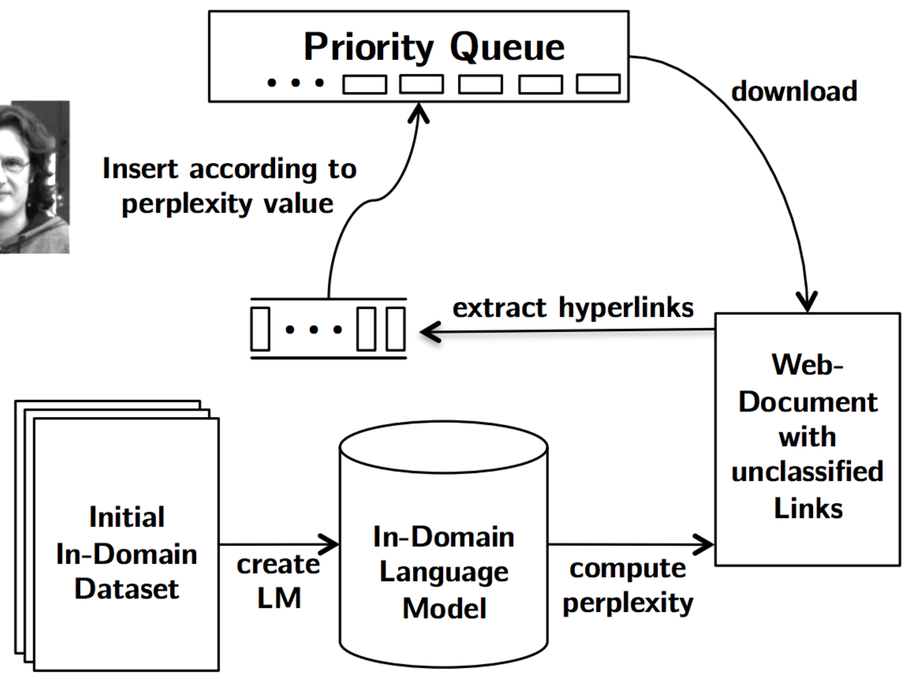
Result:

- Our system obtained the **first place** in this task.

Taxonomy Induction Method



Focused Crawling



Candidate Hypernyms via Patterns

	EN	FR	NL	IT
Wikipedia	11.0	3.2	1.4	3.0
59G	59.2	–	–	–
CommonCrawl	168000.0 ‡	–	–	–
FocusedCrawl Food	22.8	7.9	3.4	3.6
FocusedCrawl Environment	23.9	8.9	2.0	7.1
FocusedCrawl Science	8.8	5.4	6.6	5.1

Corpora sizes used in our system in GB.

	EN ‡, †,§	FR‡	NL†	IT†
General	27.6‡, 4.9†, 118.9§	3.2	2.22	0.13
Food	24.1‡	3.8	0.47	0.05
Environment	26.3‡	4.5	0.32	0.95
Science	9.3‡	2.7	0.97	0.05

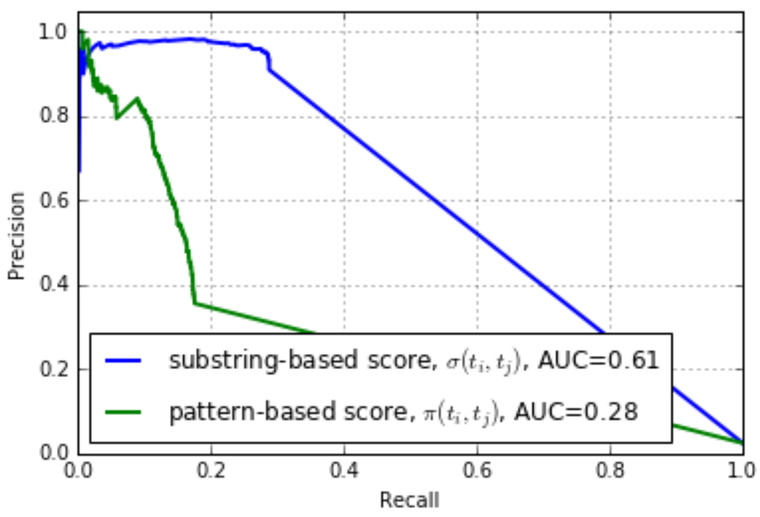
Number of hypernyms in millions of relations.
Extaction systems are denoted with ‡ for PatternSim, † for PattaMaika and § for WebISA.

Candidate Hypernyms via Substrings

Substring-based hypernymy score:

$$\sigma(t_i, t_j) = \begin{cases} \frac{length(t_j)}{length(t_i)} & \text{if } m(t_i, t_j) \wedge \neg m(t_j, t_i) \\ 0 & \text{otherwise} \end{cases}$$

- *m(t_i, t_j)* equals true if *t_i* is in *t_j* and
  - *length(t_i)* > 3
  - if EN or NL: *t_i* should match in the end of *t_j*, e.g. "natural science"
  - if (IT or FR) or ((EN or NL) and a prep. in *t_i*): *t_i* should match in the beginning of *t_j*, e.g. "algebre lineaire", "toast with bacon" or "brood van gekiemdgraan"



Performance of the substring-, and pattern-based features on the trial dataset.

Results: Gold Standard and Manual Evaluation

	Monolingual (EN)			Multilingual (NL, FR, IT)		
Measure	**Baseline**	**BestComp**	**TAXI**	**Baseline**	**BestComp**	**TAXI**
Cyclicity	*0*	**0**	**0**	*0*	**0**	**0**
Structure (F&M)	*0.005*	**0.406**	0.291	*0.009*	0.016	**0.189**
Categorisation (i.i.)	*77.67*	**377.00**	104.50	*64.28*	**178.22**	64.94
Connectivity (c.c.)	*36.83*	44.75	**1 .00**	*40.50*	34.89	**1.00**
Gold standard comparison (Fscore)	*0.330*	0.260	**0.320**	*0.009*	0.016	**0.189**
Manual Evaluation (Precision)	*n.a.*	**0.490**	0.200	*n.a.*	0.298	**0.625**

Overall scores obtained by averaging the results over domains (Environment, Science, Food) and languages (NL, FR, IT). The BestComp lists the respective best scores across all competitors.

References

- Seitner J., Bizer C., Eckert K., Faralli S., Meusel R.,Paulheim H., Ponzetto S.P. (2016): **A Large DataBase of Hypernymy Relations Extracted from the Web**. LREC 2016
- Remus, S. and Biemann, C. (2016): **Domain-Specific Corpus Expansion with Focused Webcrawling.** LREC 2016

TECHNISCHE UNIVERSITÄT DARMSTADT

Language Technology

UNIVERSITY OF MANNHEIM

UCL Université catholique de Louvain

DFG

Code & Data: http://tudarmstadt-lt.github.io/taxi