Mining Scientific Terms and their Definitions : A Study of the ACL Anthology

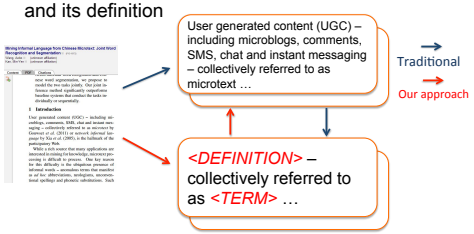
Yiping Jin, Min-Yen Kan, Jun-Ping Ng and Xiangnan He

Web IR / NLP Group
Interactive and Digital Media Institute
National University of Singapore
{yiping, kanmy}@comp.nus.edu.sg

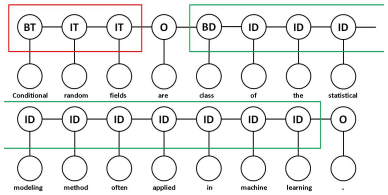
Dataset available

Introduction

- We propose to model definition extraction problem using Conditional Random Fields
- Previous works focus on glossary sentence identification (at the sentence level). We tackle the problem of obtaining the exact bounds of the term and its definition

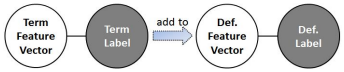


Problem Formalization



- Assign each input word w_i an annotation $a_i \in$ {(T)*erm*, (D)*efinition*, (O)*ther*}
- Recover *definitional sentences* that contains both a term and its definition

Best Solution Explored: 2-Step Serial Word-Level CRF Model



- Base classifier exploits lexical, orthography, dictionary and corpus features
- Augment with dependency parsing and shallow parsing features
- Utilize results from *term* classification and incorporate into *definition* classification

Main Experiment Results

System / Feature Class	Term			Definition			Overall*	
	P	R	F_1	P	R	F_1	F_{micro}	F_{macro}
1: Baseline (Lexical + Orthography + Dictionary + Corpus)	0.50	0.35	0.41	0.40	0.52	0.45	0.45	0.44
2: (1) + shallow parsing	0.50	0.40	**0.45**	0.42	0.42	0.47	0.47	0.47
3: (2) + dependency parsing	0.50	**0.41**	**0.45**	0.45	0.54	0.49	0.49	0.48
4: (3) + 2-stage [DefMiner]	0.50	**0.41**	**0.45**	**0.55**	**0.58**	**0.56**	**0.55**	**0.51**
5: (3) + Reverse 2-stage	0.50	0.40	0.44	0.45	0.54	0.49	0.49	0.48
6: (3) + Term Oracle	N/A	N/A	N/A	0.79	0.82	0.80	N/A	N/A

* The result is reported for experiments on our manually annotated W00 corpus. Evaluation on token (word) level.

WCL Dataset Results

Comparable to (Navigli and Velardi, 2010)

System	Term (Word Level)	Definition (Word Level)	Sentence Level
	P / R / F_1	P / R / F_1	P / R / F_1
DefMiner	.82/.78/.80	.82/.79/.81	.92/.79/.85
N&V '10	- / - / -	- / - / -	.99/.61/.77

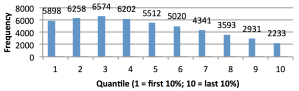
Conclusion

- We introduced DefMiner, a sequence labeling system that identifies scientific terms and their definitions
- Improved system accuracy by exploiting a small set of shallow and dependency parsing features
- Serial classification (term ⇒ definition) boosts the performance significantly
- Applied to a large corpus of scientific publications, highlighting trends and applications

Applying DefMiner to the ACL Anthology Reference Corpus

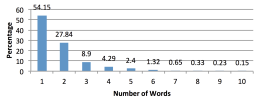
Macroscopic

Definition location distribution over 10 quantiles

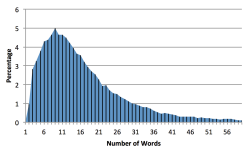


- Definition sentences tend to occur towards the beginning of documents

Term length distribution



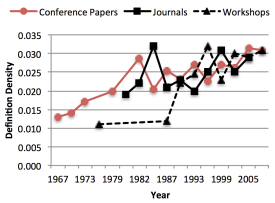
Definition length distribution



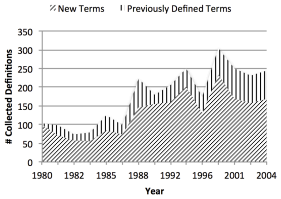
- 45% of the detected terms are multi-word terms
- Definition length is more varied. 75% are between 5-16 words

Temporal

Across publication categories



New and recurring definitions



- Density of definitions increases in workshop / conference papers over time
- Around 40% of the definitions introduced in 2004 have been seen before

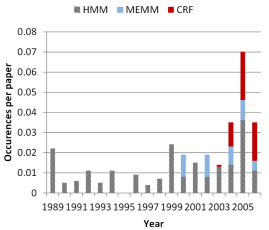
Term-Level Microscopic

Frequently occurring defined terms in the ACL Corpus

WordNet (292)	Part of Speech (45)
Precision (172)	Probabilistic CFG (43)
Recall (167)	FrameNet (38)
Noun phrase (97)	Conditional Random Field (29)
Word sense disambiguation (60)	Inverse Document Frequency (28)
Support Vector Machine (60)	PropBank (27)
Hidden Markov Model (54)	Context Free Grammar (25)
Latent Semantic Analysis (57)	Accuracy (20)

- Extracted terms can often be fit into 3 categories: resources, methodologies, and evaluation metrics

HMM, MEMM and CRF mentions in definitions over time



- Possible to see trends in comparable methodologies over time