WING-NUS at SemEval-2017 Task 10: Keyphrase Identification and Classification as Joint Sequence Labeling

Animesh Prasad and Min-Yen Kan

School of Computing, National University of Singapore

❖ Introduction

- **Tasks:**
  - Keyphrases Identification (Subtask A)
  - Typing among one of three types: Materials, Process and Task (Subtask B)
- **Challenges**:
  - Keyphrases occur more densely in the given excerpts compared against standard set of 5-25 keyphrases over an entire document
  - Keyphrases overlap significantly. e.g. equally sized blocks and sequences of optimal walks of a growing length in weighted digraph
  - Determining the keyphrase type depends on the context. e.g. oxidation test and assessment of the corrosion condition type depends on the context.

❖ Proposed Technique

- **Features**
  - Token(T), lowercased token
  - 1 to 4 character n-gram from beginning and end of the token
  - POS of the token
  - Orthographic features like capitalization, alpha/numeric?, ASCII?, quoted?, hyphenated?, math operators?
  - Occurrence in title
- **Model**
  - First Order Conditional Random Field

❖ Experiments

- **Features Ablation**
  - Model performance over different feature ablation, as evaluated on *Dev*. Best performance is **bolded**.
  - Most of the contributions come from character n-gram and previous tokens output label

	Subtask A			Subtask B		
Features	P	R	F_1	P	R	F_1
All	0.55	0.38	0.45	0.51	0.32	**0.40**
All-(T,T_{lower})	0.49	0.34	0.40	0.44	0.26	0.34
All-($T_{n\text{-}gram}$)	0.53	0.33	0.40	0.46	0.25	0.33
All-(T_{POS})	0.55	0.36	0.43	0.50	0.30	0.37
All-($T_{orthographic}$)	0.55	0.37	0.44	0.51	0.31	0.38
All-($T_{in\text{-}title}$)	0.55	0.39	**0.46**	0.51	0.32	0.39
All-($T\text{-}1_{output}$)	0.30	0.39	0.34	0.26	0.32	0.29

- **Model Configurations**
  - We explore three configurations

    Joint: Performing both Subtask A and B jointly

    Unified: Expert model for keyphrase identification (Subtask A) by collapsing all keyphrase types in one canonical type

    Individual: Expert model for each keyphrase type

  - Subtask A performance for ***Joint*** versus ***Unified*** models, as assessed on ***Dev***. Best performance is **bolded**.

Setup	P	R	F_1
Joint	0.55	0.38	**0.45**
Unified	0.49	0.40	0.44

- Subtask B performance for ***Joint*** versus ***Unified*** models, as assessed on ***Dev***. Best performance is **bolded**.

Setup	Type	P	R	F_1
Joint	Material	0.61	0.36	**0.45**
	Process	0.45	0.34	**0.39**
	Task	0.29	0.12	**0.17**
	Micro Average	0.51	0.32	**0.40**
Individual	Material	0.50	0.28	0.36
	Process	0.29	0.23	0.26
	Task	0.22	0.07	0.11
	Micro Average	0.37	0.22	0.28

- Joint modeling leverages more rich contextual information, outperforms individual expert systems

❖ Results

- **Official Scores**
  - End to end scores on ***Test***

Type	P	R	F_1
Material	0.40	0.40	0.40
Process	0.37	0.26	0.30
Task	0.13	0.07	0.09
Micro Average*	0.26	0.29	0.27

  - Subtask-wise scores on ***Test***

Subtask	P	R	F_1
A	0.51	0.42	0.46
B	0.37	0.31	0.33

  - Significant drop in F_1 for certain type with skewer test distribution

❖ Discussions

- Feature based CRF model performs close to reported best performance on precision, with a difference of **0.04**
- Lower recall by around **0.10** is caused by systematic modeling error that CRF incurs because of overlapping annotations which is further exacerbated by strict evaluation
- **Future Directions**
  - Using semantic features to learn the context dependent typing of the keyphrases
  - Using deep learning based models using word embeddings, though our primary attempt didn't give better result than feature based models, due to high class imbalance

https://github.com/animeshprasad/science_ie

wing.comp.nus.edu.sg/keyphrase

animesh@comp.nus.edu.sg