ENTITY RESOLUTION IN FINANCIAL TEXT

THE UNIVERSITY OF SYDNEY

Ka Shing Law, James R. Curran and Ben Hachey

ə-lab, School of Information Technologies
klaw0084@uni.sydney.edu.au

1. Introduction

- Entity resolution attempts to match two separate sources of information concerning a specific name reference
- Most relevant work focuses on mention-level resolution, implemented as a pipeline of discrete steps
- Pipelines are easy to implement but lead to error propagation between steps
- We explore the idea of joint candidate generation and disambiguation approach using document-level text categorisation

2. Data

- Financial text is taken from Reuters NewsScope Archive (RNA), incorporating news events from all over the world, spanning across years 2003 to 2009
- Each RNA news event is coded with extensive meta-data including "Related RICs", which are used to identify stocks, indices and tradeable instruments mentioned in a document (Figure 1)

RNA Event Text	ASX200
In Australia, resource issues includinding the major miners such as BHP Billiton and Rio Tinto, helped drive the key S&P/ASX200 index 0.21 percent higher.	BHP.AX RIO.AX

Figure 1: RNA *news event and ASX200 stocks in "Related RICs" meta-data field*

3. Background

- Previous work focuses on mention-level resolution, where mentions are individual instances of entity references in text, *e.g.* "BHP Billiton"
- These instances are first identified and then matched to an underlying knowledge base, *e.g.* the unique "BHP.AX" ticker symbol on the Australian Securities Exchange
- Mention-level approaches are often based on a pipelines consisting of:
  1. Named entity recognition, where entity mentions are identified (Figure 2)
  2. Candidate generation, where possible entities are identified for each mention
  3. Candidate disambiguation, where the best entity match is chosen for each mention (Figure 2)
- Cucerzan [2007] and Bunescu and Paşca [2006] explored pipeline approaches in general text, treating Wikipedia as a knowledge base

RNA Event Text	Candidates
In Australia, resource issues includinding the major miners such as **BHP Billiton** and **Rio Tinto**, helped drive the key S&P/ASX200 index 0.21 percent higher.	**BHP.AX** **RIO.AX**

Figure 2: *Organisational mentions from Figure 1 and their disambiguated candidates*

4. Pipeline Baseline

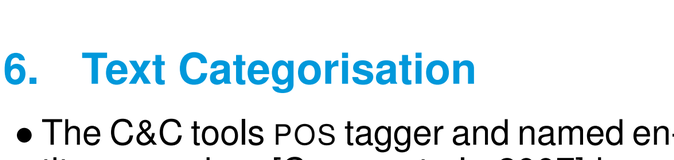

Figure 3: *Pipeline approach*

5. Regex Baseline

- Create a regular expression for each company name
- Search for name in RNA news event (Figure 4)

Figure 4: *Regular expression approach*

6. Text Categorisation

- The C&C tools POS tagger and named entity recogniser [Curran et al., 2007] is applied to the RNA news events, and making use of the pipe-delimited output format: word|POS|NER

Parsed Output

In|IN|O Australia|NNP|I-LOC ,|,|O resource|NN|O issues|NNS|O including|VBG|O the|DT|O major|JJO miners|NNS|O such|JJ|O as|IN|O BHP|NNP|I-ORG Billiton|NNP|I-ORG and|CC|I-ORG Rio|NNP|I-ORG Tinto|NNP|I-ORG ,|,|O helped|VBD|O drive|VB|O the|DT|O key|JJ|O S&P/ASX200|CD|O index|NN|O 0.21|CD|I-PCT percent|NN|I-PCT higher|JJR|O .|.|O

Figure 5: *Pipe-delimited output of news event in Figure 1*

- A **complete** and **reduced** feature sets are generated
- 346 binary classifiers are built from feature sets (Figure 6)

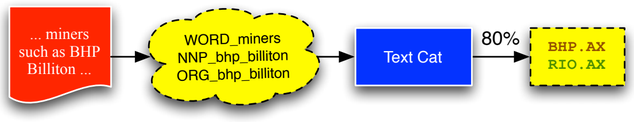

Figure 6: *Text categorisation approach*

7. Results

Empirical results for the pipeline baseline and text categorisation approach can be seen in the table below

- Pipeline baseline and text categorisation approaches were applied, with the precision, recall and F-score figures being recorded across years 2003 to 2009
- Pipeline performs significantly poorly across the board, whilst text categorisation approach results are relatively consistent
- These results backup the theory that pipeline architecture performs more poorly than joint candidate generation and disambiguation due to error propagation

Approach	Year	P	R	F
Pipeline Baseline	2003	48.10	56.14	51.81
	2004	46.79	60.15	52.64
	2005	54.54	65.69	59.60
	2006	55.31	69.17	61.47
	2007	48.67	74.60	58.90
	2008	50.60	72.94	59.75
	2009	48.47	69.06	56.96
Complete Text Cat	2003	92.28	64.81	76.15
	2004	90.50	60.39	72.45
	2005	92.95	57.20	70.82
	2006	94.03	61.85	74.62
	2007	88.95	50.98	64.82
	2008	92.09	53.52	67.70
	2009	89.81	46.92	61.64

Table 1: *Precision, recall and F-score of the text categorisation approaches compared to the pipeline architecture baseline*

Experiments were also conducted on a joint model where data was combined across all years

- Due to incorporation of more data, joint text categorisation retrieves more contextual signals, thus increasing performance, but takes significantly longer in both training and classification

Approach	P	R	F
Pipeline	50.84	67.07	57.84
Regex	74.03	71.95	72.98
Complete TC	91.07	71.58	80.16
Reduced TC	83.96	73.98	78.65

Table 2: *Precision, recall and F-score of the joint text categorisation approaches compared to the pipeline architecture baseline*

8. Future work

Recent work has been undertaken to explore a multi-class problem instead of multi-label in an attempt to increase efficiency

9. Conclusion

Through analysis of a text categorisation machine learning approach combined with empirical results show that joint candidate generation and disambiguation significantly out-performs a pipeline architecture in a document-level entity resolution task.

References

Razvan C. Bunescu and Marius Paşca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, 2006.

Siliu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, 2007.

James R. Curran, Stephen Clark, and Johan Bos. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 33–36, 2007.