Evaluation of Coreference Resolution for Biomedical Text

Miji Choi
[1]The University of Melbourne
Melbourne, Australia
[2]National ICT Australia
jooc1@student.unimelb.edu.au

Karin Verspoor
The University of Melbourne
Melbourne, Australia
karin.verspoor@unimelb.edu.au

Justin Zobel
The University of Melbourne
Melbourne, Australia
jzobel@unimelb.edu.au

ABSTRACT

The accuracy of document processing activities such as retrieval or event extraction can be improved by resolution of lexical ambiguities. In this brief paper we investigate coreference resolution in biomedical texts, reporting on an experiment that shows the benefit of domain-specific knowledge. Comparison of a state-of-the-art general system with a purpose-built system shows that the latter is a dramatic improvement.

Categories and Subject Descriptors

Computing methodologies:: artificial intelligence:: natural language processing:: information extraction, phonology/ morphology; Applied computing:: life and medical science:: health informatics.

General Terms

Algorithms, Performance, Reliability.

Keywords

Coreference resolution, domain-specific knowledge, named entity recognition.

1. INTRODUCTION

The peer-reviewed scientific literature is a vast repository of authoritative knowledge. The life sciences literature is the basis of biomedical research and clinical practice, and must be searchable to be of value. However, with around 40,000 new journal papers every month, manual discovery or annotation is infeasible, and thus it is critical that document processing techniques be robust and accurate, to enable not only conventional search, but automated discovery and assessment of knowledge such as interacting relationships (events and facts) between biomolecules such as proteins, genes, chemical compounds and drugs. Biological molecular pathways, for example, integrated with knowledge of relevant protein-protein interactions, or chemical reactions, are used to understand complex biological processes that could explain specific health conditions in human body in biomedical and pharmaceutical research.

A particular challenge is the need for lexical ambiguity resolution [1]. Lexical ambiguity is a general problem for text processing – such as for search or for event extraction – but is particularly acute in this domain, which has a vast but inconsistent technical lexicon; the domain also presents particular opportunities, because many technical terms are constructed in accordance with a set of highly standardized rules. Thus while there are particular kinds of ambiguity (genes and proteins may share names, for example) there are also deductions that can be made from name structure (for example, that a certain name must be a chemical).

A key obstacle is the low detection reliability of hidden or complex mentions of entities involving coreference expressions in natural language texts [2, 3]. Thus, coreference resolution is an essential task in information extraction, because it can automatically provide links between entities, and as well can facilitate better indexing for medical information search with rich semantic information.

For example, the following passage includes an interacting relation; the *binding* event between the anaphoric mention *the protein* and a cell entity *CD40* is implied in the text. The mention *the protein* refers to the specific protein name, *TRAF2*, previously mentioned in the same discourse.

> *… The phosphorylation appears to be related to the signalling events that are activated by TRAF2 under these circumstances, since two non-functional mutants were found to be phosphorylated significantly less than the wild-type protein. Furthermore, the phosphorylation status of TRAF2 had significant effects on the ability of* ***the protein*** ***to bind to CD40****, as evidenced by our observations …*

Such anaphoric mentions, or pronouns in texts, are mostly ignored by event extraction systems, and are not considered as term occurrences in information retrieval systems. In this brief paper, we report an initial investigation of the challenges of biomedical coreference resolution, test an existing general domain coreference resolution system on biomedical texts, and demonstrate that domain-specific knowledge can be helpful for coreference resolution for the biomedical domain.

2. EXPERIMENT

To evaluate the important of domain-specific knowledge, we compare an existing coreference resolution system, TEES, that uses a domain-specific named entity recognition (NER) module with an existing general system, CoreNLP, that does not use a domain-specific NER. The aim is to explore how domain-specific information impacts on performance for coreference resolution involving protein and gene entities. The TEES system, which includes a biomedical domain-specific NER component for protein and gene mentions [4], and the Stanford CoreNLP system, which uses syntactic and discourse information but no NER outputs [5], are evaluated on a domain-specific annotated corpus.

MedIR2014, July 11, 2014, Gold Coast, Australia.

2.1 Data Sets

We use the training dataset from the Protein Coreference Shared task at BioNLP 2011 [2] for our evaluation of existing coreference resolution systems. The annotated corpus includes 2,313 coreference relations, which are pairs of anaphors and antecedents related to protein and gene entities, from 800 Pubmed journal abstracts. As shown in Table 1, this gold standard dataset consists of coreference relations involving relative pronouns such as *which*, *that*, or *who*, or pronouns such as *it*, *its*, or *they*. Among 2,313 coreference relations, 560 relations embed one or more specific protein and gene name.

Table 1. Statistics of the annotated corpus at the coreference relation level

Anaphor	Relative pronoun	1,174 (51%)
	Pronoun	754 (32%)
	Definite Noun Phrase	346 (15%)
	Indefinite Noun Phrase	11 (0.5%)
	Proper Noun	22 (1%)
	Unclassified	6
Antecedent	Including protein/gene	560
	Including conjunction	217
	Cross-sentence	389
	Identical relation	43
	Head-word match	254

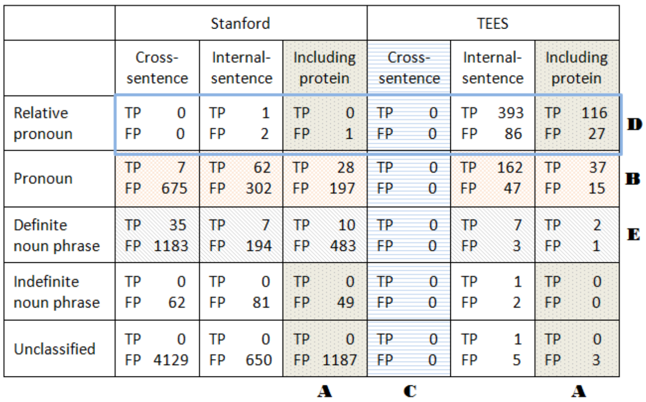
2.2 Results

Performance for identification of coreference mentions and relations of each system evaluated on the annotated corpus is compared in Table 2. The Stanford system achieved low performance with F-score 12% and 2% for the detection of coreference mentions and relations respectively, and produced a greater number of detected mentions, while the TEES system achieved better performance with F-score 69% and 37% for coreference mention and relation levels respectively, but produced smaller number of detections, which reduced system recall. Both systems demonstrate huge reduction in detection of coreference relations from the mention detection with the number of exact matched 1,006 at the mention level to 112 by the Stanford system, as well as from 2,466 to 546 by the TEES system.

Table 2. Results of evaluation of existing systems on the annotated corpus

	Stanford		TEES	
	Mention	Relation	Mention	Relation
Gold corpus	4,367	2,313	4,367	2,313
System detected	12,848	7,387	2,796	707
Exact match	1,006	112	2,466	564
Precision	0.08	0.02	0.88	0.80
Recall	0.23	0.05	0.56	0.24
F-score	0.12	0.02	0.69	0.37

Our investigation of low performance by each system at the coreference relation level is analysed in detail in Figure 1.



	Stanford			TEES			
	Cross-sentence	Internal-sentence	Including protein	Cross-sentence	Internal-sentence	Including protein	
Relative pronoun	TP 0 FP 0	TP 1 FP 2	TP 0 FP 1	TP 0 FP 0	TP 393 FP 86	TP 116 FP 27	D
Pronoun	TP 7 FP 675	TP 62 FP 302	TP 28 FP 197	TP 0 FP 0	TP 162 FP 47	TP 37 FP 15	B
Definite noun phrase	TP 35 FP 1183	TP 7 FP 194	TP 10 FP 483	TP 0 FP 0	TP 7 FP 3	TP 2 FP 1	E
Indefinite noun phrase	TP 0 FP 62	TP 0 FP 81	TP 0 FP 49	TP 0 FP 0	TP 1 FP 2	TP 0 FP 0	
Unclassified	TP 0 FP 4129	TP 0 FP 650	TP 0 FP 1187	TP 0 FP 0	TP 1 FP 5	TP 0 FP 3	
			A	C		A	

Figure 1. Analysis of performance of existing systems comparing to the annotated corpus

Several factors such as lack of domain-specific knowledge (A), bias towards selection of closest candidate of antecedent (B), limiting analysis to within-sentence relations (C), syntactic parsing error (D), and disregard of definite noun phrase (E) have been observed. The main cause, lack of domain-specific knowledge, is explored below.

The annotated corpus contains 560 coreference relations, where anaphoric mentions refer to protein or gene entities previously mentioned in a text. For those coreference relations, the TEES system outperformed the Stanford system by identifying 155 true positives – far more than the 38 identified by the Stanford system, as shown in Table 3.

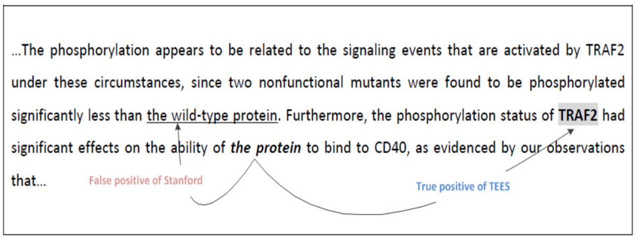
Table 3. Result of performance of existing systems for coreference relations involving protein names

	Output		Precision	Recall	F-score
Stanford	TP	38	0.02	0.07	0.03
	FP	1732			
TEES	TP	155	0.77	0.28	0.41
	FP	46			

The Stanford system also produces a large number of false positives. Even though half of the false positives are relations where anaphors are unclassified, the system links coreference relations where an anaphor and an antecedent are identical, or have a common head word (the main noun of the phrase). This is because coreference resolution systems in general domains aim to identify all mentions that refer to the same entity in a text, rather than to resolve only specifically anaphoric mentions. Considering those anaphoric mentions, inspection of individual instances (as illustrated in Figure 2) strongly suggests that lack of domain-specific knowledge is the main cause of failure.

On the other hand, the TEES system achieved 77% precision, but still only 28% recall. The main reason for the low recall is that the system is limited to identification of coreference relations where anaphors and antecedents corefer within a single sentence. Even though anaphoric coreference mentions mostly link to their antecedents across sentences, the system still identified 155 correct coreference relations by taking advantage of domain-specific information provided through recognition of proteins.

Figure 2 demonstrates how the process of NER in the biomedical domain helps to determine correct coreference relations. In the example, the anaphoric mention *the protein* is correctly identified as referring to *TRAF2* by the TEES system, but the Stanford System links it to the incorrect antecedent *the wild-type protein*.



...The phosphorylation appears to be related to the signaling events that are activated by TRAF2 under these circumstances, since two nonfunctional mutants were found to be phosphorylated significantly less than the wild-type protein. Furthermore, the phosphorylation status of **TRAF2** had significant effects on the ability of ***the protein*** to bind to CD40, as evidenced by our observations that...

Figure 2. Example of a coreference relation involving a protein entity, and results of coreference resolution performed by both the TEES and the Stanford systems

3. CONCLUSIONS

In this study, we have explored how domain-specific knowledge can be helpful for resolving coreferring expressions in the biomedical domain. The performance difference between a system using a domain-specific NER approach and a general system is substantial. In detailed analysis of individual cases of failure (not reported here) we have observed that the domain knowledge, rather than variation in methods, is the main explanation for the success of the domain-specific approach.

4. ACKNOWLEDGMENTS

This work was supported by the University of Melbourne, and by the Australian Federal and Victorian State governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA).

5. REFERENCES

[1] Krovetz, R. *Homonymy and polysemy in information retrieval*. Association for Computational Linguistics, 1997.

[2] Nguyen, N., Kim, J.-D. and Tsujii, J. i. *Overview of the protein coreference task in BioNLP shared task 2011*. Association for Computational Linguistics, 2011.

[3] Miwa, M., Sætre, R., Kim, J.-D. and Tsujii, J. i. Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology*, 8, 01 2010), 131-146.

[4] Björne, J. and Salakoski, T. *Generalizing biomedical event extraction*. Association for Computational Linguistics, 2011.

[5] Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M. and Jurafsky, D. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. Association for Computational Linguistics, 2011.