

# Sentence retrieval for extracting biomedical knowledge

Tara McIntosh and James R. Curran

School of Information Technologies

University of Sydney

NSW 2006, Australia

{tara, james}@it.usyd.edu.au

## Abstract

At present, the majority of biomedical Information Retrieval tools process abstracts rather than full-text articles. The increasing availability of full text will allow more knowledge to be extracted with greater reliability. The first step of this is to extract sentences and passages from the text which report scientific results.

We investigate the challenges of sentence retrieval, using an annotated corpus of articles cited in a Molecular Interaction Map (Kohn, 1999) developed by McIntosh and Curran (2007). From the annotated facts we generate keywords for sentence retrieval, and analyse the impact of various query relaxation strategies on performance. We also investigate the impact of *hedging* and *commitment* in the reporting of scientific results on retrieval. Finally, we look at whether linguistic properties such as anaphora and negation have an impact on retrieval performance.

## 1 Introduction

Almost all known and postulated knowledge relating to biological processes is recorded in the form of semi-structured text, the literature, and data repositories. The volume of biomedical literature rapidly becoming available makes it no longer feasible for biologists to keep abreast of their specialist fields. The standard keyword-based Information Retrieval (IR) approaches over abstracts retrieve too many articles that must be manually inspected.

There is considerable interest in incorporating NLP tools into IR systems to overcome this information bottleneck. There is also a strong focus on improving Information Extraction (IE), which attempts to identify the relationships, such as inter-

actions, between bio-entities, including genes and proteins (Bunescu et al., 2006).

The primary biomedical IR tool, the National Library of Medicine's PubMed, allows researchers to search for relevant documents using keyword-based queries over abstracts. Research systems are beginning to bring NLP techniques to bear on this task. The MEDIE system identifies biomedical relationships within individual sentences in Medline abstracts, which are first parsed using the HSPG grammar (Ohta et al., 2006). As McIntosh and Curran (2007) showed, a significant number of interactions are not contained within the abstract of a paper. There is a need for IR/NLP tools to exploit the growing number of publicly available full-text articles.

Since NLP tools are currently quite slow, an important step in this process is to identify sentences and passages that are likely to contain scientific results, which is basically an IR task at the sentence level. Also, referring scientists to particular sentences will reduce the burden of reading full-text articles to identify interactions.

In this paper, we analyse some factors influencing the sentence retrieval problem for biomedical fact extraction from full-text articles. We use a corpus of full-text articles that have been exhaustively annotated with molecular interactions (McIntosh and Curran, 2007). The corpus is based on the *Molecular Interaction Map* (MIM), constructed by Kohn (1999), and documents the process of inferring the MIM facts from full text. Since we have exhaustively identified and annotated all of the sentences supporting these specific facts, we can reliably identify all relevant and irrelevant retrieved sentences under a range of conditions.

We evaluate various keyword queries with respect to their ability to identify previously annotated positive sentences describing molecular interactions in the full-text corpus. In the first set of

experiments, simple keyword queries were generated from the description of the facts in the MIM by a domain expert. The keywords fall into various classes, e.g. the entities involved in a particular interaction or the verbs describing the interaction. We then explored various forms of query relaxation, and their impact on sentence retrieval.

Next, we investigate the impact of *hedging* and *commitment* on sentence retrieval. Hedging is used to indicate uncertainty or scepticism about a particular proposition (Hyland, 1996). For example, the following sentences express the same proposition between two proteins RPA1 and DNA-PK, however only the first does so with certainty:

1. RPA1 was sufficient to form a complex with DNA-PK.
2. Experiments were performed to test whether DNA-PK could form a protein complex with RPA1.

Hedging has been studied in the citation analysis of scientific literature (Mercer and Marco, 2004). Our study is the first we are aware of to investigate the impact of hedging on sentence retrieval. We compare the frequency of hedging and commitment words in the output of the system.

Sentences in the MIM corpus were also annotated with various linguistic properties, e.g. whether it requires the resolution of anaphora or negation for the fact to be inferred from the sentence (McIntosh and Curran, 2007). We investigate what proportion of the results contain these various phenomena.

Our conclusions are that some levels of query relaxation are required to reach acceptable levels of fact recall (although the cost in precision is quite high); that hedging and commitment terms appear frequently in both relevant and irrelevant results; and that even the most relaxed queries fail to retrieve a significant proportion of sentences requiring anaphora, negation and extra fact resolution.

## 2 Biomedical IR and NLP

At present, most biomedical IR and IE tools process abstracts rather than full-text articles (Ohta et al., 2006). This is due to the availability of abstracts (Hirschman et al., 2002) and the lack of full-text training corpora. The majority of bioNLP corpora consist of sentences or whole abstracts annotated for biomedical Named Entity Recognition (NER) and IE, such as GENIA (Kim et al.,

2003) and BioInfer (Pyysalo et al., 2007). For other bioNLP tasks, such as coreference resolution, there is very limited training data. The BioInfer corpus is also annotated with coreference expressions, however they do not annotate those which cross sentence boundaries (Pyysalo et al., 2007).

Full-text articles are becoming increasingly available to NLP researchers, who have identified the importance of processing specific full-text sections. Regev et al. (2002) developed the first bioIR system specifically focusing on limited text sections, primarily Figure legends. Their performance in the KDD Cup Challenge showed the importance of considering full-text article structure. Yu et al. (2002) showed that the Introduction defines the majority of synonyms, while Schuemie et al. (2004) and Shah et al. (2003) showed that the Results and Methods are the most and least informative, respectively. In contrast, Sinclair and Webber (2004) found the Methods useful in assigning Gene Ontology codes to articles.

Full-text articles also have the advantage of repeating facts in different contexts, increasing the likelihood of an imperfect IR system identifying them (McIntosh and Curran, 2007). This redundancy can also be used for passage validation and ranking (Clarke et al., 2001).

*Hedging* is frequently used in scientific literature to indicate a lack of commitment to a statement, scepticism, and/or open-mindedness about propositions (Hyland, 1996). For example, the terms *may*, *might*, *propose*, and *possibly* convey speculation, compared to *does* and *indicates*. Hedging occurs more frequently in citation contexts than in the text as a whole (Mercer and Marco, 2004). Light et al. (2004) introduced three levels of certainty, highly speculative, low speculative and definite, in biomedical literature. In this work, we expand the list of speculative words described in the literature to automatically analyse their impact on sentence retrieval.

## 3 Full-text MIM corpus

McIntosh and Curran (2007) manually extracted and annotated a corpus of sentences or short passages from full-text articles. Each annotated sentence, known as an *instance*, expresses a documented interaction between bio-entities. The interactions identified are based on the *Molecular Interaction Map* (MIM) constructed by Kohn (1999), which describes 203 different interactions

---

1.	N4 Main fact: <i>RPA2 binds XPA via the C-terminal region of RPA2</i>
TP	Mutant RPA that lacked the p34 C terminus failed to interact with XPA, whereas RPA containing the p70 mutant (Delta RS) interacted with XPA (Fig. 2). (Results)
TP	We found that the C-terminal domain of RPA p34 is responsible for RPA interaction with XPA. (Intro.)
FP	For this, we compared the XPA-DNA interaction of wild-type RPA with that of mutants lacking either the N-terminal or C-terminal domain of p34 (RPA:p34Delta 2-30 and RPA:p34Delta 33C, respectively) (Fig. 4). (Results)
FP	Once XPA and RPA form a stable complex on the DNA, they are thought to bring other repair proteins to the site of initiation of nucleotide excision repair. (Discussion)

---

Table 1: Example TP and FP retrieved for a Main fact matching keywords in Table 3

between bio-entities in mammalian cells. Each interaction in the MIM is associated with a description that summarises the evidence for the interaction from the literature, including the relevant citations.

Each main concept in the summaries is referred to as a *main fact* or a *subfact*, which represent part of a main fact, in the corpus. Text substantiating these facts from the cited articles were manually retrieved and annotated in the corpus. For example, instances in the corpus supporting the main fact N4 *RPA2 binds XPA via the C-terminal region of RPA2* in Table 1 are indicated with the TP (true positive) tag. The two instances in Table 1 were identified in the Results and Introduction sections.

The corpus contains 1637 annotated instances which support 77 of the 203 MIM summaries in Kohn (1999), totalling 1738 sentences from 64 full-text HTML articles. Only 92 instances consist of two or more adjacent sentences, which are all needed to substantiate a fact, e.g. because an anaphor’s referent is in the previous sentence. Each instance has been annotated with linguistic phenomena, such as coreference and negated expressions which needs to be resolved for the automatic extraction of the exact relationship mentioned in the MIM fact. Instances which also depend on additional knowledge from the full text which is necessary to infer the MIM fact are also annotated. These logical dependencies include synonyms and extra facts, which may or may not be defined in the same article. For example, the two TP in Table 1 require the resolution of the synonymous bio-entities *RPA2* and *p34*, to infer the main fact from the text. Full details of this corpus are detailed in McIntosh and Curran (2007).

The IR experiments reported here are based on the set of 64 full-text articles. The articles were converted from HTML to plain text using Lynx followed by some manual post-processing to filter out any remaining noise. We used a boundary detector based on the MXTerminator (Reynar and

Ratnaparkhi, 1997) with modifications to handle common mistaken boundaries such as *et al.* These sentences were tokenised, ensuring entities with punctuation, like *E2F-4* were not split. The 64 articles contain 16,257 sentences and 363,131 words.

## 4 Full-text IR

The experiments we describe here are based on a IR system, which retrieves all sentences, without ranking, matching a specific set of keywords in a full-text article. Each search is restricted to a particular article (or articles) which is known in advance to contain text supporting the relevant MIM facts. The task of the IR system is to retrieve just those sentences annotated as positive instances, based on keywords created from the description of the fact by a domain expert.

### 4.1 Evaluation

For a given MIM fact, any retrieved sentence that matches the keyword queries, which also appears as an annotated instance for the specific fact in the corpus is considered a relevant result, that is, a true positive (TP). Since McIntosh and Curran (2007) did an exhaustive manual search for sentences supporting a given fact, any other sentence retrieved by the system is irrelevant, that is, a false positive (FP). Finally, any annotated instances that were not retrieved by the system are false negatives (FN).

For the keyword expansion results we also report precision  $P$ , recall  $R$  and F-score  $F$ :

$$\begin{aligned}
 P &= \frac{TP}{TP + FP} \\
 R &= \frac{TP}{TP + FN} \\
 F &= \frac{2PR}{P + R}
 \end{aligned}$$

2. A5 Main fact: <i>c-Abl phosphorylates tyrosines in the C-terminal domain of RNA polymerase II</i>
FP Given the fact that Arg and Abl are highly divergent in the C-terminal region except for the CTD-interacting domain, it is possible that these two kinases may transduce different signals to mediate the tyrosine phosphorylation of RNA polymerase II. (Discussion)
3. B3 Subfact: <i>DNA-PK can bind dsDNA without Ku</i>
FP It seems likely that DNA-PK does not recognize DNA alone initially if Ku is present; rather it binds to some part of Ku or both Ku and DNA in the Ku:DNA complex. (Discussion)
4. C43 Main fact: <i>p16 associates with TFIIH and RNA pol II CTD</i>
FP The possibility that p16[INK4A] might associate with the RNA pol II, a protein substrate of the CTD kinase of TFIIH, was next examined. (Results)

Table 2: Example FP containing hedging terms

## 4.2 Keywords and Queries

For each of the facts in the corpus, we generated keyword lists for the main terms associated with the instances. These lists were created semi-automatically by first obtaining the most frequent terms from the instances, excluding stop words.

Each list was manually reduced by a domain expert to only those associated with the fact and separated into the three classes: *bio-entities* involved in the interaction including synonyms, *verbs* describing the interaction, and *extra terms* which were considered necessary by the domain expert to fully identify the entire fact. Extra terms typically refer to specific structures within the entities involved in the interaction, and in many cases no extra terms were specified by the domain expert.

For example, Table 3 shows the keyword list for the MIM fact in Table 1. The synonyms are in parentheses. The bio-entity p70 in the first TP instance for this MIM fact (Table 1) is not included in the list because it isn't part of the fact.

These classes of keywords are then used to form queries with various levels of relaxation. The verb class is relaxed in two ways, giving a total of five keyword classes:

**ent** sentences must contain all main entities

**verb** sentences must contain all main verbs associated with a MIM fact

**verb syn** as above, but sentences may contain synonyms for each main verb

**any verb** sentences must contain at least one verb from the set of all main verbs and synonyms

**extra** if a MIM fact is associated with extra terms, sentences must contain these or their synonyms

These five classes are combined together to form the query sets described in Section 5.

Terms	Keywords
entities	RPA2 (p34, RPAp34), XPA
verb	binds (associates, complex)
extra term	C-terminal (carboxyl-terminus)

Table 3: Search keywords for fact N4

## 4.3 Hedging and Commitment

Scientific literature contains various levels of certainty, ranging from doubt to complete confidence, and *hedging* is frequently used to indicate any lack of commitment to a fact (Hyland, 1996). Hedging is typically realised using modal verbs, epistemic adjectives, nouns and adverbials, lexical verbs and indefinite quantifiers (Holmes, 1988). For the analyses reported here, we experimented with terms in each of these categories identified by Holmes (1988) expressing uncertainty, and the small set of speculative terms identified by Light et al. (2004). We also constructed a list of commitment terms – these words express a high degree of certainty. These were manually identified in the corpus by a domain expert whilst they were performing an analysis of the false positives and negatives. Examples of each of these lists are shown in Table 6.

Examples 2–4 in Table 2 are FP of their respective facts which contain at least one hedging expression. In example 2, the epistemic adjective *possible* and modal verb *may* are used to express open-mindedness of the fact. Epistemic adjectives and modal verbs are also used in examples 3 and 4 respectively. In examples 2–4, the authors are indicating their research aims and hypotheses. However, this is only directly stated in example 4, indicated by the phrase *was next examined*. If we ignore the notion of hedging, only examples 2 and 3 would match the MIM facts. Although example 4 contains all entities and the exact verb used to describe their relationships (*associate*), it only

5. P21 Subject: <i>p300 acetylates p53</i>
FN Note that incubation with DNA-PK produced a new p53 isoform (labeled 3) that is phosphorylated on Ser-37 as well as on Ser-33. This isoform was preferentially acetylated by p300. (Figure legend)
6. A4 Subject: <i>c-Abl inhibits Mdm2-mediated degradation of p53</i>
FN We demonstrate that c-Abl increases the expression level of the p53 protein. The enhanced expression is achieved by inhibiting Mdm2-mediated degradation of p53. (Abstract)
7. N6 Subject: <i>XPF binds to the C-terminal region of ERCC1</i>
FN Previous mutagenesis studies showed that a ‘Rad10-like’ ERCC1 protein, with a stop at residue 214, was functionally inactive (27). This can now be explained by the inability of this protein to form a complex with XPF. (Discussion)
8. P36 Subject: <i>TBP binds to an acidic domain in central Mdm2</i>
FN We show that MDM2 binds to the general transcription factor TFIID in vivo. The C-terminal Ring finger interacts with TAF[II]250/CCG1, and the central acidic domain interacts with TBP. (Abstract)

Table 4: Example FN containing coreference expressions

substantiates *part* of the MIM fact (*p16 associates with RNA pol II*).

#### 4.4 Linguistic Phenomena

McIntosh and Curran (2007) annotated the instances in the MIM corpus with linguistic phenomena and extra fact dependencies. We use these annotations to evaluate the potential impact on sentence retrieval of preprocessing the text with NLP tools that interpret these phenomena.

*Coreference* expressions are often used in biomedical literature to make abbreviated or indirect references to bio-entities or processes. *Negated* expressions include descriptions of an abnormal condition, such as experimental mutations, and the resulting abnormal outcome, such as cancer. From these negated expressions a normal condition and outcome can be inferred.

Examples of these are shown in Table 4. Unlike the hedging examples (Table 2), each of these sentences expresses the authors’ complete confidence in the facts. Although each of these examples contains two sentences, the keyword searches are unable to identify either of the sentences. This is because no single sentence contains all of the keywords associated with the fact.

Examples 5–7 require anaphora resolution, e.g. example 6 involves linking the event anaphor enhanced expression to the function of c-Abl in the first sentence (antecedent), and then the role of c-Abl in the inhibition can be inferred.

Example 7 is complicated further by the negated expressions. The anaphor *this protein* refers to the mutated/truncated form of the protein ERCC1 (with a stop at residue 214), rather than Rad10 or normal ERCC1. Two negative expressions must then be processed. First the mutated

	P	R	F	TP	FP	FN
ent + verb + extra	34	35	34	601	1168	1137
ent + verb	28	37	32	647	1640	1091
ent + verb syn + ext.	33	63	43	1089	2188	649
ent + verb syn	26	67	37	1165	3342	573
ent + any verb	14	73	24	1274	7582	464
ent	13	75	22	1300	8932	438

Table 5: IR performance for various queries

form of ERCC1 is *inactive*, and the second negated expression states that the mutated ERCC1 is unable to bind XPF. From resolving and merging these two negatives we can infer the MIM subfact of N6.

In example 8, the first sentence states a different interaction to that in the MIM fact, and there is no specific coreference expression linking these sentences together, however to infer the fact the reference of the C-terminal ring finger needs to be associated with the bio-entity MDM2.

## 5 Results and Discussion

Table 5 shows the precision (*P*), recall (*R*) and F-score (*F*), and the distribution of TP, FP and FN results for sentences matching keyword queries with varying specificity for individual MIM facts.

The first experiment in Table 5 uses the most restrictive queries, requiring all of the keywords: bio-entities, main verbs and extra terms to be present. This query is unrealistic because it requires knowledge of the exact relationship being known in advance – the exact interaction verbs and extra terms. As a result it achieves the highest precision of 34% but the lowest recall of 35%.

Each subsequent experiment shown in Table 5 relaxes the search criteria, and so recall increases

Word list	TP	FN	FP	Examples
Epistemic adjectives	2.44	2.45	4.58	probable, possible, unlikely
Epistemic nouns	2.29	2.14	3.98	chance, claim, suggestion
Modal verbs	8.32	9.48	15.04	could, should, might
Epistemic adverbials	2.52	2.45	4.06	maybe, perhaps, presumably, surely
Indefinite quantifiers	3.59	6.73	5.27	about, generally, often, sometimes
Epistemic lexical verbs	12.67	16.21	18.15	appear, hypothesize, presume, suggest
Speculative words (Light et al., 2004)	13.05	14.07	16.08	likely, may, suggest, promise
Any hedging word	24.96	31.80	37.51	
Any positive word	40.31	48.01	36.65	demonstrate, established, indicating
Only hedging words	15.34	15.29	24.89	
Only positive words	30.69	31.50	24.03	

Table 6: Hedging and Commitment

and precision decreases. The least restrictive search, *ent*, results in the largest recall and the worse precision, returning an enormous number of FP. Unfortunately, these are exactly the kinds of queries a biologist might initially type into PubMed.

There is a significant improvement in F-score, from 32% to 37%, when the corresponding verb lists are expanded to include their synonyms (*ent* + *verb syn*), but the number of FP increases by 51%.

The best performance of 43% F-score is achieved with the *ent* + *verb syn* + *extra* queries, that is the bio-entities, the verbs or their synonyms and the extra terms. Including the extra terms significantly reduces the number of FP, however, this search unrealistically relies heavily on prior knowledge of the exact MIM fact. Unfortunately, the most realistic query setting is *ent* + *any verb*, since it is feasible to enumerate possible interaction verbs and their synonyms without prior knowledge of the type of interaction.

## 5.1 Hedging and Commitment

The results of our hedging and commitment experiments are shown in Table 6. This table shows the importance of the various categories of hedging expressions, detailed in Holmes (1988) and Light et al. (2004), as well as the words expressing certainty and definite facts (positive words). Example terms for each of these categories are also shown in Table 6.

We are interested in identifying any potential IR improvement that may be gained from recognising hedging and commitment in articles. More specifically, can hedging and or commitment be used to

help separate relevant and irrelevant sentences.

The majority of hedging categories occur in less than 5% of TP, FN and FP. The most significant class discrimination is obtained with modal verbs, with a high 15.04% of FP containing at least one, with almost 7% difference between the TP and FP. Epistemic lexical verbs and the speculative words identified by Light et al. (2004) are the least discriminative, and are also very frequently occurring in the literature. We then investigated the overall importance of hedging words by combining the hedging categories into one word list, and identified those TP, FN and FP which contain any hedging terms. These results indicate that hedging cannot be used as a form of filtering of FP, as a high proportion of both TP and FN also contain hedging words.

We then analysed terms expressing commitment in the retrieved sentences. As expected many of the TP and FN contain positive words, 40.31% and 48.01% respectively, however a large proportion of FP (36.65%) also contain positive language. These experiments do not consider the possibility of both hedging and commitment terms appearing in the same sentences. For instance, it is common for scientists to present a known fact with commitment, and then speculate about possible reasons or future studies, for example:

*While DNA-PK clearly interacts with DNA on its own, based on the DNA-PK activation by DNA alone (Fig. 1), the nature of the stimulation by Ku is still unclear.*

When we consider sentences with hedging words and no commitment words (Only hedging words) and vice versa, the FP are separated more from the TP and FN.

	Negated	Anaphora	Event Ana.	Cataphora	Extra Dep.	None
TP & FN Instances	101 ( %)	145 ( %)	34 ( %)	23 ( %)	561 ( %)	892 ( %)
ent + verb + extra	60 (59.4)	93 (64.1)	30 ( 88.2)	10 ( 43.5)	358 ( 63.8)	450 (50.4)
ent + verb	60 (59.4)	88 (60.7)	30 ( 88.2)	10 ( 43.5)	338 ( 60.2)	434 (48.7)
ent + verb syn + extra	25 (24.8)	59 (40.7)	24 ( 70.6)	6 ( 26.1)	209 ( 37.3)	120 (13.5)
ent + verb syn	24 (23.8)	53 (36.6)	24 ( 70.6)	6 ( 26.1)	177 ( 31.6)	95 (10.7)
ent + any verb	20 (19.8)	49 (33.8)	22 ( 64.7)	6 ( 26.1)	140 ( 25.0)	67 ( 7.5)
ent	18 (17.8)	46 (31.7)	21 ( 61.8)	6 ( 26.1)	135 ( 24.1)	49 ( 5.5)

Table 7: Distribution of linguistic phenomena in FN instances

Our analysis shows that hedging and commitment terms cannot be used for filtering FN on their own, however these categories, in particular the modal verbs, may be used in the development of models with other features to distinguish between TP and FP.

## 5.2 Linguistic Phenomena

We are interested in understanding why certain FN are not been identified. For each FN we investigated which linguistic phenomena may be responsible for them going undetected. Table 7 details the linguistic phenomena associated with the instances in the MIM corpus. The second row shows the total number of instances annotated with the specific phenomena. The main linguistic phenomena that are annotated in the corpus, which require resolution are extra fact dependencies (561), followed by anaphoric expressions (145). Cataphoric expressions are not common with only 23 instances requiring these to be resolved. Some instances may be annotated with multiple linguistic constructs. The majority of the instances (892) do not contain any of these characteristics.

The majority of instances which contain annotated phenomena are identified with the most relaxed search (*ent*). For instance only 46 of the FN contained at least one anaphoric expression, and this accounts for 31.7% of all instances annotated with anaphora. Many of these FN contain more than one sentence, where no single sentence mentions each of the entities. Similar results are observed with event anaphora, with 61.8% of instances going undetected. Examples of these are shown in Table 4.

The next main issue is the notion of extra fact dependencies, and many instances (135) requiring these to be resolved are also undetected with the most relaxed search. In McIntosh and Curran (2007) both synonym and extra fact dependencies

indicated the necessity for full-text processing and merging of knowledge throughout the article.

Considering the F-score associated with the two least restrictive searches, in particular the extremely large number of FP identified (Table 5) it is more realistic to look at those obtained with the *ent + verb syn* query. There is overall a slight increase in the FN requiring linguistic phenomena to be resolved.

## 6 Conclusion

In this paper, we have presented an analysis of biomedical keyword-based full-text IR, investigating the characteristics of various query relaxation strategies, hedging and commitment, and identification of linguistic phenomena which may improve relevance.

Our first experiments show that the most restrictive search with query terms matching those in the actual MIM facts is not successful. This is due to the high level of synonymous verbs in biomedical texts. Furthermore, searching only for bio-entities results in too many FP. The best F-score of 43% is achieved with the *ent + verb syn + extra* query.

Hedging has previously been reported to be most commonly used when citing in scientific literature. In this study, we investigated the impact of hedging on the reporting a new scientific findings. We find that while hedging appears frequently in irrelevant sentences, it also appears frequently in relevant sentences. The most discriminative class of hedging was the modal verbs. We also considered commitment terms used in scientific literature, however these also were commonly used across TP and FP. Therefore, hedging and commitment terms cannot be used as a means of post-processing and filtering.

Linguistic analysis of FN shows the potential improvements gained from anaphora resolution and the identification of extra fact dependencies.

Even with the most relaxed queries, relevant sentences are often missed due to these characteristics. Furthermore, with the most reliable query level, *ent + verb syn + extra*, these phenomena are significantly more important to resolve.

In summary, this paper provides guidance to developers of biomedical IR systems that operate on the sentence level. In future work, we will be implementing a system that incorporates the kind of NLP required to exploit the linguistic phenomena identified as important here.

## Acknowledgements

This work was supported by the CSIRO ICT Centre and ARC Discovery grants DP0453131 and DP0665973.

## References

- Razvan Bunescu, Raymond Mooney, Arun Ramani, and Edward Marcotte. 2006. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline. In *Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06*, pages 49–56. New York City.
- Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. 2001. Exploiting redundancy in question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 358–365. New Orleans, LA.
- Lynette Hirschman, Jong C. Park, Junichi Tsujii, Limsoon Wong, and Cathy Wu. 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics Review*, (12):1553–1561.
- Janet Holmes. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9(1):21–44.
- Ken Hyland. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics*, 17(4):433–454.
- Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):i180–i182.
- Kurt W. Kohn. 1999. Molecular interaction map of the mammalian cell cycle and DNA repair systems. *Molecular Biology of the Cell*, 10:2703–2734.
- Marc Light, Xin Ting Qui, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pages 17–24. Boston.
- Tara McIntosh and James R. Curran. 2007. Challenges for extracting biomedical knowledge from full text. In *Proceedings of the Workshop on BioNLP 2007, Biological, Translational, and Clinical Language Processing*, pages 171–178. Prague, Czech Republic.
- Robert E. Mercer and Chrysanne Di Marco. 2004. A design methodology for a biomedical literature indexing tool using the rhetoric of science. In *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pages 77–84. Boston.
- Tomoko Ohta, Yusuke Miyao, Takashi Ninomiya, Yoshimasa Tsuruoka, Akane Yakushiji, Katsuya Masuda, Jumpei Takeuchi, Kazuhiro Yoshida, Tadayoshi Hara, Jin-Dong Kim, Yuka Tateisi, and Jun'ichi Tsujii. 2006. An intelligent search engine and GUI-based efficient Medline search tool based on deep syntactic parsing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 17–20. Sydney, Australia.
- Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).
- Yizhar Regev, Michal Finkelstein-Langau, Ronen Feldman, Mayo Gorodetsky, Xin Zheng, Samuel Levy, Rosane Charlab, Charles Lawrence, Ross A. Lippert, Qing Zhang, and Hagit Shatkay. 2002. Rule-based extraction of experimental evidence in the biomedical domain - the KDD Cup 2002 (Task 1). *ACM SIGKDD Explorations*, 4(2):90–92.
- J.C Reynar and A. Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 16–19. Washington DC.
- M.J. Schuemie, M. Weeber, B.J.A. Schijvenaars, E.M. van Mulligen, C.C. van der Eijk, R.Jelier, B.Mons, and J.A Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604.
- Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. 2003. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4(20).
- Gail Sinclair and Bonnie Webber. 2004. Classification from full text: A comparison of canonical sections of scientific papers. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLBPBA)*, pages 66–69. Geneva, Switzerland.
- Hong Yu, Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, and W.John Wilbur. 2002. Automatic extraction of gene and protein synonyms from Medline and journal articles. In *Proceedings of the AMIA Symposium 2002*, pages 919–923. San Antonio, TX.