Challenges for extracting biomedical knowledge from full text

Tara McIntosh
School of IT
University of Sydney
NSW 2006, Australia
tara@it.usyd.edu.au

James R. Curran
School of IT
University of Sydney
NSW 2006, Australia
james@it.usyd.edu.au

Abstract

At present, most biomedical Information Retrieval and Extraction tools process abstracts rather than full-text articles. The increasing availability of full text will allow more knowledge to be extracted with greater reliability. To investigate the challenges of full-text processing, we manually annotated a corpus of cited articles from a Molecular Interaction Map (Kohn, 1999).

Our analysis demonstrates the necessity of full-text processing; identifies the article sections where interactions are most commonly stated; and quantifies both the amount of external knowledge required and the proportion of interactions requiring multiple or deeper inference steps. Further, it identifies a range of NLP tools required, including: identifying synonyms, and resolving coreference and negated expressions. This is important guidance for researchers engineering biomedical text processing systems.

1 Introduction

It is no longer feasible for biologists to keep abreast of the vast quantity of biomedical literature. Even keyword-based Information Retrieval (IR) over abstracts retrieves too many articles to be individually inspected. There is considerable interest in NLP systems that overcome this information bottleneck.

Most bioNLP systems have been applied to abstracts only, due to their availability (Hirschman et al., 2002). Unfortunately, the information in abstracts is dense but limited. Full-text articles have the advantage of providing more information and repeating facts in different contexts, increasing the likelihood of an imperfect system identifying them.

Full text contains explicit structure, e.g. sections and captions, which can be exploited to improve Information Extraction (IE) (Regev et al., 2002). Previous work has investigated the importance of extracting information from specific sections, e.g. Schuemie et al. (2004), but there has been little analysis of when the entire document is needed for accurate knowledge extraction. For instance, extracting a fact from the Results may require a synonym to be resolved that is only mentioned in the Introduction. External domain knowledge may also be required.

We investigated these issues by manually annotating full-text passages that describe the functional relationships between bio-entities summarised in a *Molecular Interaction Map* (MIM). Our corpus tracks the process Kohn (1999) followed in summarising interactions for the mammalian cell MIM, by identifying information required to infer facts, which we call *dependencies*. We replicate the process of manual curation and demonstrate the necessity of full-text processing for fact extraction.

In the same annotation process we have identified NLP problems in these passages which must be solved to identify the facts correctly including: synonym and hyponym substitution, coreference resolution, negation handling, and the incorporation of knowledge from within the full text and the domain. This allows us to report on the relative importance of anaphora resolution and other tasks to the problem of biomedical fact extraction.

As well as serving as a dataset for future tool development, our corpus is an excellent case study providing valuable guidance to developers of biomedical text mining and retrieval systems.

BioNLP 2007: Biological, translational, and clinical language processing, pages 171–178,
Prague, June 2007.

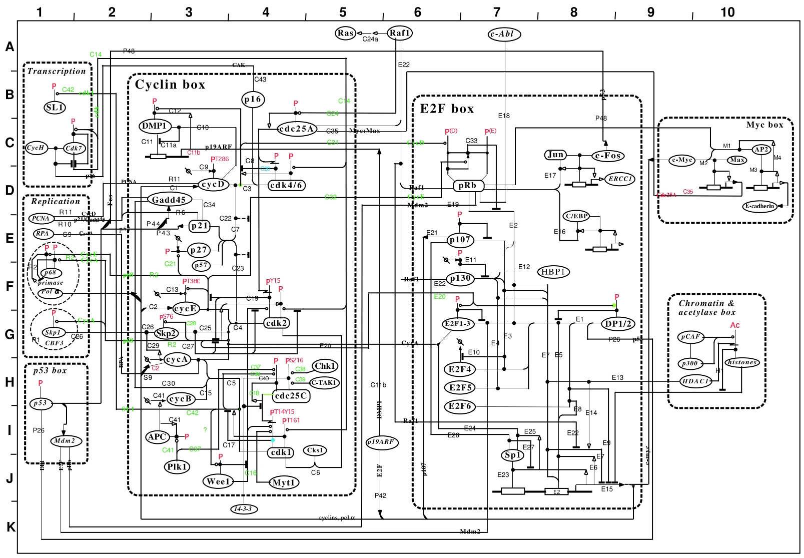

Figure 1: Map A of the Molecular Interaction Map compiled by Kohn (1999)

2 Biomedical NLP

Full-text articles are becoming increasingly available to NLP researchers, who have begun investigating how specific sections and structures can be mined in various information extraction tasks. Regev et al. (2002) developed the first bioIR system specifically focusing on limited text sections. Their performance in the KDD Cup Challenge, primarily using Figure legends, showed the importance of considering document 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.

These section specific results highlight the information loss resulting from restricting searches to individual sections, as sections often provide unique information. Furthermore, facts appearing in different contexts across various sections, will be lost. This redundancy has been used for passage validation and ranking (Clarke et al., 2001).

There are limited training resources for biomedical full-text systems. The majority of corpora consist of abstracts annotated for bio-entity recognition and Relationship Extraction, such as the GENIA (Kim et al., 2003) and the BioCreAtIvE corpora. However, due to the lack of full-text corpora, many current systems only process abstracts (Ohta et al., 2006). Few biomedical corpora exist for other tasks, such as coreference resolution (Castaño et al., 2004; Vlachos et al., 2006), and these are very small. In this paper, we estimate the importance of these tasks in bioNLP systems, which will help determine which tasks system developers should focus effort on first.

Despite limited full-text training corpora, competitions such as the Genomics track of TREC, require systems to retrieve and rank passages from full text that are relevant to question style queries.

3 Molecular Interaction Maps

Kohn (1999) constructed a *Molecular Interaction Map* (MIM) based on literature describing 203 different interactions between bio-entities, such as proteins and genes, in mammalian cells (Figure 1). Interactions in the MIM are represented as links between nodes labelled with the bio-entities. Each link is associated with a description that summarises the evidence for the interaction from the literature, including citations. For example, Table 1 contains the description passage for interaction M4 (on the right of the Myc Box at grid reference C10 in Figure 1). Although MIM interactions may be mentioned in other articles, the articles cited by Kohn (1999) document the main biomedical research leading to the discovery of these interactions.

c-Myc and pRb enhance transcription from the E-cadherin promoter in an AP2-dependent manner in epithelial cells (mechanism unknown) (Batsche et al., 1998). Activation by pRb and c-Myc is not additive, suggesting that they act upon the same site, thereby perhaps blocking the binding of an unidentified inhibitor. No c-Myc recognition element is required for activation of the E-cadherin promoter by c-Myc. Max blocks transcriptional activation from the E-cadherin promoter by c-Myc, presumably because it blocks the binding between c-Myc and AP2.

Table 1: MIM annotation M4

1. M4 Subfact: *Activation of E-cadherin by pRb and c-Myc is not additive, suggesting they act on the same site*

a) However, the precise molecular mechanisms by which `RB`, `Myc`, and AP-2 cooperate to effect transcriptional activation of E-cadherin requires further study. . . . the positive effects of `RB` and `c-Myc` were not additive. (Discussion)

Synonym: *pRb equivalent to RB* – undefined

Synonym: *c-Myc equivalent to Myc*

b) The c-myc proto-oncogene, which encodes two amino-terminally distinct Myc proteins, acts as a transcription factor. (Intro)

Table 2: Example instances depending on synonym facts

In creating our corpus we have attempted to *reverse engineer* and document the MIM creation process for many of the interactions in Kohn (1999). We exhaustively traced and documented the process of identifying passages from the cited full-text articles that substantiate the MIM interactions. This allows us to identify and quantify the amount of information that is unavailable when systems are restricted to abstracts.

4 Corpus Creation

The first stage of corpus creation involved obtaining the full text of the articles cited in the MIM descriptions. There are 262 articles cited in Kohn (1999), and we have manually extracted the text from 218 of them; we have abstracts for the other 44 which have not been included in the analysis presented here.

Currently, the annotated part of the corpus consists of passages from 101 full-text articles, supporting 95 of the 203 MIM descriptions. A biomedical expert exhaustively identified these passages by manually reading each article several times. 30% of these articles support multiple MIM descriptions and so passages from these articles may appear multiple times. We restricted the corpus to the cited articles only. This allows us to quantify the need for external resources, e.g. synonym lists and ontologies. The corpus collection involved the following:

1. Each sentence in a MIM description is a called a *main fact*.
2. For each main fact we annotated every passage (*instance*) that the fact can be derived from. These include direct statements of the fact and passages the fact can be implied from.
3. Main facts are often complex sentences, combining numerous facts from the article. Passages from which part of a fact can be derived are also annotated as instances. A *subfact* is then created to represent these partial facts. This may be repeated for subfacts.
4. Many instances cannot be directly linked to their corresponding fact, as they *depend* on additional passages within the full text or external domain knowledge. New facts are formed to represent the dependency information – *synonym* and *extra* facts. Instances of these are annotated, and a link is added between the original and dependency facts.
5. Each instance is annotated with its location within the article. Linguistic phenomena, including anaphora, cataphora, and negated expressions which must be resolved to derive the fact are identified.

Tables 1 and 2 show an example of this process. One of the main facts of interaction M4 (Table 1) is *Activation by pRb and c-Myc is not additive . . . blocking the binding of an unidentified inhibitor.* An instance supporting part of this fact, the subfact in Table 2 *Activation of E-cadherin by pRb and c-Myc is not additive* . . . , 1.a), was identified. This instance requires the resolution of two synonymy dependencies, only one of which appears in the article.

2.	E13 Main Fact: *HDAC1 binds to the pocket proteins pRb, p107 and p130 and in turn is recruited to E2F complexes on promoters*
a)	The experiments described above indicate that p107 and p130 can interact with HDAC1. We thus reasoned that they could repress E2F activity by recruiting `histone deacetylase` activity to E2F containing promoters. (Results)
	Extra: *HDAC1 is a histone deacetylase*
b)	We have previously shown that Rb, the founding member of the pocket proteins family, represses E2F1 activity by recruiting the histone deacetylase HDAC1. (Abstract)

Table 3: Example instances depending on extra facts

3.	N4 Main fact: *RPA2 binds XPA via the C-terminal region of RPA2*
	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)
4.	C9 Subfact: *Cyclin D1 degraded rapidly by phosphorylation at threonine-286*
	Although "free" or CDK4-bound cyclin D1 molecules are intrinsically unstable (t1/2 $<$ 30 min), a cyclin D1 mutant (T286A) containing an alanine for threonine-286 substitution fails to undergo efficient polyubiquitination in an in vitro system or in vivo, and it is markedly stabilized (t1/2 approximately 3.5 hr) when inducibly expressed in either quiescent or proliferating mouse fibroblasts. (Abstract)

Table 4: Example instances with negated expressions

5 Dependencies

In our corpus, an instance of a fact may depend on additional facts (*dependencies*) to allow the fact to be derived from the original instance. Dependencies may occur elsewhere in the document or may not be mentioned at all. We consider two types of dependencies: synonym facts and extra facts.

5.1 Synonym Facts

The frequent use of synonyms, abbreviations and acronyms in biomedical text is a common source of ambiguity that is often hard to resolve (Sehgal et al., 2004). Furthermore, synonym lists are difficult to maintain in rapidly moving fields like biology (Lussier et al., 2006). There has been recent interest in developing systems to identify and extract these (Ao and Takagi, 2005; Okazaki and Ananiadou, 2006).

In our corpus we group all of these synonyms, abbreviations, acronyms and other orthographic variations as *synonym facts*. For example, the synonyms (1) `E2F4`, (2) `E2F-4` and (3) `E2F1-4` in our corpus refer to the same entity `E2F4`, however term (3) also includes the entities `E2F1`, `E2F2` and `E2F3`.

In Table 2, an instance supporting subfact 1. is shown in 1.a). The bio-entity `pRb` mentioned in the subfact does not appear in this instance. Thus 1.a) depends on knowing that `pRb` is equivalent to `RB`, and so we form a new synonym fact. This synonym is undefined in the article and cannot be assumed as `RB` is also a homograph for the gene ruby (`rb`), rubidium (`Rb`) and Robertsonian (`Rb`) translocations.

Instance 1 also depends on a second synonym – `c-Myc` and `Myc` are used interchangeably, where the protein `Myc` is referred to by its gene name, `c-Myc`. Metonymy is common in biology, and an instance supporting this synonym fact was found in the article, 1.b).

5.2 Extra Facts

Extra facts include all assertions (excluding synonym definitions) which are necessary to make a valid inference from an instance to a fact or subfact. These extra facts must be found within the same article. Many extra facts are descriptions or classes of bio-entities and hyponym relationships. According to Nédellec et al. (2006), a clearer distinction between entities and their classes/descriptions is needed in bioNLP corpora.

Example 2 in Table 3 is an instance which depends on an extra fact, 2.b), to derive the main fact. The class of proteins `histone deacetylase` in sentence 2 must be linked to the specific protein `HDAC1` in sentence 1, since the sortal anaphor `they` in sentence 2 refers to the antecedents `p107` and `p130`, and does not include `HDAC1`. This extra fact is identified in the apposition `the histone deacetylase HDAC1` in instance 2.b).

5. C11b Subfact: *p19ARF induces cell cycle arrest in a p53-dependent manner*
   INK4a/ARF is perhaps the second most commonly disrupted locus in cancer cells. It encodes `two distinct tumor suppressor proteins`: `p16INK4a`, which inhibits the phosphorylation of the retinoblastoma protein by cyclin D-dependent kinases, and `p19ARF`, which stabilizes and activates p53 to promote either cell cycle arrest or apoptosis. (Intro)
6. C36 Main fact: *Cdc25C is phosphorylated by Cyclin B-cdk1*
   In this work, we examine the effect of phosphorylation on the human cdc25-C protein (Sadhu et al.,1990). We show that this protein is `phosphorylated` during mitosis in human cells and that `this` requires active cdc2-cyclin B. (Intro)

Table 5: Example instances with cataphora and event anaphora

6 Negated Expressions

To quantify the importance of lexical and logical negations we have annotated each instance involving one or more negated expressions that must be resolved to derive the fact. In biomedical literature, negated expressions are commonly used to describe an abnormal condition, such as a mutation, and its resulting abnormal outcome, such as cancer, from which the normal condition and outcome can be inferred. This typically requires two or more negated expressions to be processed simultaneously.

Table 4 shows examples of instances with negated expressions. In the subject NP of instance 3, the lexical negative form of `RPA` (`Mutant RPA`) is followed directly by a logical negative detailing the function it failed to perform. These two negative expressions support the positive in the main fact. This implicit reporting of results expressed in terms of negative experimental outcomes is very common in molecular biology and genetics.

Example 4 requires external domain knowledge. Firstly, the amino acid `alanine` cannot be phosphorylated like `threonine`. Secondly, `polyubiquitination` triggers a signal for a protein (`cyclin D1`) to be degraded. Therefore from this negated pair the positive fact from interaction C9 can be inferred.

The context surrounding potential negative expressions must be analysed to determine if it is indeed a negative. For example, not all mutations result in negative outcomes – the mutation of `p70` in instance 3 did not have a negative outcome.

7 Coreference Expressions

In biomedical literature, coreference expressions are used to make abbreviated or indirect references to bio-entities or events, and to provide additional information, such as more detailed descriptions.

To quantify the importance of coreference expressions, instances in our corpus are annotated with pronominal, sortal and event anaphoric, and cataphoric expressions, including those extending beyond one sentence. Instances 4–6 in Tables 4–5, each contain annotated pronominal or sortal anaphoric expressions. Instance 5 also involves a cataphoric expression, where `suppressor proteins` refers to `p16INK4a` and `p19ARF`

Event anaphora refer to processes and are quite common in biomedical text. We have annotated these separately to pronominal and sortal anaphora. Our event anaphora annotations are different to Humphreys et al. (1997). They associate sequential events, while we only refer to the same event.

An example is shown in instance 6 (Table 5) where the additional sortal anaphor complicates resolving the event anaphor. The third `this` refers to the phosphorylation event, `phosphorylated`, and not the protein `cdc25-C` like the second `this`.

8 Locating Facts

The key facts and results are generally repeated and reworded in various contexts within an article. This redundancy can be used in two ways to improve system precision and recall. Firstly, the redundancy increases the chance of an imperfect system identifying at least one instance. Secondly, the redundancy can be used for fact validation. By annotating every instance that supports a fact we are able to measure the degree of factual redundancy in full-text articles.

We have also annotated each instance with its location within the article: which section (or structure such as a title, heading or caption) it was contained within and the number of the paragraph. Using this data, we can evaluate the informativeness of each section and structure for identifying interactions.

Using our detailed dependency annotations we can also determine how many instances need addi-

Location	Main Fact	Subfact	Synonym	Extra
Title	3.3 (0.2)	1.9 (0.7)	0.0 (0.0)	0.8 (0.8)
Abstract	19.1 (10.1)	9.3 (5.1)	36.2 (21.7)	25.8 (14.8)
Introduction	11.3 (5.2)	8.3 (3.4)	30.4 (17.4)	17.2 (7.8)
Results	31.0 (13.8)	37.6 (16.1)	20.3 (15.9)	32.0 (12.5)
Discussion	21.8 (7.3)	19.5 (6.6)	2.9 (1.4)	9.4 (3.1)
Figure Heading	5.0 (0.6)	10.7 (3.8)	1.4 (1.4)	2.3 (0.0)
Figure Legend	3.1 (1.3)	4.8 (2.0)	0.0 (0.0)	7.0 (4.7)
Table Data	0.0 (0.0)	0.2 (0.0)	0.0 (0.0)	0.0 (0.0)
Methods	0.2 (0.0)	0.1 (0.1)	0.0 (0.0)	4.7 (0.8)
Conclusion	0.6 (0.4)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)
Footnotes	0.0 (0.0)	0.0 (0.0)	5.8 (2.9)	0.0 (0.0)
Headings	4.8 (0.6)	7.5 (2.7)	2.9 (1.4)	0.8 (0.8)
Full-text	100.0 (39.4)	100.0 (40.6)	100.0 (62.3)	100.0 (45.3)

Table 6: Instances found excluding (including) all dependencies

Fact Type	# Created	# Found	# Instances
Main Fact	170	156	523
Subfact	251	251	1196
Synonym	155	62	69
Extra	152	87	128
Total	728	556	1916

Table 7: Distribution of fact types in corpus

Instances	Synonym	Extra
Main Fact	46.8 (10.9)	26.2 (18.9)
Subfact	36.9 (8.2)	26.7 (15.4)
Synonym	8.7 (2.9)	7.2 (4.3)
Extra	25.0 (0.0)	13.3 (10.9)

Table 8: Instances with (all found) dependencies

tional knowledge outside of the current section to support a particular fact. This demonstrates how important full-text processing is.

9 Corpus Analysis

Having described the corpus annotation we can now investigate various statistical properties of the data. Table 7 shows the distribution of the various annotated fact types within the corpus. There are a total of 728 different facts identified, with 556 (76%) found within the documents. We have annotated 1916 individual passages as instances, totally 2429 sentences. There were 14 main facts that we found no instances or subfact instances for.

The most redundancy occurs in main facts and subfacts, with on average 3.35 and 4.76 instances each respectively, whilst synonym facts have almost no redundancy. Also, a large proportion of synonym and extra facts, 60% and 43% respectively, do not appear anywhere in the articles (Table 7).

This high level of redundancy in facts demonstrates the significant advantages of processing full text. However, the proportion of missing synonym and extra facts shows the importance of external resources, such as synonym lists, and tools for recognising orthographic variants.

9.1 Locating Facts

Table 6 shows the percentage of instances identified in particular locations within the articles. The best sections for finding instances of facts and subfacts were the Results and Discussion sections, whereas synonym and extra facts were best found in the Abstract, Introduction and Results. The later sections of each article rarely contributed any instances. Interestingly, we did not find the Figure headings or legends to be that informative for main facts. Figure headings are restricted in length and thus are rarely able to express main facts as well as subfacts.

The proportion of main facts and subfact instances found in the abstract is quite small, further demonstrating the value of full-text processing.

If we take into account the additional dependency information, and restrict the instances to those fully supported within a given section, the results drop dramatically (those in parentheses in Table 6). In

Depth	Fact	Subfact	Synonym	Extra
0	35.2	45.1	87.0	64.8
1	53.9	44.2	13.0	26.6
2	9.6	9.5	0.0	7.0
3	1.3	0.9	0.0	1.6
4	0.0	0.3	0.0	0.0

Table 9: Maximum depth of instance dependencies

Breadth	Fact	Subfact	Synonym	Extra
0	35.2	45.1	87.0	64.8
1	36.5	35.5	7.2	29.7
2	22.6	15.7	5.8	4.7
3	4.6	2.9	0.0	0.8
4	0.8	0.6	0.0	0.0
5	0.2	0.2	0.0	0.0

Table 10: Breadth of instance dependencies

total, the number of instances drops to 39.4% and 40.6%, for main facts and subfacts, respectively. This again demonstrates the need for full-text processing, including the dependencies between facts found in different sections of the article.

9.2 Dependencies

Our corpus represents each of the facts and subfacts as a dependency graph of instances, each which in turn may require support from other facts, including synonym and extra facts.

Table 8 shows the percentage of instances which depend on synonym and extra facts in our corpus. 46.8% of main fact instances depend on at least one synonym fact, but only 10.9% of main fact instances which depend on at least one synonym were completely resolved (i.e. all of the synonyms were found as well). Interestingly, synonym and extra facts often required other synonym and extra facts.

Our corpus contains more synonym than extra fact dependencies, however more extra facts were defined in the articles. The large proportion of main facts and subfacts depending on synonyms and extra facts demonstrates the importance of automatically extracting this information from full text.

Since the inference from an instance to a fact may depend on other facts, long chains of dependencies may occur, all of which would need to be resolved before a main fact could be derived from the text.

Expressions	Instances
Negated	4.3
Anaphora	13.2
Event Anaphora	6.6
Cataphora	2.7

Table 11: Distribution of annotated expressions

Table 9 shows the distribution of maximum chain depth in our dependency graphs. The maximum depth is predominately less than 3. Table 10 shows the distribution of the breadth of dependency graphs. Again, most instances are supported by fewer than 3 dependency chains. Most instances depend on some other information, but luckily, a large proportion of those only require information from a small number of other facts. However, given that these facts could occur anywhere within the full text, extracting them is still a very challenging task.

9.3 Negated & Coreference Expressions

Table 11 shows the percentage of instances annotated with negated, anaphoric and cataphoric expressions in our corpus. We have separated event anaphora from pronominal and sortal anaphora. There are fewer cataphoric and negated expressions than anaphoric expressions. Therefore, we would expect the greatest improvement when systems incorporate anaphora resolution components, and little improvement from cataphoric and negated expression analysis. However, negated expressions provide valuable information regarding experimental conditions and outcomes, and thus may be appropriate for specific extraction tasks.

10 Conclusion

This paper describes a corpus documenting the manual identification of facts from full-text articles by biomedical researchers. The corpus consists of articles cited in a Molecular Interaction Map developed by Kohn (1999). Each fact can be derived from one or more passages from the citations. Each of these *instances* was annotated with their location in the article and whether they contained coreference or negated expressions. Each instance was also linked with other information, including synonyms and extra knowledge, that was required to derive the particular interaction. The annotation task was quite complex and as future work we will increase the reliability of our corpus by including the annotations of other domain experts using our guidelines, and use this resource for tool development. The guidelines and corpus will be made publicly available.

Our corpus analysis demonstrates that full-text analysis is crucial for exploiting biomedical literature. Less than 20% of fact instances we identified were contained in the abstract. Analysing sections in isolation reduced the number of supported facts by 60%. We also showed that many instances were dependent on a significant amount of other information, both within and outside the article. Finally, we showed the potential impact of various NLP components such as anaphora resolution systems.

This work provides important empirical guidance for developers of biomedical text mining systems.

Acknowledgements

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

References

Hiroko Ao and Toshihisa Takagi. 2005. ALICE: An algorithm to extract abbreviations from Medline. *Journal of the American Medical Informatics Association*, 12(5):576–586.

J. Castaño, J. Zhang, and J. Pustejovsky. 2004. Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution in NLP*, Alicante, Spain.

Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. 2001. Exploiting redundancy in question answering. In *Proc. 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.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Proc. of the ACL/EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, Spain.

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.

Yves Lussier, Tara Borlawsky, Daniel Rappaport, Yang Liu, and Carol Friedman. 2006. PHENOGO: Assigning phenotypic context to gene ontology annotations with natural language processing. In *Proc. of the Pacific Symposium on Biocomputing*, volume 11, pages 64–75, Maui, HI.

Clair Nédellec, Philippe Bessières, Robert Bossy, Alain Kptoujanksy, and Alain-Pierre Manine. 2006. Annotation guidelines for machine learning-based named entity recognition in microbiology. In *Proc. of the ACL Workshop on Data and Text for Mining Integrative Biology*, pages 40–54, Berlin.

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 *Proc. of the COLING/ACL Interactive Presentation Sessions*, pages 17–20, Sydney, Australia.

Naoaki Okazaki and Sophia Ananiadou. 2006. A term recognition approach to acronym recognition. In *Proc. the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 643–650, Sydney, Australia.

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 SIGKKD Explorations*, 4(2):90–92.

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.

Aditya K. Sehgal, Padmini Srinivasan, and Olivier Bodenreider. 2004. Gene terms and english words: An ambiguous mix. In *Proc. of the ACM SIGIR Workshop on Search and Discovery for Bioinformatics*, Sheffield, UK.

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 *Proc. of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 66–69, Geneva, Switzerland.

Andreas Vlachos, Caroline Gasperin, Ian Lewin, and Ted Briscoe. 2006. Bootstrapping the recognition and anaphoric linking of named entities in drosophila articles. In *Proc. of the Pacific Symposium on Biocomputing*, volume 11, pages 100–111, Maui, HI.

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 *Proc. of the AMIA Symposium 2002*, pages 919–923, San Antonio, TX.