Postnominal Prepositional Phrase Attachment in Proteomics

Jonathan Schuman and **Sabine Bergler**
The CLaC Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montreal, Canada
{j_schuma,bergler}@cs.concordia.ca

Abstract

We present a small set of attachment heuristics for postnominal PPs occurring in full-text articles related to enzymes. A detailed analysis of the results suggests their utility for extraction of relations expressed by nominalizations (often with several attached PPs). The system achieves 82% accuracy on a manually annotated test corpus of over 3000 PPs from varied biomedical texts.

1 Introduction

The biomedical sciences suffer from an overwhelming volume of information that is growing at explosive rates. Most of this information is found only in the form of published literature. Given the large volume, it is becoming increasingly difficult for researchers to find relevant information. Accordingly, there is much to be gained from the development of robust and reliable tools to automate this task.

Current systems in this domain focus primarily on abstracts. Though the salient points of an article are present in the abstract, much detailed information is entirely absent and can be found only in the full text (Shatkay and Feldman, 2003; Corney et al., 2004). Optimal conditions for enzymatic activity, details of experimental procedures, and useful observations that are tangential to the main point of the article are just a few examples of such information.

Full-text articles in enzymology are characterized by many complex noun phrases (NPs), usually with chains of several prepositional phrases (PPs). Nominalized relations are particularly frequent, with arguments and adjuncts mentioned in attached PPs. Thus, the tasks of automated search, retrieval, and extraction in this domain stand to benefit significantly from efforts in semantic interpretation of NPs and PPs.

There are currently no publicly available biomedical corpora suitable for this task. (See (Cohen et al., 2005) for an overview of currently available biomedical corpora.) Therefore, statistical approaches that rely on extensive training data are essentially not feasible. Instead, we approach the task through careful analysis of the data and development of heuristics. In this paper, we report on a rule-based postnominal PP attachment system developed as a first step toward a more general NP semantics for proteomics.

2 Background

Leroy *et al.* (2002; 2003) note the importance of noun phrases and prepositions in the capture of relational information in biomedical texts, citing the particular significance of the prepositions *by*, *of*, and *in*. Their parser can extract many different relations using few rules by relying on closed-class words (e.g. prepositions) instead of restricting patterns with specific predefined verbs and entities. This bottom-up approach achieves high precision (90%) and a claimed (though unquantified) high recall. However, they side-step the issue of prepositional attachment ambiguity altogether. Also, their system is targeted specifically and only toward relations. While relations do cover a considerable portion of the most relevant information in biomedical texts, there is also much relevant lower frequency information (particularly in enzymology) such as the conditions under which these relations are expressed.

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 82–89,
New York City, June 2006.

Hahn *et al.* (2002) point out that PPs are crucial for semantic interpretation of biomedical texts due to the wide variety of conceptual relations they introduce. They note that this is reflected in their training and test data, extracted from findings reports in histopathology, where prepositions account for about 10% of all words and more than 25% of the text is contained in PPs. The coverage of PPs in our development and test data, comprised of varied texts in proteomics, is even higher with 26% of the text occurring in postnominal PPs alone.

Little research in the biomedical domain addresses the problem of PP attachment proper. This is partly due to the number of systems that process text using named-entity-based templates, disregarding PPs. In fact, the only recent BioNLP system found in the literature that makes any mention of PP attachment is Medstract (Pustejovsky et al., 2002), an automated information extraction system for Medline abstracts. The shallow parsing module used in Medstract performs "limited" prepositional attachment—only *of* prepositions are attached.

There are, of course, several PP attachment systems for other domains. Volk (2001) addresses PP attachment using the frequency of co-occurrence of a PP's preposition, object NP, and possible attachment points, calculated from query results of a web-based search engine. This system was evaluated on sentences from a weekly computer magazine, scoring 74% accuracy for both VP and NP attachment. Brill & Resnik (1994) put transformation-based learning with added word-class information from WordNet to the task of PP attachment. Their system achieves 81.8% accuracy on sentences from the Penn Treebank Wall Street Journal corpus.

The main concerns of both these systems differ from the requirements for successful PP attachment in proteomics. The main attachment ambiguity in these general texts is between VP and NP attachment, where there are few NPs to choose from for a given PP. In contrast, proteomics texts, where NPs are the main information carriers, contain many NPs with long sequences of postnominal PPs. Consequently, the possible attachment points for a given PP are more numerous. By "postnominal", we denote PPs following an NP, where the attachment point may be within the NP but may also precede it. In focusing on postnominal PPs, we exclude here PPs that trivially attach to the VP for lack of NP attachment points and focus on the subset of PPs with the highest degree of attachment ambiguity.

3 Approach

For this exploratory study we compiled two manually annotated corpora[1], a smaller, targeted development corpus consisting of sentences referring to enzymes in five articles, and a larger test corpus consisting of the full text of nine articles drawn from a wider set of topics. This bias in the data was set deliberately to test whether NPs referring to enzymes follow a distinct pattern. Our results suggest that the compiled heuristics are in fact not specific to enzymes, but work with comparable performance for a much wider set of NPs.

As our goal is semantic interpretation of NPs, only postnominal PPs were considered. A large number of these follow a very simple attachment principle—right association.

Right association (Kimball, 1973), or late closure, describes a preference for parses that result in the parse tree with the most right branches. Simply stated, right association assumes that new constituents are part of the closest possible constituent that is under construction. In the case of postnominal PPs, right association attaches each PP to the NP that immediately precedes it. An example where this strategy does fairly well is given below.

> *The effect of hydrolysis of the hemicelluloses in the milled wood lignin on the molecular mass distribution was then examined...*

Notice that, except for the last PP, attachment to the preceding NP is correct. The last PP, *on the molecular mass distribution*, modifies the head NP *effect*.

Another frequent pattern in our corpus is given below with a corresponding text fragment. In this pattern, the entire NP consists of one reaction fully described by several PPs that all attach to a nominalization in the head NP. Attachment according to this pattern is in direct opposition to right association.

```
<ACTION> <PREPOSITION> <PRODUCT>
         <PREPOSITION> <SUBSTRATE>
         <PREPOSITION> <ENZYME>
         <PREPOSITION> <MEASUREMENT>
```

[1]There was a single annotator for both corpora, who was also the developer of the heuristics.

...the release of reducing sugars from carboxymethylcellulose by cellulase at 37 ^{o}C, pH 4.8...

In general, the attachment behavior of a large percentage of PPs in the examined literature can be characterized by either right association or attachment to a nominalization. The preposition of a PP seems to be the main criterion for determining which attachment principle to apply. A few prepositions were observed to follow right association almost exclusively, while others show a strong affinity toward nominalizations, defaulting to right association only when no nominalization is available.

These observations were implemented as attachment heuristics for the most frequently occurring PPs, as distinguished by their prepositions (see Table 1 for frequency data). These rules, as outlined below, account for 90% of all postnominal PPs in the corpus. The remaining 10%, for which no clear pattern could be found, are attached using right association.

	Devel. Corpus			Test Corpus		
Prep	Freq	Syst	Base	Freq	Syst	Base
of	50.0	99.0	99.0	53.4	98.2	98.2
in	11.9	74.8	55.6	11.7	67.0	54.6
from	8.3	87.0	87.0	3.67	71.8	71.8
for	4.5	81.1	81.0	5.1	56.1	56.0
with	4.5	83.8	75.7	4.7	70.8	65.2
between	4.2	68.6	68.6	1.2	84.2	84.2
at	3.3	81.5	18.5	4.0	68.3	40.7
on	3.1	84.6	57.7	2.1	80.0	53.9
by	2.5	95.2	23.8	2.4	76.7	45.2
to	2.3	63.2	63.2	5.0	51.6	51.6
as	1.8	66.7	46.7	0.7	40.9	36.4

Table 1: Frequency of prepositions with corresponding PP attachment accuracy for the implemented heuristics and the baseline (right association) on development and test set.

Right Association (of, from, for)

PPs headed by *of*, *from*, and *for* attach almost exclusively according to right association. In particular, no violation of right association by *of* PPs has been found. The system, therefore, attaches any PP from this class to the NP immediately preceding it.

Strong Nominalization Affinity (by, at)

In contrast, *by* and *at* PPs attach almost exclusively to nominalizations. Only rarely have they been observed to attach to non-nominalization NPs. In most cases where no nominalizations are present in the NP, a PP of this class actually attaches to a preceding VP. Typical nominalization and VP attachments found in the corpus are exemplified in the following two sentences.

...the formation of stalk cells by culB^{-} pkaR^{-} cells decreased about threefold...

...xylooligosaccharides were not detected in hydrolytic products from corn cell walls by TLC analysis.

This attachment preference is implemented in the system as the heuristic for strong nominalization affinity. Given a PP from this class, the system first attempts attachment to the closest nominalization to the left. If no such NP is found, the PP is assumed to attach to a VP.

Weak Nominalization Affinity (in, with, as)

In, *with*, and *as* PPs show similar affinity toward nominalizations. In fact, initially, these PPs were attached with the strong affinity heuristic. However, after further observation it became apparent that these PPs do often attach to non-nominalization NPs. A typical example for each of these possibilities is given as follows.

...incubation of the substrate pullulan with protein fractions.

The major form of beta-amylase in Arabidopsis...

Here, the system first attempts nominalization attachment. If no nominalizations are present in the NP, instead of defaulting to VP attachment, the PP is attached to the closest NP to its left that is not the object of an *of* PP. This behavior is intuitively consistent since *in* PPs are usually adjuncts to the main NP (which is usually an entity if not a nominalization) and are unlikely to modify any of the NP's modifiers.

"Effect on"

The final heuristic encodes the frequent attachment of *on* PPs with NPs indicating effect, influence, impact, etc. While this relationship seems intuitive and likely to occur in varied texts, it may be disproportionally frequent in proteomics texts. Nonetheless, the heuristic does have a strong basis in the examined literature. An example is provided below.

... the effects of reduced β-amylase activity on seed formation and germination...

The system checks NPs preceding an *on* PP for the closest occurrence of an "effect" NP. If no such NPs are found, right association is used.

4 System Overview

There are three main phases of processing that must occur before the PP attachment heuristics can be applied. These include preprocessing and two stages of NP chunking. Upon completion of these three phases, the PP attachment module is executed.

The preprocessing phase consists of standard tokenization and part-of-speech tagging, as well as named entity recognition (and other term lookup) using gazetteer lists and simple transducers. Recognition is currently limited to enzymes, organisms, chemicals, (enzymological) activities, and measurements. A comprehensive enzyme list including synonyms was compiled from BRENDA[2] and some limited organism lists[3], including common abbreviations, were augmented based on organisms found in the development corpus. For recognition of substrates and products, some of the chemical entity lists from BioRAT (Corney et al., 2004) are used. Activity lists from BioRAT, with several enzyme-specific additions, are also used.

The next phase of processing uses a chunker reported in (Bergler et al., 2003) and further developed for a related project. NP chunking is performed in two stages, using two separate context-free grammars and an Earley-type chart parser. No domain-specific information is used in either of the grammars; recognized entities and terms are used only for improved tokenization. The first stage chunks base NPs, without attachments. Here, the parser input is segmented into smaller sentence fragments to reduce ambiguity and processing time. The fragments are delimited by verbs, prepositions, and sentence boundaries, since none of these can occur within a base NP. In the second chunking stage, entire sentences are parsed to extract NPs containing conjunctions and PP attachments. At this stage, no attempt is made to determine the proper attachment structure of the PPs or to exclude postnominal PPs that should actually be attached to a preceding VP—any PP that follows an NP has the potential to attach somewhere in the NP.

The final phase of processing is performed by the PP attachment module. Here, each postnominal PP is examined and attached according to the rule for its preposition. Only base NPs within the same NP are considered as possible attachment points. For the strong nominalization affinity heuristic, if no nominalization is found, the PP is assumed to attach to the closest preceding VP. For both nominalization affinity heuristics, the UMLS SPECIALIST Lexicon[4] is used to determine whether the head noun of each possible attachment point is a nominalization.

5 Results & Analysis

The development corpus was compiled from five articles retrieved from PubMed Central[5] (PMC). The articles were the top-ranked results returned from five separate queries[6] using BioKI:Enzymes, a literature navigation tool (Bergler et al., 2006). Sentences containing enzymes were extracted and the remaining sentences were discarded. In total, 476 sentences yielding 830 postnominal PPs were manually annotated as the development corpus.

Attachment accuracy on the development corpus is 88%. The accuracy and coverage of each rule is summarized in Table 2 and discussed in the following sections. Also, as a reference point for performance comparison, the system was tested using only the right association heuristic resulting in a baseline accuracy of 80%. The system performance is contrasted with the baseline and summarized for each preposition in Table 1.

	Devel. Corpus		Test Corpus	
Heuristic	Freq	Accuracy	Freq	Accuracy
Right Association	62.8	96.2	62.1	93.3
Weak NA	18.2	76.2	17.1	67.0
Strong NA	5.8	87.5	6.4	71.4
"Effect on"	3.1	84.6	2.1	80.0
Default (RA)	10.1	60.7	12.3	49.5

Table 2: Coverage and accuracy of each heuristic.

[2] http://www.brenda.uni-koeln.de

[3] Compiled for a related project.

[4] http://www.nlm.nih.gov/research/umls/

[5] http://www.pubmedcentral.com

[6] Amylase, CGTase, pullulanase, ferulic acid esterase, and cellwallase were used as the PMC search terms and a list of different enzymes was used for scoring.

To measure heuristic performance, the PP attachment heuristics were scored on manual NP and PP annotations. Thus all reported accuracy numbers reflect performance of the heuristics alone, isolated from possible chunking errors. The PP attachment module is, however, designed for input from the chunker and does not handle constructs which the chunker does not provide (e.g. PP conjunctions and non-simple parenthetical NPs).

5.1 Right Association

The application of right association for PPs headed by *of*, *for*, and *from* resulted in correct attachment in 96.2% of their occurrences in the development corpus. Because this class of PPs is processed using the baseline heuristic without any refinements, it has no effect on overall system accuracy as compared to overall baseline accuracy. However, it does provide a clear delineation of the subset of PPs for which right association is a sufficient and optimal solution for attachment. Given the coverage of this class of PPs (62.8% of the corpus), it also provides an explanation for the relatively high baseline performance.

Of PPs are attached with 99% accuracy. All errors involve attachment of PP conjunctions, such as "*... a search of the literature and of the GenBank database...*", or attachment to NPs containing non-simple parenthetical statements, such as "*The synergy degree (the activities of XynA and cellulase cellulosome mixtures divided by the corresponding theoretical activities) of cellulase...*". Sentences of these forms are not accounted for in the NP chunker, around which the PP attachment system was designed. Both scenarios reflect shortcomings in the NP grammars, not in the heuristic.

For and *from* PPs are attached with 81% and 87% accuracy, respectively. The majority of the error here corresponds to PPs that should be attached to a VP. For example, attachment errors occurred both in the sentence "*... this was followed by exoglucanases liberating cellobiose from these nicks...*" and in the sentence "*... the reactions were stopped by placing the microtubes in boiling water for 2 to 3 min.*"

5.2 Strong Nominalization Affinity

The heuristic for strong nominalization affinity deals with only two types of PPs, those headed by the prepositions *by* and *at*, both of which occur with relatively low frequency in the development corpus. Accordingly, the heuristic's impact on the overall accuracy of the system is rather small. However, it affords the largest increase in accuracy for the PPs of its class. The heuristic correctly determines attachment with 87.5% accuracy.

While these PPs account for a small portion of the corpus, they play a critical role in describing enzymological information. Specifically, *by* PPs are most often used in the description of relationships between entities, as in the NP "*degradation of xylan networks between cellulose microfibrils by xylanases*", while *at* PPs often quantitatively indicate the condition under which observed behavior or experiments take place, as in the NP "*Incubation of the enzyme at 40 ^{o}C and pH 9.0*".

The heuristic provides a strong performance increase over the baseline, correctly attaching 95.2% of *by* PPs in contrast to 23.8% with the baseline. In fact, only a single error occurred in attaching *by* PPs in the development corpus and the sentence in question, given below, appears to be ungrammatical in all of its possible interpretations.

> *The TLC pattern of liberated cellooligosaccharides by mixtures of XynA cellulosomes and cellulase cellulosomes was similar to that caused by cellulase cellulosomes alone.*

A few other errors (e.g. typos, omission of words, and grammatically incorrect or ambiguous constructs) were observed in the development corpus. The extent of such errors and the degree to which they affect the results (either negatively or positively) is unknown. However, such errors are inescapable and any automated system is susceptible to their effects.

Although no errors in *by* PP attachment were found in the development corpus, aside from the given problematic sentence, one that would be processed erroneously by the system was found manually in the GENIA Treebank[7]. It is given below to demonstrate a boundary case for this heuristic.

> *... modulation of activity in B cells by human T-cell leukemia virus type I tax gene...*

Here, the system would attach the *by* PP to the closest nominalization *activity*, when in fact, the cor-

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

rect attachment is to the nominalization *modulation*. This error scenario is relevant to all of the PPs with nominalization affinity. A possible solution is to separate general nominalizations, such as *activity* and *action*, from more specific ones, such as *modulation*, and to favor the latter type whenever possible. An experiment toward this end, with emphasis on *in* PPs, was performed with promising results. It is discussed in the following section.

For *at* PPs, 81.5% accuracy was achieved, as compared to 18.5% with the baseline. The higher degree of error with *at* PPs is indicative of their more varied usage, requiring more contextual information for correct attachment. An example of typical variation is given in the following two sentences, both of which contain *at* PPs that the system incorrectly attached to the nominalization *activity*.

> *The amylase exhibited maximal activity at pH 8.7 and 55 °C in the presence of 2.5 M NaCl.*
>
> *. . . Bacillus sp. strain IMD370 produced alkaline α-amylases with maxima for activity at pH 10.0.*

While both sentences report observed conditions for maximal enzyme activity using similar language, the attachment of the *at* PPs differs between them. In the first sentence, the activity was *exhibited at* the given pH and temperature (VP attachment), but in the second sentence, the enzyme was not necessarily *produced at* the given pH (NP attachment)—production may have occurred under different conditions from those reported for the activity maxima.

For errors of this nature, it seems that employing semantic information about the preceding VP and possibly also the head NP would lead to more accurate attachment. There are, however, other similar errors where even the addition of such information does not immediately suggest the proper attachment.

5.3 Weak Nominalization Affinity

The weak nominalization affinity heuristic covers a large portion of the development corpus (18.2%). Overall system improvement over baseline attachment accuracy can be achieved through successful attachment of this class of PPs, particularly *in* and *with* PPs, which are the second and fourth most frequently used PPs in the development corpus, respectively. Unfortunately, the usage of these PPs is also perhaps the hardest to characterize. The heuristic achieves only 76.2% accuracy. Though noticeably better than right association alone, it is apparent that the behavior of this class of PPs cannot be entirely characterized by nominalization affinity.

Accuracy of *in* PP attachment increased by 19.2% from the baseline with this heuristic. A significant source of attachment error is the problem of multiple nominalizations in the same NP. As mentioned above, splitting nominalizations into general and specific classes may solve this problem. To explore this conjecture, the most common (particularly with *in* PPs) general nominalization, *activity*, was ignored when searching for nominalization attachment points. This resulted in a 3% increase in the accuracy for *in* PPs with no adverse effects on any of the other PPs with nominalization affinity.

Despite further anticipated improvements from similar changes, attachment of *in* PPs stands to benefit the most from additional semantic information in the form of rules that encode containment semantics (i.e. which types of things can be contained in other types of things). Possible containment rules exist for the few semantic categories that are already implemented; enzymes, for instance, can be contained in organisms, but organisms are rarely contained in anything (though organisms can be said to be contained in their species, the relationship is rarely expressed as containment). Further analysis and more semantic categories are needed to formulate more generally applicable rules.

With and *as* PPs are attached with 83.8% and 66.7% accuracy, respectively. All of the errors for these PPs involve incorrect attachment to an NP when the correct attachment is to a VP. Presented below are two sentences that provide examples of the particular difficulty of resolving these errors.

> *The xylanase A . . . was expressed by E. coli with a C-terminal His tag from the vector pET-29b. . .*
>
> *The pullulanase-type activity was identified as ZPU1 and the isoamylase-type activity as SU1.*

In the first sentence, the *with* PP describes the method by which xylanase A was expressed; it does not restrict the organism in which the expression occurred. This distinction requires understanding the semantic relationship between C-terminal His tags, protein (or enzyme) expression, and E. coli. Namely, that His tags (polyhistidine-tags) are amino

acid motifs used for purification of proteins, specifically proteins expressed in E. coli. Such information could only be obtained from a highly domain-specific knowledge source. In the second sentence, the verb to which the *as* PP attaches is omitted. Accordingly, even if the semantics of verbs were used to help determine attachment, the system would need to recognize the ellipsis for correct attachment.

5.4 "Effect on" Heuristic

The attachment accuracy for *on* PPs is 84.6% using the "effect on" heuristic, a noticeable improvement over the 57.7% accuracy of the baseline. The few attachment errors for *on* PPs were varied and revealed no regularities suggesting future improvements.

5.5 Unclassified PPs

The remaining PPs, for which no heurisitics were implemented, represent 10% of the development corpus. The system attaches these PPs using right association, with accuracy of 60.7%. Most frequent are PPs headed by *between*, which are attached with 68.6% accuracy. A significant improvement is expected from a heuristic that attaches these PPs based on observations of semantic features in the corpus. Namely, that most of the NPs to which *between* PPs attach can be categorized as binary relations (e.g. bond, linkage, difference, synergy). This relational feature can be expressed in the head noun or in a prenominal modifier. In fact, more than 25% of *between* PPs in the development corpus attach to the NP *synergistic effects* (or some similar alternative), where *between* shows affinity toward the adjective *synergistic*, not the head noun *effects*, which does not attract *between* PP attachment on its own.

6 Evaluation on Varied Texts

To assess the general applicability of the heuristics to varied texts, the system was evaluated on a test corpus of an additional nine articles[8] from PMC. The entire text, except the abstract and introduction, of each article was manually annotated, resulting in 1603 sentences with 3079 postnominal PPs. The system's overall attachment accuracy on this test data is 82%, comparable to that for the development enzymology data. The accuracy and coverage of each rule for the test data, as contrasted with the development set, is given in Table 2. The baseline heuristic achieved an accuracy of 77.5%. A comparative performance breakdown by preposition is given in Table 1.

Overall, changes in the coverage and accuracy of the heuristics are much less pronounced than expected from the increase in size and variance of both subject matter and writing style between the development and test data. The only significant change in rule coverage is a slight increase in the number of unclassified PPs to 12.3%. These PPs are also more varied and the right-associative default heuristic is less applicable (49.5% accuracy in the test data vs. 60.7% in the development data). The largest contribution to this additional error stems from a doubling of the frequency of *to* PPs in the test corpus. Preliminary analysis of the corresponding errors suggests that these PPs would be much better suited to the strong nominalization affinity heuristic than the right association default. The error incurred over all unclassified PPs accounts for 1.4% of the accuracy difference between the development and test data. The larger number of these PPs also explains the smaller overall difference between the system and baseline performance.

For PPs were observed to have more frequent VP attachment in the test data. In particular, *for* PPs with object NPs specifying a duration (or other measurement), as exemplified below, attach almost exclusively to VPs and nominalizations.

The sample was spun in a microfuge for 10 min...

This behavior is also apparent in the development data, though in much smaller numbers. Applying the strong nominalization affinity heuristic to these PPs resulted in an increase of *for* PP attachment accuracy in the test corpus to 75.8% and an overall increase in accuracy of 1.0%.

A similar pattern was observed for *at* PPs, where the pattern `<CHEMICAL> at <CONCENTRATION>` accounts for 25.6% of all *at* PP attachment errors and the majority of the performance decrease for the strong nominalization affinity heuristic between the two data sets. The remainder of the performance decrease for this heuristic is attributed to gaps in the

[8] PMC query terms: metabolism, biosynthesis, proteolysis, peptidyltransferase, hexokinase, epimerase, laccase, ligase, dehydrogenase.

UMLS SPECIALIST Lexicon. For instance, the underlined head nouns in the following examples are not marked as nominalizations in the lexicon.

> *The double mutant inhibited misreading by paromomycin . . .*
>
> *. . . the formation of stalk cells by culB$^-$ pkaR$^-$ cells. . .*

In our test corpus, these errors were only apparent in *by* PP attachment, but can potentially affect all nominalization-based attachment.

Aside from the cases mentioned in this section, attachment trends in the test corpus are quite similar to those observed in the development corpus. Given the diversity in the test data, both in terms of subject matter (between articles) and writing style (between sections), the results suggest the suitability of our heuristics to proteomics texts in general.

7 Conclusion

The next step for BioNLP is to process the full text of scientific articles, where heavy NPs with potentially long chains of PP attachments are frequent. This study has investigated the attachment behavior of postnominal PPs in enzyme-related texts and evaluated a small set of simple attachment heuristics on a test set of over 3000 PPs from a collection of more varied texts in proteomics. The heuristics cover all prepositions, even infrequent ones, that nonetheless convey important information. This approach requires only NP chunked input and a nominalization dictionary, all readily available from on-line resources. The heuristics are thus useful for shallow approaches and their accuracy of 82% puts them in a position to reliably improve both, proper recognition of entities and their properties and bottom-up recognition of relationships between entities expressed in nominalizations.

References

Sabine Bergler, René Witte, Michelle Khalifé, Zhuoyan Li, and Frank Rudzicz. 2003. Using knowledge-poor coreference resolution for text summarization. In *On-line Proceedings of the Workshop on Text Summarization, Document Understanding Conference (DUC 2003)*, Edmonton, Canada, May.

Sabine Bergler, Jonathan Schuman, Julien Dubuc, and Alexandr Lebedev. 2006. BioKI:Enzymes - an adaptable system to locate low-frequency information in full-text proteomics articles. Poster abstract in *Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology (BioNLP'06)*, New York, NY, June.

Eric Brill and Philip Resnik. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*.

Kevin Bretonnel Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases (BioLINK)*, pages 38–45, Detroit, MI, June. Association for Computational Linguistics.

David P.A. Corney, Bernard F. Buxton, William B. Langdon, and David T. Jones. 2004. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213.

Udo Hahn, Martin Romacker, and Stefan Schulz. 2002. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. In *Proceedings of the 7th Pacific Symposium on Biocomputing*, pages 338–49, Hawaii, USA.

John Kimball. 1973. Seven principles of surface structure parsing in natural language. *Cognition*, 2:15–47.

Gondy Leroy and Hsinchun Chen. 2002. Filling preposition-based templates to capture information from medical abstracts. In *Proceedings of the 7th Pacific Symposium on Biocomputing*, pages 350–361, Hawaii, USA.

Gondy Leroy, Hsinchun Chen, and Jesse D. Martinez. 2003. A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, 36:145–158, June.

James Pustejovsky, José Castaño, Roser Sauri, Anna Rumshisky, Jason Zhang, and Wei Luo. 2002. Medstract: Creating large-scale information servers for biomedical libraries. In *ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, PA.

Hagit Shatkay and Ronen Feldman. 2003. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–855.

Martin Volk. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja, editors, *Proceedings of Corpus Linguistics*, pages 601–606, Lancaster, England, March.