Parsing Biomedical Literature*

Matthew Lease and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP),
Brown University, Providence, RI USA
{mlease, ec}@cs.brown.edu

Abstract. We present a preliminary study of several parser adaptation techniques evaluated on the GENIA corpus of MEDLINE abstracts [1,2]. We begin by observing that the Penn Treebank (PTB) is lexically impoverished when measured on various genres of scientific and technical writing, and that this significantly impacts parse accuracy. To resolve this without requiring in-domain treebank data, we show how existing domain-specific lexical resources may be leveraged to augment PTB-training: part-of-speech tags, dictionary collocations, and named-entities. Using a state-of-the-art statistical parser [3] as our baseline, our lexically-adapted parser achieves a 14.2% reduction in error. With oracle-knowledge of named-entities, this error reduction improves to 21.2%.

1 Introduction

Since the advent of the Penn Treebank (PTB) [4], statistical approaches to natural language parsing have quickly matured [3,5]. By providing a very large corpus of manually labeled parsing examples, PTB has played an invaluable role in enabling the broad analysis, automatic training, and quantitative evaluation of parsing techniques. However, while PTB's Wall Street Journal (WSJ) corpus has historically served as the canonical benchmark for evaluating statistical parsing, the need for broader evaluation has been increasingly recognized in recent years. Furthermore, since it is impractical to create a large treebank like PTB for every genre of interest, significant attention has been directed towards maximally reusing existing training data in order to mitigate the need for domain-specific training examples. These issues have been most notably explored in parser adaptation studies conducted between PTB's WSJ and Brown corpora [6,7,8,9].

As part of our own exploration of these issues, we have been investigating statistical parser adaptation to a novel domain: biomedical literature. This literature presents a stark contrast to WSJ and Brown: it is suffused with domain-specific vocabulary, has markedly different stylistic constraints, and is often written by non-native speakers. Moreover, broader consideration of technical literature shows this challenge and opportunity is not confined to biomedical literature

* We would like to thank the National Science Foundation for their support of this work (IIS-0112432, LIS-9721276, and DMS-0074276), as well as thank Sharon Goldwater and our anonymous reviewers for their valuable feeback.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 58–69, 2005.

alone, but is also demonstrated by patent literature, engineering manuals, and field-specific scientific discourse. Through our work with biomedical literature, we hope to gain insights into effective techniques for adapting statistical parsing to technical literature in general.

Our interest in biomedical literature is also motivated by a real need to improve information extraction in this domain. With over 15 million citations in PubMed today, biomedical literature is the largest and fastest growing knowledge domain of any science. As such, simply managing the sheer volume of its accumulated information has become a significant problem. In response to this, a large research community has formed around the challenge of enabling automated mining of the literature [10,11]. While the potential value of parsing has often been discussed by this community, attempts to employ it thus far appear to have been limited by the parsing technologies employed. Reported difficulties include poor coverage, inability to resolve syntactic ambiguity, unacceptable memory and speed, and difficulty in hand-crafting rules of grammar [12,13]. Perhaps the most telling indicator of community perspective came in a recent survey's bleak observation that efficient and accurate parsing of unrestricted text appears to be out of reach of current techniques [14].

In this paper, we show that broad, accurate parsing of biomedical literature is indeed possible. Using an off-the-shelf WSJ-trained statistical parser [3] as our baseline, we provide the first full-coverage parse accuracy results for biomedical literature, as measured on the GENIA corpus of MEDLINE abstracts [1,2]. Furthermore, after showing that PTB is lexically impoverished when measured on various genres of scientific and technical writing, we describe three methods for improving parse accuracy by leveraging lexical resources from the domain: part-of-speech (POS) tags, dictionary collocations, and named-entities. Our general hope is that lexically-based techniques such as these can provide alternative and complementary value to treebank-based adaptation methods such as co-training [9] and sample selection [15]. Our lexically-adapted parser achieves a 14.2% reduction in error over the baseline, and in the case of oracle-knowledge of named-entities, this reduction improves to 21.2%.

Section 2 describes the GENIA corpus in detail. In Section 3, we present unknown word rate experiments which measure the coverage of PTB's grammar on various genres of scientific and technical writing. Section 4 describes our methods for lexical adaptation and their corresponding effects on parse accuracy. Section 5 concludes with a discussion challenges and opportunities for future work.

2 The GENIA Corpus

The GENIA corpus [1,2] consists of MEDLINE abstracts related to transcription factors in human blood cells. Version `3.02p` of the corpus includes 1999[1] abstracts (18,545 sentences, 436,947 words) annotated with part-of-speech (POS)

[1] The reported total of 2000 abstracts includes repetition of article ID 97218353.

tags and named-entities. Named-entities were labelled according to a corpus-defined ontology, and the POS-tagging scheme employed is very similar to that used in PTB (see Section 4.1).

Using these POS annotations and PTB guidelines [16], we hand-parsed 21 of these abstracts (215 sentences) to create a pilot treebank for measuring parse accuracy. We performed the treebanking using the `GRAPH`[2] tool developed for the Prague Dependency Treebank. Initial bracketing was performed without any form of automation. Following this, our baseline parser [3] was used to propose alternative parses. In cases where hand-generated parses conflicted with those proposed by the parser, hand-parses were manually corrected, or not corrected, according to PTB bracketing guidelines. Our pilot treebank is publicly available[3].

Subsequent to this, the Tsujii lab released its own beta version treebank, which includes 200 abstracts (1761 sentences) from the original corpus. This treebanking was performed largely in accordance with PTB guidelines (perhaps the most significant difference being constituent labels `NAC` and `NX` were excluded in favor of `NP`). Because there is no redundancy in the coverage of the Tsuijii lab's treebank and our own pilot treebank (and by chance, `NAC` and `NX` do not occur in our pilot treebank either), we have combined the two treebanks to maximize our evaluation treebank (see Table 3).

An additional note is required regarding our use of named-entities (Section 4.3). Entity annotations (not available in the treebank) were obtained from the earlier `3.02p` version of the corpus. Any sentences that did not match between the two versions of the corpus (due to differences in tokenization or other variations) were discarded. The practical impact of this was negligible, as only 25 sentences had to be discarded[4].

3 Unknown Words

Casual reading of technical literature quickly reveals a rich, field-specific vocabulary. For example, consider the following sentence taken from GENIA:

> The study of NF-kappaB showed that oxLDLs led to a decrease of activation-induced p65/p50 NF-kappaB heterodimer binding to DNA, whereas the presence of the constitutive nuclear form of p50 dimer was unchanged.

To quantitatively measure the size and field-specificity of domain vocabulary, we extracted the lexicon contained in WSJ sections 2-21 and evaluated the unknown word rate (by token) for various genres of technical literature. Results are given in Table 1.

[2] http://quest.ms.mff.cuni.cz/pdt/Tools/Tree_Editors/Graph

[3] http://www.cog.brown.edu/Research/nlp

[4] Because our preliminary use of named-entities assumes oracle-knowledge, this experiment was carried out on the development section only, thus only the development section was reduced in this way.

Table 1. Unknown word rate on various technical corpora given WSJ 2-21 lexion

Corpus	Unknown Word Rate
WSJ sect. 24	2.7
Brown-DEV	5.8
Brown sect. J	7.3
CRAN	10.0
CACM	10.7
DOE	16.7
GENIA	25.5

Brown-DEV corresponds to a balanced sampling of the Brown corpus (see Table 4). Section J of Brown contains "Learned" writing samples and demonstrated the highest rate of any single Brown section. CRAN contains 1400 abstracts in the field of aerodynamics, and CACM includes 3200 abstracts from Communications of the ACM [17]. DOE contains abstracts from the Department of Energy, released as part of PTB. GENIA here refers to 333 abstracts (IDs 97449161-99101008) not overlapping our treebank. As this table shows, unknown word rate clearly increases as we move to increasingly technical domains. Annecdotal evaluation on patent literature suggests its unknown rate lies somewhere between that of DOE and GENIA.

While these results appear to indicate WSJ is lexically impoverished with respect to increasingly technical domains, it was also necessary to consider the possibility that the results were simply symptomatic of technical domains having very large lexicons. If such were the case, we would expect to see these domains demonstrate high unknown word rates even in the presence of a domain-specific lexicon. To test this hypothesis, we contrasted unknown word rates on GENIA using lexicons extracted from WSJ sections 2-21, Brown (training section from Table 4), and from GENIA itself (1,333 abstracts: IDs 90110496-97445684)[5]. Results are presented in Table 2.

Table 2. Unknown word rate on GENIA using lexicons extracted from WSJ, Brown, and GENIA

Lexicon	Size	Unknown Word Rate
Brown	25K	28.2
WSJ	40K	25.5
Brown+WSJ	50K	22.4
GENIA	15K	5.3
Brown+WSJ+GENIA	60K	4.6

[5] While this set of abstracts does overlap the Tsujii treebank, this experiment was run prior to the treebank's release.

Although the unknown word rate in the presence of in-domain training for GENIA (5.3%, Table 2) is nearly twice that of out-of-domain training (2.7%, Table 1), suggesting a larger lexicon does indeed exist, it is also strikingly clear that WSJ and Brown provide almost no lexical value to the domain: expanding GENIAs lexicon by 45,000 new terms found in WSJ and Brown produced only a meager 0.7% reduction in unknown word rate. Contrast this with the enormous reduction achieved through using GENIA's lexicon instead of the WSJ or Brown lexicons (Table 2).

4 Parser Adaptation

In this section, we present three methods for parser adaptation motivated by the results of our unknown word rate experiments (Section 3). The goal of these adaptations is to help an off-the-shelf PTB-trained parser compensate for the large amount of domain-specific vocabulary found in technical literature, specifically biomedical text. To accomplish this without depending on in-domain treebank data, we consider three alternative (and less expensive) domain-specific knowledge sources: part-of-speech tags, dictionary collocations, and named-entities. We report on the results of each technique both in isolation and in combination.

We adopt as our baseline for these experiments the publicly available Charniak parser [3] trained on WSJ sections 2-21 of the Penn Treebank. Our division of the GENIA corpus into development and test sets is shown in Table 3. Analysis was carried out on the development section, and the test section was reserved for final evaluation. Parse accuracy was measured using the standard PARSEVAL metric of bracket-bracket scoring, assuming the usual conventions regarding punctuation [18]. Statistical significance for each experiment was assessed using a two-tailed paired t-test on sentence-averaged f-measure scores. Since our evaluation treebank excludes `NX` and `NAC` constituent labels in favor of `NP` (Section 2), for all experiments

Table 3. Division of the GENIA combined treebank into development and test sections

Source	Section	Abstract IDs	Sentences
Pilot	Development	99101510-99120900	215
Tsujii	Development	91079577-92060325	732
Tsujii	Test	92062170-94051535	1004

Table 4. Brown corpus division. Training and evaluation sections were obtained from Gildea [7]. The development (and final training) section was created by extracting every tenth sentence from Gildea's training corpus.

	POS-Train	Development	Test
Sentences	19637	2181	2425

Table 5. PARSEVAL f-measure scores on the GENIA development section using the adaptation methods described in Section 4. Statistical significance of individual adaptations are compared against no adaptation, and combined adaptations are compared against the best prior adaptation. As the p values indicate, all of the adaptions listed here produced a significant improvement in parse accuracy.

Adaptation	F-measure	Error reduction	Significance
none	78.3	–	–
lexicon	78.6	1.4	$p = 0.002$
no NNP	79.1	3.7	$p = 0.002$
train POS	80.8	11.5	$p < 0.001$
entities	80.9	12.0	$p < 0.001$
no NNP, train POS	81.5	14.7	$p = 0.043$
no NNP, train POS, entities	82.9	21.2	$p < 0.001$

Table 6. Final PARSEVAL f-measure results on GENIA compared with scores on Brown and WSJ sect. 23. In all cases, the parser was trained on WSJ sect. 2-21 with the over-parsing parameter set to 21x over-parsing. Adapted GENIA results includes POS adaptations only (oracle-type entity adaptation was not used). Adapted Brown results use POS re-training on Brown train section.

Corpus	F-measure	Error reduction	Significance
GENIA-unadapted	76.3	–	–
GENIA-adapted	79.6	14.2	$p < 0.001$
Brown-unadapted	83.4	–	–
Brown-adapted	84.1	4.1	$p = 0.002$
WSJ	89.5	–	–

(including baseline) we post-processed parser output to collapse these label distinctions[6]. Results from our various experiments are summarized in Table 5.

Final results of our adapted parser are given in Table 6. For comparison with standard benchmarks, parser performance was also evaluated on WSJ section 23 and on Brown. Table 4 shows our division of the Brown corpus.

4.1 Using POS Tags

Part-of-speech tags provide an important data feature to statistical parsers [3,5]. Since technical and scientific texts introduce a significant amount of domain-specific vocabulary (Section 3), a POS-tagger trained only on everyday

[6] While PTB examples could be similarly pre-processed prior to training, thereby reducing the search space while parsing, the reduction would be minor and would mean giving up a potentially useful distinction in syntactic contexts.

English is immediately at a disadvantage for tagging such text. Indeed, our off-the-shelf PTB-trained parser achieves only 84.6% tagging accuracy on GENIA. Consequently, our simple first adaptation step was to retrain the parser's POS-tagger on the 1,778 GENIA abstracts not present in the combined treebank (in addition to WSJ sections 2-21). This simple fix raised tagging accuracy to 95.9%. Correspondingly, parsing accuracy improved from 78.3% to 80.8% (Table 5).

While such POS-retraining is a direct remedy to learning appropriate tags for new vocabulary, it is only a partial fix to a larger problem. In particular, the trees found in PTB codify a relationship between PTB POS tags and constituent structure, and any mismatch between the tagging schemata used in PTB and that used by our new corpus could result in misapplication or underutilization of the bracketing rules acquired by the parser during training. To overcome this, it is necessary to introduce an additional mapping step which converts between the two POS tagging schemata. For closely related schemata, this mapping may be trivial, but this cannot be assumed without a carefully analysis of tag distribution and usage across the two corpora.

In the case of GENIA, the tagging guidelines used were based on PTB and only subsequently revised (to improve inter-annotator agreement), so while differences do exist, the problem is much less significant than the general case of arbitrarily different schemata. Reported differences include treatment of hyphenated, partial, and foreign terms, and most notably, the distinction between proper (NNP) and common (NN) nouns [2]. In order to quantitatively assess the degree to which these and other revisions were made to the tagging scheme, we extracted the POS distribution for 333 GENIA abstracts (as used in our unknown word rate experiments from Section 3). From this distribution, we learned that NNP almost never occurs in GENIA. This meant that our PTB-trained parser would be unable to leverage PTB's constituent structure examples examples that involved proper nouns.

As a preliminary remedy, we simply relabeled all proper nouns as common in PTB and re-trained the parser. This improved tagging accuracy to 96.4% and parsing accuracy to 81.5% (Table 5). We should note, however, that this solution is not ideal. While it does allow use of PTB's NNP-examples, it does so at the cost of confusing legitimate differences in the syntactic distribution of common and proper nouns in English (as reflected by a 0.7% loss in accuracy on WSJ evaluation when using this NN-NNP conflated training data). Clearly it would be better if GENIA's nouns could be re-tagged to preserve this distinction while preserving inter-annotator agreement. A first step in this direction would be to perform this re-tagging automatically based on determiner usage and GENIA's entity annotations, with success measured by the corresponding impact on parse accuracy. This, along with a more careful analysis of tagging differences, remains for future work.

We have also evaluated parser performance under the oracle condition of perfect tags. This was implemented as a soft constraint so that the parser's joint probability model could overrule the oracle tag for cases in which no parse could be found using it (cases of annotator error or data sparsity). Using the oracle tag 99.8% of

the time (in addition to other POS adaptations) had almost no impact on parse accuracy, suggesting that further POS-related improvements in parse accuracy will only come from the sort of careful analysis of the tagging schemata discussed above.

4.2 Using a Domain-Specific Lexicon

Another strategy we employed for lexical adaptation was the use of a domain-specific dictionary. For biomedicine, such a dictionary is available from the National Library of Medicine: the Unified Medical Language System (UMLS) SPECIALIST lexicon [19]. Covering both general English as well as biomedical vocabulary, the SPECIALIST lexicon contains over 415,000 entries (including orthographic and morphological variants). Entries are also assigned one of eleven POS categories specified as part of the lexicon.

Given our finding from Section 4.1 that even oracle POS tags would do little to improve upon our re-trained POS tagger, we did not make use of lexicon POS tags. Instead, we restricted our use of the lexicon to extracting collocations. We then added a hard-constraint to the parser that these collocations could not be cross-bracketed and that each collocation must represent a flat phrase with no internal sub-constituents. This approach was motivated by a couple of observations. On one hand, we observed cases where the parser would be confused by long compound nouns; in desperation to find the start of a verb phrase, it would sometimes use part of the compound to head a new verb phrase. Unfortunately, WSJ sections 2-21 contain approximately 500 verb phrases headed by present-participle verbs mistagged as nouns, thus making this bizarre bracketing rule statistically viable. A second observation was the frequency with which we saw the terms "in vivo" and "in vitro" (treebanked as foreign adjverbial or adjectival collocations) mis-analyzed. Even in biomedical texts, "in" appears far more often as a preposition than as part of such collocations, and as such, is almost always mis-parsed in these collocational contexts to head a prepositional phrase. Our hope was that by preventing such collocations from being cross-bracketted, we could prevent this class of parsing mistakes.

We found use of lexical collocations did yield a small (0.3%) but statistically significant improvement in performance over the unmodified parser (Table 5). However, when combined with either POS or entity adaptations, the lexicon's impact on parsing accuracy was statistically insignificant. Our interpretation of this latter result is that the primary limitation of the lexicon is coverage, despite its size. That is, when either of the other adaptations were used, the lexicon did not offer much beyond them. It is not surprising that oracle-knowledge of entities (Section 4.3) provided greater coverage than the generic dictionary, and the improvement in tagging from POS adaptation (sharper tag probabilities) helped somewhat in preventing the verb-ification of some of the long compound nouns. While the lexicon was the only adaptation to correctly fix "in vivo" type mistakes, these phrases alone were not sufficiently frequent to provide a statistically significant improvement in parse accuracy on top of other adaptations. As such, the primary value of this method would be in cases where such a lexicon is available but POS tags and labelled entities are not.

4.3 Using Named-Entities

The primary focus of the GENIA corpus is to support training and evaluation of automatic named-entity recognition. As such, a variety of biologically meaningful terms have been annotated in the corpus according to a corpus-defined ontology. Given the availability of these annotations, we were interested in considering the extent to which they could be used as a source of lexical information for parser adaptation.

Given the problems described earlier with regard to lexical collocations being cross-bracketted by our off-the-shelf PTB-trained parser (Section 4.2), our hope was that named-entities could be used similarly to lexical collocations in helping to prevent this class of mistakes. To put it another way, we hoped to exploit the correlation between named-entities and noun phrase (NP) boundaries. A common preprocessing step in detecting named-entities is to use a chunker to find NPs. Our approach was to do the reverse: to use named-entities as a feature for finding NP boundaries.

Our initial plan was to use the same strategy we had used with dictionary collocations: to add a hard-constraint to the parser that a named-entity could not be cross-bracketed and had to represent a flat phrase with no internal sub-constituents. However, we found upon closer inspection that the entities often did contain substructure (primarily parenthetical acronyms), and so we relaxed the flat-constituent constraint and enforced only the cross-bracketing constraint.

As a preliminary step, we evaluated the utility of this method using oracle-knowledge of named-entities. By itself, this method was roughly equivalent to POS re-training in improving parsing accuracy from 78.3% to 80.9% (Table 5). But when combined with POS adaptations, use of named-entities provided another significant improvement in performance, from 81.5% to 82.9%. Clearly this is a promising avenue for further work, and it will be interesting to see how much of this benefit from the oracle case can be realized when using automatically detected entities.

5 Discussion

We have found only limited use of parsing reported to date for biomedical literature, thus it is difficult to compare our parsing results against previous work in parsing this domain. To the best of our knowledge, only one other wide-coverage parser has been applied to biomedical literature: Grover et al. report 99% coverage using a hand-written grammar with a statistical ranking component [20]. We do not know of any quantitative accuracy figures reported for this domain other than those described here.

For those interested in mining the biomedical literature, the next important step will be assessing the utility of PTB-style parsing compared to other parsing models that have been employed for information extraction. There has been promising work in using PTB-style parses for information extraction by inducing predicate-argument structures from the output parses [21]. It will be interesting to see for the biomedical domain how these predicate-argument structures compare to those induced by other grammar formalisms currently in use, such as HPSG [22].

The next immediate extension of our work is to evaluate use of detected named-entities in place of the oracle case described in Section 4.3, replacing the current hard-constraint with a soft-constraint confidence term to be incorporated into the parser's generative model. Performance of named-entity recognition on GENIA was recently studied as part of a shared task at BioNLP/NLPBA 2004. The best system achieved 72.6% f-measure [23], though note that this task required both detection and classification of named-entities. As our usage of entities does not require classification, this number should be considered a lower-bound in the context of our usage model. We expect this level of accuracy should be sufficient to improve parse scores, though how much of the oracle benefit we can realize remains to be seen.

There are also interesting POS issues meriting further investigation. As discussed in Section 4.1, we would like to find a better solution to the lack of proper noun annotations in GENIA, perhaps by detecting proper nouns using determiners and labelled entities. More careful analysis of the differences between the PTB and GENIA tagging schemata is also needed. Additionally, there are interesting issues regarding how POS tags are used by the parsing model. Whereas the Collins' model [5] treats POS tagging as an external preprocessing step (a single best tag is input to the parsing model), the Charniak model [3] generates tag hypotheses as part of its combined generative model, and thus considers multiple hypotheses in searching for the best parse. The significance of this is that other components of the generative model can influence tag selection, and Charniak has reported adding this feature to his simulated version of the Collins model improved its accuracy by 0.6% [24]. However, this result was for in-domain evaluation; the picture becomes more complicated when we begin parsing out-of-domain. If we have an in-domain trained POS-tagger, we might not want a combined model trained on out-of-domain data overruling our tagger's predictions. One option may be introducing a weighting factor into the generative model to indicate the degree of confidence assigned to our tagger relative to the other components of the combined model.

Another issue for further work is the parsing of paper titles. In the GENIA development section, only 28% of the titles are sentences whereas 71% are noun phrases. This distribution is radically different than the rest of the corpus, which is heavily dominated by sentence-type utterances. As headlines are even more rare in our WSJ training data than titles are in GENIA (since WSJ contains full article text), our parser performs miserably at utterance-type detection (i.e. correctly labelling the top-most node in the parse tree): 58.6%. Correspondingly, parse accuracy on titles is only 69.1%, which represents a statistically significant decrease in accuracy in comparison to the entire development section ($p = 0.038$). In investigating this, we noticed an oddity in GENIA in that most titles were encoded in the corpus with an ending period that did not exist in the original papers the corpus was derived from. By removing these periods, we improved utterance-type detection to 77.9%. While parse accuracy rose to 72.0%, this was statistically insignificant ($p = 0.082$). The solution we would like to move towards is to respect the legitimate distributional differences between title and

non-title utterances and parameterize the parser differently for the two cases. Generally speaking, such "contextual parsing" might allow us to improve parsing accuracy more widely by parameterizing our parser differently based on where the current utterance fits in the larger discourse. This example of period usage in titles also highlights a broader issue that seemingly innocuous issues in corpus preparation can have significant impact when parsing. As a further example of this, the choice to (at times) separately tokenize term-embedded parentheses in GENIA creates unnecessary attachment ambiguity in the resulting parenthetical phrases. For example, in the phrase "C3a and C3a(desArg)", "C3a(desArg)" is tokenized as "C3a (desArg)", which produces ambiguity as to whether the parenthetical should attach low (to the latter "C3a") or high (to the compound "C3a and C3a"). Issues such as these remind us to be mindful of the relationship between corpus preparation and parsing, as well as downstream processing, and that some issues which appear difficult to resolve while parsing might be handled more easily at another stage in the processing pipeline.

We view biomedical and other technical texts as providing an interesting set of challenges and questions for future parsing research. An interesting introduction to some of these challenges, supported by examples drawn from the domain, can be found in [25]. A significant question for consideration is the degree to which these challenges are related to domain knowledge vs. stylistic norms of the genre. For example, [2] reports that whereas POS determination required domain expertise, prepositional phrase (PP)-attachment could be largely determined even by non-biologists. Our own treebanking experience left us with the opposite impression. For example, in the phrase "gene expression and protein secretion of IL-6", should the PP attach high (IL-6 *gene expression and protein secretion*) or low (gene expression and IL-6 *protein secretion*)? Domain knowledge appears to be necessary here for correct resolution. In contrast to this, POS tags appear to be a distributional rather than a semantic concern. Issues like this highlight how little we really understand currently about the parameters of corpus variation. How do the frequencies of different syntactic constructions vary by genre, and are there key structural variations at work? How do we effectively adapt parsers in response? These issues remain important topics for future investigation.

References

1. Kim, J.d., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpus - a semantically annotated corpus for bio-textmining. Bioinformatics (Supplement: Eleventh International Conference on Intelligent Systems for Molecular Biology) **19** (2003) i180–i182
2. Tateisi, Y., Ohta, T., dong Kim, J., Hong, H., Jian, S., Tsujii, J.: The genia corpus: Medline abstracts annotated with linguistic information. In: Third meeting of SIG on Text Mining, Intelligent Systems for Molecular Biology (ISMB). (2003)
3. Charniak, E.: A maximum-entropy-inspired parser. In: Proc. NAACL. (2000) 132–139
4. Marcus, M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19** (1993) 313–330

5. Collins, M.: Discriminative reranking for natural language parsing. In: Proc. ICML. (2000) 175–182
6. Ratnaparkhi, A.: Learning to parse natural language with maximum entropy models. Machine Learning **34** (1999) 151–175
7. Gildea, D.: Corpus variation and parser performance. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. (2001) 167–202
8. Roark, B., Bacchiani, M.: Supervised and unsupervised pcfg adaptation to novel domains. In: Proceedings of HLT-NAACL. (2003) 205–212
9. Steedman, M., Hwa, R., Clark, S., Osborne, M., Sarkar, A., Hockenmaier, J., Ruhlen, P., Baker, S., Crim, J.: Example selection for bootstrapping statistical parsers. In: Proceedings of HLT-NAACL. (2003) 331–338
10. de Bruijn, B., Martin, J.: Literature mining in molecular biology. In: Proceedings of the European Federation for Medical Informatics (EFMI) Workshop on Natural Language Processing in Biomedical Applications. (2002)
11. Hirschman, L., Park, J., Tsujii, J., Wong, L., Wu, C.: Accomplishments and challenges in literature data mining for biology. Bioinformatics **18** (2002) 1553–1561
12. Yakushiji, A., Tateisi, Y., Miyao, Y., Tsujii, J.: Event extraction from biomedical papers using a full parser. In: Pacific Symposium on Biocomputing. (2001) 408–419
13. Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., Mazo, I.: Extracting human protein interactions from medline using a full-sentence parser. Bioinformatics **20** (2004) 604–611
14. Shatkay, H., Feldman, R.: Mining the biomedical literature in the genomic era: An overview. Journal of Computational Biology **10** (2003) 821–855
15. Hwa, R.: Learning Probabilistic Lexicalized Grammars for Natural Language Processing. PhD thesis, Harvard University (2001)
16. Bies, A., Ferguson, M., Katz, K., MacIntyre, R.: Bracketting Guideliness for Treebank II style Penn Treebank Project. Linguistic Data Consortium. (1995)
17. Buckley, C.: Implementation of the smart information retrieval system. Technical Report 85-686, Cornell University (1985)
18. Goodman, J.: Parsing inside-out. PhD thesis, Harvard University (1998)
19. McCray, A.T., Srinivasan, S., Browne, A.C.: Lexical methods for managing variation in biomedical terminologies. In: Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care (SCAMC). (1994) 235–239
20. Grover, C., Lapata, M., Lascarides, A.: A comparison of parsing technologies for the biomedical domain. Journal of Natural Language Engineering (2002)
21. Surdeanu, M., Harabagiu, S., Williams, J., Aarseth, P.: Using predicate-argument structures for information extraction. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03). (2003) 8–15
22. Miyao, Y., Ninomiya, T., Tsujii, J.: Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. In: Proc. of IJCNLP-04. (2004) 684–693
23. Zhou, G., Su, J.: Exploring deep knowledge resources in biomedical name recognition. In: Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04). (2004)
24. Charniak, E.: Statistical parsing with a context-free grammar and word statistics. In: Proceedings of the Fourteenth National Conference on Artificial Intelligence, Menlo Park, AAAI Press/MIT Press (1997)
25. Park, J.C.: Using combinatory categorical grammar to extract biomedical information. IEEE Intelligent Systems **16** (2001) 62–67