PERFORMANCE OF THE CHARNIAK-LEASE PARSER ON BIOLOGICAL TEXT USING DIFFERENT TRAINING CORPORA

ALISON CALLAHAN
Department of Biology
Carleton University, 1125 Colonel By Drive
Ottawa ON K1S5B6 Canada

MICHEL DUMONTIER*
Department of Biology, School of Computer Science
Carleton University, 1125 Colonel By Drive
Ottawa ON K1S5B6 Canada

September 18, 2008

POS tagging is used as the first step in many NLP workflows, although the accuracy of tag assignment frequently goes unchecked. We hypothesize that changing the training corpora for a parser will affect its POS tagging of a target corpus. To this end we train the Charniak-Lease parser on the WSJ corpus and two biomedical corpora and evaluate its output to MedPost, a POS tagger with a reported 97% accuracy on biomedical text. Our findings indicate that using biomedical training corpora significantly improves performance, but that minor differences in the biomedical training corpora have a significant effect on the correctness of POS tagging. Specifically, the tagging of hyphenated words and verbs was affected. This work suggests that the choice of training corpora is crucial to domain targeted NLP analysis.

1. Introduction

The majority of natural language processing workflows utilize a parts-of-speech tagger and/or syntactic parser to identify the complex syntactical structure of corpora (*e.g.* [1][2]). POS tagging makes it possible to establish dependencies between words in a phrase by virtue of their syntax, making information extraction more tractable. As more of these tools become available, and as they become tailored to specific domains such as biology, it is of interest to compare their performance against one another to address their usability.

Despite an apparent awareness of the problem, it remains a common practice to present novel methods for extracting information from text that amalgamate a variety of simple tools whose individual performance goes unchecked. As an example, the methodology of Rindflesch and Fiszman [3] relies on accurate syntactic structure assignment in order to perform semantic interpretation. Unlike intrinsic evaluation, in which the performance of an NLP

* Corresponding author: michel_dumontier@carleton.ca

tool is evaluated in its ideal application, the methodology of performance evaluation in different domains [4] remains an open problem. While overall NLP performance is commonly evaluated, the accuracy of the part-of-speech (POS) component is frequently disregarded in the analysis. Nonetheless, it has been recognized that the challenge of successful parsing lies in the requirement of correct consecutive 'decisions' to be made by the parser [4]. Since POS tagging is one of the first 'decisions' to be made, this is a critical aspect of information extraction. However, little work has been done to evaluate the correctness of POS tagging by parsers trained on different corpora.

A well known POS tagger is the Charniak-Lease Parser [5]. The Charniak-Lease parser is by default trained on the Wall Street Journal (WSJ) corpus, but users have the option to train the parser on a different corpus to better match a specific domain. MedPost [6] is another parser trained specifically on a biological corpus derived from MEDLINE abstracts that has a reported 97% accuracy.

In this work, we compare POS tagging of MedPost to the Charniak-Lease parser trained on two different biomedical corpora, the PennBioIE oncology corpus and MedPost training corpus, and the non-science related Wall Street Journal corpus. We explore how domain specific and general terminology are treated and how this information may serve to guide the choice of POS components used as part of NLP workflows.

2. Methods

An overview of the analysis workflow is illustrated in

Figure 1. Briefly, the Charniak-Lease Parser was trained using three training corpora, and each was subsequently used to assign POS to a domain-specific gold standard corpus. The resulting POS tagging from each was then compared to that generated by MedPost.

2.1 Training the Charniak-Lease parser

We developed corpus-trained parsers using the Charniak-Lease parser (CLP) framework. The first parser, **WParser**, was the result of training the CLP with the Wall Street Journal corpus [7], termed WCorpus, and this is the default training of the CLP as available. The second parser, **PParser**, was the result of training CLP with the PennBioIE oncology corpus (release 0.9) [8], termed PCorpus. Grammatical function suffixes (e.g. NP-**SBJ)** were removed from PCorpus because they are not part of CLP's default tag set. The third parser, **MParser**, was the result of training CLP with 5700 (manually) tagged sentences used to train MedPost[1], termed MCorpus. Since MCorpus has parts-of-speech assigned but not syntactical structure, we processed the MCorpus using the Collins parser [9], which generates parse trees from POS-tagged text. Minimal processing was required to correct minor bracketing errors (e.g. double brackets ((NP...).

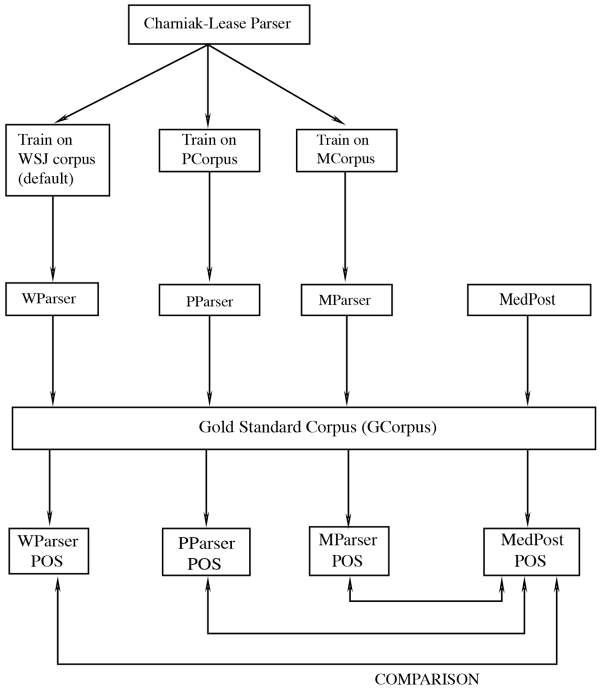


Figure 1. An overview of our workflow.

2.2 A reference gold standard corpus

A corpus of 2465 abstracts, termed GCorpus, was constructed from the PubMed database using the search terms 'HIF', 'HIF-1', 'HIF-1alpha', 'hypoxia', and 'hypoxia-inducible factor'. Hypoxia-inducible factor 1 (HIF-1) is a transcription factor that plays an essential role in cellular and systemic homeostatic responses

[1] MedPost and its corpus of training sentences are available at http://www.ncbi.nlm.nih.gov/staff/lsmith/MedPost.html

to hypoxia. The GCorpus was tagged using MedPost and used to evaluate (by comparison) the output of the trained CLP parsers. While MedPost may be used with different POS tag sets, the PTB tag set was selected as this is the only set supported by CLP.

2.4 Evaluating parser output

Each of the trained CLP parsers was used to generate POS and syntactic phrase structure (*i.e.* noun phrases and verb phrases) for the GCorpus. Since MedPost only assigns POS, the syntactic information was not examined. The abstract number, word, and tag for each parser were extracted from the result files and stored in a MySQL database. Queries against the MySQL database were designed to identify differences in tagging.

3. Results

All POS tagging of the GCorpus by the trained parsers was compared to MedPost tagging. The results presented focus on the difference such that a reported percentage of 10.0% means that 10 out of 100 tags assigned by the trained parser differed from the tag assigned by MedPost while the remainder (90/100) was identically tagged. 'Words' refer to words in a corpus, while 'tags' refer to the assigned POS tag (such as NN for noun).

3.1 Characteristics of the biomedical domain training corpora

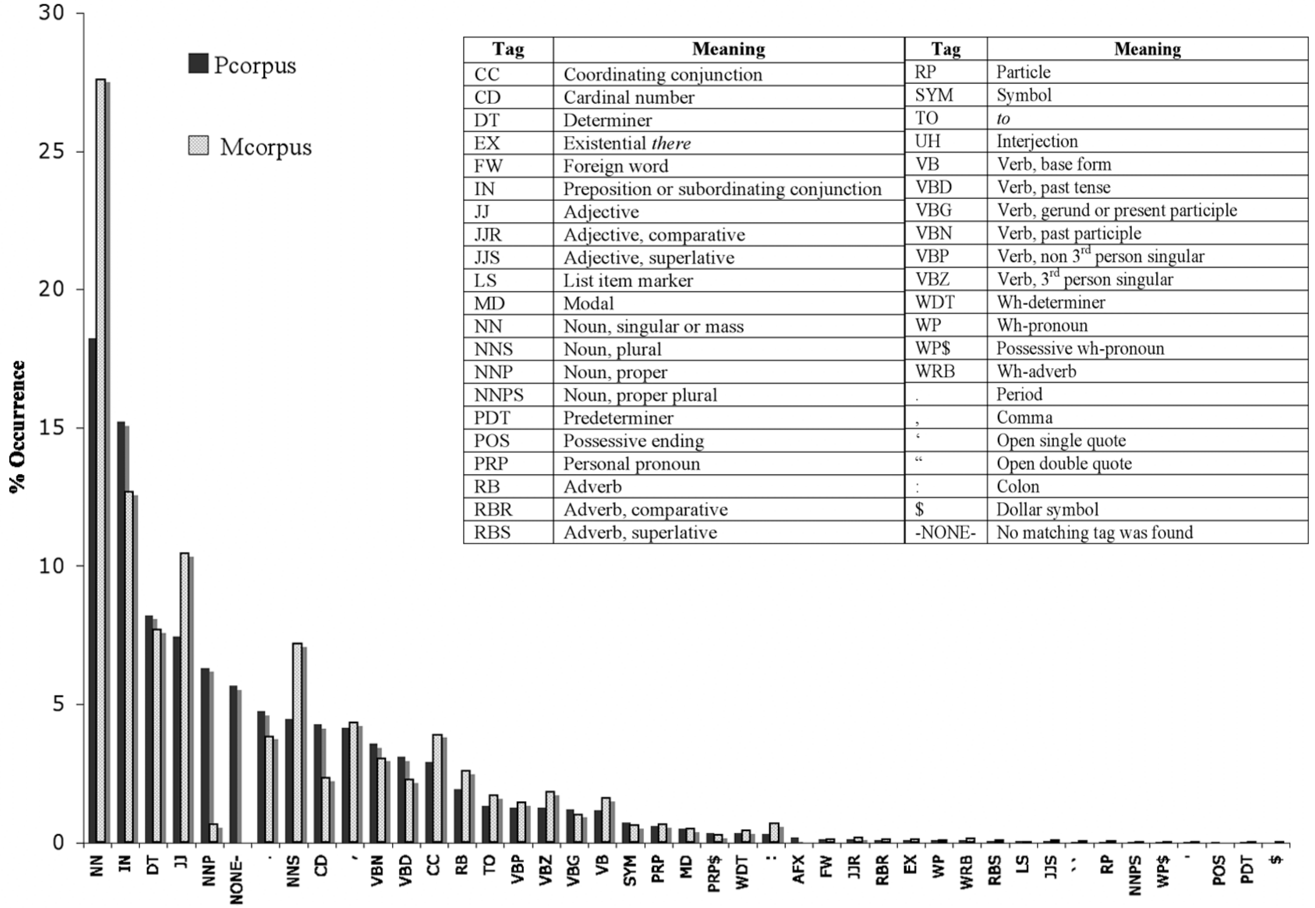
MCorpus contains more words tagged as nouns, adjectives, plural nouns, coordinating conjunctions, TO, 3rd person singular verbs, and non 3rd person singular verbs than the PCorpus (Figure 2).

Figure 3 shows the differences between the two biomedical corpora in terms of word occurrence and uniqueness. MCorpus, with 8925 unique words, is more diverse than PCorpus, with 5968 unique words. 55.4% of words occur only once in the MCorpus, while 50.3% of words in PCorpus occur once. Slightly more than 14% of words in both corpora occur with a frequency of 2. A larger percentage of words occur with a frequency of 3-5 and 6-10 in the PCorpus (15.9% and 8.5%) than the MCorpus (14.1% and 7.0%).

The WCorpus was available only by purchase, and so a similar analysis was not possible.

3.2 Training corpora affects **domain-specific** *POS tagging*

POS tagging by CLP trained with the WCorpus, PCorpus and MCorpus differed particular, the WParser generated significantly more tagging differences than



Tag	Meaning	Tag	Meaning
CC	Coordinating conjunction	RP	Particle
CD	Cardinal number	SYM	Symbol
DT	Determiner	TO	*to*
EX	Existential *there*	UH	Interjection
FW	Foreign word	VB	Verb, base form
IN	Preposition or subordinating conjunction	VBD	Verb, past tense
JJ	Adjective	VBG	Verb, gerund or present participle
JJR	Adjective, comparative	VBN	Verb, past participle
JJS	Adjective, superlative	VBP	Verb, non 3rd person singular
LS	List item marker	VBZ	Verb, 3rd person singular
MD	Modal	WDT	Wh-determiner
NN	Noun, singular or mass	WP	Wh-pronoun
NNS	Noun, plural	WP$	Possessive wh-pronoun
NNP	Noun, proper	WRB	Wh-adverb
NNPS	Noun, proper plural	.	Period
PDT	Predeterminer	,	Comma
POS	Possessive ending	‘	Open single quote
PRP	Personal pronoun	“	Open double quote
RB	Adverb	:	Colon
RBR	Adverb, comparative	$	Dollar symbol
RBS	Adverb, superlative	-NONE-	No matching tag was found

Figure 2. The distribution of POS tags in PCorpus and MCorpus, and their meaning.

significantly with respect to the tags assigned by MedPost (Table 1). In both PParser and MParser, with at least twice the number of discrepancies in every examined case except ‘hypoxia-inducible’. In the case of ‘hypoxia-inducible’, WParser exhibited far less disagreement with MedPost overall, but exhibited a larger number of mismatches for hypoxia/HIF terminology. In contrast, MParser yielded the least different tagging for hypoxia (0.16%) and HIF1-alpha (0.77%) as compared to 12.5% and 6.2% for PParser**.**

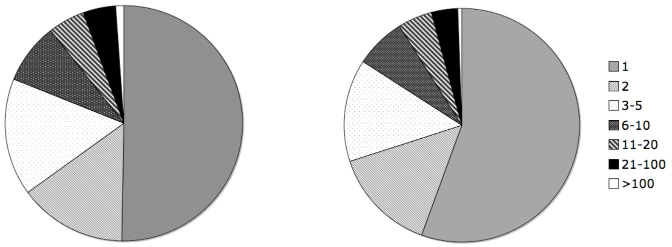


Figure 3. The distribution of unique words in the PCorpus (left) and MCorpus (right), representing the percentage of words that occur with the specified frequency.

Table 1: Differences in POS assignment by trained CLP to MedPost: percentage of tags that differ (actual counts) and total word frequency.

Word set	**WParser**	**PParser**	**MParser**	**Word Count**
All words	**24.1%** (121,247)	**12.0%** (60,235)	**12.6%** (63,529)	502,854
'hypoxia'	**36.7%** (1,336)	**12.5%** (455)	**0.2%** (6)	3,643
‘HIF’	**66.4%** (509)	**6.1%** (47)	**3.7%** (28)	766
‘HIF1’	**92.8%** (65)	**8.6%** (6)	**12.9%** (9)	70
‘HIF-1’	**86.5%** (2225)	**4.6%** (119)	**3.8%** (97)	2,571
‘HIF1alpha’	**56.9%** (116)	**2.5%** (5)	**1.5%** (3)	204
‘HIF-1alpha’	**18.1%** (799)	**6.2%** (273)	**0.8%** (34)	4406
‘hypoxia-inducible’	**1.4%** (38)	**17.4%** (446)	**11.8%** (304)	2,570

In the 6 cases in which the tag assigned to ‘hypoxia’ by MParser differed with respect to the correctly tagged MedPost noun (NN) assignment, MParser tagged it as the following: NNS (2), IN (2), PRP (1), VBP (1) (see Figure 2 for tag meanings).

3.3 Gold standard contains incorrect POS

We manually verified a limited number of cases in which the GCorpus contained incorrect POS. We found 12 cases (0.4%) in which MParser correctly assigned the word 'hypoxia-inducible' as an adjective, but PParser incorrectly assigned 4 of these, and MedPost incorrectly assigned these as a noun. While we expect that further cases of incorrect POS are contained within the gold standard, these are to likely occur at a level that should not greatly change our overall results.

3.4. Parser performance on the general biological domain

We determined the similarity of tagging by the trained parsers for words that are found in the training corpora and are widely used in the general biological domain. We found highly variable tagging of domain specific words by the trained parsers. In some cases, POS did not appear to be affected by differences between biomedical training corpora. For instance, there was complete agreement between the PParser, MParser and MedPost for the words ‘transcription’, ‘cell’, ‘culture’, ‘tissue’ and ‘experimental’. The WParser exhibited complete agreement with MedPost in all of the above cases **except for ‘transcription’**, for which it differed from MedPost in 291 of 1,374 cases. In these cases, WParser tagged ‘transcription’ as a foreign word, a proper noun, a non 3rd person singular present verb, or an adjective while MedPost tagged it as a noun. In contrast to the above agreement between PParser and MParser, tagging differences did exist for certain words, including those listed in Table 2. Of these, MParser exhibited greater agreement for the word ‘binding’, whereas the PParser exhibited greater agreement with MedPost for the words ‘induced’, ‘conditions’, ‘factor’ and ‘results’. WParser exhibited intermediate agreement for ‘conditions’, ‘factor’ and ‘results’ and a greater number of tagging differences for both ‘binding’ and ‘induced’.

Table 2: Differences in POS for words in the biological domain by trained CLP with MedPost.: percentage of tags (actual count) and the total occurrence.

Word	WParser	PParser	MParser	Total
‘binding’	**75.2%** (519)	**57.0%** (393)	**28.8%** (199)	690
‘induced’	**40.2%** (311)	**24.2%** (187)	**27.0%** (209)	774
‘conditions’	**0.55%** (2)	**0%** (0)	**0.89%** (3)	366
‘factor’	**0.91 %** (29)	**0%** (0)	**1.2%** (39)	3202
‘results’	**9.5%** (75)	**6.7%** (54)	**12.1%** (95)	787

The presence of special characters such as hyphens also affected POS assignment. Table 3 shows the differences in tagging by each of the three differently trained parsers for hyphenated words such as the selected non-hypoxia-related terms ‘up-regulates’, ‘helix-loop-helix’ and ‘wild-type’ (see Table 1 for other examples including ‘HIF-1’). MParser had a significantly higher tagging agreement to MedPost than PParser and WParser for hyphenated words, including ‘HIF-1’ and ‘HIF-1alpha’. While the WParser had a much larger number of tagging differences for hyphenated words overall, the agreement between its tagging of ‘up-regulates’ and that of MedPost was better than both PParser and MParser. These results indicate that the presence of a hyphen affects tagging.

Table 3: Differences in POS tagging for hyphenated words: percentage of tags (actual counts) and total word occurrence.

Word set/ term	**WParser**	**PParser**	**MParser**	**Total**
hyphenated words	**40.35%** (10,812)	**14.5%** (3,886)	**9.8%** (2,625)	26,796
‘up-regulates’	**62.2%** (23)	**100%** (37)	**70.3%** (26)	37
‘helix-loop-helix’	**70.6%** (48)	**100%** (68)	**1.47%** (1)	68
‘wild-type’	**93.8%** (167)	**100%** (180)	**0%** (0)	180

MParser tagged the hyphenated terms ‘up-regulates’, ‘helix-loop-helix’ and ‘wild-type’ as nouns (NN) or plural nouns (NNS) in the majority of cases (and occasionally as an adverb (RB)), while PParser consistently tagged ‘up-regulates’ as a singular 3rd person present tense verb (VBZ) and ‘helix-loop-helix’ and ‘wild-type’ as adjectives (JJ). WParser again displayed an intermediate performance, tagging more than half of the instances of each word correctly and in agreement with PParser, and the remainder incorrectly and in agreement with MedPost. Table 4 summarizes this information.

Table 4: POS assignment for sample hyphenated words.

Word	**WParser**	**PParser**	**MParser**	**MedPost**
‘up-regulates’	JJ (8) NNS (13) VBZ(16)	VBZ (37)	NN (22) NNS(12) RB(2) JJ (1)	NNS (32) JJ (3) NN (1) RB (1)
‘helix-loop-helix’	JJ (47) NN (20) NNP (1)	JJ (68)	NN (68)	NN (67) JJ (1)
‘wild-type’	JJ (162) NN(13) FW(4) NNP (1)	JJ (180)	NN (180)	NN (180)

4. Discussion

POS tagging by a trained parser is intimately related to its training corpus. The differences in tagging agreement between the WParser trained on non-biomedical text, and the two parsers PParser and MParser trained on biomedical corpora provide evidence for the necessity of domain-specific training corpora for parsers. The differences in tagging performance between PParser and MParser indicate that less extreme differences in domain-specific corpora can signficantly affect tagging performance. For example, identical words occurring in similar phrases may result in identical tag assignment, even though the assignment is incorrect. Closer examination of the POS assigned to words in the hypoxia domain by a trained parser **as compared to** MedPost POS provided insight into their differences. For example, the two cases in which MParser incorrectly tagged 'hypoxia' as a preposition occurred in similar phrases where the preceding word was a variation of 'mimic':

*Case 1: "Induction generated by the addition of cobalt ion (this treatment mimics **hypoxia**) was also inhibited by SNP (IC50 = 2.5 microM)."*

*Case 2: "We show here in both pancreatic and prostate carcinoma cell lines cobalt chloride (used to mimic **hypoxia**) -induced VEGF expression requires Src activation and leads to increased steady-state levels of HIF-1alpha and increased phosphorylation of signal and transducer of transcription 3 (STAT3)."*

In contrast, 'hypoxia' was correctly tagged as a noun by PParser, except where it was tagged as a foreign word (FW) in Case 2. This indicates that the tagging inconsistencies are due to a difference in the training corpora as opposed to the context in which the term is situated in a parsed sentence.

In the two cases where 'hypoxia' was incorrectly tagged as a plural noun, it was situated between a pair of brackets and preceded by an adjective:

Case 3: "(chemical hypoxia)"
Case 4: "(relative hypoxia)"

In the case where 'hypoxia' was tagged as a personal pronoun, we identified a bracketing error (double bracket), which when removed led to the correct tag assignment.

In the majority of cases MParser tagging was most similar to that of MedPost. Moreover, the correctness of the its POS assignment was increased. For instance, in agreement with MedPost, 'hypoxia' was most consistently tagged as a noun by MParser (unlike PParser and WParser).

We observed a discrepancy in the tagging of hyphenated words which appears related to the training corpus. PParser consistently and correctly tagged 'up-regulates' as a present tense 3rd person singular verb (VBZ), whereas MParser **and** MedPost both incorrectly tagged 'up-regulates' as a plural noun, an adjective or an adverb. WParser correctly tagged 'up-regulates' as a verb in more than half of cases, but also tagged it incorrectly as a plural noun or an adjective. WParser's behaviour was similar for 'helix-loop-helix' and 'wild-type', assigning both correct and incorrect tags. The consistent tagging of 'up-regulates' as a noun by MParser indicates that hyphenated words were generally tagged as a noun in the MCorpus. In fact, none of the words tagged as 'VBZ' are hyphenated in either the MCorpus or the PCorpus (Table 5). It may be the case that a different set of word-association probabilities govern the correct assignment of verb tags, such as the 'es' ending in conjunction with being followed by a noun. The percentage of non-verb hyphenated words is also greater in MCorpus than PCorpus. This is also true of words tagged as singular 3rd person present tense verb (VBZ), however the difference is greater for hyphenated words than words tagged as 'VBZ'. These values alone are not responsible for the behaviour of the parser, however, as a quick check reveals that the majority of hyphenated words are tagged as some form of noun in MCorpus, while only slightly more than half of hyphenated words are tagged as nouns in the PCorpus. It is more likely that the tags assigned in the vicinity of the word of interest play the most significant role in tagging. The difference in tagging between 'HIF1' and 'HIF-1', however, indicates that the presence of a hyphen affects tagging. Although this is but one example, in a more general sense these results indicate that great care must be taken when selecting a training corpus to either be consistent with the type of text that is to be parsed, or that is general enough to be accurate for the majority of grammatical cases encountered.

Table 5: Words tagged 'VBZ' and hyphenated words in the two training corpora

Training corpus	**Total # words**	**Singular 3rd person present tense verb**	**Hyphenated words**
PCorpus	58,427	1.23% (718)	2.16% (1,261)
MCorpus	64,941	1.81% (1,177)	3.50% (2,275)

Although the MCorpus contains a broad terminology covering the biological domain, 9 instances of 'hypoxia' exist out of the 326,242 words. In contrast, neither the PCorpus nor the WCorpus contain the term 'hypoxia' or any highly relevant words. This supports the idea that the presence of a target word in the training corpus may have a dramatic effect in proper POS tagging by the parser. The performance of the WParser, which disagreed with MedPost with greater frequency and magnitude and was trained on a corpus in the economic/business domain, also supports this conclusion.

4.2 The relevance of our analysis to the application of text mining tools in the biological domain

Several research projects have already used the CLP or MedPost in the analysis of text or as part of natural language processing workflows. MedPost has been used to process text prior for more sophisticated evaluation of dependency parsers and of behaviour of CLP [10]. However, no explicit evaluation and presentation of results regarding the POS tagging accuracy was reported. While the authors recognized that incorrect POS tags affected the performance of CLP the effect of training corpus (they relied on a subset of the GENIA corpus [11]), this issue was not addressed. The results of our work suggest that the output of both CLP trained on different corpora and MedPost can be both variable and incorrect. Similarly, MedScan [12], is used in extracting information on human protein interactions from MEDLINE abstracts and utilizes a syntactic parser as a central component of the text processing system. However, there is no evaluation of its accuracy or reliability. Our results suggest that MedScan accuracy over a target corpora will be related to the presence and correct tagging of relevant terminology in MedPost's training corpora.

4.3 Conclusions

In this work, we demonstrate how POS tagging is affected by training corpora. Parsers trained biomedical corpora performed significantly better than a parser trained on text from the Wall Street Journal, justifying the use of domain-specific training corpora. Surprisingly, we find that significant differences exist between the POS tagging of MedPost and CLP using the same training corpora. Differences in performance are related to both punctuation (given the example of hyphenation) and word type. Furthermore, our analysis revealed that MedPost itself makes tagging errors, largely based on omissions in the training corpora. As such, at the very least, the output of every processing step in an NLP pipeline must be subject to evaluation for the purpose of ensuring accurate output and/or accounting for potential errors. This is especially important given the increasing use of these tools for information and relation extraction *e.g.* [13], and the relevance of this process to complex and increasingly prevalent tools that bridge the gap between syntax and semantic meaning.

4.4 Future Directions

CLP is only one example of an NLP tool that has the potential to affect the output of text mining workflows in the biological domain. Other parsers include the Bikel [14] and the Collins [9] parser, both of which could be subject to a similar analysis to that presented here. Moreover, our analysis only touched upon the performance of the embedded POS tagger of CLP, and not the quality of the parse trees themselves. This may prove to be the subject useful additional analysis that would elucidate the relationship between POS tagging and syntactic parsing. A more thorough examination of the MedPost performance and how closely it is related to changes in training corpora is an appropriate goal given the tagging inconsistencies revealed by our analysis. The degree of ‘enrichment’ afforded by the use of domain-specific training corpora as compared to general corpora, and subsequently how to predict ideal training corpora, could also be subject to further study.

4.5 Acknowledgements

We would like to thank Leo Ferres for valuable discussions. This work is supported by an NSERC USRA to AC and an NSERC DG to MD.

5. References

[1] K. Fundel, R. Kuffner, and R. Zimmer, *Bioinformatics* **23(3)**, 365 (2007).

[2] M. Huang, X. Zhu, and M. Li, *International Journal of Medical Informatics* **75**, 443 (2006)

[3] T. Rindflesch and M. Fiszman, *Journal of Biomedical Informatics* **36**, 462 (2003)

[4] T. Kakkonen and E. Sutinen, *Lecture Notes in Artificial Intelligence* **4139**, 704 (2006)

[5] M. Lease and E. Charniak, *Lecture Notes in Artificial Intelligence* **3651**, 58 (2005)

[6] L. Smith, T. Rindflesch, and W. Wilbur *Bioinformatic*s **20(14)**, 2320 (2004)

[7] J. Garofalo, D. Graff, D. Paul, and D. Pallet (http://www.ldc.penn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S6A)

[8] “Mining the Bibliome” PennBioIE Release 0.9 (http://bioie.ldc.upenn.edu/publications/latest_release/)

[9] M. Collins, PhD University of Pennsylvania (1999)

[10] A. Clegg and A. Shepherd, *BMC Bioinformatics* 8:24 (2007)

[11] GENIA Treebank Beta Version http://wwwtsujii.is.s.u-tokyo.ac.jp/GENIA

[12] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin and I. Mazo, *Bioinformatics* **20(5)**, 604 (2004)

[13] C. Ramakrishnan, K. Kochut, and A. Sheth, *Lecture Notes in Computer Science* **4273**, 583 (2006)

[14] D. Bikel (http://www.cis.upenn.edu/~dbikel/software.html)