Lexical Adaptation of Link Grammar to the Biomedical Sublanguage: a Comparative Evaluation of Three Approaches

Sampo Pyysalo, Tapio Salakoski, Sophie Aubin, Adeline Nazarenko

Lexical Adaptation of Link Grammar to the Biomedical Sublanguage: a Comparative Evaluation of Three Approaches

Sampo Pyysalo and **Tapio Salakoski**
Turku Centre for Computer Science (TUCS)
and University of Turku
Lemminkäisenkatu 14 A,
FIN-20520 Turku, Finland

Sophie Aubin and **Adeline Nazarenko**
LIPN
Université Paris 13 & CNRS UMR 7030
99, av. J.-B. Clément,
F-93430 Villetaneuse, France

Abstract

We study the adaptation of Link Grammar Parser to the biomedical sublanguage with a focus on domain terms not found in a general parser lexicon. Using two biomedical corpora, we implement and evaluate three approaches to addressing unknown words: automatic lexicon expansion, the use of morphological clues, and disambiguation using a part-of-speech tagger. We evaluate each approach separately for its effect on parsing performance and consider combinations of these approaches. In addition to a 45% increase in parsing efficiency, we find that the best approach, incorporating information from a domain part-of-speech tagger, offers a statistically significant 10% relative decrease in error. The adapted parser is available under an open-source license at `http://www.it.utu.fi/biolg`.

1 Introduction

In applying general parsers to specific domains, adaptation is often necessary to achieve high parsing performance (see e.g. (Sekine, 1997)). Sublanguage is defined by Grishman (2001) as a specialized form of a natural language that is used within a particular domain or subject matter. It is characterized by specialized vocabulary, semantic relationships, and in many cases syntax.

In this paper, we study lexical adaptation, that is, adaptation addressing the specialized vocabulary. This is an important part of the process of customizing a general parser to a sublanguage. Among other issues, the unknown word rate increases dramatically when moving from general language to increasingly technical domains such as that of biomedicine (Lease and Charniak, 2005). This can lead to increased ambiguity, reduced parsing performance, and errors in establishing the correct relationships between words for semantic mining (Pyysalo et al., 2006).

Until recently, Information Extraction (IE) systems for mining semantic relationships from texts of technical sublanguages avoided full syntactic parsing. The quality of parsing has a well-established effect on the performance of IE systems, and the accuracy of general parsers in technical domains is comparatively low. Additionally, many domain-specific parsers lack portability to a new domain. Finally, the time required for full parsing is also a problem for IE systems. But the biomedical IE community now faces limitations in pattern-matching (Blaschke et al., 1999) and shallow parsing (Pustejovsky et al., 2002) methods that are inefficient in the processing of long distance dependencies and complex sentences. Recent advances in parsing techniques have further created an increased interest in the adaptation of full parsers.

Here, we consider the lexical adaptation of a full parser, the Link Grammar Parser[1] (LGP) of Sleator of Temperley (1991). The choice of parser addresses the recent interest in LGP in the biomedical IE community (Ding et al., 2003; Szolovits, 2003; Ahmed et al., 2005; Alphonse et al., 2004). Our evaluation is performed using two corpora of sentences from Medline abstracts with a focus on protein-protein interactions, the identification of which is the key aim of most biomedical IE systems.

Recently, two approaches addressing unknown

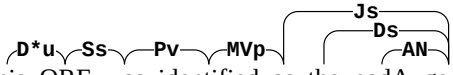
[1] http://www.link.cs.cmu.edu/link/

words in applying LGP to the biomedical domain have been proposed. Szolovits (2003) introduced a method for heuristically mapping terminology between lexicons and applied this mapping to augment the LGP dictionary with terms from the UMLS Specialist Lexicon[2]. Based on an analysis of a domain corpus, two of the authors have proposed an extension of the morpho-guessing system of LGP for disambiguating domain terms based on their suffixes (Aubin et al., 2005). The effect of the proposed extensions on parsing performance against an annotated reference corpus was not evaluated in these two studies.

Here we analyze the effect of these lexical extensions using an annotated biomedical corpus. We further propose, implement and evaluate in detail a third approach to resolving unknown words in LGP using information from a part-of-speech (POS) tagger.

2 Link Grammar Parsing

The link grammar formalism is closely related to dependency formalism. It is based on the notion of typed *links* connecting words. The result of parsing is one or more ordered parses, termed *linkages*. A linkage consists of a set of links connecting the words of a sentence so that links do not cross, no two links connect the same two words, and the types of the links satisfy the *linking requirements* given to each word in the lexicon. An example linkage is given below.



Since the link grammar is rule-based and the parser makes no use of statistical methods, LGP is a good candidate for adaptation to new domains where annotated corpus data is rarely available. The lexical adaptation approaches we evaluate further require only a light linguistic analysis of domain language.

LGP has three different methods applied in a cascade to handle vocabulary: dictionary lookup, morpho-guessing and unknown word guessing. The LGP dictionary enumerates all words, including inflected forms, and grammar rules are encoded through the linking requirements associated with the words. Some unknown words are assigned linking requirements based on their morphological features, such as the suffix *-ly* for adverbs. This system is termed *morpho-guessing* (MG). Finally, words that are neither found in the parser dictionary nor recognized by its morpho-guessing rules are assigned all possible combinations of the generic verb, noun and adjective linking requirements. This general approach is, in principle, always capable of generating the correct combination of linking requirements for unknown words. However, with an increasing number of unknown words in a sentence, the approach leads to a combinatorial explosion in the number of possible linkages and a rapid increase in parsing time and decrease in parsing performance. The parser is also time-limited: when a sentence cannot be parsed within a user-specified time limit, LGP attempts parses using more efficient, but restricted settings, leading to reduced parse quality.

When parsing sublanguages that contain many words that are not in the lexicon, it is therefore beneficial to attempt to resolve unknown words to reduce ambiguity in parsing.

3 Lexical adaptations

We evaluate three approaches to lexical adaptation: lexicon extension, morphological clues, and POS tagging. The approaches primarily involve open-class words and use linking requirements from the original LGP. Closed-class words, such as prepositions are considered domain-independent and expected to appear in the original lexicon, and modification of the existing linking requirements (grammar adaptation) is outside the scope of this study.

3.1 Extension of the lexicon

The extension of the lexicon with external domain-specific knowledge is the most frequent approach to adaptation, provided that the resources are available for the domain. This can be done either manually or with automatic mapping methods.

Here, we evaluate the heuristic lexicon mapping proposed by Szolovits (2003). This mapping can be used to automatically add domain-specific terminology from an external specialized lexicon to the lexicon of a parser. Words are mapped from a source lexicon (e.g. the domain lexicon) to a target lexicon (e.g. the parser lexicon) based on their lexical descriptions. As these descriptions typically differ between lexicons, they cannot be transferred directly from one lexicon to another. Instead, the mapping operates with sets of words that have the

[2] http://specialist.nlm.nih.gov/

Suffix	POS	examples	Suffix	POS	examples
-ase	noun	synthetase, kinase	*-in*	noun	actin, kanamycin
-ity	noun	chronicity, hypochromicity	*-ion*	noun	septation, reguion
-on	noun	replicon, intron	*-ol*	noun	glycosylphosphatidylinositol
-ose	noun	isomaltotetraose, isomaltotriose	*-or*	noun	cofactor, repressor/activator
-yl	noun	hydroxyethyl, hydroxymethyl	*-ine*	noun	5-(hydroxymethyl)-2'-deoxyuridine
-ide	noun	iodide, oligodeoxynucleotide	*-i*	noun	casei, lactococci, termini
-ic	adjective	glycolytic, ribonucleic, uronic	*-al*	adjective	ribosomal, ribsosomal
-ive	adjective	nonpermissive, thermosensitive	*-ar*	adjective	intermolecular, intramolecular
-ble	adjective	inducible, metastable	*-ous*	adjective	exogenous, heterologous
-ae	latin adj.	influenzae, tarentolae	*-us*	latin adj.	pentosaceus, luteus, carnosus
-um	latin adj.	japonicum, tabacum, xylinum	*-is*	latin adj.	brevis, israelensis
-fold	adjective/adverb	10-fold, 4.5-fold, five-fold			

Table 1: Biomedical suffixes involved in the extension of the morpho-guessing rules

exact same lexical description in their respective lexicons.

To assign a lexical description to a word w not in the target lexicon, the mapping finds words that have the exact same lexical description as w in the source lexicon, and that further have a description in the target lexicon. Overlap in sets having the same descriptions is then used to select one of these target lexicon descriptions to assign to w.

Szolovits applied the introduced mapping to extend the lexicon of LGP with terms from the UMLS Specialist Lexicon and observed that the mapping heuristic chose poor definitions for some smaller sets, for which the definitions were manually modified. The created UMLS dictionary extension contains 121,120 words that do not appear in the original LGP dictionary.

Szolovits observed that many of the phrases included in the extension "bear no specific lexical information in Specialist that is not obvious from their component words". Additionally, phrases are parsed using the LGP idiom system, which does not assign internal structure to the phrases, complicating comparison against a reference corpus. For these reasons, we evaluate the `no-phrases` version of the extension[3]. The effect of this extension has also been considered by Pyysalo et al. (2006).

3.2 Morphological clues

Morphological clues can be exploited by LGP to predict the morpho-syntactic classes (hence syntactic behaviour) of unknown words. Specific domains are an interesting application for this type of adaptation because a great part of technical lexicons presents regular morphological features, which, according to Mikheev (1997), obey morphological regularities of the general language. We observe that this assumption holds only partially because of the presence of foreign words in specialized texts and argue that a minimal morphological study of the corpus is necessary. Such studies have been performed, on the biomedical domain by Spyns (1994) and Aubin et al. (2005).

While many POS taggers employ morphological features to tag unknown words, domain extension of a rule-based approach such as the LGP morpho-guessing system can be preferable in lexical adaptation to domains where resources such as tagged corpora are not available for training taggers. Further, the MG extension allow assigning specific rules at a greater granularity than POS tags.

We have implemented and evaluated the extension of the LGP morpho-guessing rules proposed by Aubin et al. (2005). This extension of 23 new suffixes for the biomedical domain is presented in Table 1. Aubin et al. (2005) further identified in the corpus a small number of exceptions to these rules (*"wherein"*, *"kcal/mol"*, *"ultrafine"*, etc.), which were manually added to the dictionary.

3.3 POS tagging

We finally propose to provide the parser with an input sentence enriched with POS tags. In order to retain the decision-making power of the parser and to avoid inconsistencies between tagged words and their entry in the parser lexicon (see Grover et al. (2005)), we restrict the use of POS tags to unknown words only.

We modified LGP so that POS information can be passed to the parser by appending POS tags to input words (e.g. *actin/NN*). We further modified the parser so that when an unknown word is given a POS tag, the parser assigns linking requirements to the words based on a given mapping from POS tags to LGP dictionary entries. We defined such

[3] http://www.cdm.csail.mit.edu/projects/text/

Tag	Description	LGP rule
NN	common noun, sing.	words.n.4
NNS	common noun, pl.	words.n.2.s
NNP	proper noun, sing.	CAPITALIZED-WORDS
NNPS	proper noun, pl.	PL-CAPITALIZED-WORDS
JJ	adjective, base	UNKNOWN-WORD.a
JJR	adjective, comparative	words.adj.2
JJS	adjective, superlative	words.adj.3
VB	verb, base	words.v.6.1
VBD	verb, past tense	words.v.6.3
VBZ	verb, present 3d pers.	S-WORDS.v
VBP	verb, present non-3d	words.v.6.1
VBG	verb, gerund	ING-WORDS
VBN	verb, past participle	ED-WORDS
CD	number	NUMBERS
RB	adverb, base	words.adv.1

Table 2: POS tags mapping to LGP rules

a mapping, presented in Table 2, for Penn tagset POS categories corresponding to content words. FW (foreign words) and SYM (symbols) tags were not mapped due to their syntactic heterogeneity. Existing LGP rules were used to define the behavior of POS-mapped words, and the most generic applicable rule was chosen in each case. For instance, words tagged "NN" map the rule for nouns that can be either mass or countable, so that there is no constraint on determiners.

To evaluate the effect of using both a general and a domain tagger, the experiments were made using two taggers: the Brill tagger[4] trained on the Wall Street Journal (general language) and the GENIA Tagger[5] (Tsuruoka et al., 2005) trained on the biomedical corpus GENIA. A detailed evaluation and error analysis of GENIA Tagger is given in (Tsuruoka et al., 2005), finding 98% accuracy on two biomedical corpora. On this basis, we estimate the tagging accuracy at 81% for the Brill tagger and 97% for GENIA Tagger. This estimate was performed by manually checking tagging divergences between the two taggers on one of our corpora.

4 Evaluation protocol

4.1 Corpora

Two corpora are used for the present evaluation: "interaction" and "transcript", both built in the context of IE in biomedical texts. Both corpora were tokenized and cleared of bibliographic references in a preprocessing step.

Interaction contains 542 sentences (16,874 tokens) annotated for dependencies using the Link Grammar annotation scheme. 600 sentences were initially selected randomly from Pubmed[6] with the condition that they contain at least two proteins for which a known interaction was entered into the DIP database[7]. 58 sentences consisting only of a nominal phrase were then excluded as LGP does not, by design, parse them[8]. Each sentence was separately annotated by two annotators, and differences were resolved by discussion. Links to punctuation were excluded, and link types were not annotated. A total of 14,242 links were annotated in these sentences.

The transcript corpus is made of 16,989 sentences (438,390 tokens) consisting of the result for the query *"Bacillus subtilis transcription"* on Pubmed. It was not annotated.

Both corpora are used to characterize the vocabulary coverage by the different methods applied in LGP. The annotated interaction corpus is also used as the reference corpus for the evaluation of parsing performance.

4.2 Evaluation criteria

We first evaluate *vocabulary coverage* in the original and extended versions of LGP. We present the contribution of each method (dictionary, morpho-guessing, POS-mapping and unknown words) implemented in LGP to handle vocabulary. Results are given separately for types (i.e. distinct forms) and tokens (i.e. occurrences) in the corpus.

We assess the *ambiguity of the parsing* process with two criteria: parsing time and linkage numbers. Parsing time is immediately relevant to applications of the parser to systems where large corpora must be parsed. Linkage numbers are a more direct measure of the ambiguity of parsing a sentence. For each sentence, the parser enumerates the total number of linkages allowed by the grammar. By taking the ratio of the number of linkages allowed by two versions of the parser, we can estimate the relative increase or decrease in ambiguity. We report the per-sentence averages of both parsing time and linkage number ratios.

To determine the *parsing performance* of the extensions of LGP, we used each of the extensions to parse the interaction corpus sentences and compared the produced linkages against the reference

[4]http://research.microsoft.com/users/brill/

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

[6]http://www.pubmed.com/

[7]http://dip.doe-mbi.ucla.edu/

[8]This limitation could be overcome by modification of the grammar, but here we decided to avoid grammar adaptation and evaluate the parser with respect to its intended coverage.

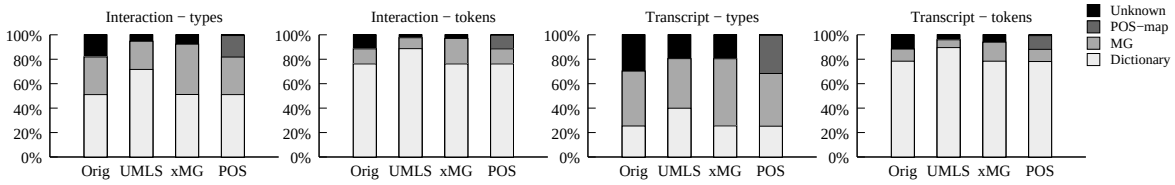


Figure 1: Vocabulary handling in the interaction and transcript corpora: the fraction of words and types covered by each method in the original LGP and the three adaptations. Coverage for the POS adaptation is shown only for GENIA Tagger as the coverage of the Brill tagger was essentially identical.

corpus. For each sentence, we determine the recall, i.e. the fraction of links in the reference corpus that were present in parses returned by LGP[9]. We report average recall for both the *first linkages* as ordered by the LGP heuristics and, to separate the effect of the heuristics from parser performance, also the *best linkages*, that is, the linkages with the most annotated links recovered. We further separately evaluate overall performance and performance for the subset of sentences where no timeouts occurred in parsing.

Experiments were performed on a 2.8GHz Intel Xeon with parameter values `timeout=60sec`, `limit=1000`, `islands-ok=true`. Default values were used for other parameters. The statistical significance of differences between the original parser and each of the modifications is assessed using the Wilcoxon signed-ranks test for overall first linkage performance, using the Bonferroni correction for multiple comparisons.

5 Results

In this section we present the evaluation results for the original LGP (Orig), LGP with the UMLS dictionary extension (UMLS), LGP with the morpho-guessing extension (xMG) and LGP with the POS extension, evaluated with the two taggers, Brill and GENIA tagger (GT).

5.1 Vocabulary coverage

Figure 1 shows the proportion of vocabulary covered by each method on the interaction and transcript corpora .

The comparison of the results on types and tokens shows that the dictionary has a good recognition rate on frequent types for both the original and the UMLS versions. By contrast, the MG and POS-map methods contribute for the recognition of a great number of types (particularly in transcript) but few tokens. In addition, the discrepancy on types between the two corpora for the dictionary method in all versions reflects the increasing presence of low frequency non-canonical words with the growing size of the corpus. Interestingly, we find that the reduction in unknown words (black part in the charts) due to the UMLS and xMG extensions is roughly similar, despite the former containing over 100,000 new words and the latter only 23 new rules. The POS extension, as expected, reduces the part of unknown words to almost null.

The remaining unknown words are of different nature for the extensions. Quite surprisingly, UMLS lacks a great number of species names (numerous in transcript) and frequent gene or protein names (e.g. *lacZ*, 78 occurrences in transcript). In addition, the Specialist Lexicon version used here contains no complex terms which prevents from detecting words like *vitro* and *vivo* used in the frequent terms *in vitro* and *in vivo*. The evaluated xMG extension cannot handle gene/protein names either, and also misses frequent technical terms that have no specific morphological features, such as *sigma*, *mutant* and *plasmid*.

To assess lexicon coverage, we measured the *contribution*[10] and the *recognition*[11] of the UMLS dictionary extension. We find that while the contribution of the UMLS dictionary extension is very low, with 0.54% on interaction and 2.3% on transcript, the recognition of the dictionary method is augmented significantly by the UMLS extension (51% to 71% for interaction and 25% to 40% for transcript). Nevertheless, as the size of the dictionary does not significantly penalize the parsing time with LGP, even a generic resource that con-

[9] Note that for connected, acyclic dependency graphs, precision equals recall: for each missing link, there is exactly one extra link. While there are some exceptions to connectedness and acyclicity in both LGP linkages and the annotation, we believe recall can be used as a fair estimate of overall performance.

[10] proportion of types of the resource found in the corpus

[11] proportion of types of the corpus found in the resource

	Orig	UMLS	Δ	xMG	Δ	Brill	Δ	GT	Δ
All, first linkage	74.2	75.4	4.7	76.0	7.0	75.4	4.7	76.8	10.1
All, best linkage	82.7	83.5	4.6	84.5	10.4	83.7	5.8	85.3	15.0
NT, first linkage	78.0	78.1	0.5	78.9	4.1	78.0	0.0	79.4	6.4
NT, best linkage	87.4	86.9	-4.0	88.0	4.8	86.7	-5.6	88.3	7.1
p	N/A	$p \approx 0.06$		$p < 0.01$		$p \approx 0.07$		$p < 0.01$	

Table 3: Performance. First linkage denotes the linkage ordered first by the parser heuristics and best linkage the best performance achieved by any linkage returned by the parser. Results marked NT are for the subset of sentences where no timeouts occurred for any of the modifications. Δ columns give relative decrease in error with respect to the original LGP, and p values are for "All, first linkage" performance.

tributes relatively little can be beneficial.

5.2 Ambiguity

The results of measuring the effect of the various extensions on ambiguity are given in Table 4.

Metric	Orig	UMLS	xMG	Brill	GT
Time	15.4s	9.9s	10.8s	8.8s	8.6s
Lkg. ratio	1	0.67	0.68	0.70	0.66

Table 4: Ambiguity. Time is average parsing time per sentence, linkage ratio is average of per-sentence linkage number ratios.

The reduction in the number of unknown words for the UMLS and xMG extensions is coupled with a roughly 30% reduction in both parsing time and linkage numbers. Although the POS extension essentially eliminates unknown words, it only gives a decrease in parsing time and linkage numbers that roughly mirrors the effect of the UMLS and xMG extensions.

None of the extensions achieves more than 35% reduction in linkage numbers or more than 45% reduction in parsing time. This may reflect structural ambiguity in the language and suggest a limit on how much ambiguity can be controlled through these lexical adaptation approaches.

5.3 Performance

The evaluation results are presented in Table 3. We find that in addition to increased efficiency, all of the extensions offer an increase in overall parsing performance compared to the original LGP for both the first and best linkages. Remarkably, this increase occurs even with the Brill tagger, which was trained on general English. In overall performance, the UMLS extension and the POS extension with the Brill tagger are roughly equal. The xMG extension outperforms both, and the POS extension with GENIA Tagger has the best performance of all considered extensions.

The positive effect of the extensions on parsing performance is linked to the reduced number of timeouts that occurred when parsing. Effects not related to time limitations can be studied on sentences where no timeouts occurred (NT). Here the effects of the extensions diverge: for the first linkage, performance with the UMLS extension and the POS extension with the Brill tagger essentially matches that of the unmodified LGP, while performance with xMG and GENIA Tagger remains better. For the best linkage, we observe a negative effect from the UMLS extension, indicating that for some words the unknown word handling mechanism of LGP finds correct links that are not allowed by the linking requirements given to those words in the extended dictionary. This suggest that some errors have occurred in the automatic mapping process[12]. We similarly observe the expected decrease in performance for the Brill tagger for the best linkage, reflecting tagging errors.

Even for the best linkage in sentences where no timeouts occurred, the performance with the xMG extension and the POS extension with GENIA Tagger is better than that of the original LGP. These extensions can thus assign more appropriate linking requirements for some words than the unknown word system of LGP. This indicates high tagging accuracy for GENIA Tagger as well as an appropriate choice of linking requirements for both extensions, and suggests some limitation in the unknown word system of LGP.

Despite significant improvements in parsing performance, the best performance achieved by any LGP extension is 88%. This may again suggest a limit on what performance can be achieved through the lexical adaptation approaches.

[12] An example of one such error is in the mapping of abbreviations (e.g. *MHC*) to countable nouns, leading to failures to parse in the absence of determiners.

Metric	Orig	UMLS & xMG	Δ	xMG & POS	Δ	UMLS & POS	Δ	All 3	Δ
All, first linkage	74.2	75.7	5.8	76.8	10.1	76.0	7.0	76.1	7.4
All, best linkage	82.7	83.7	5.8	85.3	15.0	84.2	8.7	84.2	8.7
NT, first linkage	78.0	78.4	1.8	79.3	5.9	78.6	2.7	78.7	3.2
NT, best linkage	87.4	87.0	-3.2	88.2	6.3	87.2	-1.6	87.1	-2.4
p	N/A	$p < 0.05$		$p < 0.01$		$p < 0.01$		$p < 0.01$	

Table 5: Performance for combinations of the extensions.

5.4 Combinations of the Extensions

The UMLS, xMG and POS tagging extensions are to some extent complementary as their coverage of the corpus vocabulary does not completely overlap. The dictionary extension provides the most frequent domain-specific lexicon while the xMG extension has the advantage of being able to handle non-canonical (e.g. *mutation/deletion*, *DNA-regions*) and rare words and misspellings. The POS extension can benefit from the context-sensitiveness of the tagger to disambiguate words.

We evaluated all possible combinations of the three extensions. In these experiments we only used GENIA Tagger for the POS extension. The results are given in tables 5 and 6.

Metric	Orig	UMLS & xMG	xMG & POS	UMLS & POS	All 3
Time	15.4s	9.5s	8.7s	8.3s	8.4s
Lkg. ratio	1	0.67	0.59	0.62	0.66

Table 6: Ambiguity for combinations of the extensions.

On ambiguity, we observe small advantages for many of the combinations, but rarely more than a 10% reduction for either metric compared to the simple extensions. The effect of the combinations on overall performance is mixed. While all combinations outperform the original LGP, combinations involving the UMLS extension appear to perform worse than those that do not, while combinations involving the xMG and POS extensions perform better. For sentences where no timeouts occurred the effect is simple: for the best linkage, all combinations involving the UMLS extension perform worse than the original LGP; only the combination of the xMG and POS extensions is better.

The performance of the best combination approach essentially matches that of the POS extension with GENIA Tagger alone, suggesting that no further benefit can be derived from combinations when an accurate domain tagger is available.

6 Conclusions and Future Work

We have studied three lexical adaptation approaches addressing biomedical domain vocabulary not found in the lexicon of the Link Grammar Parser: automatic lexicon expansion, surface clue based morpho-guessing, and the use of a POS tagger. We found that in a time-limited setting, any approach resolving unknown words can improve efficiency and overall performance. In more detailed evaluation, we found that the automatic dictionary extension and the use of a general English POS tagger can reduce performance, while the morpho-guessing approach and the use of a domain-specific POS tagger had only positive effects. We found no further benefit from combinations of the three approaches.

Generally, our results suggest that when available, a high-quality domain POS tagger is the best solution to unknown word issues in the domain adaptation of a general parser, here providing an overall 10% relative reduction in error combined with a 45% decrease in parsing time. In the absence of such a resource, the use of a general POS tagger is a poor substitute, and can lead to decreased performance. The use of heuristic methods for lexicon expansion carries the risk of mapping errors and should be accompanied by an evaluation of the effect on parsing performance. Conversely, surface clues can provide remarkably good coverage and performance when tuned to the domain, here using as few as 23 new rules.

Our implementation of the adaptations to LGP combines the morpho-guessing extension with the capability to use information from a POS tagger. Thus, the adapted parser is faster and more accurate than the unmodified LGP in parsing biomedical texts both when used as such and when used together with a domain POS tagger. Further, both extensions are implemented so that defining other morpho-guessing rules and POS-mappings is straightforward, facilitating adaptation of the

modified parser also to other domains. The adapted LGP is available under an open-source licence at `http://www.it.utu.fi/biolg`.

While we found that the considered approaches can significantly improve efficiency and parsing performance, our results also indicate some limitations for lexical adaptation. As future work, complementary approaches addressing grammar adaptation, text preprocessing, handling of complex terms, improved parse ranking and named entity recognition can be considered to further improve the applicability of LGP to the biomedical domain.

Acknowledgements

The work of Sampo Pyysalo has been supported by Tekes, the Finnish Funding Agency for Technology and Innovation.

References

[Ahmed et al.2005] Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu, and Chitta Baral. 2005. Intex: A syntactic role driven protein-protein interaction extractor for bio-medical text. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 54–61, Detroit, USA.

[Alphonse et al.2004] Erick Alphonse, Sophie Aubin, Philippe Bessières, Gilles Bisson, Thierry Hamon, Sandrine Laguarigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau, and Davy Weissenbacher. 2004. Event-Based Information Extraction for the biomedical domain: the Caderige project. In N. Collier, P. Ruch, and A. Nazarenko, editors, *COLING NLPBA/BioNLP Workshop*, pages 43–49, Geneva, Switzerland.

[Aubin et al.2005] Sophie Aubin, Adeline Nazarenko, and Claire Nédellec. 2005. Adapting a General Parser to a Sublanguage. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, editors, *Proceedings of the International Conference RANLP'05*, pages 89–93, Borovets, Bulgaria.

[Blaschke et al.1999] Christian Blaschke, Miguel A. Andrade, Christos A. Ouzounis, and Alfonso Valencia. 1999. Automatic extraction of biological information from scientific text: Protein-protein interactions. In Th. Lengauer, R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes, and R. Zimmer, editors, *ISMB*, pages 60–67.

[Ding et al.2003] Jing Ding, Daniel Berleant, Jun Xu, and Andy W. Fulmer. 2003. Extracting biochemical interactions from medline using a link grammar parser. In B. Werner, editor, *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 467–471. IEEE Computer Society, Los Alamitos, CA.

[Grishman2001] Ralph Grishman. 2001. Adaptive Information Extraction and Sublanguage Analysis. In B. Nebel, editor, *Proceedings of the Workshop on Adaptive Text Extraction and Mining at the 17th International Joint Conference on Artificial Intelligence (IJCAI'01)*, Seattle, USA.

[Grover et al.2005] Claire Grover, Maria Lapata, and Alex Lascarides. 2005. A Comparison of Parsing Technologies for the Biomedical Domain. *Journal of Natural Language Engineering*, 11(1):27–65.

[Lease and Charniak2005] Matthew Lease and Eugene Charniak. 2005. Parsing biomedical literature. In R. Dale, K. F. Wong, J. Su, and O. Y. Kwong, editors, *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 58–69, Korea. Springer-Verlag GmbH.

[Mikheev1997] Andrei Mikheev. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

[Pustejovsky et al.2002] James Pustejovsky, José Castaño, Jason Zhang, Maciej Kotecki, and Brent Cochran. 2002. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 362–373.

[Pyysalo et al.2006] Sampo Pyysalo, Filip Ginter, Tapio Pahikkala, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2006. Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions. *Special edition of the International Journal of Medical Informatics on Natural Language Processing in Biomedicine*. To appear.

[Sekine1997] Satoshi Sekine. 1997. The Domain Dependence of Parsing. In *Proceedings of the Applied Natural Language Processing (ANLP'97)*, pages 96–102, Washington D.C., USA.

[Sleator and Temperley1991] Daniel D. Sleator and Davy Temperley. 1991. Parsing English with a link grammar. Technical Report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

[Spyns1994] Peter Spyns. 1994. A robust category guesser for Dutch medical language. In *Proceedings of ANLP 94 (ACL)*, pages 150–155.

[Szolovits2003] Peter Szolovits. 2003. Adding a medical lexicon to an english parser. In M. Musen, editor, *Proceedings of the 2003 AMIA Annual Symposium*, pages 639–643. American Medical Informatics Association, Bethesda, MD.

[Tsuruoka et al.2005] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In P. Bozanis and E. N. Houstis, editors, *10th Panhellenic Conference on Informatics*, volume 3746, pages 382–392.