Adapting a general parser to a sublanguage

Sophie Aubin*, Adeline Nazarenko* and **Claire Nédellec****

(*) LIPN, University of Paris 13 & CNRS UMR 7030
99, av. J.B. Clément, F-93430 Villetaneuse, France
`{sophie.aubin,nazarenko}` at `lipn.univ-paris13.fr`
(**) Unité Mathématique Informatique et Génome (MIG, INRA)
Domaine de Vilvert, F-78350 Jouy en Josas Cedex, France
`claire.nedellec` at `jouy.inra.fr`

Abstract

In this paper, we propose a method to adapt a general parser (Link Parser) to sublanguages, focusing on the parsing of texts in biology. Our main proposal is the use of terminology (identification and analysis of terms) in order to reduce the complexity of the text to be parsed. Several other strategies are explored and finally combined among which text normalization, lexicon and morpho-guessing module extensions and grammar rules adaptation. We compare the parsing results before and after these adaptations.

1 Introduction

Most available NLP tools are developed for general language while processing technical texts, *i.e.* sublanguages, becomes a necessity for various applications like extracting information from biological texts (see (Grishman 01),(Pyysalo *et al.* 04), (Grover *et al.* 04) and (Akane *et al.* 05)). In order to assist the biologists in their daily bibliographical work, the ExtraPloDocs project[1] develops the natural language processing and machine learning tools that enable to build focused information extraction systems in genomics (gene-protein interaction, gene fonctionalities, gene homologies, etc.) at a reasonable cost. Beyond keyword and statistics based approaches, extracting such relational information must be based on syntax to achieve good precision and coverage (see for instance (Ding *et al.* 03)). We therefore need a reliable syntactic parsing of the texts dealing with genomics.

Instead of redeveloping new parsers for each sublanguage, we try to define a method for adapting a general parser to a specific sublanguage. This paper presents a strategy to adapt the Link Parser (LP) (Sleator & Temperley 91) to parse Medline abstracts dealing with genomics.

---

[1] ExtraPloDocs website : http://www-lipn.univ-paris13.fr/RCLN/Extra/ExtraPloDocs/
These results are also exploited for the development of specialized search engines in the ALVIS project (STREP) : http://cosco.hiit.fi/search/alvis.html

In this paper, we first discuss the question of sublanguages and the different strategies that can be adopted to parse technical texts. Section 3 presents the context of the adaptation of the LP to the biological domain. In section 4, we analyse several cases of parsing failure along with the solutions we propose to adapt the parser. We finally present the evaluation of the modifications we made on the LP grammar and lexicon.

2 Previous works

Sublanguages have been studied for a long time even though it remains a rather confidential part of linguistic and NLP studies. It is noticeable that in specific domains of knowledge, among certain communities and in particular types of texts, people have their own way of writing. These specific languages are called either sublanguages (Harris *et al.* 89; Grishman & Kittredge 86), restricted or specialized languages depending on the fact that one focuses on the continuity or the gap between these languages and the "usual language". In fact, a sublanguage is a restricted (fewer lexicon items and semantic classes) as well as a deviant language (original lexicon items and phrasings). This is also noticeable from a distributional point of view. As Harris noticed it, a sublanguage can be characterized by its selectional restrictions and more generally by the distribution of lexicon items and syntactic patterns.

(Sekine 97) has argued that parsing should be domain dependent. Three alternative approaches can be considered. Several NLP teams have decided to develop a specialized parser for a given sublanguage (see for instance the String project (Sager *et al.* 87) or (Pustejovsky *et al.* 02)) but this approach is considered too expensive for many applications. A second track consists in training a grammar from a specialized corpus, which requires annotated corpora that are rare in specialized domains. An intermediate approach aims at manually adapting a parser as proposed

in (Pyysalo *et al.* 04). This is our approach. This work can be considered as a preliminary work to evaluate the potentialities of automating this adaptation.

Two different approaches have been explored for the parsing evaluation. The first is linguistically oriented and based on test suites, a set of sentences that illustrates the various syntactic structures that a parser is supposed to analyse like in TSNLP (Lehman 96). The second approach, more pragmatic and more common, consists in evaluating the performances of a parser on a given corpus supposed to be representative of the textual data to parse. We will show in the following that we adopted a mixed approach.

As we will see below, one of the main problems in parsing sublanguages is the ambiguity of prepositional attachment.

3 Context

3.1 The corpora

Three different corpora were built from Medline[2] abstracts (in English) dealing with transcription in *Bacillus subtilis*. As recommended by (Prasad & Sarkar 00) and (Srinivas *et al.* 98), we mixed the two evaluation standards by randomly selecting 212 sentences that we organized according to their linguistic specificities. Despite its relatively small size, the MED-TEST corpus is a good sample of the sublanguage of genomics. We also used a larger corpus of full abstracts (TRANSCRIPT, 16,981 sentences, 434,886 words) and the GIEC corpus made of 160 sentences expressing gene/protein interactions. The GIEC corpus was built and used as a benchmark corpus in the context of the Genic Interaction Extraction Challenge[3] joint to the ICML 2005.

3.2 The initial parser choice

In the context of our IE task, and particularly for the ontology acquisition, we need reliable and precise syntactic relations between the words of the whole sentence (except empty words). For those reasons, a symbolic dependency-based parser seemed to be the most adequate.

LP presents several advantages among which the robustness, the good quality of the parsing, the adequation of the dependency technique and representation with our IE task and the

declarative format of its lexicon. From the results of the evaluation that we did on different parsers with the MED-TEST corpus, it turned out that dependency-based parsers have better results on long and complex sentences, particularly with coordinations. This conclusion is shared by (Ding *et al.* 03) who also worked on Medline abstracts. Other experiments, in the context of the ExtrAns project (Mollá *et al.* 00), showed that 76% of 2,781 sentences from a Unix manpage corpus were completely parsed by LP with no regard to the parsing quality, while we reach only 54% on the biological corpus. When looking at the quality of the parses, we noticed different kinds of errors depending either on the biological domain or on more general linguistic difficulties like ambiguous constructions. We propose three solutions to address these issues, the text normalization, the use of terminology and the adaptation of the lexicon/grammar of LP.

4 Diagnosis and adaptation

Our analysis of the performance of the Link grammar on the biological corpus confirms previous works. The main problems can be classified along the following axes.

4.1 "Textual noise"

Scientific texts present particularities that we chose to handle in a normalization step prior to the parsing. First, the segmentation in sentences and words was taken off from the parser and enriched with named entities recognition and rules specific to the biological domain. We also delete some extratextual information that alter the parsing quality. Finally, we use dictionaries and transducers to replace genes and species names by two codes, which prevents from extending the LP dictionary too much.

4.2 Unknown words

In the TRANSCRIPT corpus, we identified 6,005 out-of-lexicon forms (45,804 occurences) among 12,584 distinct words, *i.e.* 47.72%. They are mostly latin words, numbers, DNA sequences, gene names, misspellings and technical lexicon.

However, LP includes a module that can assign a syntactic category to an unknown word. It is based on the word suffix. Modifying the morphoguessing (MG) module seemed a better strategy than extending the dictionary since biological objects differ from an organism to another. We then

created 19 new MG classes for nouns (*-ase, -ity,* etc.) and adjectives (*-al, -ous,* etc.) along with their rule.

In the same time, we added about 500 words of the biological domain to the LP lexicon in different classes, mainly nouns, adjectives and verbs.

4.3 Specific constructions

Some words already defined in the LP lexicon present a specific usage in biological texts, which implied some modifications including moving words from one class to another and adaptating or creating rules.

The main motivation for moving words from one class to another is that the abstracts are written by non-native English speakers. This point was also raised by (Pyysalo *et al.* 04). One way to allow the parsing of such ungrammatical sentences is to relax constraints by moving some words from the countable to the mass-countable class for instance.

Some very frequent words present idiosyncratic uses (particular valency of verbs for instance), which induced the modification or creation of rules. Numbers and measure units are omnipresent in the corpus and were not necessarily well described or even present in the lexicon/grammar. Other minor changes were made that are not mentioned in this paper.

4.4 Structural ambiguity

We identified two cases of ambiguity that can be partially resolved by using terminology.

Prepositional attachment is a tricky point that is often fixed using statistical information from the text itself (Hindle & Rooth 93; Fabre & Bourigault 01), a larger corpus (Bourigault & Frérot 04), the web (Volk 02; Gala Pavia 03) or an external resources such as WordNet (Stetina & Nagao 97). The second major ambiguity factor is the attachment of series of more than two nouns. As shown in Figure 1, neither a parallel attachment (lp) nor a serial one (lp-bio) seem to be satisfying. We noticed that such cases often appear inside larger nominal phrases often corresponding to domain specific terms. For this reason, we decided to identify terms in a pre-processing step and to reduce them to their syntactic head. If needed, the internal analysis of terms is added to the parsing result for the simplified sentence (see lp-bio-t). The strategy proposed by (Sutcliffe *et al.* 95) that consists in the linkage of the words
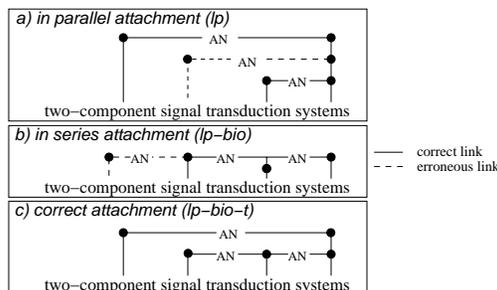

Figure 1: Series of nouns dependencies

contained in a compound (for instance *"sporulation_process"*) was excluded. It makes the lexicon size augment and does not reduce complexity for reasons due to the implementation of LP.

Figure 2 shows the influence of the adaptation on the parsing with the fixing of a segmentation error and the disambiguation of prepositional and nominal attachements.

Before practically integrating the use of terminology in our processing suite, we made a simulation of this simplification of terms.

5 Evaluation

We performed a two-stage evaluation of the modifications in order to measure the respective contribution of the LP adaptation on the one hand and of the term simplification on the other hand.

5.1 Corpus and criteria

We used a subset (10 files[4]) of the MED-TEST corpus but, contrary to the first evaluation (choice of a parser), we wanted to look at the quality of the whole parse and not only to specific relations.

Table 1 (for the MED-TEST subset) shows the way that out-of-lexicon words (OoL), i.e. unknown (UW) and guessed (GW) words, are handled by giving the percentage of incorrect morphosyntactic category assignments with the original resources (lp), those adapted to biology (lp-bio) and finally the latter associated with the simplification of terms (lp-bio-t).

In Table 2, five criteria inform on the parsing time and quality for each sentence : the number of linkages (NbL), the parsing time (PT) in seconds, the fact that a complete linkage is found or not (CLF), the number of erroneous links (EL) and the quality of the constituency parse (CQ). (NbW) is the average number of words in a sen-
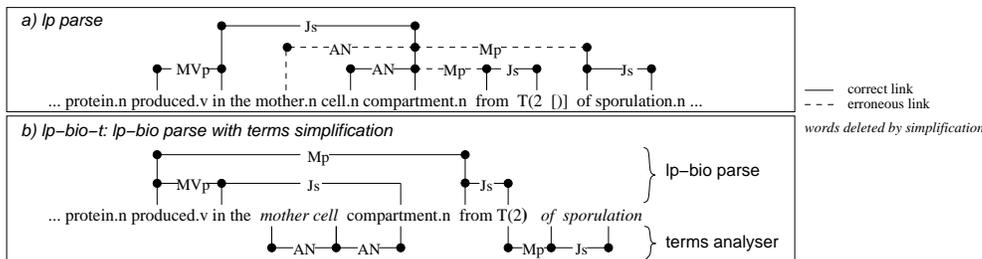
---

[4]141 sentences, 2630 words

Figure 2: Example of parsing

	lp		lp-bio		lp-bio-t	
	a	b	a	b	a	b
UW	244	41.4%	53	52.8%	26	19.2%
GW	24	4.2%	72	0%	31	0%
OoL	268	38%	125	22.4%	57	8.8%

a : total MS assignations, b : % of incorrect assignations

Table 1: Incorrect MS category assignments

crit.	lp	lp-bio		lp-bio-t	
	avg	avg	%/lp	avg	%/lp
NbW	24.05	24.05	100%	18.9	78.6%
NbL	190,306	232,622	122.2%	1,431	0.75%
PT	37.83	29.4	77.7%	0.53	1.4%
CLF	0.54	0.72	133%	0.77	142.6%
EL	2.87	1.91	66.5%	1.15	40.1%
CQ	0.54	0.7	129.6%	0.8	148.1%

Table 2: Parsing time and quality

tence which varies with the term simplification. The results are given for each one of the three versions of the parser.

UW, GW, NbL, PT and CLF are objective data while EL and CQ necessitate a linguistic expertise. The CQ evaluation consisted in the assignation of a general quality score to the sentence.

5.2 Results and comments

The **extension of the MG module** reduced the number of erroneous morpho-syntactic category assignations (see Table 1) from 38% to 22.4%. 61% of the sentences where one or more assignation error was corrected by the MG module actually have better parsing results (15% have been degraded). More generally, the increase of guessed forms makes the category assignation more reliable.

The **extension of the lexicon** and the **normalization of genes and species names** discharged the two modules from 143 assignations out of 268, 50 of which were wrong. 64% of the sentences where one or more assignation error was corrected by the extension of lexicon have better parsing results (18% of the sentences were degraded).

The effect of the **rules modification and creation** is difficult to evaluate precisely though it is certain to play a part in the parsing improvement, especially the relaxing of constraints on determiners and inserts.

The most obvious contribution to the better parsing quality is the one of the **term simplification**. The drastic reduction in parsing time and number of linkages gives an idea of the reduction of complexity. It is not only due to the smaller number of words since the number of erroneous links is reduced of 60% while the number of words is reduced of only 21.4%. This confirms previous similar studies that showed a reduction of 40% of the error rate on the main syntactic relations with a French corpus.

Remaining errors are mainly due to four different phenomena. First, the normalization step, prior to the parsing, needs to be enhanced. Concerning LP, there are still lexicon gaps, wrong class assignations and a still unsatisfactory handling of numerical expressions. In addition, and like (Sutcliffe *et al.* 95), we identified a weakness of LP regarding coordination. A specific study of the coordination system in LP and in the biological texts may be necessary. Finally, some ambiguous nominal and prepositional attachments still remain in spite of the term simplification. These may be resolved in a post-processing step like in ExtrAns that uses a corpus based approach to retrieve the correct attachment from the different linkages given by LP for a sentence.

Other questions like the feeding of LP with a morpho-syntactically tagged text or the ameliora-

tion of the parse ranking in LP were not discussed in this paper but are interesting issues that we intend to study.

6 Conclusion

Since parsing is domain and language dependent, a general parser must be adapted to each given sublanguage. In the context of an IE project in biology, we have adapted the Link Parser to analyse the specific language of Medline abstracts in genomics. Our initial diagnosis mainly raised two different problems which are traditional in sublanguage analysis: the lack of lexical coverage and the structural ambiguity, especially in the cases of prepositional phrase attachments.

We showed that the lexical problem can be manually handled by introducing new words in the lexicon and by extending the morpho-guessing module. We also proposed to distinguish and combine terminological and syntactic analysis. In the same way as the morpho-syntactic tagging should be considered independently from the parsing, we argue that the terminology analysis must be handled separately. This represents the main automated part of the adaptation task. The use of terminology to alleviate the parsing task is relevant and applicable in the context of domain specific texts processing since terminology tools and lists of terms are generally available. It also reduces the part of effective modification of the lexicon/grammar of the parser. This first evaluation has shown promising results.

This work has been developed as part of the ExtraPloDocs (extraction of gene-protein interactions in Medline abstracts) and ALVIS projects. We have shown that combining the terminological and syntactic analysis has an important impact on the resulting parses because the terminological analysis simplifies the parser input.

7 Bibliography

References

(Akane *et al.* 05) Y. Akane, Y. Miyao, Y. Tateisi, and J. Tsujii. Biomedical Information Extraction with Predicate-Argument Structure Patterns. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine*, pages 60–69, 2005.

(Bourigault & Frérot 04) D. Bourigault and C. Frérot. Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène. In *Actes des 11mes journées sur le Traitement Automatique des Langues Naturelles, Fès, Maroc*, 2004.

(Ding *et al.* 03) Jing Ding, Daniel Berleant, Jun Xu, and Andy W. Fulmer. Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser. In *15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*, 2003.

(Fabre & Bourigault 01) C. Fabre and D. Bourigault. Linguistic clues for corpus-based acquisition of lexical dependencies. In *Proceedings of the Corpus Linguistics 2001 Conference, UCREL Technical Papers*, volume 13, pages 176–184. Lancaster University, 2001.

(Gala Pavia 03) N. Gala Pavia. *Un modèle d'analyseur syntaxique robuste basé sur la modularité et la lexicalisation de ses grammaires, Thèse de Doctorat.* Unpublished PhD thesis, Université Paris XI, Orsay, 2003.

(Grishman & Kittredge 86) Ralph Grishman and Richard Kittredge. *Analyzing Language in Restricted Domains. Sublanguage Description and Processing.* Lawrence Erlbaum Ass., Hillsdale, NJ, USA, 1986.

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

(Grover *et al.* 04) Claire Grover, Maria Lapata, and Alex Lascarides. A Comparison of Parsing Technologies for the Biomedical Domain. *Journal of Natural Language Engineering*, 2004.

(Harris *et al.* 89) Zellig Harris, Michael Gottfried, Thomas Ryckman, Paul Mattick, Jr., Anne Daladier, T.N. Harris, and S. Harris. *The Form of Information in Science: Analysis of an Immunology Sublanguage.* Reidel, Dordrecht, 1989.

(Hindle & Rooth 93) Donald Hindle and Mats Rooth. Structural Ambiguity and Lexical Relations. In *Meeting of the Association for Computational Linguistics*, pages 229–236, 1993.

(Lehman 96) Sabine Lehman. TSNLP-test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, Budapest, 1996.

(Mollá *et al.* 00) Diego Mollá, Gerold Schneider, Rolf Schwitter, and Michael Hess. Answer Extraction Using a Dependency Grammar in ExtrAns. *Traitement Automatique de Langues (T.A.L.), Special Issue on Dependency Grammars*, 2000.

(Prasad & Sarkar 00) R. Prasad and A. Sarkar. Comparing test-suite based evaluation and corpus-based evaluation of a wide-coverage grammar for English. In *Using Evaluation within Human Language Technology Programs: Results and Trends. LREC'2000 Satellite Workshop*, pages 7–12, 2000.

(Pustejovsky *et al.* 02) J. Pustejovsky, J. Castano, and J. Zhang. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 362–373, 2002.

(Pyysalo *et al.* 04) S. Pyysalo, P. Ginter, T. Pahikkala, Boberg J., Järvinen J., T. Salakoski, and J. Koivula. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proceedings of the international Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 15–21, 2004.

(Sager *et al.* 87) Naomi Sager, Carol Friedman, and Margaret S. Lyman. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1987.

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

(Sleator & Temperley 91) D. Sleator and D. Temperley. Parsing English with a Link Grammar. Technical report, Carnegie Mellon University, 1991.

(Srinivas *et al.* 98) B. Srinivas, A. Sarkar, C. Doran, and B.A. Hockey. Grammar and Parser Evaluation in the XTAG Project. In *Workshop on the Evaluation of Parsing Systems*, 1998.

(Stetina & Nagao 97) J. Stetina and M. Nagao. Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In J. Zhou and K. W. Church, editors, *Proceedings of the Fifth Workshop on Very large Corpora*, pages 66–80, Beijing, China, 1997.

(Sutcliffe *et al.* 95) R. F. E. Sutcliffe, T. Brehony, and A. McElligott. The Grammatical Analysis of Technical Texts using a Link Parser. In *Second Conference of the Pacific Association for Computational Linguistics, PACLING'95*, 19-22 April 1995.

(Volk 02) Martin Volk. Using the Web as Corpus for Linguistic Research. In Renate Pajusalu and Tiit Hennoste, editors, *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim.* Publications of the Department of General Linguistics 3. University of Tartu, Estonia, 2002.