UNIVERSITE D'ORLEANS

LABORATOIRE D'INFORMATIQUE FONDAMENTALE D'ORLEANS

4 rue Léonard de Vinci
BP 6759
F-45067 Orléans Cedex 2
FRANCE
http://www.univ-orleans.fr/lifo

Rapport de Recherche

Unsupervised Dependency Parsing, a new PCFG

Marie Arcadias, Guillaume Cleuziou,
Edmond Lassalle, Christel Vrain

LIFO, Université d'Orléans
Orange Labs, Lannion

Rapport n^o **RR-2014-03**

Unsupervised Dependency Parsing, a new PCFG

Marie Arcadias
Edmond Lassalle
Orange Labs / 2 avenue Marzin
22300 Lannion, FRANCE
firstname.surname@orange.com

Guillaume Cleuziou
Christel Vrain
LIFO / Université d'Orléans
Rue Léonard de Vinci
45067 Orléans cedex 2, FRANCE
firstname.surname@univ-orleans.fr

Abstract

Dependency learning aims at building a model that allows transforming textual sentences into trees representing a syntactical hierarchy between the words of the sentence. We present an intermediate model between full syntactic parsing of a sentence and bags of words. It is based on a very light probabilistic context free grammar, allowing to express dependencies between the words of a sentence. Our model can be tuned a little depending on the language. Experimentally, we were able to surpass the scores of the DMV reference on attested benchmarks for five over ten languages, such as English, Portuguese or Japanese. We give the first results on French corpora. Learning is very fast and parsing is almost instantaneous.

1 Introduction and state of the art

The dependency structure (DS) of a sentence shows a syntactic hierarchy between the words, allowing then to infer semantic information. Among other applications, dependency structures are used in language modeling (Chelba et al., 1997), textual entailment (Haghighi et al., 2005), question answering (Wang et al., 2007), information extraction (Culotta and Sorensen, 2004), lexical ontology induction (Snow et al., 2004) and machine translation (Quirk et al., 2005).

The DS of a sentence (cf. Figure 1) is a tree, the nodes of which are labelled by the words of the sentence. One of the words is defined as the root of the tree (most of the time, the main verb). Then subtrees, covering contiguous parts of the sentences, are attached to the root. In other words, a dependency tree is made of directed relations between a syntactically strong word (called head) and a weaker word (called dependent). The dependency model is an interesting compromise between the full syntactic analysis and a representation as a "bag-of-words".

A large amount of manually annotated examples are necessary for supervised dependency learning. It is a very long and tedious task, and it requires deep linguistic knowledge. Furthermore, it has to be done anew for each kind of text to analyze. This explains why the amount of annotated text is poor compared to the abundance of different types of text available on the web. In this paper, we suggest an unsupervised approach demanding only a shallow knowledge on the language and on the type of the text. The framework is therefore Unsupervised Dependency Learning (UDL).

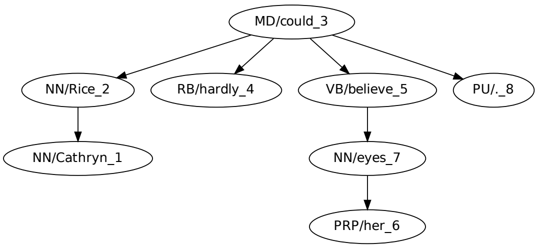

Figure 1: Dependency Tree given by the treebank for the sentence "Cathryn Rice could hardly believe her eyes".

The Penn Treebank, an American tagged corpus of newspaper articles, offers a dependency version, giving the DS of each sentence. Klein and Manning (2004) were the first to obtain significant results in UDL. They got better scores, on sentences of under 10 words, than the basic attachment of each word to its next right neighbor. They called their model Dependency Model of Valence (DMV).

Sentences are coded as sequences of parts-of-speech (POS) tags and are used as inputs for the learning and the parsing algorithms. DMV is a generative model based on the valence of the POS, i.e. their ability to generate children (i.e. dependents), their number and type of POS. The root of the sentence is first probabilistically chosen. Then, this root generates recursively its children among the other words of the sentence, and the subtree of each child is built, depending on their POS and relative position (left or right). The estimation of probabilities includes the type of preferred dependencies (verb over noun rather than noun over verb for example). Starting with initial probabilities tuned manually based on linguistic knowledge, an expectation-maximization step learns the probabilities of the model.

This is a rich and interesting model, but the parameters initialization is a full and complex problem. It demands both technical innovation from a machine learning expert and a strong linguistic background from an expert of the syntax of the studied language.

2 Learning a probabilistic context free grammar

The originality of our contribution is the choice of a simple context free grammar which can express dependencies between the words of a sentence. Our approach is then decomposed into two parts: learning this probabilistic context free grammar (PCFG) by the Inside-Outside algorithm (Lari and Young, 1990), parsing based on the learned PCFG, using a probabilistic version of CYK algorithm (Jurafsky and Martin, 2009). Finally, formal trees are transformed into dependency trees. For the definition of formal grammars, like PCFG, we suggest the reading of Jurafsky and Martin (2009).

Inside-Outside is a generative model that can be considered as an extension of hidden Markov models (HMM). Whereas HMM are limited to learning regular grammars, Inside-Outside can deal with context free grammars. While HMM use calculations on subsequences before and after a position t to obtain the probabilities of the derivation rules of the grammar, Inside-Outside algorithm calculates it from subsequences inside and outside two positions t_1 and t_2. The probabilistic version of CYK choses the most probable parse among all possible analysis.

DGdg formalism As already written, the originality of our contribution is the choice of a simple context free grammar which can express dependencies between the words of a sentence. For example, in the sentence “Cathryn Rice could hardly believe her eyes.”, “could” is a dominant to which “Rice” is attached to the left, and “hardly”, “believe” and the full stop are attached to the right. The dependency tree is represented in Figure 1. Our model classifies each word (represented by its POS tag) as a dominant or a dominated item beside its neighbors. Then, to parse a sentence, the model combines, thanks to intermediate symbols, the groups of words until each word finds a position in the dependency tree, as we can see in figure 2.

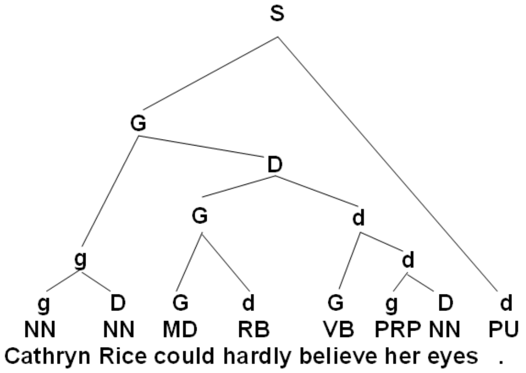

Figure 2: Parse of the sentence by the context free grammar DGdg.

To do that, we consider 5 non terminal symbols (nt) : the start symbol S, two symbols G and D representing respectively left and right dominants and two symbols g and d for left and right dominated items. The terminals represent the POS tags; they can differ depending on the language and on the tagger. Here are universal tags used by McDonald et al. (2013) (e.g. DET for determiner). $\Sigma = \{ADJ, ADP, ADV, CONJ, DET, NOUN, NUM,$ $PRON, PRT, PUNC, VERB, X\}$

The production rules are in Chomsky normal form (this is compulsory for Inside-Outside and CYK). Our constraints are :

- The uppercase non terminal dominates the lowercase non terminal it is associated with by a production rule. E.g. $G \rightarrow G\ d$ means that a left dominant splits into another left dominant and a right dominated symbol.
- A left non terminal g (respectively G) is associated to the left with D (resp. d). $nt \rightarrow G\ d$ or $nt \rightarrow g\ D$.

The meaning we impose to the non terminals forbids many rules, and thus limits the size of the grammar, while keeping its dependency state of mind.

The first type of rule in Chomsky normal form ($nt \to nt\ nt$) builds the internal construction of the sentences. We call them structure rules. The second type of rules in Chomsky normal form ($nt \to terminal$) expresses information whether a POS can (or cannot) dominate its right or left neighbors. For example, in English, we will forbid rules as $nt \to DET$ for all $nt \neq g$ because a determiner is always dominated by the next right noun.

The variants Depending on the structure of the studied language, the structure rules may not fit the deep split of the sentence. The grammar we presented before is called 4bin because it contains, in addition to the start symbol, 4 non terminals (D, G, d and g) and the structure rules are written in a binary way, according to Chomsky normal form.

The meaning of these 4 non terminals attests the fundamental difference between POS which would dominate from left, those which would dominate from right and those which would be dominated from right and left. For English, we can see cases where it is useful. But sometimes, the difference can be irrelevant. For example, in the noun phrase “the last president elect”, the two adjectives “last” and “elect” are both dominated by the noun “president”, and in the same way. Therefore, we consider a 3bin version with only 3 non terminals (in addition to S). We maintain in this variant the fact that g and d are left and right, but keep only one uppercase dominant symbol, called N (for neutral), which had no side information meaning.

Because of Chomsky normal form, the splits of the sentences are binary. However, translating ternary rules into a binary form allows us to use ternary structures suggested by sentences as : *subject* (left), *verb* (middle), *object* (right).

Following this idea, and keeping the directed dominants D and G, but allowing also a centered domination, we use again the neutral symbol N in variants 5ter and 5ter+. These last versions differers because 5ter forbids a recursive use of N while 5ter+ accepts it, leading to more complex structures. Table 1 sums up the differences between the variants.

Tuning phase All UDL models are tuned according to the language of the corpus. For our model, it consists in selecting only the rules of type $nt \to terminal$ which are linguistically relevant. We already give an example illustrating selection of such rules for the determiner. In the following experimentations, we tune the models observing for each language some trees given as a reference in the dependency treebanks.

3 Experiments and results

CONLL 2006 + FTB	3bin	4bin	4ter	5ter	5ter+
Bulgarian	17.7%	**23.9 %**	22.9%	22.8 %	23.6 %
Danish	**26.7%**	20.5%	14.0%	13.9 %	13.6 %
Dutch	29.7%	**34.6%**	30.3%	27.0%	27.0 %
English	15.3%	29.0%	26.4%	38.1%	**39.0%**
French	29.8%	32.9%	42.1%	37.4%	**42.2 %**
German	20.9%	**33.1 %**	20.9%	31.5%	31.7%
Japanese	31.2%	32.4 %	32.2%	33.4%	**64.7%**
Portuguese	30.0%	**54.0%**	42.0%	37.8%	34.1 %
Slovene	12.2%	21.5%	**23.2%**	21.6%	21.8%
Spanish	20.8%	39.2 %	39.2%	30.0%	**40.2%**
Swedish	21.2%	18.9 %	21.6%	**21.8%**	**21.8 %**

Table 3: Scores for all DGdg methods

The French Treebank (Abeillé and Barrier, 2004) gives the constituent structures (noun phrases, verb phrases...) as well as the syntactic functions (subject, object...) of many sentences from Le Monde newspaper. From 2009, this treebank was converted into dependency trees (Candito et al., 2010). We compare the trees learned by our models to those given as a reference in Candito et al. (2010) treebank. We compute the *Unlabeled Attachment score* (UAS) which gives the rate of correct dependency (without punctuations).

The scores are quite different according to the variant used. We obtained for 3bin: 29.8%, for 4ter : 42.1%, for 5ter : 37.4% and 5ter+ : 42.2 %; in order to assess the quality of these scores, we randomly generate trees and measure their UAS, obtaining only 14.2%. We notice that the two variants allowing ternary recursive rules (4ter and 5ter+), with a central group of words dominating one group on each side gives almost identical scores, much higher than the other variants. This would imply that the underlying structure of these journalistic sentences, quite elaborated, would be better captured by more complex models. As far as we know, we are the first to apply UDL for French.

On the other hand, this task was widely processed for English, and for other languages since

	4bin	3bin	4ter	5ter	5ter+
The main differences between the variants	First grammar	The dominant has no side meaning	Ternary rules are allowed	4ter + dominant non terminal centered N	5ter + recursive rules for N
The structure rules	$nt \to G\ d$, $nt \to g\ D$,	$nt \to N\ d$, $nt \to g\ N$,	$nt \to G\ d$, $nt \to g\ D$, $nt(\neq N) \to g$ CAP d,	$nt \to G\ d$, $nt \to g\ D$, $nt \to g$ CAP d	$nt \to G\ d$, $nt \to g\ D$, $nt \to g$ CAP d,
Non terminals	S, D, G, d, g	S, N, d, g	S, D, G, d, g	S, D, N, G, d, g	S, D, N, G, d, g

Table 1: The variants of the context free grammar DGdg (CAP represent non terminals in capital letters (G, D or N)).

CONLL 2006 + FTB	Random	DMV soft-EM	DGdg score	Variant used	Time for learning	Nb of corpus words	Nb of distinct categories
Bulgarian	16.1%	**39.1%**	23.9%	4bin	23 min	190 217	12
Danish	14.7%	**43.5%**	26.7%	3bin	35 min	94 386	10
Dutch	14.8%	21.3%	**34.6%**	4bin	12 min	195 069	13
English	13.4%	38.1%	**39.0%**	5ter+	99 min	937 545	23
French	14.2%	no ref.	42.2%	5ter+	320 min	278 083	15
German	13.1%	**33.3%**	33.1%	4bin	196 min	699 331	52
Japanese	20.7%	56.6%	**64.7%**	5ter+	19 min	151 461	21
Portuguese	15.3%	37.9%	**54.0%**	4bin	46 min	206 490	16
Slovene	13.7%	**30.8%**	23.2%	4ter	16 min	28 750	12
Spanish	13.3%	33.3%	**40.2%**	5ter+	35 min	89 334	15
Swedish	14.8%	**41.8%**	21.8%	5ter,5ter+	80 min	191 467	15

Table 2: Best scores

the conference CONLL 2006 (Buchholz and Marsi, 2006). We compare our model to the DMV reference. Table 2 summarizes the results, as well as the variant which achieves the best score. As we can see in table 3, results can be very different depending on the variants, showing that the choice of a variant must be wisely done according to the language and the type of text. These results shows that for some language, as for instance Bulgarian, all our methods badly modelize the shape of the corpus sentences, between 17.7 and 23.9% for Bulgarian. On the other hand, for some other languages, such as Japanese, one of our variant (here 5ter+) strongly outperforms all the others.

Table 2 gives the best UAS compared to the reference soft-EM given in (Spitkovsky et al., 2011). The dependency treebanks come from for English : (Marcus et al., 1993), for French (Candito et al., 2010; Abeillé and Barrier, 2004), for the other languages CONLL 2006 that is Bulgarian : (Simov et al., 2002), Danish : (Kromann, 2003), Dutch : (Van der Beek et al., 2002), German : (Brants et al., 2002), Japanese : (Hinrichs et al., 2000), Portuguese : (Afonso et al., 2002), Slovene : (Dzeroski et al., 2006), Spanish : (Civit and Martì, 2004) and Swedish : (Nilsson et al., 2005). The references are obviously the same for 3. The treebanks are provided already split into test and training sets of sentences.

4 Discussion and conclusion

The learning times depend strongly on the volume of the data and weakly on the number of syntactic categories. The little number of structure rules of the grammar leads to a reasonable learning time, even very fast for little corpora. Once the grammar is learned, parsing is almost instantaneous (a few seconds for thousands of sentences). This shows the flexibility and the speed of our model. This is why we can say that it is portable and efficient. Some complementary tests show that we can obtain better scores with more fine grained categories, even though the learning time is then a bit less fast.

To improve our model, we think about integrating lexical information to be able to make a difference between two sequences with the same POS tags, but which should have different dependency trees.

Acknowledgments

We thank Mark Johnson for his codes of Inside-Outside and CYK and Marie Candito for delivering her dependency treebank as well as for her advices.

References

[Abeillé and Barrier2004] Anne Abeillé and Nicolas Barrier. 2004. Enriching a french treebank. In *Proc. of LREC 2004, Lisbon.*

[Afonso et al.2002] Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sint (c) tica: A treebank for portuguese. In *LREC*.

[Brants et al.2002] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, page 24 41.

[Buchholz and Marsi2006] Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

[Candito et al.2010] Marie Candito, Benot Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *LREC 2010*, page 1840 1847.

[Chelba et al.1997] Ciprian Chelba, David Engle, Frederick Jelinek, Victor Jimenez, Sanjeev Khudanpur, Lidia Mangu, Harry Printz, Eric Ristad, Ronald Rosenfeld, and Andreas Stolcke. 1997. Structure and performance of a dependency language model. In *EUROSPEECH*.

[Civit and Martì2004] Montserrat Civit and MaAntnia Martì. 2004. Building Cast3LB: a spanish treebank. *Research on Language and Computation*, 2(4):549–574.

[Culotta and Sorensen2004] Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423.

[Dzeroski et al.2006] Saso Dzeroski, Tomaz Erjavec, Nina Ledinek, Petr Pajas, Zdenek Zabokrtsky, and Andreja Zele. 2006. Towards a slovene dependency treebank. In *Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation (LREC)*.

[Haghighi et al.2005] Aria D. Haghighi, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust textual inference via graph matching. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 387 394, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hinrichs et al.2000] Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. 2000. The VERBMOBIL treebanks. In *KONVENS*, page 107 112.

[Jurafsky and Martin2009] Dan Jurafsky and James H. Martin. 2009. *Speech and Language Processing*. Prentice Hall.

[Klein and Manning2004] Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting on ACL*, page 478.

[Kromann2003] Matthias Trautner Kromann. 2003. The danish dependency treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*, page 217.

[Lari and Young1990] K. Lari and S.J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech & Language*, 4(1):35 56, January.

[Marcus et al.1993] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313 330.

[McDonald et al.2013] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, and Oscar Tckstrm. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*.

[Nilsson et al.2005] Jens Nilsson, Johan Hall, and Joakim Nivre. 2005. MAMBA meets TIGER: reconstructing a treebank from antiquity.

[Quirk et al.2005] Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 271 279.

[Simov et al.2002] Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of bulgarian: the BulTreeBank. In *LREC*.

[Snow et al.2004] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*.

[Spitkovsky et al.2011] Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2011. Lateen EM: unsupervised training with multiple objectives, applied to dependency grammar induction. In *EMNLP*, page 1269 1280.

[Van der Beek et al.2002] Leonoor Van der Beek, Gosse Bouma, Rob Malouf, and Gertjan Van Noord. 2002. The alpino dependency treebank. *Language and Computers*, 45(1):8 22.

[Wang et al.2007] Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for QA. In *EMNLP-CoNLL*, volume 7, page 22 32.