The Relevance of Collocations for Parsing

Eric Wehrli
LATL-CUI
University of Geneva
Eric.Wehrli@unige.ch

Abstract

Although multiword expressions (MWEs) have received an increasing amount of attention in the NLP community over the last two decades, few papers have been dedicated to the specific problem of the interaction between MWEs and parsing. In this paper, we will discuss how the collocation identification task has been integrated in our rule-based parser and show how collocation knowledge has a positive impact on the parsing process. A manual evaluation has been conducted over a corpus of 4000 sentences, comparing outputs of the parser used with and without the collocation component. Results of the evaluation clearly support our claim.

1 Introduction

Collocations and more generally multiword expressions (MWEs) have received a large and increasing amount of attention in the NLP community over the last two decades, as attested by the number of workshops, special interest groups, and –of course– publications. The importance of this phenomenon is now clearly recognized within the NLP community.

It is fair to say that collocation extraction has been the main focus of attention, and a great deal of research has been devoted to developing techniques for collocation extraction from corpora (Church & Hanks, 1990; Smadja, 1993; Evert, 2004; Seretan & Wehrli, 2009, among many others). Much less attention has been paid to the interaction between collocations and the parsing process[1]. In this paper, we will argue (i) that collocation detection should be considered as a component of the parsing process, and (ii) that contrary to a common view, collocations (and more generally MWEs) do not constitute a problem or a hurdle for NLP (cf. Green et al., 2011; Sag et al., 2002), but rather have a positive impact on parsing results.

Section 2 shows how collocation identification has been integrated into the parsing process. An evaluation which compares the results of the parse of a corpus **with** and **without** the collocation identification component will be discussed in section 3.

2 Parsing collocations

That syntactic information is useful – indeed necessary – for a proper identification of collocations is widely acknowledged by now. More controversial, however, is the dual point, that is

[1] Preprocessing, that is, the detection of MWEs during tokenisation (ie. before parsing) is used in several systems – for instance, ParGram (Butt et al., 1999), or more recently, Talismane (Urieli, 2013). However, this technique can only be successfully applied to MWEs whose components are adjacent (or near-adjacent), leaving aside most of the cases that will be discussed below.

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 26–32,
Gothenburg, Sweden, 26-27 April 2014. ©2014 Association for Computational Linguistics

that collocation identification is useful for parsing.

Several researchers (cf. Seretan et al., 2009; Seretan, 2011, and references given there) have convincingly argued that collocation identification crucially depends on precise and detailed syntactic information. One main argument supporting that view is the fact that in some collocations, the two constituents can be far away from each other, or in reverse order, depending on grammatical processes such as extraposition, relativization, passive, etc. Based on such considerations, we developed a collocation extraction system based on our Fips multilingual rule-based parser(cf. Wehrli, 2007; Wehrli et al., 2010). Although quite satisfactory in terms of extraction precision, we noticed some shortcomings in terms of recall, due to the fact that the parser would not always return the most appropriate structure. A closer examination of some of the cases where the parser failed to return the structure containing a collocation – and therefore failed to identify it – showed that heuristics had (wrongly) favoured an alternative structure. Had the parser known that there was a collocation, the correct structure could have received a higher score.

These observations led us to revise our position and consider that parsing and the identification of collocations are in fact interrelated tasks. Not only does collocation identification rely on syntactic dependencies, and thus on parsed data, but the parser can fruitfully use collocational knowledge to favour some analyses over competing ones. A new version of the Fips parser has since been developed, in which collocations are identified as soon as the relevant structure is computed, that is as soon as the second term of the collocation is attached to the structure.

The collocation identification process is triggered by the (left or right) attachment of a lexical element marked [+partOfCollocation][2]. Governing nodes are iteratively considered, halting at the first node of major category (noun, verb, adjective, adverb). If that second node is itself marked [+partOfCollocation], then we check whether the two terms correspond to a known collocation.

Consider first some simple cases, as illustrated in (1).

(1)a. He had no **loose change**.

b. Paul **took up** a new **challenge**.

The collocation *loose change* in sentence (1a) is identified when the adjective *loose* is (left-) attached to the noun *change*. Both elements are lexically marked [+partOfCollocation], the procedure looked up the collocation database for a [$_{\mathrm{NP}}$ [$_{\mathrm{AP}}$ loose] change] collocation. In the second example (1b), the procedure is triggered by the attachment of the noun *challenge* to the determiner phrase (DP) *a*, which is already attached as direct object subconstituent of the verb *took (up)*. As pointed out above, the procedure checks the governing nodes until finding a node of major category – in this case the verb. Both the verb and the noun are marked [+partOfCollocation], so that the procedure looks up the database for a collocation of type verb-direct object.

Let us now turn to somewhat more complex cases, such as the ones illustrated (2):

(2)a. Which **record** did Paul **break**?

b. The **record** Paul has just **broken** was very old.

c. This **record** seems difficult to **break**.

d. This **record**, Paul will **break** at the next Olympic Games.

[2] The collocation identification process only concerns lexicalized collocations, that is collocations that we have entered into the parser's lexical database.

e. Which **record** did Paul consider difficult to **break**?

f. The **record** will be **broken**.

g. The **record** is likely to be **broken**.

h. Ce **défi**, Jean le considère comme difficile à **relever**.
"This **challenge**, Jean considers [it] as difficult to **take up**"

Sentence (2a) is a *wh*-interrogative clause, in which the direct object constituent occurs at the beginning of the sentence. Assuming a generative grammar analysis, we consider that such preposed constituents are connected to so-called canonical positions. In this case, the fronted element being a direct object, the canonical position is the typical direct object position in an English declarative sentence, that is a postverbal DP position immediately dominated by the VP node. The parser establishes such a link and returns the structure below, where [$_{\mathrm{DP}}$ e]$_i$ stands for the empty category (the "trace") of the preposed constituent *which record*.

(3) [$_{\mathrm{CP}}$ [$_{\mathrm{DP}}$ which record]$_i$] did [$_{\mathrm{TP}}$ [$_{\mathrm{DP}}$ Paul] break [$_{\mathrm{DP}}$ e]$_i$]

In such cases, the collocation identification process is triggered by the insertion of the empty constituent in the direct object position of the verb. Since the empty constituent is connected to the preposed constituent, such examples can be easily treated as a minor variant of case (1b).

All so-called *wh*-constructions[3] are treated in a similar fashion, that is relative clause (2b) and topicalization (2c). Sentence (2d) concerns the *tough*-movement construction, that is constructions involving adjectives such as *tough, easy, difficult*, etc. governing an infinitival clause. In such constructions, the matrix subject is construed as the direct object of the infinitival verb. In dealing with such structures, the parser will hypothesize an abstract *wh*-operator in the specifier position of the infinitival clause, which is linked to the matrix subject. Like all *wh*-constituents, the abstract operator will itself be connected to an empty constituent later on in the analysis, giving rise to a chain connecting the subject of the main clause and the direct object position of the infinitival clause. The structure as computed by the parser is given in (4), with the chain marked by the index *i*.

(4) [$_{\mathrm{TP}}$ [$_{\mathrm{DP}}$ this record]$_i$ seems [$_{\mathrm{AP}}$ difficult [$_{\mathrm{CP}}$ [$_{\mathrm{DP}}$ e]$_i$ [$_{\mathrm{TP}}$ to [$_{\mathrm{VP}}$ break [$_{\mathrm{DP}}$ e]$_i$]]]]]

Finally, examples (2f,g) concern the passive construction, in which we assume that the direct object is promoted to the subject position. In the tradition of generative grammar, we could say that the "surface" subject is interpreted as the "deep" direct object of the verb. Given such an analysis of passive, the parser will connect the subject constituent of a passive verb with an empty constituent in direct object position, as illustrated in (5).

(5) [$_{\mathrm{TP}}$ [$_{\mathrm{DP}}$ the record]$_i$ will [$_{\mathrm{VP}}$ be [$_{\mathrm{VP}}$ broken [$_{\mathrm{DP}}$ e]$_i$]]]

The detection of a verb-object collocation in a passive sentence is thus triggered by the insertion of the empty constituent in direct object position. The collocation identification procedure checks whether the antecedent of the (empty) direct object and the verb constitute a (verb-object) collocation.

2.1 Why collocations help

The parser can benefit from collocation knowledge in two ways. The improvement comes either from a better choice of lexical element (in

[3] See Chomsky (1977) for a general analysis of *wh*-constructions.

case of ambiguous words), or from a more felicitous phrase attachment. Both cases are illustrated below, by means of examples taken from our evaluation corpus. Consider first collocations of the noun-noun type containing syntactically ambiguous words (in the sense that they can be assigned more than one lexical category) as in (6):

(6)a. balancing act
eating habits
nursing care
living standards
working conditions

b. austerity measures
opinion polls
tax cuts
protest marches

As illustrated by Chomsky's famous example *Flying planes can be dangerous*, *-ing* forms of English transitive verbs are quite systematically ambiguous, between a verbal reading (gerund) and an adjectival reading (participle use). The examples given in (6a) are all cases of collocations involving a present participle modifying a noun. All those examples were wrongly interpreted as gerunds by the parser running without the collocation identification procedure. The noun-noun collocations in (6b) all have a noun head which is ambiguous between a nominal and a verbal reading. Such examples were also wrongly interpreted with the verbal reading when parsed without the identification procedure.

The second way in which collocational knowledge can help the parser has to do with structural ambiguities. This concerns particularly collocations which include a prepositional phrase, such as the noun-preposition-noun collocations, as in (7):

(7) bone of contention
state of emergency
struggle for life
flag of convenience

The attachment of prepositional phrases is known to be a very difficult task for parsers (cf. Church & Patil, 1982). So, knowing that a particular prepositional phrase is part of a collocation (and giving priority to such analyses containing collocations over other possible analyses) is an effective way to solve many cases of PP attachments.

3 Evaluation

To evaluate the effect of collocational knowledge on parsing, we compared the results produced by the parser **with** and **without** the collocation identification procedure. The corpus used for this evaluation consists of 56 articles taken from the magazine *The Economist*, corresponding to almost 4000 sentences. We first compared the number of complete analyses achieved by both runs, with the results in Figure 1[4]:

with collocations	without collocations
70.3%	69.2%

Figure 1: Percentage of complete analyses

Although the number of complete parses (sentences for which the parser can assign a complete structure) varies very slightly (a little more than a percent point better for the version with collocation identification, at 70.3%), the content of the analyses may differ in significant ways, as the next evaluation will show.

A manual evaluation of the results was conducted over the corpus, using a specific user interface. To simplify the evaluation, we selected the POS-tagging mode of the parser, and further

[4] By complete analysis, we mean a single constituent covering the whole sentence. When the Fips parser fails to achieve a complete analysis, it returns a sequence of chunks (usually 2 or 3) covering the whole sentence.

diff.	diff N vs V	with coll.	without coll.
416	148	116	32

Figure 3: Differences with and without collocation

restricted the output to the triple (word, pos-tag, position)[5]. For the POS tagset, we opted for the universal tagset (cf. Petrov et al., 2012). Both output files could then easily be manually compared using a specific user interface as illustrated in figure 2 below, where differences are displayed in red.

Notice that in order to facilitate the manual evaluation, we only took into account differences involving the NOUN and VERB tags. In the screenshot the two result files are displayed, on the left the results obtained by the parser with (W) the collocation identification component, on the right the results obtained with the parser without (WO) the collocation identification component. For each file, one line contains the input lexical item (simple word or compound), its tag, and its position with respect to the beginning of file (article). Differences (restricted here to NOUN vs VERB tags) between the two files are indicated in red. For each difference, the user selects the best choice, using the **Better left** or **Better right** button or the **Skip** button if the difference is irrelevant (or if neither tag is correct). After each choice, the next difference is immediately displayed.

The results are given in figure 3. Column 1 gives the total number of differences, column 2 the number of differences for the NOUN vs VERB tags, columns 3 and 4 show how many times the result (NOUN / VERB) is better with the collocation component (column 3) or without it (column 4).

This manual evaluation clearly shows that the quality of the parses improves significantly when the parser ”knows” about collocations, that is when collocation detection takes place during the parse. The comparison of the results obtained with and without collocation knowledge shows a total 416 differences of POS-tags, of which 148 concern the difference between Noun vs Verb tags. In 116 cases (nearly 80%) the choice was better when the parser had collocational knowledge, while in 32 cases (approx. 21%) the choice was better without the collocational knowledge.

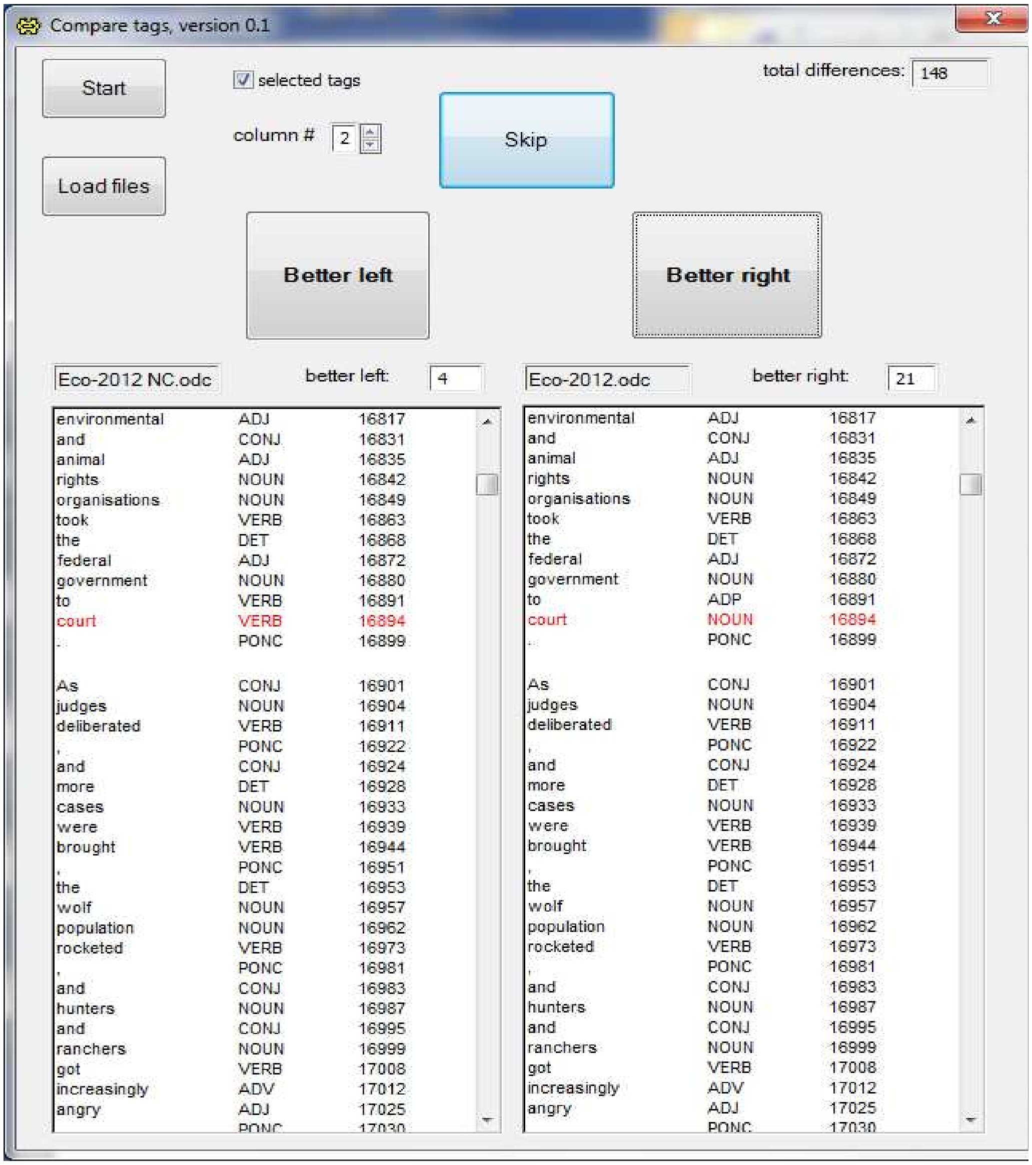
The fact that in a little over 20% of the cases the parser makes a better choice without collocational knowledge may seem a bit odd or counter-intuitive. Going through several such cases revealed that in all of them, the parser could not achieve a full parse and returned a sequence of chunks. It turns out that in its current state, the Fips parser does not use collocational knowledge to rank chunks. Nor can it identify collocations that spread over two chunks. Clearly something to be updated.

4 Concluding remarks and future work

In this paper, we have argued that collocation identification and parsing should be viewed as interrelated tasks. One the one hand, collocation identification relies on precise and detailed syntactic information, while on the other hand the parser can fruitfully use collocation knowledge in order to rank competing analyses and, more interestingly, to disambiguate some otherwise difficult cases.

This preliminary study focused primarily on the NOUN vs VERB ambiguity, an ambiguity which is very common in English and which may have a devastating effect when the wrong reading is chosen. For instance, in a translation task, such mistakes are very likely to lead to incomprehensible results.

[5] Using Fips in POS-tagging mode only means that the output will restricted to word and POS-tags. The analysis itself is identical whether we use Fips in parsing mode or in Pos-tagging mode.



☑ selected tags

column # 2

total differences: 148

Eco-2012 NC.odc better left: 4

Eco-2012.odc better right: 21

environmental	ADJ	16817
and	CONJ	16831
animal	ADJ	16835
rights	NOUN	16842
organisations	NOUN	16849
took	VERB	16863
the	DET	16868
federal	ADJ	16872
government	NOUN	16880
to	VERB	16891
court	VERB	16894
.	PONC	16899
As	CONJ	16901
judges	NOUN	16904
deliberated	VERB	16911
,	PONC	16922
and	CONJ	16924
more	DET	16928
cases	NOUN	16933
were	VERB	16939
brought	VERB	16944
,	PONC	16951
the	DET	16953
wolf	NOUN	16957
population	NOUN	16962
rocketed	VERB	16973
,	PONC	16981
and	CONJ	16983
hunters	NOUN	16987
and	CONJ	16995
ranchers	NOUN	16999
got	VERB	17008
increasingly	ADV	17012
angry	ADJ	17025
	PONC	17030

environmental	ADJ	16817
and	CONJ	16831
animal	ADJ	16835
rights	NOUN	16842
organisations	NOUN	16849
took	VERB	16863
the	DET	16868
federal	ADJ	16872
government	NOUN	16880
to	ADP	16891
court	NOUN	16894
.	PONC	16899
As	CONJ	16901
judges	NOUN	16904
deliberated	VERB	16911
,	PONC	16922
and	CONJ	16924
more	DET	16928
cases	NOUN	16933
were	VERB	16939
brought	VERB	16944
,	PONC	16951
the	DET	16953
wolf	NOUN	16957
population	NOUN	16962
rocketed	VERB	16973
,	PONC	16981
and	CONJ	16983
hunters	NOUN	16987
and	CONJ	16995
ranchers	NOUN	16999
got	VERB	17008
increasingly	ADV	17012
angry	ADJ	17025
	PONC	17030

Figure 2: Manual evaluation user interface

In future work, we intend (i) to perform a evaluation over a much larger corpus, (ii) to take into account all types of collocations, and (iii) to consider other languages, such as French, German or Italian.

5 References

Butt, M., T.H. King, M.-E. Niño & F. Segond, 1999. *A Grammar Writer's Cookbook*, Stanford, CSLI Publications.

Church, K. & P. Hanks, 1990. "Word association norms, mutual information, and lexicography", *Computational Linguistics* 16(1), 22-29.

Church, K. & R. Patil, 1982. "Coping with Syntactic Ambiguity or How to Put the Block in the Box on the Table", *American Journal of Computational Linguistics*, vol. 8, number 3-4, 139-150.

Chomsky, N. 1977. "On Wh-Movement", in Peter Culicover, Thomas Wasow, and Adrian Akmajian, eds., *Formal Syntax*, New York: Academic Press, 71-132.

Evert, S., 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, PhD dissertation, IMS, University of Stuttgart.

Green S., M.-C. de Marneffe, J. Bauer & Ch.D. Manning, 2011. ”Multiword Expression Identification with Tree Substitution Grammars: A Parsing *tour de force* with French", *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 725-735.

Petrov, S., D. Das & R. McDonald, 2012. "A Universal Part-of-Speech Tagset", *Proceedings of LREC-2011.*

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002), ”Multiword Expressions: A Pain in the Neck for NLP", Proceedings of Cicling 2002, Springer-Verlag.

Seretan, V., 2011. *Syntax-Based Collocation Extraction*, Springer Verlag.

Seretan, V. & E. Wehrli, 2009. "Multilingual Collocation Extraction with a Syntactic Parser", *Language Resources and Evaluation* 43:1, 71-85.

Smadja, F., 1993. "Retrieving collocations from text: Xtract", *Computational Linguistics* 19(1), 143-177.

Urieli, A., 2013. *Robust French Syntax Analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*, PhD dissertation, University of Toulouse. [http://redac.univ-tlse2.fr/applications/talismane/biblio/URIELI-thesis-2013.pdf]

Wehrli, E., 2007. "Fips, a deep linguistic multilingual parser" in *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic, 120-127.

Wehrli, E., V. Seretan & L. Nerima, 2010. "Sentence Analysis and Collocation Identification" in *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications* (MWE 2010), Beijing, China, 27-35.