Anaphora Resolution, Collocations and Translation

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

Luka Nerima
LATL-CUI
University of Geneva
Luka.Nerima@unige.ch

Abstract

Collocation identification and anaphora resolution are widely recognized as major issues for natural language processing, and particularly for machine translation. This paper focuses on their intersection domain, that is verb-object collocations in which the object has been pronominalized. To handle such cases, an anaphora resolution procedure must link the direct object pronoun to its antecedent. The identification of a collocation can then be made on the basis of the verb and its object or its antecedent. Preliminary results obtained from the translation of a large corpus will be discussed.

1 Introduction

Collocation identification and anaphora resolution (henceforth AR) are widely recognized as major issues for natural language processing, and particularly for machine translation. An abundant literature has been dedicated to each of those issues (see in particular Mitkov (2002) for AR, Wehrli *et al.* (2010) and Seretan (2011) for collocation identification), but to the best of our knowledge their intersection domain – a collocation in which the base term has been pronominalized – has hardly been treated yet. This paper intends to be a modest contribution towards filling this gap, focusing on the translation from English to French of collocations of the type verb-direct object, with and without pronominalization of the complement. The paper is organized as follows. The next section will give a brief overview of the translation problems with respect to both collocations and anaphors. We will also show how current MT systems fail to handle successfully such cases. In section 3 our treatment of collocations and anaphora resolution will be presented, along with some preliminary results. Finally, in section 4, we will try to address the issue of the frequency of those phenomena, presenting the results of our collocation extraction system over a corpus of approximately 10'000 articles from the news magazine *The Economist* totalizing over 8'000'000 words.

2 Collocations in Translation

The importance of collocations in translation has long been recognized, both by human translators and by developers of MT systems. For one thing, collocations tend to be ubiquitous in natural languages. Furthermore, it is often the case that they cannot be translated literally, as illustrated below. One of the characteristic features of collocations is that the choice of the collocate may be quite arbitrary and therefore cannot be safely derived from the meaning of the expression, and for that matter be translated literally. Consider, for instance, the examples in (1)-(2):

(1)a. heavy smoker

b. French
*lourd fumeur
gros/grand fumeur "big/large smoker"

c. German
*schwerer Raucher
starker Raucher "strong smoker"

(2)a. John broke a record.

b. French
John a battu un record
"John has beaten a record"

Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, September 3, 2013.
Monti, J., Mitkov, R., Corpas Pastor, G., Seretan. V., eds.

The adjective *heavy* in the collocation (1) *heavy smoker* cannot be translated literally into French or into German. Both of those languages have their own equivalent collocation, which in turn could not be translated literally into English. Similarly, the verbal collocate in a verb-object collocation can usually not be translated literally, as illustrated in (2). In most cases, a literal translation, though sometimes understandable, would be felt as "non idiomatic" or "awkward" by native speakers. Even though this state of affair does not apply to all collocations, it is widespread across languages and requires a proper treatment of collocations. Commercial MT systems usually have a good handling of collocations of the type "noun-with-spaces", such as adjective-noun, noun-noun, noun-preposition-noun, and the like. With respect to collocations which display a certain amount of syntactic flexibility and in which the two constituents can be arbitrarily far away from each other, commercial MT systems do relatively poorly, as illustrated in the few examples given at the end of the next section.

2.1 Translating collocations with Its-2

In this section, we describe how collocations are handled in the Its-2 translation system (cf. Wehrli et al. 2009a, 2009b), which is based on the Fips multilingual parser (cf. Wehrli, 2007). The proposed treatment relies on the assumption that collocations are "pervasive" in NL (cf. Jackendoff, 1997; Mel'cuk, 2003), which calls for a "light" and efficient treatment – perhaps in contrast to true idiomatic expressions, which are far less numerous and may require and justify a much heavier treatment[1].

Let us first consider again example (2), which involves a verb-object collocation, both in the source language (*break-record*) and in the target language (*battre-record* "beat record")

The structure assigned to this sentence by the Fips parser is identical to the structure of a non-collocational sentence such as

(3) Jean a mangé un biscuit
"Jean has eaten a cookie"

Ideally, therefore, we would like to say that the only difference between the two examples boils down to a lexical difference: the verb and the object head noun correspond to a collocation in (2), but not in (3). Based on this observation, we will strive to develop a transfer and generation process which will be identical for the two cases, except for the lexical transfer.

The general transfer algorithm of Its-2 recursively traverses the syntactic tree structure generated by the parser in the following order: head, left sub-constituents, right sub-constituents. Lexical transfer occurs during the transfer of a non-empty head. At that time, the bilingual dictionary is consulted and the target language item with the highest score among all the possible translations of the source language lexical item is selected. If a collocation is identified in the source sentence, as in our example, the lexical item associated with the verb *break* will also specify that collocation. In such a case, lexical transfer occurs on the basis of the collocation and not on the basis of the lexeme.

This procedure yields encouraging results, as illustrated by the following simple example of translation, which we compare with outputs from some commercial MT systems, both statistical and rule-based[2]. A few more examples, with sentences taken from the magazine *The Economist*, are given in the last section:

(4)a. The record that Paul set is likely to be broken.

b. Its-2
Le record que Paul a établi est susceptible d'être battu.

c. Google translate
L'enregistrement qui Paul ensemble est susceptible d'être rompu.

d. Systran
Le disque que l'ensemble de Paul est susceptible d'être cassé.

e. Reverso
Le rapport(record) que Paul met va probablement être cassé.

Example (4) contains two collocations, *to set a record* and *to break a record*. The first one occurs in a relative clause, while the latter is in the

[1] See Sag et al. 2002 for a thorough and enlightening discussion of multiword expressions.

[2] The commercial MT systems are Google-Translate (translate.google.fr), Systran (www.systranet.com) and Reverso (www.reverso.net), accessed between August 22 and August 29, 2012.

passive voice. As a result, in neither of them the direct object follows the verb. For that reason, the three commercial MT systems that we considered fail to identify the presence of those collocations and, thus, yield a poor translation. Its-2, thanks to the Fips parser, is quite capable of identifying verb-object collocations even when complex grammatical processes disturb the canonical order of constituents and correctly translate them by means of the equivalent French collocations *établir un record* and *battre un record*.

(5)a. The world record will be broken.

b. Its-2
Le record du monde sera battu.

c. Google translate
Le record du monde sera brisé.

d. Systran
Le record mondial sera cassé.

e. Reverso
Le record du monde sera cassé.

Example (5) also exhibits two collocations, *world record* and *to break a record*. The first one is of the "noun-with-spaces" variety, and therefore is well-translated by all the systems. The second one is in the passive form and, as in the previous example, commercial systems fail to recognize it.

2.2 Anaphora resolution

As a first step towards a proper treatment of anaphora, we have developed a simple procedure that allows the Fips parser to handle personal pronouns, by far the most widespread type of anaphora[3], restricted to 3rd person[4]. A second limitation of our AR procedure is that it only covers cases of anaphoric pronouns with antecedent within the same sentence or within the preceding sentence. As reported by Laurent (2001) on the basis of a French corpus, these two cases cover nearly 89% of the cases (67% and 22%, respectively). Roughly speaking, our AR procedure adopts the Lappin and Leass (1994) algorithm, adapted to the grammatical representations and other specificities of the Fips parser.

First, the AR procedure must distinguish between anaphoric and non-anaphoric occurrences. For English, this concerns mainly the singular pronoun *it*, which can have an impersonal reading, as in (6). Identifying impersonal pronouns is achieved by taking advantage of the rich lexical information available to our parser.

(6)a. It is raining.

b. It turned out that Bill was lying.

c. To put it lightly.

d. It is said that they have been cheated.

The next step concerns anaphors in the stricter sense of Chomsky's binding theory (cf. Chomsky, 1981), that is reflexive and reciprocal pronouns, which must be bound in their governing category. Our somewhat simplified interpretation of principle A of the binding theory states that a reflexive/reciprocal pronoun must be linked to (ie. agrees with and refers to) the subject of its minimal clause[5].

Finally, in the third step, we consider referential pronouns, such as personal pronouns (*he, him, it, she, her, them, etc.*), still using the insight of binding theory, which states according to principle B that pronouns must be free (ie. not bound) in their governing category. Here again, our simplified interpretation of principle B prevents a pronoun from referring to any noun phrase in the same minimal clause.

Note that the binding theory is not an AR method per se, in the sense that it does not say what the antecedent of a pronoun is. What it does, though, is to filter out possible, but irrelevant candidates. To illustrate, consider the simple sentences in (7), where the indices represent the coindexing relation between a pronominal element and its antecedent.

(7)a. Peter$_i$ watches himself$_i$ in the mirror.

b. Peter$_i$ watches him$_k$ in the mirror.

c. *Peter$_i$ watches him$_i$ in the mirror.

[3]According to Tutin (2002), personal pronouns range from 60% to 80% of anaphoric expressions, based on a large, well-balanced French corpus. Russo et al. (2011) report relatively similar results for English, Italian, German and French.

[4]First and second person pronouns are left out, since they do not have any linguistic antecedent. Rather, their interpretation is usually set by the discourse situation.

[5]The minimal clause containing a constituent X is the first sentential node (tensed or untensed) which dominates X in the phrase structure.

Sentence (7a) is well-formed because the anaphor *himself* is bound by the subject *Peter*. Given principle A of binding theory, we can conclude that the only possible antecedent of *himself* is *Peter*. Following the same reasoning, binding theory validates (7b) and rules out (7c). Since *him* is a pronoun, it cannot be bound (ie. find its antecedent) within the same minimal clause. Therefore, it cannot refer to *Peter*.

Our implementation of a simple but efficient AR procedure makes use of a stack of noun phrases, restricted to argument noun phrases, that the parser stores for each analysis and maintains across sentence boundaries. When a pronoun is read, the parser first determines whether it is a reflexive/reciprocal pronoun, in which case by virtue of principle A it must co-refer to the subject of its minimal clause, or a 3rd person pronoun. In the latter case, the parser will distinguish between referential and non-referential *it*, as discussed above. As we mentioned, that distinction can be made on the basis of the lexical and grammatical information available to the parser, in connection with the grammatical environment of the pronoun. For referential 3rd person personal pronouns, the procedure selects all the noun phrases stored on the stack which agree in person, number and gender with the pronoun. If more than one is selected, preference goes first to the subject arguments and second non subject arguments, a heuristic inspired in part by the Centering theory (cf. Grosz et al., 1986, 1995; Kibble, 2001). Needless to say, the procedure sketched above is merely a first attempt at tackling the AR problem.

3 Results and final remarks

The examples discussed above are all simple sentences constructed for the purpose of the present research. Let us now turn to "real" sentences taken respectively, from the July 2, 2002 and from the February 7, 2004 issues of *The Economist*.

Consider the English collocation *to make a case*, as illustrated by the examples (8-9). A literal translation into French of this collocation would give something like *faire un cas*, which is hardly understandable and certainly fails to convey the meaning of that collocation. A more appropriate translation would use the collocation *présenter un argument*. In the first example, the collocation occurs in a *tough*-movement construction, a peculiar grammatical construction in which an adjective of the *tough*-classs (*tough, difficult, easy, hard, fun, etc.*) governs an infinitival complement whose direct object cannot be lexically realized, but is understood as the subject of the sentence – in our example the phrase *such a case*[6]. Following a standard generative linguistics analysis of that construction, we assume that the direct object position of the infinitival verb is occupied by an abstract anaphoric pronoun linked to the subject noun phrase.

We can observe that Google-translate chooses a literal translation of the collocation (8a), while Its-2 correctly identifies the presence of the collocation and translates it appropriately with the corresponding French collocation *présenter un argument*.

(8)a. Such a case would not be at all difficult to make.

b. Google-translate
Un tel cas ne serait pas du tout difficile à faire.

c. Its-2
Un tel argument ne serait pas du tout difficile à présenter.

In our second example (9), the collocation *make a case* occurs twice (*making this case*, *makes it*). Notice that in the second occurrence, the base term of the collocation has been pronominalized, with its antecedent in the previous sentence. Thanks to the AR procedure, Its-2 correctly identifies the collocation and translates it appropriately (9c), which is not the case for Google-translate (9b).

(9)a. Every Democrat is making this case. But Mr Edwards makes it much more stylishly than Mr Kerry.

b. Google-translate
Chaque démocrate rend ce cas. Mais M. Edwards, il est beaucoup plus élégant que M. Kerry.

c. Its-2
Chaque démocrate présente cet argument. Mais M. Edwards le présente beaucoup plus élégamment que M. Kerry.

[6] See Chomsky (1977) for a detailed analysis of this construction.

To measure the accuracy of our collocation identification procedure as well as the impact of the anaphora resolution algorithm, we parsed a corpus taken from *The Economist* totalizing over 8'000'000 words (463'173 sentences). 14'663 occurrences (tokens) of verb-object collocations were identified, corresponding to 553 types[7]. In 68 cases, the direct object had been pronominalized, as in the next two examples, where the source sentence(s) is given in the (a) section in which both the collocation (verb + pronoun) and the antecedent of the pronoun are emphasized. The (b) section gives the Its-2 translation with the anaphora procedure turned off, the (c) section the Its-2 translation with the AR procedure turned on, and the (d) section, the translation obtained with Google-translate.

(10)a. The golden **rule** also turns slithery under close inspection.
On an annual basis, the government is **breaking it**.

b. [-AR] Sur une base annuelle, le gouvernement **le casse**.

c. [+AR] Sur une base annuelle, le gouvernement **l'enfreint**.

d. [Google] Sur une base annuelle, le gouvernement est **le casser**.

The best result is (c), the only one where the collocation *break-rule* is correctly identified thanks to the AR procedure which connects the direct object pronoun to the subject of the preceding sentence *golden rule*. The translation of that collocation yields the French verb *enfreindre* rather than *casser*.

(11)a. In Spain the **target** is mainly symbolic, since companies will not face financial penalties if they do not **meet it**.

b. [-AR] En Espagne la cible est principalement symbolique, depuis que les sociétés n'affronteront pas des pénalités financières si ils ne **le rencontrent** pas.

c. [+AR] En Espagne la cible est principalement symbolique, depuis que les sociétés n'affronteront pas des pénalités financières si elles ne **l'atteignent** pas.

d. [Google] En Espagne, la cible est surtout symbolique, puisque les entreprises ne seront pas passibles de sanctions financières si elles ne **répondent** pas.

In that last example, the source sentence contains two pronouns, *they* referring to *companies* and *it* referring to *target*. In (c), both of them have been correctly handled by the AR procedure and with the latter the collocation *meet-target* has been identified, yielding the correct collocation translation *atteindre(-cible)*.

Although not very frequent, collocations with a direct object pronoun should not be overlooked if one aims at a high-quality translation, as illustrated by the examples (10-11). Extending the collocation lexicon and the AR procedure to a larger set of pronouns, as we intend to do in future work is likely to increase the number of pronominalized collocations detected by the system.

[7]The most frequent collocations are *to take place* (529 occurrences), *to make sense* (407), *to play a role* (323), *to make money* (304) and *to make a difference* (266). Among the collocations with pronominalized objects, the most frequent are *to spend money* (7) and *to solve a problem* (5).

Acknowledgements

Part of the research described in this paper has been supported by a grant from the Swiss National Science Foundation (grant No. 100012-113864/1).

4 References

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

Chomsky, N. 1981. *Lectures on Government and Binding*, Foris Publications.

Grosz, B.. A. Joshi & S. Weinstein, 1995. "Centering: A Framework for Modeling the Local Coherence of Discourse", *Computation Linguistics*, 21:2, 203-225.

Grosz, B. & C. L. Sidner, 1986. "Attention, intention, and the structure of discourse", *Computational Linguistics*, 12:3, 175?204.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*, Cambridge, Mass., MIT Press.

Kibble, R. 2001. "A Reformulation of Rule 2 of Centering Theory", in *Computational Linguistics*, 27:4, Cambridge, Mass., MIT Press.

Lappin, Sh. & H. Leass, 1994. "An Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics* 20:4, 535-561.

Laurent, D. 2001. *De la résolution des anaphores*, Rapport interne, Synapse Développement.

Mel'cuk, I. 2003. "Collocations : définition, rôle et utilité", in F. Grossmann and A. Tutin, eds., *Les collocations : analyse et traitement*, Amsterdam, De Werelt, pp. 23-32.

Mitkov, R. 2002. *Anaphora Resolution*, Longman.

Russo, L., Y. Scherrer, J.-Ph. Golman, S. Loaiciga, L. Nerima & E. Wehrli, 2011. "Etudes inter-langues de la distribution et des ambiguités syntaxiques des pronoms", Montpellier, TALN.2011.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger, 2002. "Multiword expressions: A pain in the neck for NLP", in *Proccedings of Computational Linguistics and Intelligent Text Processing (CICLing-2002)*, Lecture Notes in Computer Science, 2276, 1-15.

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

Tutin, A. 2002. "A Corpus-based Study of Pronominal Anaphoric Expressions in French", in *Proceedings of DAARC 2002*, Lisbonne, Portugal.

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

Wehrli, E., Nerima, L., and Scherrer Y., 2009a. "Deep linguistic multilingual translation and bilingual dictionaries", *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 90-94, Athens, Greece.

Wehrli, E., Seretan, V., Nerima, L., and Russo, L., 2009b. "Collocations in a rule-based MT system: A case study evaluation of their translation adequacy", *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pp. 128-135, Barcelona, Spain.

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