Collocations and Anaphora Resolution in Machine Translation

Eric Wehrli
CUI-LATL
University of Geneva

October 22, 2015

Introduction

- Collocation identification and anaphora resolution (AR) are recognized as major issues for MT and several methods and algorithms have been proposed to handle them.
- In this talk, I will focus on the intersection domain between collocations and AR and show how such sentences can be handled by our multilingual Its-2 system.
- To the best of our knowledge, no other system can cope with collocations (say verb-object collocations) in which the direct object has been pronominalized.

Importance of collocations for NLP

- collocation - a definition
  Recurrent and conventional association of two lexical units (not counting grammatical words) in a specific grammatical configuration (adjective-noun, verb-object, noun-prep-noun, etc.)
- collocations are ubiquitous in natural languages
  cf. Jackendoff (1997), Mel'čuk (2003)
- it is often the case that collocations cannot be translated literally
  - *heavy smoker* – **lourd fumeur* vs. *gros fumeur ("big smoker")*
  - *to make an appointment* – **faire un rdv* vs. *prendre un rdv*
  - *loose change*, *dead loss*, *vain joy*

Collocations cont'd

- In a narrow sense, collocations are groups of two items (not counting grammatical words), the base and the collocate. While the base term keeps (one of) its usual meaning, the collocate is chosen in an arbitrary manner, with a meaning that may diverge quite significantly from its usual meaning.
- In a wider sense, collocations are simply arbitrary and recurrent associations of two lexical units (not counting grammatical words) in a specific syntactic environment.
  - verb + particle, such as *get off*, German : *an*kommen, *ab*fahren, etc.
    Der Zug fährt bald ab
  - *pierre d'achoppement* (*stumbling block*),*bone of contention*
  - *fish and chips*

Syntactic flexibility

Several types of collocations display a high level of syntactic flexibility (e.g., verb-object, verb-prepObject)
$\rightarrow$ difficult to identify them in a sentence

(1)a. The scheme addresses one of America's prickliest problems.

b. The problem –that poor children do not get the chances that rich ones do– is a real one, but needs to be addressed earlier.

c. ...éprouver, comme pour d'autres entités plus grandes, ou moins européennes dans leurs caractéristiques, de grandes difficultés.

d. Particular attention will have to be paid to them.

Collocations and MT

- most systems can cope with collocations of the "word-with-spaces" type (*heavy smoker*, *peace process*).
- verb-object collocations are handled correctly only when the object follows directly (or almost directly) the verb, in all other situations they are translated literally.

Examples

(2)a. Paul broke the long jump record.

b. Systran
Paul a cassé le disque de long saut.

c. Bing/Google
Paul a battu le record de saut en longueur.

(3)a. The long jump record that Paul broke was very old.

b. Google
Le record de saut en longueur que Paul a éclaté était très vieux.

c. Bing
Le record de saut en longueur qui a battu Paul était très vieux.

(4)a. The long jump record seemed difficult to break.

b. Bing/Google
Le record de saut en longueur semblait difficile à briser.

The Its-2 MT system

- Its-2 - MT system based on the Fips parser.
- Transfer between source and target languages is made on the basis of the (normalized) syntactic structures built by the parser, and it produced equivalent target structures.
- Transfer algorithm
  recursively traverse the source language syntactic structure in the order : head, left sub-constituents, right sub-constituents ;
  lexical transfer occurs upon the transfer of a non-empty head.

Translating collocations with Its-2

- goal –> light and efficient treatment
- ideally, a collocation should be handled just like any regular syntactic structure of the same type, modulo the lexical transfer – lexical transfer is made on the basis of the collocation rather than on the basis of the lexeme associated with the head.

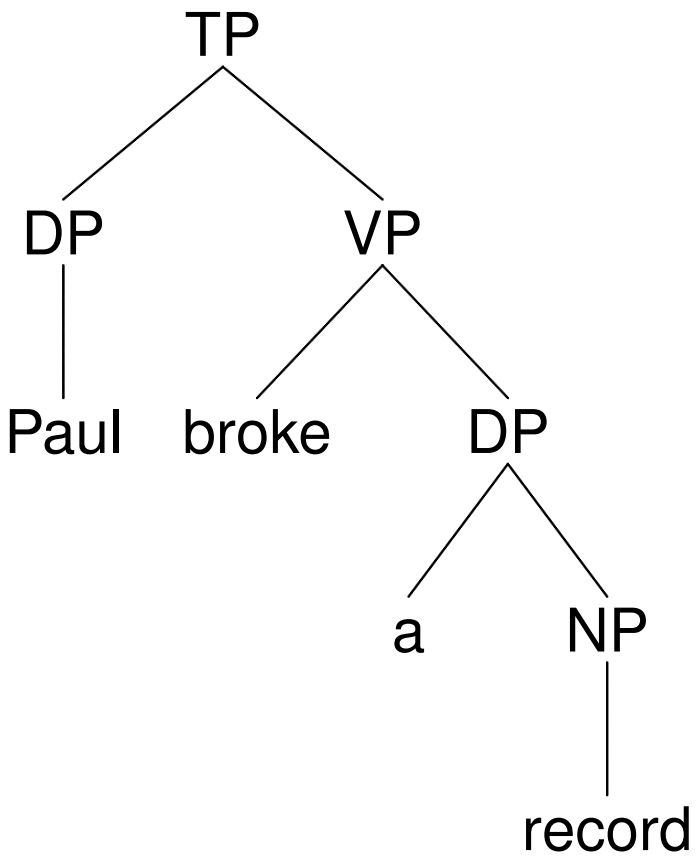
Collocation identification with the Fips parser

- collocation identification is best performed on the basis of analyzed data ;
- occurs during parsing, after the application of a right (or left) attachment rule ;
- governing nodes are iteratively considered, halting at the first node of major category (N, V, Adj, Adv) ;
- consider the pair [governing item + governed item] – check whether it constitutes an entry in the collocation database ;
- verify the (optional) restrictions associated with the collocation, eg.
  to take steps (prendre des mesures) vs. *to take a step* (faire un pas/avancer).

Collocation identification - a simple example

(5)a. Paul broke a record

b. [$_{TP}$ [$_{DP}$ Paul] [$_{VP}$ broke [$_{DP}$ a [$_{NP}$ record]]]]



Collocation identification - more complex examples

- *wh*-interrogatives
  Which record did Paul break ?
  $[_{\mathrm{CP}}\ [\ _{\mathrm{DP}}$ which record$]_i$ did $[_{\mathrm{TP}}$ Paul $[_{\mathrm{VP}}$ break $[\ _{\mathrm{DP}}$ e$]_i$ $]\ ]\ ]$
- relative clauses
  the record that Paul has just broken was very old
- *tough*-movement
  this record seems difficult to break
- *wh*-interrogative + *tough*-movement
  Which record did Paul consider difficult to break ?

Referential pronouns

- Since pronouns usually agree in gender and number with their antecedent and because grammatical gender do not correspond across languages,
- referential pronouns cannot be translated without knowledge of their antecedents.
- Most MT systems use an AR procedure to try to identify antecedents, with mixed results.

Anaphora resolution (1)

- We have designed an AR algorithm, which takes advantage of the specificities of the syntactic structures that the Fips parser creates, as well as of the general architecture of the system and of its implementation.
- The algorithm is very similar to the one proposed by Lappin & Leass (1994).

Anaphora resolution (2)

1. first, identify impersonal 3rd person pronouns, based on lexical and syntactic information (e.g. weather verbs, impersonal constructions, etc.), which should not undergo AR ;
2. second, for all referential pronouns, consider the preceding NPs agreeing with the pronoun (number and gender) and verify whether the (much simplified) binding rules are satisfied :
   - reflexive/reciprocal pronouns refer to the subject of the minimal clause (principle A) ;
   - referential pronouns cannot refer to an NP within the minimal clause (principle B) ;
3. if more than one potential antecedent remains, preference is given to subject NPs, then to other argument NPs.

Examples

(6)a. He set a new record last year and will probably break it at the Olympics in Brazil

b. Its-2
Il a établi un nouveau record l'an dernier et le battra probablement aux Jeux olympiques au Brésil

c. Google
Il a établi un nouveau record l'année dernière et va probablement casser lors des Jeux olympiques au Brésil

(7)a. Besides, the problem might be solving itself.

b. Brazil now knows it has a race problem, but is wondering how to solve it.

c. Clearly, a problem exists. But it is unlikely to be solved by driving people out of Rome for others to deal with.

- Collocation information is useful for word sense disambiguation...

(8)a. the record is old, but I think that Paul is likely to break it.

b. Its-2– Le record est vieux, mais je pense que Paul est susceptible de le battre.

c. Google– Le dossier est vieux, mais je pense que Paul est susceptible de casser.

d. Bing– L'enregistrement est vieux, mais je pense que Paul est susceptible de le casser.

e. Reverso– Le rapport(record) est vieux, mais je pense que Paul va probablement le casser.

A “real" example from The Economist

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

b. Systran
Chaque Démocrate fait ce cas. Mais M. Edwards le fait beaucoup plus élégamment que M. Kerry.

c. Bing
Tout démocrate rend cette affaire. Mais M. Edwards, il est beaucoup plus élégant que M. Kerry.

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

Some figures

corpus : approx. 9,900 articles from *The Economist* (1995-2013).

non-pronominal in situ	13,248
passive	965
wh-interrogative	41
relative	432
pronoun	33
total	14,719
total empty object+pronoun	1,471

TABLE: Number and types of verb-object collocations

More figures

number of collocations (types, token)	5,642	227,321
number of V-O collocations (types, token)	438	14,719

TABLE: Number of lexicalized collocations found in corpus

number of collocations	English	French
all types	9,149	16,135
verb-object	554	1,210

TABLE: Number of collocations in English and French lexicons

Remarks

- We have shown that AR and collocation identification interact in an interesting way. Specifically, to handle our verb-object examples, AR must crucially apply before collocation identification. Since AR relies on the structural representation of sentences, collocation identification cannot occur before parsing, as sometimes suggested.
- On the other hand, since collocation identification can be very helpful for the parser (for instance to resolve ambiguities or to rank competing analyses), applying both AR and collocation identification on parsed material is also ruled out.
- System architecture : both AR and CI should be integrated within the parsing process, in that order.

Future work

- Addition of more collocations to the lexical database.
- Extension to other types of collocations (e.g. verb-prepObject) and to other languages (German, Italian, Spanish, Portuguese, etc.).
- Evaluation of the AR procedure (precision and recall) ;
- Evaluation of the claim that collocational knowledge helps the parser.

Contact and demos

- contact : Eric.Wehrli@unige.ch
- demo Fips parser, Its-2 translator :

  `http://latlapps.unige.ch`