[From: *Tralogy*, Paris, 3-4 March 2011]

Johan Segura and Violaine Prince

Using Alignment to detect associated multiword expressions in bilingual corpora

Abstract

Translating multiword expressions from a language to another needs to recognize them as such. Bilingual multiword expressions are an issue when they are not the exact word-to-word translation of each other. The following examples are provided for a French-English translation task: (1) Phrasal verbs such as « *to call in on* » becoming « *rendre visite* », (2) « *sorry to hear that* », that a human translator translates into the simple 'désolé que" (3) most of adverbial locutions like « *such as* », equivalent to « *de telle façon que* », or « *de manière à* », etc. Thus, Machine Translation (MT) either requires a rich multiword bilingual database, or tends to create or enrich a first set of associated multiword expressions. Most of the time, existing resources are incomplete, and an interesting way to enhance covering is to provide a tool detecting 'associable' multiword expressions in parallel corpora. The latter are sets of texts that are translations of each others. There is an extensive literature in alignment techniques trying to link sentences from a text in a language, named the source, to one or many in the other language, seen as target. Sentence alignment is the basic preliminary task that underlies all others, more fine-grained. Word-to-word alignment has largely been dealt with by statistical systems. Multiword expressions have a granularity that lies between word and sentence. They are mostly phrasal, and sometimes with a rather strong syntactical and lexical divergence. With the improvement of parsers, alignment methods using syntax have emerged. Syntax allows the translation task, among others, to focus on relevant phrase fragments and to link multiword units together. For instance, Ozdowska's AliBi system is based on dependencies structures. The Groves', Hearnes' and Way's system uses syntactic trees with internal node alignments. Bilingual terminology, consisting in recognizing equivalent groups of words, also relies on syntax to extract patterns, such as Noun-Verb, Adjective-Noun, Prepositional Noun Phrase, etc...(e.g. Claveau, 2009). Most of these multiword expressions could be reduced down to collocations. A collocation is a multiword expression, naturally translated with quite strong constraints (e.g., « *to show respect* »-» *faire preuve de respect* »). Seretan's method [Seretan, V. (1999)] recognizes numerous equivalent pairs of collocations throught bilingual alignments which POS-tags are equivalents or close (even with distant words). But it only retrieves two-words collocations. Thus, there is a need for systems that might detect longer collocations, and more divergent ones. The method proposed in this article is an alignment process between pairs of sentences, strongly based on syntax. It relies on is a rule-based system combining partial alignments from a database through a non-iterative graph-theory based process. Multiword expressions patterns built on examples help providing alignments with a good coverage, which in turn detect new multiword expressions, and enrich the database. The article sketches the state-of-the art in alignment, focusing on syntactic

oriented systems, describes the designed system as well a corpus run experiment with promising results.

Introduction

Automatic sub-sentential alignment is one of the basic tasks preceding machine translation (MT), performed to enhance its efficiency, by increasing translation memories and resources with human translated data ([Groves, D. (2004)], [Ozdowska, S. (2006)]). It is seen as a cornerstone in MT, and several works, especially in statistical MT, do not separate alignment from MT ([Brown, F. B. (1990)], ,[Yamada, K. (2001)].

Sub-sentential alignment needs parallel bilingual corpora, i.e., a set of two corpora in two different languages, being a translation of each other, in which sentences are aligned (their order number provides a natural alignment) and constitute pairs of parallel sentences.

It aims at automatically providing translation links between sentences *constituents*, i.e., words or multiwords expressions, smaller than a sentence, within a pair of parallel sentences. Two items are particularly crucial in such a task: Alignment relevance and alignment requirements (paradigm, methods, resources). Both are related.

Alignment relevance is often seen as a depending on the constituent granularity, and talking about "*alignment units*" is often substituted to the pure relevance issue. Several works have focused on word-to word (e.g.,[Vogel, S. (1996)], [Melamed, D. (2000)]), word-to-phrase (e.g., [Brown, F. B. (1993)]), and phrase-to-phrase (e.g., [Hearne, M. (2003)]) alignments. Alignment between two items of different size (mostly in the last two cases) is an indication of a divergence between languages. The divergence origin could be either lexical, if a word *w* in a language *l1* has no direct equivalent in language *l2* and happens to be translated by a multiword expression, or syntactic, if building rules in *l1* are not the same as those in *l2*, leading to expressions of different size, and also of different linguistic properties.

Divergence can appear for different reasons: Colloquial expressions, idiomatic expressions, inconsistent choices from the human translator, exceptional phenomena,... The method presented here is said to be *example-based* or memory-based (EBMT). EMBT allows recognizing stored patterns and using them again when necessary, though a human-like analogy reasoning. This approach, defended by [Nagao, M. (1984)] is suitable for treating this kind of divergence. Nagao raised main issues such as matching fragments from an examples base, selecting the relevant ones and combining them to obtain the final result.

In this paper, alignment scope will be flexible considering the different aspects of divergence. Two multiword expressions translation from each other with significative syntactic divergence shall be linked entirely rather than trying to respect a word-to-word alignment paradigm (which will result in a non-covering set of links). For instance, we will totally align the phrasal verb "*to call in on*" with the expression "*rendre visite*".

The method uses an 'alignment memory', consisting in a learnt set of good alignments, as well as a rule-based process that asynchronously combines alignment constraints in order maximize coverage. The method is partly supervised: Present rule-based systems all introduce a learning feature in their information acquisition process. The learnt rules rely on contiguity in elements. Their shape allows a light combinatorial solution finding procedure. The method paradoxically suffers from getting too much rules proposed, the algorithmic effort is mainly about filtering and selecting a possibly optimal set of rules. Rules combination is seen as a graph-based problem to solve rules compatibilities.

Alignment relevance does not depend on a given granularity but on a successful application of rules.

Aligning Multiword Expressions: State-of-the Art

Literature in automatic alignment and translation is extensive. So, to browse the state of the art in a relevant fashion, we first focus on the impact of syntactic information in phrasal alignment, setting syntactic items as the basic requirements for the multiword expressions alignment task. Then, we present the issue of translating collocations that can be a task similar to translating expressions, with possible different sizes in both languages, which are not necessarily translatable word by word. Finally, a short overview of the example-based machine translation (EBMT) is given and as well as a discussion about the benefits this approach could provide to multiword expressions detection and translation.

Syntactic Issues in Bilingual Alignment

Literature on alignment is abundant, and some of the major works have already been mentioned in introduction. The founding work in alignment is accorded to Brown and al. from IBM [Brown, F. B. (1990)], [Brown, F. B. (1993)]. GIZA++ [Och, F. J. (2003)] system which is based on this IBM models, has evolved through time from a pure lexical to a sophisticated tool relying on a complex language model to account for translation divergence.

Syntactic trees as elements of the alignment process have appeared with [Yamada, K. (2001)]. Their model assumes that the syntactic structure of the target sentence is obtained as a set of transformations of the source sentence tree.

Cherry and Lin [Cherry, C. (2003)] propose a statistical approach in which syntax determines the training parameters by selecting a relevant neighborhood according to syntactic dependencies, thus rejecting bad alignments violating dependencies constraints. Rule-based alignment methods tend to favor structural information in defining their parameters, and determining their alignment unit.

Menezes and Richardson [Menezes, A. (2003)] define a tree-to-tree alignment process. Dependency trees are required for both languages, and conditional rules are iteratively applied until either all nodes are covered, or all rules have been applied. Alignment is initiated by rules requiring a lexicon, which help propagating links. Rules depend on a government structure where parent-nodes subsume the alignment of their children. A notable aspect is that structure is used for both alignment propagation and learnt rules extension.

[Hearne, M. (2003)], [Groves, D. (2004)] have extended this approach by using a constituents parsing. Their work differs by the way the collected alignment data are processed for annotation. The authors update tree-to-tree alignments, which are too general when only dependency relations are at stake, in order to get sub-tree alignments (thus closer to a 'phrase-to-phrase' granularity requirement). The translation memory is enhanced with several smaller alignments, which are in turn, combined to create new alignment structures, called *derivations*.

This brief overview of a fruitful research trend indicates that syntactic structural information could be helpful in different fashions by:

(1) Preventing alignments violating linguistic structural properties (e.g.,[Cherry, C. (2003)])

(2) Propagating alignments according to parent-child links (e.g., ([Menezes, A. (2003)] [Ozdowska, S. (2006)])

(3) Predicting an alignment with a POS tag, when the lexicon does not provide information ([Cherry, C. (2003)])

(4) Generating structures that accelerate the rules base building process in a data driven approach (e.g. [Hearne, M. (2003)] [Groves, D. (2004)])

These four elements represent the basic requirements to obtain a syntactically relevant alignment.

However, if these methods focus on alignment as the main asset, they all need, at various levels, either important resources (e.g., lexicons, dependency parsers) or an important processing effort (especially when tackling both dependency and constituents). The aim of this work is to try to estimate the ability of discarding lexicons in the alignment effort, and how much reducing the syntactic process effort could be maintained without a major damage in the alignment stage efficiency. Moreover, alignment is not so much the final goal, as it is a way of detecting plausible candidates for translation divergence, mainly at a phrasal granularity. Thus, a crucial perspective of this work, that constraints its model and method detailed in next section, is to enrich a translation memory as a sort of 'super' lexicon of equivalent expressions, involving stylistic idiosyncrasies of both languages.

Detection of collocations and multiword expressions through bilingual ressources

A collocation being at least a two words expression, it naturally belongs to the class of multiword expressions. A study from Yarowsky [Yarowsky, D. (1993)] shows that *an* *ambiguous word has only one sense in a given collocation with a probability of 90-99%.* Collocations can present a lexical divergence with strong syntactic similarities (e.g.,"se présenter à un examen"-"to take an examination") but can also be intrinsically divergent (e.g.,"perdre son intérêt"-"to stale"). [Nesi, P. (1996)] asserts that they are susceptible to appear in each sentence of the language.

Although extensive work has been carried out with collocations, only a few approach them through the scope of translation divergence. The models we shortly describe here have strong motivations in detecting collocational expressions with their translations which are supposed to be also collocations. The type of collocations these different approaches are interested in, can be very different. They can be seen as an alignment between source and target languages in parallel sentence-aligned corpora but only a partial one, focused on multiword expressions.

[Smadja, F. (1992)] asserts that alignment and translation failure of a stochastic model (as [Brown, F. B. (1990)]) are essentially due to the presence of ambiguous words and collocations. His work uses *Xtract* [Smadja, F. (1990)] that extracts a collocational list of words from a source text. *Xtract* tags an extracted collocation as **rigid** when no element could be inserted in the expression, or **flexible** if it is susceptible of polymorphism. This model requires no syntactical information and uses a mutual information measure to detect words strongly correlated with the words of the source collocation. Then, the translated collocation is built from this list of candidate words by filtering them. The hypothesis made here is that words from a collocation shall be strongly correlated.

In 1996, Smadja presents a program called *Champollion* which produces translations of collocations from a parallel corpus. It still relies on Xtract to detect collocations from source text and uses a Dice coefficient instead of the mutual information measure.

In 1993, Kupiec is the first to use syntax to detect collocations and translations [Kupiec, (1993)]. His method focuses on noun phrases thanks to syntactic parsing carried on both source and target texts. This approach lacks to detect flexible expressions unless there is a contiguous occurrence of them.

[Lü, Y. (2004)] states that resources can be an issue depending on the pair of languages studied, and proposes a collocation translation detection based on monolingual resources. The method detects collocations as couples of dependent words, and the aimed translation is a collocation of the same syntactic type. Selecting good translation candidates is a stochastic process (EM algorithm) based on two monolingual texts and a bilingual translation dictionary.

Seretan [Seretan, V. (1999)] describes a method to extract flexible collocations and translation equivalents from bilingual parallel corpora. This complete model uses deep syntactic analysis, sentence alignment, and lexical measures on bilingual parallel corpora.

The collocations targeted by this method are two word expressions with direct syntactic transfer. We will focus on this work since it is close, in its features, to our working paradigm.

The first step collects collocation candidates which are couple of words. These couples must respect some conditions to be good candidates such as a *syntactic configuration* (Noun-Adj, or Adj-Noun, or Verb-Noun etc...) and a weak lexical association measure. Beforehand, morphological variants are grouped for a more significant statistical study. For each selected collocation, a context is activated in source text and then in target text by a sentence level alignment. This target context will be the area where the translation equivalent will be looked for. At this point, a strong hypothesis is made: Each equivalent shall share the same *syntactic configuration* (as it is for the previous example "break a record"-"battre un record"). The selection among candidates uses frequency measures. It also relies on the following assumption: If equivalents in both languages share a lexical substring (see the word *record*, where it is completely the same) then the 'similar' equivalent plays the role of the collocational basis. Another example: *tarte à la crème*, giving *cream pie*. *crème* and *cream* play the role of the basis.

The approaches described in this part are mostly motivated with creating lexical resources as an improved bilingual dictionary. They stand out of the scope of transfer rules. Detected collocations forms can be rigid or flexible, which is an important linguistic information but rarely exceed the two-word expressions in both languages. Finally when syntactical divergence is important, the collocation won't be recognized as such.

The idiomatic form of collocations and their omnipresence in natural language makes one consider that example-based machine translation would be crucial for their automatic translation/alignment. Next subsection briefly sketches the outcomes of EBMT, and its possible benefits for multiword expressions alignment in general, and divergent collocations alignment in particular

Example-Based Machine Translation

EBMT tries to imitate the human translation by analogy. It is an intuitive approach which process consists in storing pieces of translations already met in the past, getting the relevant ones in a new translation request situation, then combining the pieces to obtain a solution.

These methods can generally be part of a wider process and be a complement of a statistical or a rule-based approach. They raise some issues that can lead to different kinds of work. Some outputs can appear in supervised translation tools (generally referred to as a translation memory) while others may be part of an entirely automatic process. Most of the time, examples are segmented (in order to give numerous elementary generic patterns) inherent in the approach. Some techniques use relatively arbitrary segmentation or chunks parsers or sub-trees from dependency or constituents analysis (also partial analysis, avoiding errors or integrating them...). In each case it can result in contiguous fragments or flexible ones. These fragments can be more or less generalized from the examples: They can consist in a list of words, of POS-tags, and sometimes of both. At last, filtering candidates may take very different aspects, with positive or negative discrimination (keeping or removing segments); Different measures can appear for this purpose (lexical, syntactic, coverage, divergence, frequencies, ...).

The first appearance of EBMT' issues is accorded to Makato Nagao in [Nagao, M. (1984)]. He clearly defined the three important steps of an EBMT process evoked above: Matching fragments from a database, filtering and combining. These three steps will be seen in every approach detailed here, but in our method, which consists more in a pure alignment process, the two last steps will be as one. Nagao claims that a human translation process doesn't lay on a deep analysis structure but on analogy with generic fragments. This idea motivates the whole example-based approach.

A few approaches are presented here, which cannot represent an exhaustive description of the huge literature in EMBT or translation memory (we can see for a more complete one [Somers, H. L. (1992)]) and are mostly motivated with issues appearing in our system).

Every key of an EBMT can stick to different paradigms, but each of them is intimately related to the other ones. In [Cranias, L. (1994)], the authors deal with measuring the relevance of translation propositions for an input sentence, and start with a discussion about the form and the size the fragments should have. An answer is given to the size issue: The sub-sentential level should be preferred for 'genericness' reasons (exact same sentences occur only rarely) but raise in turn the issue of recombining the fragments in a way that preserve the language structure and the sense ([Sato, S. (1990)]). The notion of a good matching measure is also tackled in [Nirenburg, S. (1993)]. He observes that the simplest metric is a complete match and proposes a heuristic: "*quality of a match is proportional to a measure of contiguity of matching*" (This notion appears in the method presented later). He describes a relaxed matching which extends the available examples from the database by allowing close matches. For example, a rule can be extended according to morphological variations, or synonymy.

A different way to deal with the recombining issue is presented by Hiroshi Maruyama in [Maruyama, H. (1992)]; He explores the problem with a system using a database of tree fragments. The algorithm he describes precisely, is a minimum weight set-cover program (A deterministic fashioned recombining one is a corner-stone in our system); The weight function is not specified but can be adapted to a wide field of linguistic information. The complexity depends on the maximum node degrees in trees, which is constant (2 for binary trees), the size *a* of the largest fragment, *p* the number of fragments and *N* the number of nodes: $O(Nap2_a)$ which is exponential in *a*. This approach tries to favour the combination among small fragments, giving some control over the complexity.

The choice of relevant previous examples in EBMT is still an important issue to deal with. Filtering methods based on coverage, linguistic measures are frequently studied. In [Gotti, F. (2006)], the reused examples are favoured when found in a close context. It's necessary to remind that dealing with context is of great importance in translation. Still a word to word alignment is used based on the GIZA++ toolbox and segmentation is not linguistically motivated. Gotti and Coulombe [Gotti, F. (2005)] discussed the importance of segmentation in EBMT and were driven to prove that linguistically motivated patterns are of a benefit.

The shape the patterns should take is motivated by a correct reuse. Efforts are done to make the fragments as generic as possible without losing sight of the fact that these fragments should be consistent with the recombining process. The generalization of the evoked before patterns is a solution. One can mention for instance Brown's method [Brown, R. D. (1999)], which uses syntactic analysis to raise the detected segments above the simple lexical information. The words of an analysed sentence are replaced with their classes or categories to generalize them with a much wider set of applications. For example, "*blue*" in the segment "*a blue sky*", shall be replaced by the tag "*[COLOR]*". In other cases, the system can deal with dates, places, or even use syntactic information as part-of-speech classes (e.g., noun, adverb, etc...) This approach tends to emphasize the gain of generalization by showing an accelerating efficiency in the treatment.

This brief overview has exhibited some important topics inherent to EBMT. Before tackling the model presentation, one has to keep in mind the issues evoked above: The quality of matching, patterns generalization and the more or less linguistically motivated segmentations. Our model can be seen as a subclass of EBMT that only deals with alignment and detection issues. *The main difference occurs in the filtering problem which is not a choice between proposed translations but a choice between alignments*. It is a much more constrained (not necessarily more difficult) purpose in the sense that the translated sentence is part of the input.

The Model: Syntactic Patterns for Multiword Expressions

We present here the elements of the method, and the method itself. The pair of considered languages are respectively French and English (available parsing resources). The parsing of the French source sentence is carried out by SYGFRAN [Chauché, J. (1984)] which provides a deep syntactic tree. TreeTagger [Schimd, H. (1994)] is used for English POS tagging task. TreeTagger has not been used for French since it does not offer enough syntactic information (no deep tree structure).

The system looks like an EBMT, with the possibility for the user to correct the proposed alignments or to create new ones. Corrections made by the user enrich the database with new relevant information (An adapted interface was designed for this purpose). The database model should be referred to as an **alignment memory**. Each provided alignment is divided into several pieces which will be called *patterns* and then are memorized. These patterns will be used in an alignment process and, individually, concern a phrase-level scope.

The acquisition part is followed by two data processes resulting in a specific kind of generated rules:

- The **segmentation** of the bi-sentence into relevant small pieces.
- The **generalization** of the alignment will allow general rules above the word level, considering entire classes or words categories.

Segmentation of the aligned bi-sentence lays on both tree analysis from source sentence and the links from the alignment which reflects transfer. First, the source sentence will be divided into phrases along the sub-trees from parsing. It will result into strongly justified groupings of contiguous words (from a syntactic point of view). The target words linked to the source words will be grouped together as well. See for example figure 1 where pieces are delimited with blue rectangles.

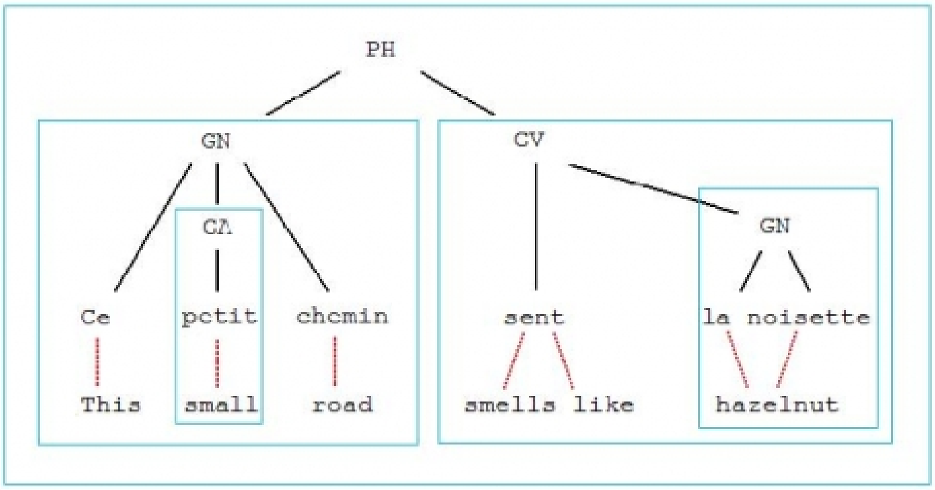


Figure 1: Selecting Sub-sentential Rules

When a much more divergent aligned bi-sentence is proposed, it is not rare to obtain many words to many words links. This will lead to a completely linked set of words. For example, the word "*Désolé*" in "*Désolé que* ... " could be entirely linked with "*Sorry to hear*" in "*Sorry to hear that...*". Not every alignment will be considered as acceptable in our database fragments. For instance, let us consider a bi-sentence with words s_1 and s_2 in the source sentence and c_1 and c_2 in the target. Suppose the following links exist: (s_1,c_1), (s_1,c_2) and (s_2,c_1). Then, the obtained alignment is considered as *acceptable* only if (s_2,c_2) exists. This last condition allows us to create alignments consisting of non-intersecting completely linked set of words, that we assume to be a rather natural definition beyond which the notion of alignment would be meaningless. Some examples of minimal correct alignments are provided in figure 2. Upper (*Red*) nodes are source sentence tags. Lower (*Blue*) nodes are target sentence tags. An edge is provided if a mapping is possible between an upper and a lower node, or a set of lower nodes.

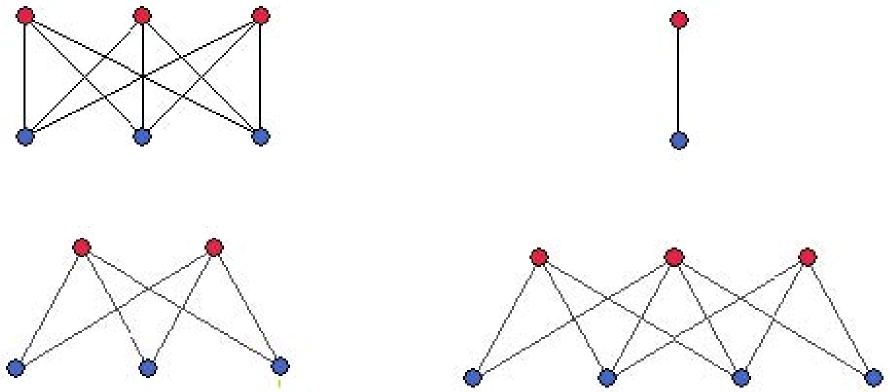


Figure 2: Correct Minimal Alignments

The segments obtained from the tree structure in the previous example, are made of pairs of **contiguous** words in source and target sentences. This is because here, the alignment links respect the scope of each sub-tree. Of course, it is not a general behaviour: Translation process through divergence will tend to separate some words. Moreover the approach depicted here stick to the usage of contiguous patterns. So, when a provided alignment presents many crossing links, the smaller element should be captured as a contiguous set of words. We will operate by extending the fragment until contiguity is respected. For instance, phrasal verbs in English may allow a pronoun to separate the verb from its particle which is not true in French. The simple example from figure 3 shows how a bi-sentence is divided when crossing links happen: the couple "*woke up*"-"*a réveillé*" is not selected as a fragment. The pronominal couple *"her"-"l'"* has to be added in order to respect the contiguous form needed for this method.

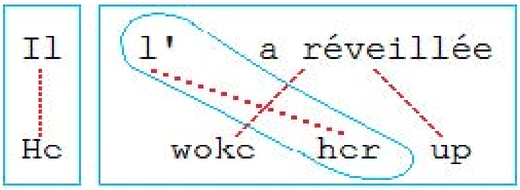


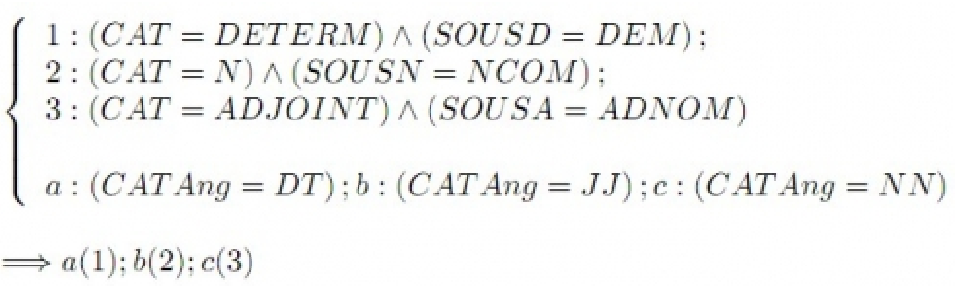
Figure 3: Selecting Sub-sentential Rules presenting crossing links

As a result this constraint formation of elementary patterns will result in contiguous lists of words in both source and target part of it.

The **contiguity** hypothesis plays an important role in the method of recombination from fragments we describe afterwards. As an example, if the alignment memory contains the pattern *"un ciel bleu"-"a blue sky"* and tries to match it with an input bi-sentence in which appears the pair *"un ciel très bleu"-"a very blue sky"*, it will fail. At this point, it can seem unthinkable to constraint the method that much with the contiguity hypothesis, thus creating a useless amount of data, which fragments can be used in very few cases. The **generalization** process is here to bypass this bad effect, and to accelerate the acquisition of useful patterns, matching in various cases.

Generalization is a common process in EBMT (see [Brown, R. D. (1999)]). The fragments collected before are already analysed for segmentation, so every word is labelled with linguistic information such as POS tags, and is replaced by this information set, providing the skeleton of a transfer rule. In the previous segmentation example from figure 1, the rule obtained from the first *GN*∗chunk is:

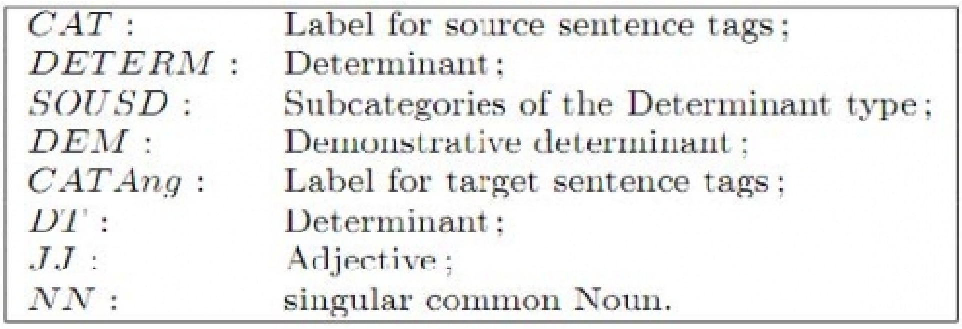
Letters label the source sentence tags and figures, the target sentence tags.



$$\begin{cases} 1 : (CAT = DETERM) \wedge (SOUSD = DEM) ; \\ 2 : (CAT = N) \wedge (SOUSN = NCOM) ; \\ 3 : (CAT = ADJOINT) \wedge (SOUSA = ADNOM) \\ \\ a : (CATAng = DT) ; b : (CATAng = JJ) ; c : (CATAng = NN) \end{cases}$$

$$\Longrightarrow a(1); b(2); c(3)$$

Here is an explanation about the tag-set



CAT :	Label for source sentence tags ;
$DETERM$:	Determinant ;
$SOUSD$:	Subcategories of the Determinant type ;
DEM :	Demonstrative determinant ;
$CATAng$:	Label for target sentence tags ;
DT :	Determinant ;
JJ :	Adjective ;
NN :	singular common Noun.

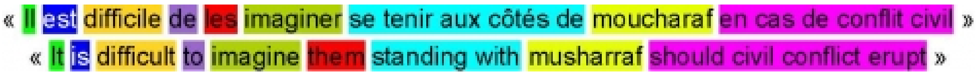
As we said, the patterns consider only POS tags, and lexical resources are never used. This approach tends to rapidly create general rules applicable in many cases. One could object that the contiguity hypothesis weakens the rules generality, making it difficult to represent phenomena such as the French negation "*ne...pas*", but the rules shape has a precise algorithmic purpose and non-contiguous linguistic entities can be covered not by one, but by many rules, or also be pre-processed in a way which does not impede the alignment process.

Recombining the fragments is the process next step. Only a few technical details will be delivered here but the main idea of this alignment approach is to look for a compromise between linguistic model and algorithmic *feasibility*. By feasibility one shall understand 'with reasonable process cost'. The pattern definition is the result of that trade-off. A similar approach with flexible (non-contiguous) patterns would lead to a combinatorial issue for which any polynomial solution is known (in fact it's an NP-hard problem known as the *biclique cover*). Moreover, one can assume that memorizing numerous and short flexible patterns would lead to a very noisy and highly ambiguous database. The contiguous form of the patterns leads to known polynomial recombining solutions. A process for insertion in contiguous linguistically motivated fragments is considered but still not achieved. Let us give some comments about the recombining process: First, given an input bisentence (which can be partially aligned), a set of compatible patterns is extracted from the database. Each of them can be individually applied to the bi-sentence, thus creating new links. The main issue of the recombining process is the incompatibility between patterns giving inconsistent informations when intersecting. Indeed, applying a bad pattern generally prevents the application of at least **two** good patterns. This observation has to be merged with the assumption that a good application of the patterns leads necessarily to a maximum covering alignment. If one assumes that every pattern necessary to build the final and correct alignment is in the database (unfortunately far from true so far), then the correct set of patterns among others (which can be seen as noise) can be obtained by selecting the maximum covering sets of patterns. Of course, the covering condition makes it possible to obtain several sets as a result. The method relies on the contiguous form of patterns to solve this problem in a deterministic fashion with any linguistic measure (lexical, syntactic, distance, mutual information...). The critic raised to these measures ,cut out from observation, is about their discriminative behaviour upon rare (but omnipresent) phenomena such as multiword or collocational expressions with rather syntactic divergence. The approach presented here doesn't try to get rid of them but to provide a progressive treatment where measure should be used after a non-destructive filtering. A simple example where multiple outputs could be proposed is the famous locution "*l'homme est un loup pour l'homme*"-"*man is a wolf to man*": the words *man* and *wolf* are both common nouns and the method presented here doesn't discriminate them with sole POS tags (in addition, "*man*" and "*homme*" appear twice which doesn't help the disambiguating process). If one supposes that the fragments database doesn't contain fragments allowing to match "*l'homme est un loup*" with "*man is a wolf*" or "*un loup pour l'homme*" with "*a wolf to man*" (which is very bad luck), but only very short fragments allowing to link a source common noun with a target common noun. Then, we shall obtain six different alignment results. This kind of ambiguity appears often with small fragments and can't be solved with the simple maximum covering condition (so filtering them before or after the covering process doesn't affect the result). The combinatorial cost can be important in time and in space, that's why the use of other ways can't be denied. The method presented here works fine by itself but takes much more serious sense in being part of a larger process. For instance, an input bi-sentence can be treated by the covering process also if it's partly aligned from the beginning. Thus we can think of a light, partial but precise pre-alignment based on POS information as first reinforcement. But as we saw on the *wolf* example, the lack of large patterns can still induce ambiguities even with lexical post treatment. Because this kind of consideration is fundamental but out of the scope of this paper, we did the hypothesis of a weak divergence between languages and the closest fragment was chosen.

Some examples

We present here a few examples with the segments used in the covering process. The database currently contains 1092 patterns linking 453 French sequences with 404 English ones. As said before, short patterns are error prone and long ones tend to create partial alignments. So, in order to test our method, for the moment, rather short rules have been preferred. As a future improvement, we'll tend to collect larger patterns and, among proposed solutions, to give a better score to alignments containing larger patterns.

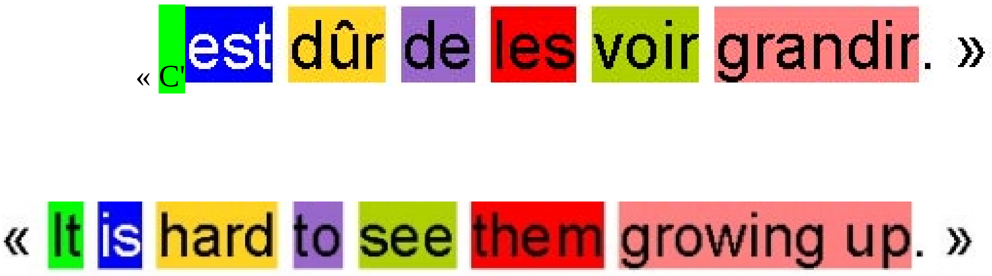
The next example has been used as training information. It is a moderately long sentence from a journalistic corpus. As said before, only short patterns are memorised from this bi-sentence. The only complex pattern (with crossing links) is the couple « les imaginer »-» imagine them ». Other patterns can be recognised from their colour.



« Il est difficile de les imaginer se tenir aux côtés de moucharaf en cas de conflit civil »

« It is difficult to imagine them standing with musharraf should civil conflict erupt »

Multiword expressions skeleton patterns are memorized. « *se tenir aux côtés de* »-» *standing with* », shall lead to a pattern which will also match « *s'asseoir en face de* »-» *sitting with* », « *se passer aux allentours de* »-» *happening near* », … These related expressions wouldn't appear together very often, but when they do, knowing that they are possible translations from each other is a valuable information. Once segmentation done, it enriches the database with 8 segments plus a total segment, kept for evaluation reasons. Then, when the same bi-sentence is proposed as an input, the total pattern will be find and a perfect alignment is proposed. We can also choose to avoid every total pattern when exploring the database in order to test the recombining process. For this example, we have a perfect recombination: The 8 patterns are selected among 71 others based only on the maximum coverage condition. With a pre-aligning process based on cognates (words which strings are the same or almost through translation) the result is the same but the compatible patterns number falls to 43. When a close new input is proposed, some common syntactic information is reused:



« C'est dûr de les voir grandir. »

« It is hard to see them growing up. »

This bi-sentence is from two English en French movie subtitles corpus. The previous example segmentation helps the system to align this new sentence even if it is shorter. So aligning long bi-sentences doesn't exclusively relies on short previous examples. Aligning « *grandir* » with « *growing up* » didn't need lexical information but was allowed for, with the help of a previous training example involving a phrasal verb. The 6 patterns here were discriminated among 20 others. We encountered and tried to train the system with more divergent examples, such as sentences presenting crossovers through natural translations. For instance:

« Il a traversé la rivière à la nage » « He swam across the river »

Divergence of opinions can lead to different alignments. Users could cross links in order to group sense units together: « nage » with « swim » and « traversé » with « swam ». But one could think that we are precisely in a case of a multiword expression and link each word with every other:

« Il a traversé la rivière à la nage » « He swam across the river »

In both cases, the segmentation process capture the minimal pattern respecting the contiguity hypothesis that is « *a traversé la rivière à la nage* »-» *swam across the river* ». The links into the two patterns will be conflictual but both solutions are consensual and the contextual recombining process will be the same. Other simple crossover examples are treated the same way: « *La foule s'écarte à reculons* »-» *The crowd backed away* » or « *Il traversa la pièce à toute allure.* »-» *He hurried through the room* ». Many linguistic phenomena present a multiword expression aspect which this method try to capture and reuse.

Conclusion

In this paper, we have described an example-based method that exclusively uses syntactic information during the different steps of the process of aligning phrases in a bilingual corpus. The basic incentive was to study the impact of syntax as a back-side information set in this task, and thus we needed to isolate it from semantic or lexical information existing in dictionaries, basing our motivation on the following conclusion: If automatic lexical transfer has improved a lot, the syntactic shape of automatically translated sentences still demonstrate a poor quality. The issue was to understand how syntactic information had to be accounted for in order to improve syntactic transfer.

In the same way, we separated alignment from translation, since we consider alignment as a learning task for translation, which might enrich translation memories with non obvious information. Among these 'non obvious' phenomena, multiword expressions, that are divergent in source and target languages, have been spotted as an interesting issue which has not been directly dealt with by the abundant domain literature. We suspected that syntax played a part in multiword expressions alignment, and a method was designed to tackle this issue. Deep syntactic analysis was used to separate and collect fragments from examples provided by the user. Then again, these fragments from bi-sentences were generalised using POS-tags. Their constrained form (the contiguous hypothesis) played an important role in the recombining effort based on maximising the coverage. In benefits, the method was able to align precisely bi-sentences at a phrasal level when relevant information was found in the database. The fragments genericness allowed the method to successfully align phrases which syntactical structure was met in a previous treatment. We could collect multiword expressions and their translations over the alignment process but, for the moment, did not endeavour to separate those relying on generic transfer rules or collocational type from those inherent to inconsistent translations. Measuring this information has become a goal. We also need to adjust the level of grammatical information harvested from POS-tagging to relevance, which has a direct impact on the genericness of patterns. Then, with sufficient amount of data, we shall measure the evolution of quality matching in the database size. This approach leads to an important issue concerning the contiguous form of the concerned fragments. Indeed, their lack of flexibility makes one unable to match a relevant pattern when polymorphisms or insertions occur. A solution could be to extend requested patterns by modifying their syntactic structure before mining the fragments database and so, allowing a fuzzy matching which is already an issue at stake in EBMT literature.

Bibliography

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D., Lafferty, Robert L. Mercer, and Roossin Paul S. (1990) « A statistical approach to machine translation » . In *Computational Linguistics*, *16.*

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) « A statistical approach to machine translation » in *Computational Linguistics, 19.*

Ralf D. Brown. (1999) « adding linguistic knowledge to a lexical example-based translation system » in *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation.*

Jacques Chauché (1984) « Un outil multidimentionnel de l'analyse du discours » In *COLING*.

Colin Cherry and Dekang Lin (2003) « A probability model to improve word alignment » in *41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp 88–95.

Lambros Cranias, Harris Papageorgiou, and Stelios Piperidis (1994) « A matching technique in example-based machine translation » in *COLING*, pp 100–104.

Fabrizio Gotti, Philipphe Langlais, and Claude Coulombe (2006) « Vers l'intégration du contexte dans une mémoire de traduction sous-phrastique : détection du domaine de traduction » in *TALN.*

Fabrizio Gotti, Philipphe Langlais, Elliott Macklovitch, Didier Bourigault, Benoit Robichaud, and Claude Coulombe (2005) « 3gtm: A third-generation translation memory. » in *3rd computational Linguistics in the North-East (CLiNE) Workshop.*

Declan Groves, Mary Hearne, and AndyWay (2004) « Robust sub-sentential alignment of phrase-structure trees. » in *COLING.*

Mary Hearne and Andy Way (2003) « Seeing the wood for the trees: Data-oriented translation. » in *MT Summit IX*, pp 165–172.

P.and Nesi H. Howarth (1996) « The teaching of collocations in eap » *Technical report, University of Leeds.*

Julian Kupiec (1993) « An algorithm for finding noun phrase correspondances in bilingual corpora » in *Proceedings of the 31nd Meeting of the Association for Computational Linguistics.*

Yajuan Lü and Ming Zhou (2004) « Collocation translation acquisition using monolingual corpora » in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04).*

Hiroshi Maruyama and Hideo Watanabe (1992) « Tree cover search algorithm for example-based translation » in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92),* pp173–184.

Dan Melamed (2000) « Models of translational equivalence among words » in *Computational Linguistics 26*, pp 221–249.

Arul Menezes and Stephen D. Richardson (2003) « A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora » in *DDMR Workshop, ACL.*

Makoto Nagao (1984) « A framework of a mechanical translation between japanese and english by analogy principle » in *Artificial and Human Intelligence: Edited Review Papers Presented at the International NATO Symposium on Artificial and Human Intelligence*, pp 305–332.

Sergei Nirenburg (1993) « Two approaches to matching in example-based machine translation » in *Proceedings of TMI'93.*

Franz Josef Och and Hermann Ney (2003) « A systematic comparison of various statistical alignment models » in *Computational Linguistics,* 29(1), pp19–51.

Sylvia Ozdowska (2006) « ALIBI, un système d'ALIgnement BIlingue à base de règles » in *PhD thesis, Université de Toulouse 2.*

Satoshi Sato and Makoto Nagao (1990) « Toward memory-based translation » in *COLING,* pp 247–252.

Helmut Schmid (1994) « Probabilistic part-of-speech tagging using decision trees » in *International Conference on New Methods in Language Processing*, pp 44–49.

Violeta Seretan (2009) « Extraction de collocations et leurs équialents de traduction à partir de corpus parallèles » in *TAL, 50* pp305–332.

Frank Smadja. Smadja (1992) « How to compile a bilingual collocational lexicon auomatically » in *proceeding of the AAAI Workshop on Statistically-Based NLP Techniques.*

Frank Smadja and Kathleen R. McKeown (1990) « Automatically extracting and representing collocations for language generation » in *proceeding of the 28th Annual Meeting of the Association for Computational Linguistics.*

Harold L. Somers (1999) « Review article: Example-based machine translation. Machine Translation », 14(2):113–157.

Stephan Vogel, Hermann Ney, and Christoph Tillmann (1996) « Hmm-based word alignment in statistical translation » in *COLING' 96: The 16th international Conference on Computational Linguistics*, pp836–841.

Kenji Yamada and Kevin Knight (2001) « A syntax-based translation model » in *39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp523–530.

David Yarowsky (1993) « One sense per collocation » in *ARPA, Human Language Technology Workshop.*