Alignment Memories: A Useful Tool to Handle Phrase Alignment Bottleneck

Johan Segura, Violaine Prince

Two Memory-Based Methods for Phrase Alignement

SEGURA Johan
segura@lirmm.fr

PRINCE Violaine
prince@lirmm.fr

***Abstract*—This document presents two bilingual phrase-based alignment methods handling syntactic constituents (sub-sentential components) of parallel sentences. The methods relie on an asymmetrical parsing of both languages: Light part-of-speech tagging for the target language, syntactic tree building for the 'source' language and the complexity of each is studied. One of their benefits is that they do not require lexical knowledge for granting alignment. Another is that they align constituents of variable length and structure, thus providing information about divergent translations. Their originality rely on the fact that parsing of the supposed source language is reused both in resource building and alignment process. The models and methods can be seen as a subclass of Example Based Machine Translation. The information acquisition process, partly supervised is embbeded in an online graphical human interface which accelerate the construction of golden corpora by one or many users.**

I. 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. It is seen as a cornerstone in MT. Sub-sentential alignment needs parallel bilingual corpora. It aims at automatically providing translation links between sentences *constituents*, i.e., words or multiword 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. Classical models, still representative, focuse on word-to-word alignements. Late researches in alignment tend to privilege a granularity bigger than the single word (e.g. [6], [8]). Detecting relevant phrases for alignment can motivate the use for syntactical information. In representative rule-based systems, rules are either applied in a pre-ordained fashion, or in a "first best-value" approach (statistically based, thus mixing statistical and symbolic methods). In different cases, rules overlapping conflicts are differently solved. Most of the time, such process relevance is less discussed than rule shapes. We will try here to discuss the role of rules shape through both problematics of tractability and linguistic relevance. We will also discuss the choices of the shapes proposed for the rules and the role they play in the alignment process. The methods described hereafter are example-based methods that use an 'alignment memory', which is a learned set of segments. These segments can be seen as bilingual phrase couples presenting internal links. The process asynchronously combines alignment constraints in order to maximize coverage (in an EBMT style) . The method is partly supervised: The present system introduces a learning feature. The information acquisition process is facilitated by a graphical human interface. One of the original features of this method is that the process can align word segments as well as syntactic patterns. It relies on an asymmetrical effort in syntactic processing: A constituent and dependence analyzer is used for the source, and a POS tagger for the target. No dictionary or lexical resources are *a priori* required. The next section details related alignment methods. Section 3 presents our model, and section 4, a first experiment with some results. Conclusion will shed light on the work extensions and further developments.

II. Related Works: Tree-based Alignment Methods

The literature on alignment is abundant, and some works have already been mentioned in the introduction. The founding work in alignment is attributed to Brown et al. at IBM [2]. The GIZA++ system [16] which is based on these 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. It's still widely used in alignment literature (e.g. in [11], [6]). Syntactic trees as elements of the alignment process have appeared with [21]. Since then, hybrid systems, embedding syntactical information in a statistical model emerged as well as purely symbolic approaches. The use of structural information brought by syntax is claimed to be helpful for different reasons among which we can quote :

1) Preventing alignments violating linguistic structural properties (e.g.,[5])
2) Propagating alignments according to parent-child links (e.g., [13] [17])
3) Predicting an alignment with a POS tag, when the lexicon does not provide information [5]
4) Generating structures that accelerate the rule-base building process in a data driven approach (e.g. [12]).

Another aspect of this model tries to take advantage of the syntax: it is an example-based alignment model sharing common issues with example-based MT (EBMT). EBMT tries to imitate the human translation by analogy. It is an intuitive approach consisting in storing pieces of translations already met in the past, getting the relevant ones in a new situation, then combining the pieces to obtain a solution. The first suggestion of EBMT issues is attributed to Makato

Nagao in [14]. He clearly defined the three important steps of an EBMT process: Matching fragments from a database, filtering and combining. Nagao claims that a human translation process doesn't involve a deep analysis structure but relies on analogy with generic fragments. This idea motivates the whole example-based approach. EBMT litterature agrees that fragments size must be at the sub-sentential level for 'genericness' reasons [7] (exact same sentences occur only rarely) but raise in turn the issue of recombining fragments in a way that preserves the language structure and the sense [18]. Furthermore, it is know that linguistically motivated patterns are of a benefits [11]. For these reasons we thought it was necessary to resort to deep syntactic informations to tackle the segmentation part. Then, when recombining from examples, one must choose a good matching measure. In [15], he author 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 classical argument in EBMT can also be found in SMT phrase-based methods like in deNero [8]. Our method sticks to this approach, although recombing effort in an alignment method is quite different from an EBMT as we'll see in the next section. Finally, the shape the patterns should take in EBMT is also motivated by a correct reuse. Efforts must be done to make the fragments as generic as possible without losing consistency with the recombining process. The pattern generalization of Brown's method [3], which uses syntactic analysis to replace some words with their classes or categories, generalizes them to a much wider set of applications. This approach emphasizes the gain of generalization by showing an accelerating efficiency in the treatment. The methods detailled hereafter make extensive use of the generalization thanks to POS tags informations.

The aim of this work is to try to evaluate the viability of an original aligner close to EBMT paradigms, deterministic and asynchronous. As an early experiment, we reduced the use of lexical information to a strict minimum, then allowing to handle non-compositional translations and accelerating fragment acquisition. It is certain that, in some future work, a word to word alignment based on lexical information will be considered since the model is meant to be embedded in some larger process. Thus, a crucial perspective of this work is to enrich a translation memory as a sort of 'super' lexicon of equivalent expressions, involving stylistic idiosyncrasies of both languages.

III. Model and Method

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 [4] which provides a deep syntactic tree. TreeTagger [19] 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). Therefore the method is asymmetrical. 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 fragments or patterns and then are memorized. These patterns will be used in an alignment process and, individually, concern a phrase-level scope.

A. Elements of the model

The model relies on a set of fragments, which implement assumptions. Basically, a fragment can be seen as a phrase couple of contiguous POS-tags presenting internal links. We describe them hereafter as alignment rules since each can implement different pairs of word phrases. A fragment is divided into two parts:

- The condition part: A condition on a word is a formula without negation involving POS-tags values.
- The application part: A set of alignment actions based on the condition checking (the internal links).

1) Admissible conditions: Let $(K_n)_{0\leq n\leq N_K}$ be a finite set of categories for source language and an other one for the target $(K'_m)'_{0\leq m\leq N_K}$. They can be instantiated by values from the two sets: $(v_n)_{n\in\mathbb{N}}$ and $(v'_n)_{n\in\mathbb{N}}$.
A **condition** (recognized by the model) on a source term will be :

$$(K_{k_1} = v_{k_1,1} \vee \ldots \vee v_{k_1,n_1}) \wedge \ldots \wedge \left(K_{k_p} = v_{k_p,1} \vee \ldots \vee v_{k_p,n_p}\right)$$

A condition on a target term with the set K' and its values v' is defined in the same way.

Example:
SYGFRAN has a set of POS tags, among which: CAT signifies *POS category*, N is for *noun*, $SOUSN$ means *nominal subcategory*. $NCOM$ stands for *common noun*, and $NPRO$ for *proper noun*. SYGFRAN tags are numerous and a choice was made from the begining to use only some of them. The condition below represents a word which analysis could not determine whether it is a common or a proper name :

$$(CAT = N) \wedge (SOUSN = NCOM \vee NPRO)$$

The admissible conditions recognized by the model deal with both the source and target sentences. A **well-formed condition** is when both source and target conditions are realized in the **bi-sentence** on contiguous terms. A bi-sentence is a pair of source and target sentences, aligned on the sentence level only.

Let $\Gamma S_1, \ldots, \Gamma S_n$ be a list of **conditions** for source terms and $\Gamma T_1, \ldots, \Gamma T_m$ for target terms. An **admissible condition** will be noted as follow :

$$\begin{cases} 1 : \Gamma S_1; \ldots; n : \Gamma S_n \\ 1 : \Gamma T_1; \ldots; m : \Gamma T_m \end{cases}$$

This condition will be **realized** in a bi-sentence if a contiguous list of terms from the source sentence respect each condition ΓS_i in the right order **and** if a list of contiguous terms from

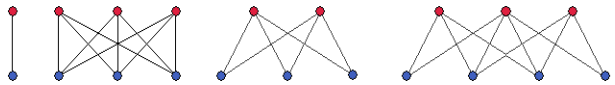

Fig. 1. Correct Minimal Alignments Represented as Bicliques

the target sentence respect each condition ΓC_j, also in the right order.

Example: The condition below would be realized on the pair *"un ciel bleu"-" blue sky"*, with the following definitions for the used POS tags.

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

Tags belonging to SYGFRAN:
$DETERM$: Determinant;
$SOUSD$: Subcategories of the Determinant type
$ARTD$: Determinate article (e.g. 'the' in English)
$ARTI$: Indeterminate article (e.g. 'a' in English)
$ADJOINT$: Adjoint type (adjectives, adverbs)
$SOUSA$: Subcategories of the adjoint type
$ADNOM$: Adjectives qualifying a noun

Tags belonging to TreeTagger:
JJ: Adjective
NN: Common Noun, singular

The **contiguity** hypothesis plays an important role in our method. The previous condition won't be realized on the pair *"un ciel très bleu"-"a very blue sky"*, that will be implemented in a larger pattern. So, the phrases concerned by the patterns:

- have an arbitrary length
- contains only contiguous words

2) Application of a rule: If a condition part is realized on a contiguous part of the bi-sentence, the application part provides a way of linking each term concerned by the condition. The *application part* of a rule respects the natural *biclique* shape of the links in order to create correct alignments built from non-intersecting bicliques. A **biclique** is a special kind of bipartite graph where every vertex of the first set is connected to every vertex of the second set. One can see figure 1 for examples of bicliques representing minimal correct alignments. 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.
This last condition allows us to create alignments consisting in non-intersecting bicliques, that we assume to be a rather natural definition beyond which the notion of alignment would be meaningless. A rule can be applied if it doesn't violate a link already present in the bi-sentence.

Example:
An admissible rule to be applied on the pair: "*à la Cour*", "*at Court*" could be written as such:

$$\begin{cases} 1 : (CAT = PREP) \wedge (CATPREPSIMPLE = A)\,; \\ 2 : (CAT = DETERM) \wedge (SOUSD = ARTD)\,; \\ 3 : (CAT = N) \wedge (SOUSN = NCOM \vee NPRO)\,; \\ \\ a : (CATAng = IN)\,; b : (CATAng = NP) \end{cases}$$

The paradigm alignment we stand on is larger than the word-to-word one.

Tags belonging to SYGFRAN:	
$PREP$: Preposition	$CATPREPSIMPLE$: Simple Preposition

Tags belonging to TreeTagger:	
IN:Preposition	NP:Proper Noun, singular

B. Saving the rules

Rules are learned from hand-aligned or semi-automatically[1] aligned data. Rules violating the conditions seen above are skipped or adapted. Moreover, a tool preventing the user from creating 'degenerated' alignments is used: A graphical human interface facilitate this work (fig. 7) for one or many annotators. Thus, the human alignment process can start from scratch, from another user work or from an automatically provided alignment. The amount of collected data cannot compete with statistical models but aims to devellop golden corpora of quality. Segmentation of the aligned bi-sentence lays on both tree analysis from source sentence and the links from the alignment which reflects transfer. 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. Thus, the syntactic structure of the parsed source sentence allows one to cut out the total alignment into several relevant aligned bi-phrases producing valid rules. For instance, Figures 2 and 3 show a constituent tree for a source sentence, its leaves being source words, and how the target sentence words could be aligned according to a subtree division of the basic syntactic tree. The segmentation is done along the inner nodes. The root node provides us with the less generic segment while the deeper ones (included in the *root-segment*) shall be the most commonly reused. The rule obtained from the first GN in figure 2 chunk is (GN being a SYGFRAN tag for 'Noun Phrase'):

[1]Aligned with a computational tool and approved by human judges.

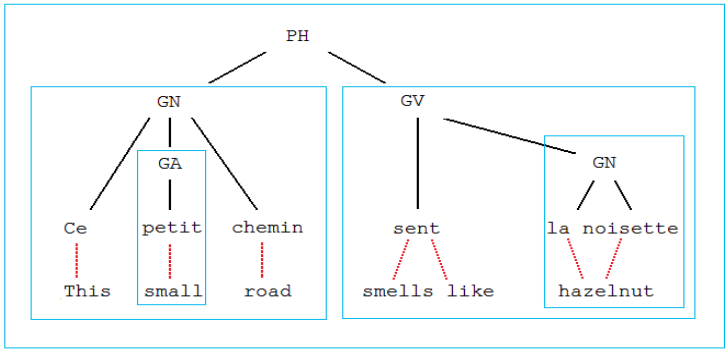

Fig. 2. Selecting Sub-sentential Rules

Ce	: CAT=DETERM;SOUSR=DEM;
petit	: CAT=ADJOINT; SOUSA=ADNOM;
chemin	: CAT=N; SOUSN=NCOM;
sent	: CAT=V;
la	: CAT=DETERM; SOUSD=ARTD;
noisette	: CAT=N; SOUSN=NCOM;
This	: CATAng=DT;
small	: CATAng=JJ;
road	: CATAng=NN;
smells	: CATAng=VBS;
like	: CATAng=VB;
hazelnut	: CATAng=NN;

Fig. 3. Labels From French and English Trees Leaves: Note That the tagging for 'like' is wrong!!

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

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

The rules consider only POS tags: 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 that does not impede the alignment process. For instance, to be fully taken into account, "*ne...pas*" should be handled by segments such as : "ne [Verb] pas", "ne [Verb] [Adverb] pas", and so on.A fragment can include several phrases when divergence is too high. In the next part, we comment two different segmentation paradigms we used, with different recombining treatments.

C. Combining fragments

Let us give some comments about the recombining process: First, given an input bi-sentence (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, and so on. 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 methods rely on the contiguous form of patterns to solve this problem in a deterministic fashion. Thus, both methods are based on solving a maximum-covering graph problem. Two ways of combining fragments are here presented: Each of them depends on a different choice in segmentation. The complexity of proposed resolutions is also an asset. From a new bi-sentence to be aligned at the sub-sentential level, one has to collect compatible candidates in the partial rules database (the saving memory). *Combining them to obtain an optimum alignment consists in selecting a maximum covering set of compatible rules.* Many maximum-independent-set issues are known to be NP-hard (such as many alignment problems [8]). As an example, if we were to extend our set of rules to only complete sets of connected nodes, which is the most general possible shape (cf fig.1), the combining process would lead to the NP-hard biclique decomposition problem*([10]).

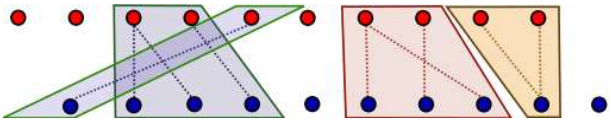

Fig. 4. Patterns presenting *a crossing* and *following* configuration

*1) Contiguous fragments*** : The fragments contiguous shape results from a need to use a lighter recombining process in a graph approach. Indeed, working among this less general class of segments weakens the problem. Let us reformulate the problem in a graph situation: Each rule is a weighted node of our graph, which weight is the coverage of the rule. There is an edge (non-oriented) between two nodes if the associated rules are compatible. *Building an optimal alignment is seen as finding a maximum-weight clique in the compatibility graph.* Actually, even in this specific framework, we encounter algorithmic difficulties which we shall detail here.

Two independant rules of this model can clearly be in a crossing of a following configuration as illustrated in figure 4. Looking for a maximal independent set of rules which are

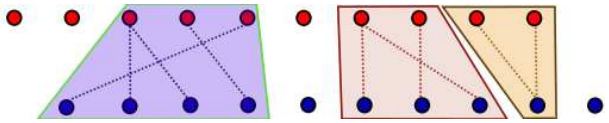

Fig. 5. Rules including *a crossing* configuration

pairwise either crossing or following is still an NP-hard issue known as the "maximum weighted independent set of axis parallel rectangles" [1]. So far, it seams that the expressiveness of the segment shape is still too high to reduce significantly the recombining process cost: Apart their linguistic justification, their impact on the recombination effort is substantial. As a first experiment, we wanted to evaluate a heuristic recombining process over this approach (with the contiguous rules described before). A maximum covering set of *following patterns* (fig 4) could be built in a polynomial time ([20]). Doing so, the solution proposed by the system will not present any crossing link. It means that every needed crossing configuration in the final alignment would lead, with this method, to a choice between links, thus generating holes, and then errors. An alternative, we chose here, consists in a pre-treatment among the set of compatible fragments to deal with crossing links. When two fragments present crossing links and form a larger contiguous pattern, as in figure 4, they merge to form a new rule added to the database. This pre-treatment, when the number of contiguous-crossing configurations is reasonable, can lead to an exact recombining procedure, but most of the time, the combinatorial effort of dealing with crossing configurations is too heavy and one has to use a threshold or at least a filter. In this last case, (the most common one on long sentences), the method is an approximation.

*2) Contiguous fragments capturing crossing configurations ***:* As an alternative and a second experiment, we proposed another segmentation process, extending the first one and leading to a tractable compatibility resolution algorithm: We decided to capture the crossing configuration during the segmentation process to avoid the combinatorial cost of dealing with them. Indeed, compatible fragments will be in *following configuration*. One can observe the new segment shape in figure 5. Of course, segments will provide us with less generic features (especially with great divergence). Recombining alignment among patterns thus formed is known as the maximum independent set problem for trapezoid graphs. Light combinatorial algorithms exist to solve this problem: $O(n\ log\ n)$ where n is the number of compatible fragments [9]. Unless divergent behavior is pre-treated with, for example, a word-to-word alignment, this method will tend to favor a left to right alignment.

IV. An overview of the tool

The alignment tool (fig. 7) which is currently in development will allow to constitute single-handedly, or in collaboration with partners, an aligned corpus of quality as it is necessary for a reference corpus. Entirely parsed parallel

Minimal segment*	Recombining cost
biclique	NP-hard
*contiguous union of bicliques***	NP-hard
*contiguous and non-crossing with context****	$nlog(n)$

Fig. 6. Complexities summary

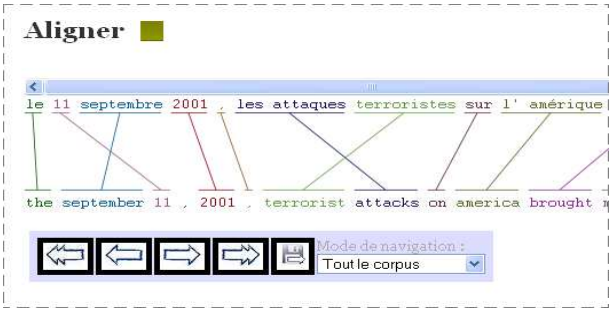

Fig. 7. A part of the human graphical interface

corpora are provided upon which we can built new sub-sentencial alignements. The tool make it possible to navigate a corpus from a bi-sentence to another through different options. The display of each bi-sentence with its sub-sentencial links is intuitive and interactive, allowing one to modify them easily with the mouse and with hotkeys. Thus different automatic tools shall be embedded, including a classical lexical alignment process based on dictionnaries as well as the syntactic pattern based methods at question here.

3) Navigation: Once the corpus is chosen, one can navigate through one's own or a group of users' alignments. Every modification can be saved and shall remain so for the user. Links are visible so it's possible for the user to easily judge its quality. Since one can navigate alignement produced by other chosen members, the work can be in a collaborative fashion. Selecting alignments from members whom links choices shall be an option so users can discuss the differences in an appropriate space. It shall possible at any moment to convert a collaborative work into an XML file for some further purpose.

4) Interaction with the user: Both parts of the bisentence are arranged in two parallel lines regarless of their length. Each will be separately scrollable so links can be finely observed. Pre-parsing the corpora is notably useful to segment properly words and take account of phenomenon such as composed-words. Thus, basic clickable elements will be words. Modifications can start from another user's alignment, thus making work collaborative. If one save an alignment involving several users, it will not affect other people's work, only the current user is concerned. One can select words to align with the mouse, a right-click shall validate and draw the links. Right-click is also used to erase links when modifying an alignment. Saving an alignment serves two purposes: First, user's current work is recorded, then the alignment memory is enriched. Segmentation based on previous parsing of source language

will produce syntactic patterns which can be used immediatly in the automatic process depicted here. Therefore, the tool can be said semi-automatic in the sense that human judgement is in the loop.

5) Current state of the tool: Currently, both graphical tool and automatic methods are implemented, but are still not merged as a single complex tool. Experiences were led with a former tool not depicted here.

V. A First Evaluation

A first set of $67,941$ pairs of sentences was extracted from WMT'09 workshop journalistic corpus [2]. French sentences present an average length of 27 words and English ones 23. In this section, we call *"contiguous fragments"*, patterns obtained from the first segmentation described in figure 4. *"crossing fragments"* will refer to the second ones from the extended segmentation including crossing configurations (figure 5). We hand-aligned 100 bi-sentences as a first training. The system fragmented each memorized pair into generalized patterns. We ran several experiments. First, we tested our recombining process by trying to align those same 100 bi-sentences: The total alignment was memorized, but we inhibited large patterns during the database mining phase, susceptible to align in one shot, in order to test the recombination among short patterns (knowing every needed pattern was effectively in the base). The idea was to evaluate the recombining process over the two segmentation stages. In the table summarizing results, R_{100} stands for the recombining process over the *contiguous fragments* and RX_{100} over the *crossing fragments*. We consolidated this alignment by using a pre-treatment cognates detection: Short patterns can lead to syntactical ambiguities sometimes quite frustrating when aligning a proper noun with an omitted uppercase first letter, with a common noun. Cognates detection was based on a Levenshtein measure and we noted an average number of 4 cognate pairs per bi-sentence (e.g.: *"musharraf"-"moucharraf", "judges"-"juges", "unpopularity"-"impopularité"*,...). $R_{100}^{Cog.}$ and $RX_{100}^{cog.}$ are the same experiments using cognates alignment as reinforcement. Then results are much better when cognates are used as anchor. No mistakes were found in the cognate detecting process. Results have shown that recombining experiments are quite successful with the *crossing fragments*, since the process has been tailored for their needs. Capturing crossing links during segmentation, has engulfed the main liability of the alignment process, thus leaving to recombination a minimal effort. It amounts to searching a database of already saved patterns and looking for following configurations. Then, we tried to align 100 fresh bi-sentences which were not from the training set. This time, alignment was performed with cognates reinforcement. Al_{100} and AlX_{100} designate the aligning experiment on the 101-nd to 200-nd bi-sentences based on the training over the first 100, respectively for the contiguous and the crossing fragments. Of course, the amount of data is insufficient to draw strong conclusion or to give predictions for the future evolution of the system, but we observe that the lack of generic features we feared for the *crossing fragments*, does not impede the recombining process to reach results which quality is equivalent to the experiment with the *contiguous fragments*.

The two methods lead to an identical F-score, but the second method seems to have a lesser recall, thus corrupting its performance. This can be explained by the fact that the recombining process maximizes the bi-sentence coverages with following positions fragments. This tends to create holes when two followings are not adjacent. The first method, a heuristic, tried to maximize coverage among the fragments in following or crossing configurations. When adding crossing fragments in the process, the second method reduces recall.

In order to measure the alignment quality, we had then to hand-align this 100 bi-sentences which provided us with an enriched database, so we ran over the first experiment consisting in evaluating the recombining process over the 200 bi-sentences already aligned with and without cognates thus trying to observe improvement or on the contrary a degradation. There were no significant differences. These experiments are referred to as R_{200} and RX_{200}, $R_{200}^{Cog.}$ and $RX_{200}^{cog.}$ in the results table. As a comparison, we gave Giza++ model results (from French to English) on the same pieces of the corpus (trained on the $67,941$ bi-sentences with the IBM model 4) although the two systems are definitely different: The sizes of needed training corpora, information used, and theories are hard to compare. No additional heuristic was used, the results are here as a baseline reference. In the table below, "P" stands for "precision", "R" stands for "recall", and "F" for the classically used F-measure. Let us note the very high values of the F-measure for the "X" based experiments, except "Al", which is invariant.

	R_{100}	$R_{100}^{Cog.}$	R_{200}	$R_{200}^{cog.}$	Al_{100}	Giza
P.	84%	92%	85%	91%	77%	75%
R.	82%	86%	83%	88%	52%	60%
F.	0.83	0.89	0.84	0.89	0.62	0.67

	RX_{100}	$RX_{100}^{Cog.}$	RX_{200}	$RX_{200}^{cog.}$	AlX_{100}
P.	98.7%	99.5%	97.9%	98.9%	82.3%
R.	97.2%	97.7%	97.1%	97.8%	49.9%
F.	0.98	0.99	0.98	0.98	0.62

VI. Conclusion

In this paper, we have described an example-based aligning method that almost exclusively uses syntactic information during the different steps of the process (the use of cognates is the only recourse to lexical information). Deep syntactic analysis was used to separate and collect fragments from examples provided by users. Then again, these fragments from bi-sentences were generalized using POS-tags. Alignment was performed between fragments recognized from a database filled with syntactic correspondences. Two databases were built, suitable for two different process using different segmentations: The first, producing *contiguous fragments* was

[2]http://www.statmt.org/wmt08/

followed by a heuristic recombining process, while the second, providing (*crossing fragments*), led to an exact solution. The constrained form of the two segmentation processes we used, played an important role in the recombining effort based on coverage maximization. The shape of the memorized fragments seems to play an important role in the recombining process. In such an approach, a trade-off should be found between the fragments genericness and their combinatorial weight: The exact process using "crossing patterns" showed an almost perfect recombination of information. It pointed out the difficulty of crossing configurations in the alignment process, which should be carefully studied as a future work. Also, different heuristic resolutions should be tested on the *contiguous fragments* memory. This first evaluation showed promising results, while quite a good precision was reached after a light training on only a hundred bi-sentences. With sufficient amount of data, the evolution of quality matching in the database size could be measured, and a more precise difference between the two approaches would then be better observed.

REFERENCES

[1] V. Bafna, B. Narayanan, and R Ravi. Nonoverlapping local alignments (weigted independant sets of axis parallel rectangles). *Discrete Applied Mathematics*, 71:41–53, 1996.

[2] 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. A statistical approach to machine translation. *Computational Linguistics*, 16, 1990.

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

[4] Jacques Chauché. Un outil multidimentionnel de l'analyse du discours. In *Coling*, 1984.

[5] Colin Cherry and Dekang Lin. A probability model to improve word alignment. In *41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 88–95, July 2003.

[6] Colin Cherry and Dekang Lin. Inversion transduction grammar for joint phrasal translation modeling. In *NAACL-HLT*, 2007.

[7] Lambros Cranias, Harris Papageorgiou, and Stelios Piperidis. A matching technique in example-based machine translation. In *COLING*, pages 100–104, 1994.

[8] John DeNero and Dan Klein. The complexity of phrase alignment problems. In *ACL (Short Papers)*, pages 25–28, 2008.

[9] S. Felsner, L. Mller, and L Wernisch. Trapezoid graphs and generalizations, geometry and algorithms. *Discrete Applied Mathematics*, 74:13–32, 1997.

[10] H Fleischner, E Mujuni, D Paulusma, and S Szeider. Covering graphs with few complete bipartite subgraphs. In *27th FSTTCS, volume 4855 of Lecture Notes un computer Science*, 2007.

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

[12] Mary Hearne and Andy Way. Seeing the wood for the trees : Data-oriented translation. In *MT Summit IX*, pages 165–172, 2003.

[13] Arul Menezes and Stephen D. Richardson. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. *DDMR Workshop, ACL*, 2003.

[14] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence: Edited Review Papers Presented at the International NATO Symposium on Artificial and Human Intelligence*, pages 305–332, 1984.

[15] Sergei Nirenburg. Two approaches to matching in example-based machine translation. In *Proceedings of TMI'93*, 1993.

[16] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[17] Sylvia Ozdowska. *ALIBI, un système d'ALIgnement BIlingue base de règles*. PhD thesis, Université de Toulouse 2, 2006.

[18] Satoshi Sato and Makoto Nagao. Toward memory-based translation. In *COLING*, pages 247–252, 1990.

[19] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[20] Stéphane Vialette. On the computational complexity of 2-interval pattern matching problems. *Theor. Comput. Sci.*, 312(2-3):223–249, 2004.

[21] Kenji Yamada and Kevin Knight. A syntax-based translation model. In *39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 523–530, July 2001.