ROSETTA: Rhetorical and semantic environment for text alignment

Hatem Ghorbel, Afzal Ballim, Giovanni Coray
LITH-MEDIA group
Swiss Federal Institute of Technology
IN Ecublens, 1015 Lausanne, Switzerland
Phone:+41-21-693 52 83 Fax:+41-21-693 52 78
{hatem.ghorbel, afzal.ballim, giovanni.coray}@epfl.ch

Abstract

In the framework of machine translation of multilingual parallel texts, the technique of alignment is based on statistical models and shallow linguistic parsing methods. When addressing the problem to corpora where different versions are derived or interpreted from the same source, we need further criteria to consider the forms of disparity between these versions. In this article we propose a content-driven approach based on the semantic and pragmatic structure of texts to aid in the process of alignment and comparison between their versions.

1. Introduction

Alignment is the process of establishing the relationship between the different subparts of two or more comparable documents. Much of the early work on alignment is still used as the basis for more advanced systems. These methods are mainly based on statistical models of translated texts, with some based on word or character frequencies, others on string occurrences. Whereas previous work in alignment has viewed texts as essentially a flat stream of characters or words, other approaches have incorporated some structural properties of the documents as further criteria. For example the logical structure of the documents (e.g., sections, chapters, titles, etc.) is often used. In this paper, we propose an extension to the existing methods where we consider the semantic and pragmatic structure of texts as the basic criteria to aid in the process of forming a correspondence between a parallel pair.

We are focusing on parallel corpora where different versions are derived from the same material (intra or inter lingual documents). The practical goal of alignment is to give experts, students or ordinary users a tool that facilitates the on-line comparative analysis of ancient texts and to navigate through the various components of the different versions. Some experiments on ancient manuscripts of medieval French have shown the limits of statistical alignment due to the considerable variation of these versions, which exhibit omissions, insertions and substitutions that range from words to sentences and sometimes to larger spans of texts. Despite these lexical and syntactic variations, the semantic content i.e the meaning is kept invariable. This presumes that these versions have very close semantic structure. We argue this hypothesis with the fact that each text holds a semantic content within its lexicon and its linguistic structure. Generally, this content remains invariable when the text is translated or when a further version is reproduced. The variation would be at the lexical level or at the level of details added or omitted. The main ideas and the intentions of the writer are however kept the same. Based on this hypothesis, if we manage to model the semantic content by means of an appropriate discourse structure, we can compare the content from the perspective of such a structure.

As a very end goal for alignment, we intend to provide an environment to facilitate the analysis of ancient texts and to enable data access of contextually relevant materials such as slides of original books, commentaries from rare books, annotations added by domain experts. This meta-information is often divorced from the original versions; thus making it easily available to users provides a powerful educational mechanism for specialized studies.

2. Background

When the idea of fully automated high quality translation was given up by most of researchers, semi-automatic and assisted translations became the goal of most of the current projects of translation. To support such approaches many linguistic resources (machine-readable dictionaries, thesauri, tagged corpora, etc.) and knowledge bases (Wordnet) were built. Parallel corpora were important bases of

multilingual resource that complements previous classical tools of translation either for human or for machine. Basing on translated segments in parallel corpora, some approaches of *memory-based* or *example-based* machine translation were developed (Sato & Nagao 1990, Sumita et al. 1990). Parallel corpora were used then for other natural language applications, for instance in translation consistency checking, cross-language information retrieval, document comparison and other multilingual applications. The problem of alignment and establishing correspondences between segments of documents became an issue and an end in itself.

3. The alignment problem

3. 1. Alignment of same-text versions

Same-text versions are different presentations of texts derived or based on the same original document. These versions represent different same-language or cross-language translations preserving semantic content i.e. meaning. Usually these versions are interpretations of epic stories that tell about gods or about the adventures of great heroes. They often represent the values and ideals of an entire country or people. An automatic alignment among this kind of texts establishes a mapping that relates the content of one text to the content of the other, wherein the subject of the mapping are interpretations of the same span of the original text. This is an interesting problem that is closely related to the alignment of different language translations of a common source. Nevertheless, in the latter case, the translation is usually so regular and homogeneous that a great deal of success is achieved using existing techniques (see state of the art). Unfortunately, most of this work has focused on cross-language translations and very few projects have considered alignment of different versions of ancient text. Own et al. (1998) have focused on the alignment of Iliad, Odyssey, and bible versions in the framework of the HEARER HOMER project.

Our motivation to perform this sort of alignment is to introduce a new tool that supports

1. comparative navigation within a corpus of parallel versions of ancient texts enabling a comparative analysis of the translation style, the linguistic and the geo-linguistic features and the dialectal properties of the language.
2. an elaboration of the comprehension of the documents by grouping parallel segments from various versions that can be found in another textual or graphical form.
3. random access of information through the different textual streams and retrieval of content and meta-data attached to the translation content from the different versions, basing on a search of an original document or an alternate translation.

3. 2. Case study: medieval texts

In this project we are interested in French medieval manuscripts, in particular the manuscripts produced between the XIIth and the XVth century. These manuscripts are sets of different conversions over the time of the same original texts. They are written by authors -often unknown- having different cultures and skills. Each manuscript reflects, thus, its own cultural and geo-linguistic features that depend on a particular civilization. As a first step, we worked with extracts from the same versions of the manuscript "Ovide moralisé" (versions of Geneva, Paris, Lyon, and Rouen) as well as versions translated to Latin and modern French by domain specialists. While manipulating these manuscripts we have noticed that:

1. the structure is very different from one version to the other. Although there seems to be a certain isomorphism in the structure within one class of versions (verse or prose), the variation is striking when comparing two versions from different classes. These irregularities in structure make it quite difficult even to perform a manual alignment.
2. Sentence and paragraph structure is difficult to detect in many versions due to the loss of punctuation or uncertainty about the sentence structure.
3. Apart from the variation in the structure between the different classes of documents, the variation in the linguistic level contains the following attributes:

- Morphological variation is quite abundant in all the versions. For instance the word "plusieurs" can be found as "plusiours". The conjugated verbs particularly cover a great number of variants of dialectal, orthographic or analogue (to modern French) origins.
- Semantic variations are less abundant but quite important, particularly among versions from different classes. These variations are mainly the use of synonyms or other group of words having the same realization (e.g. mal / mauvaise, prouffit / enseignement).
- The pragmatic and discourse structures are very close even if in some versions we find more elaborated parts than in others, or in some cases more detail is given, particularly when it comes to a description or an argumentation. Despite this disparity in the expressiveness of the language, the core of the content is kept invariant among all the converted texts.

After some experiments, we found that the alignment among same class versions can be done using standard techniques, particularly for texts of the verse class. This is due to the following two reasons:

1. The document structure is very close or in some cases identical: A poem is composed of titles (to announce a beginning of a division), verses (organized line by line) and annotations in the margins (often transcribed within the verse structure with a special markup). The graphical objects are inserted randomly in the document without a prior order. The most significant variation in this level is the omission or the insertion of a title, or a block of annotation or at most a block of verse that does not exceed a couple of lines in the worst cases.
2. The linguistic variation is limited to the morphological level and in many cases to the semantic level where a simple substitution of a word by its synonym has taken place.

As previously noted, in this particular case we are dealing with a classical problem of alignment where statistical approaches have proved their success. The TALCC aligner (Ballim et al. 1998) has been used and achieved about 90% success when respecting a same generic model of documents. The TALCC aligner is based on the Gale and Church algorithm with the addition of the logical structure of documents as a further criterion of alignment (see state of the art).

3. 3. The alignment problem with verse/prose documents

The first problem that we were faced with when comparing verse and prose versions was the document structure. The structure of a verse manuscript is composed of titles and verses organized line by line. Each verse may be a clause or in rare cases a whole sentence. Punctuation does not exist originally in the material but is introduced by experts. Annotations on the margins are transcribed within the verse structure with a special markup. Instead, a prose is a block of text organized in a paragraph structure each one starting with a title that describes that content. Sentences within the paragraphs are usually separated by original punctuation though in some cases it is difficult to detect (figure 1).

Linguistic variation

Whereas verses are mainly segments of sentences expressed in an artistic style enclosing a form of rhyme, the content in sentences in a prose is more compact and expressed with a lighter descriptive style. Obviously, the linguistic construction is different in many ways; the lexicon submits to the usual morphological variation that can exist between two versions as described before, however the order is not preserved. Syntactically the whole construction is different: a sentence in the prose version can correspond to one or more sentences in the verse version. A word itself in one version can be elaborated to a whole verse or even a block of verses. Usually the prose version is more compact than the verse one with a ratio of nearly 3/4.

Pragmatic variation

When it comes to the pragmatic structure, the difference is noticed on the level of depth of explanation and expressiveness. It has been noticed that the verse versions are enriched with more elaboration, restatements, argumentation, details (temporal, location, manner, circumstances, etc) and descriptions (comparison, causative events). The comments and personal interpretations are often more abundant in the verse versions. This explains the difference in size between the two manuscripts. Nevertheless, it has also been noted that the salient segments are kept invariant and submits only lexical variations or synonymy substitution.

Verses in medieval French	Prose in medieval French
Se l'escripture ne me ment Tout est pour nostre enseignement Quant qu'il a es livres escript Soient bien ou mal li escript	***T***outes escriptures soient bonnes et mauvaises sont pour nostre prouffit et doctrine faittes.

Figure 1: Example of alignment of verse and prose versions in medieval French.

4. State of the art

The problem of alignment seems to have first been raised when Brown (1988) and his colleagues tried to build a probabilistic model for automatic translation. Debili (1992) faced the same problem when he planned to set up dictionaries of bilingual expression transfers and synonyms. The alignment problem was then treated as only second or peripherical. Many authors now set the alignment problem in a more global framework. For instance, Warwick (1989) places the alignment in the context of the implementation of lexicographic tools for linguists and translators, or, more recently, as an aid to the evaluation of translation quality.

A good deal of work has already been done on alignment (Brown et al. 1991), (Gale and Church 1991) and (Simard et al. 1992). Since then several other approaches have been used, both for sentences, word and character alignment (Papageorgiou et al. 1994; McEnery et al. 1995). All these methods are mainly based on statistics, some based on word frequencies, others on characters occurrences.

Gale & Church's character-based algorithm propose a method which uses only internal information and does not consider any hypothesis on the lexical content of the sentences. Authors started from the observation that the length of sentences in the source text and its translation in the target text are strongly correlated: short sentences tend to be translated into short sentences and long sentences into long sentences. Furthermore, it seems that there exists a rather constant ratio between the length of sentences from a language to another in terms of number of characters. This method has been tested on a bilingual corpus of 15 economic reports published by the Swiss Banks Union in English, French and German, for a total of 14'680 words, 725 sentences and 188 paragraphs in English and their corresponding numbers in the two other languages. This method makes it possible to correctly align the total amount of sentences, except for 4 % of them. The same number of errors has been found in the English-French and English-German alignments, showing that the method is relatively language-independent. The model proposed by Gale & Church has also been tested on a much more important sampling, of 90 million words, taken from the official Canadian Hansards corpus.

Brown's word-aligner algorithm combines sentences according to the number of words included in each sentence. This algorithm is described by its author as a development of Gale & Church's algorithm, which computes the length of sentences from the number of their characters. The data of the official Hansards (McEnery et al. 1996) corpus of the Canadian parliament official decrees have first been converted into a unique English corpus and another French corpus. Each of these corpora has been fragmented into token and these tokens have been combined into groups called sentences. Moreover, auxiliary data such as the numbering of Parliamentary Sessions, the name of speakers, the time index and the ordering of questions were used to add comments throughout the text. Each of these comments can be used as an anchor point in the alignment process. The alignment of anchor points is made in two passes, first for the main anchor points, then for the secondary ones.

Chen's alignment model (Chen 1996) is built from a sample of data aligned in two languages, English and French, and tested on samples taken from the Hansards. This model, conceived on an ad hoc basis in the framework of Bayes' paradigm allows to take into account frequent cases where sentences don't align in a uniform way, in a 1:1 ratio, but rather in a 2:1 ratio. The search strategy used, which is that of dynamic programming, allows a linear search in the length of the corpus. This strategy includes a separated mechanism to process a large number of suppressed sentences in one or the other version of a bilingual corpus.

Martin Kay's method (Kay, 1993) is based on an EM (Estimation and Maximization) type of algorithm. Here, the alignment of sentences depends on the alignment of words. That of words depends on the similarity of their distribution. The method proposed by Kay is relatively more complicated to implement than the other methods already mentioned.

Simard and his colleagues give a simple method, which attempts to align texts on the basis of related words (cognates). This method seems to give satisfactory results but its drawback is that it can only operate on rather closely related language pairs. Cognates are words, which are almost identical in both languages (artificial/artificiel). This thus implies that both languages must be written with the same alphabet (Simard et al. 1992).

There have been some innovative works that incorporated further criteria in alignment such as the linguistic knowledge and the structural properties of the documents. The use of linguistic knowledge covers mainly the process of parsing (Dagan et al. 1996; Matsumoto et al. 1993) and tagging (Van der Eijk, 1993). *Kupiec* (1993) proposes an algorithm for finding nominal syntagms matching each other in a bilingual corpus. In this algorithm, syntagms are thus recognized with the aid of a specific program and the correspondences between these syntagms are determined with an algorithm based on simple statistical techniques. The use of external linguistic resources mainly bilingual dictionaries is quite efficient in identifying lexical anchors (Catizone et al. 1989; Warwick & Russel 1990; Debili & Sammouda 1992).

Structure-driven methods consider the text as structured flow of information and manipulate this meta-information about the organization of the text structure to aid in the process of alignment. Ballim and al. (1998) developed an aligner which takes advantage of the global structure that many documents have (e.g., sections, chapters, titles, etc.) This structural information is integrated with other similarity metrics such as: number of characters in PCDATA, cognates, bilingual terms and parts of speech to decide the correspondence between parallel segments. Tests and evaluations have showed that the structure-driven alignment is efficient with isomorphic documents having the same generic logical structure. However it was much more difficult to deal with non-isomorphic documents although referring to the same generic logical structure. In the same framework of structure-driven alignment Romary & BonHomme (2000) have used the TEI annotation guidelines to calculate the best alignment pairs from the multilingual texts at division, paragraph and sentence level.

5. Semantic and pragmatic approach

The success of one approach or another depends strongly on the nature of documents. When aligning the Hansard corpus for instance, statistical approaches were enough the reach a great performance (nearly 97% with the Chen algorithm). This is mainly due to the specificity of the translation between French and English, which is often located in the word and sentence level. When it comes to structured documents, the logical structure is a further hint to guide the process of alignment. That what has been shown for instance by Ballim et al. in some experiments with Systematic Corpus of federal Laws from the Swiss federal Chancellery. However, when we address the problem of alignment with other types of documents having different properties, we must find further criteria of similarities. Still little work has focused on the linguistic properties of the documents because they are difficult and complex to compute and to deal with. Nevertheless in some cases alignment becomes a task that is coupled with a semantic and pragmatic understanding of the content of the texts. It is the case of the different versions of ancient texts where even a manual alignment needs experts' competence.

Thus our semantic and pragmatic approach looks at texts from the perspective of their discourse structure rather than from the textual flow of data. The discourse structure is the way text is organized to insure coherence and cohesion of its content. Most of the discourse theories (Hobbs 1985; Grosz & Sidner, 1986; Mann & Thompson 1988) consider texts as a set of segments related by means of pragmatic relations. Parallel texts express similar semantic content i.e. share a same meaning, therefore their discourse and pragmatic structure must be very close. Hence this structure could be considered as a further criterion to detect similarities where classical approaches failed.

5. 1. Description of the discourse structure

In our hypothesis we have adopted the Rhetorical Structure Theory (RST) (Mann & Thompson 1988) as a model of discourse organization. RST is a descriptive theory about the organization of natural texts, characterizing their structure basically in terms of a closed set of relations called rhetorical relations that may hold between their parts. The term rhetorical is not limited to the relations that have a rhetorical sense but can be extended to other kinds of relations such as semantic, pragmatic, logical or even very special domain-dependent relations. Texts are decomposed into non-overlapping units called discourse segments. Each segment is related to a span of segments by means of a relation and is called a nucleus or a satellite (there are a few exceptions to this rule: some relations can join two nucleus

segments, they are called multinuclear relations). The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's purpose than the satellite, and that the nucleus of a relation is comprehensible independent of the satellite, but not vice versa. Text coherence in RST is assumed to arise from a set of constraints applied on the nucleus, on the satellite and on their combination. For example in the following sentence:

Although we obediently ate everything our mother prepared, my sister and I much preferred to eat our fruit crisp.

We detect a concession relation, the situation described in the nucleus (second clause of the example) is in contrast to that presented in the satellite. It is about a violated expectation.

The model of discourse structure we are using obeys the constraints put forth by Mann and Thompson (1988) and Marcu (1996). It is a *binary tree* whose terminal nodes represent the *elementary units* and non-terminal nodes represent the relations holding between spans of texts (In figure 2 Arrows are pointing to the nucleus spans).

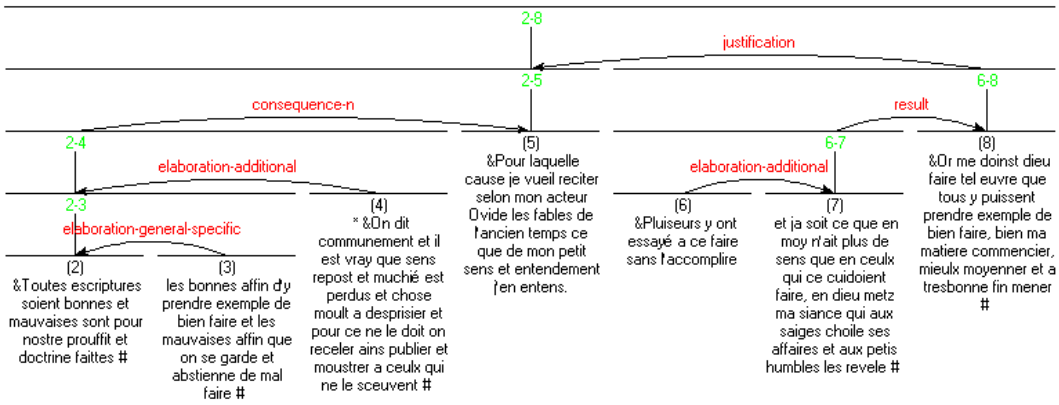

Figure 2: RST model of discourse structure.

5. 2. Semantic-pragmatic annotation

Recent developments in computational linguistic have created the means for the automatic derivation of rhetorical structures of unrestricted texts. Marcu (1996) suggested an algorithm that uses cue phrases and a simple notion of semantic similarity in order to hypothesize rhetorical relations among the elementary units. Nevertheless, these algorithms are still domain dependent and the efficiency is their main drawback.

To structure our texts, we proceeded by a manual annotation of a sample corpus (versions of Geneva and Paris) to evaluate the complexity of the task. We have fixed a taxonomy of relations where each class is composed of subclasses of more specific relations. We distinguish three main classes, semantic relations, inter-personal and textual relations. Semantic or informational relations are mainly relations used to describe how information is conveyed for instance elaboration, comparison, circumstance, condition and causative. Inter-personal or planning relations are relations that hold a pragmatic intention for example interpretation, evidence, explanation and argumentation. Textual relations are rather relations that have an influence on logical structure of the text for instance list, conclusion, disjunction, conjunction, summary, joint, topic-drift and sequence. Such classification gives more freedom to annotators to choose the relations according to their own understanding, and permits to build a similarity measure in the process of comparison (Ghorbel 2001). The first task of the annotation is the process of segmentation. Unlike previous work where segmentation is basically situated in the clause level, we focused on a more global view; the sentence level and in some cases on larger blocks of texts. This kind of macro segmentation allows us to define the *elementary units* of the discourse structure and eventually the units of the alignment process. The larger the segments are, the easier the computation of correspondence is, but the less precise the alignment is. On the other hand, considering very short segments we will end up with very large trees and the problem of complexity becomes important.

The second task of annotation consists of grouping the elementary units together by means of either a mononuclear or a multinuclear relation. This process will create spans of texts or discourse segments

related in the form of an ordered *binary tree* (figure2). Within this tree we can detect certain paths formed by the nuclear nodes. This path structure will play an important role in the alignment process. Unlike previous work (Marcu for summarization (1998), for automatic translation (2000), for essay scoring (2000) and Cristea (1998) for anaphora resolution) where the whole text is represented as a single tree, and since we are working with long texts, we found it more appropriate to consider the texts as a forest of trees. Separation between trees is viewed as a topic shift in the texts. Still this concept of separation between trees is subjective as it depends on the annotator, but it does not have adverse effects since in the alignment process a tree from the source text can be aligned with more than one tree from the target text.

6. Multi-criteria alignment

6. 1. semantic-pragmatic structure alignment

We have shown that the semantic and pragmatic structure of the discourse is very important to find anchor points and hints to align segments of similar texts. We propose in this section some algorithmic approaches to use the knowledge stored in the suggested model to detect similarities and make correspondence between segments.

Problem formulation

Consider a document D^1 to be aligned with D^2. D^1 and D^2 are respectively represented as forests of n and m trees as follows:

$D^1 = \{T_1^1, T_2^1, \ldots, T_n^1\}$

$D^2 = \{T_1^2, T_2^2, \ldots, T_m^2\}$

Each tree T_i^d (the i^{th} tree of the document d) is composed of text spans structured in a binary tree structure. The terminal units of T_i^d are text segments ordered from left to right as they appear as text sentences in the document. We denote these segments as follows:

$T_i^d = \{ s_{\vec{i}}^d, s_{\vec{i}+1}^d, \ldots, s_{\overleftarrow{i}-1}^d, s_{\overleftarrow{i}}^d \}$ where $\vec{i}$ and $\overleftarrow{i}$ are respectively the left and the right boundaries of the tree T_i^d.

We call the *salient path (SP)* in a tree the path followed when we navigate from the root to the terminal elementary units and choosing the nucleus nodes when coming through a relation node. In a tree there exists only one SP if and only if all the chosen relations are mononuclear and ($\overleftarrow{i} - \vec{i} + 1$) SPs if all chosen relations are multinuclear. In the general case the number of SPs, let's denote this number $\overline{p}$, $\overline{p}$ then ranges between 1 and ($\overleftarrow{i} - \vec{i} + 1$). Each SP points to a terminal elementary text segment. Let's denote by $\hat{T}_i^d$ the set these terminal units in a tree, then we have:

$\hat{T}_i^d = \{ s_j^d / \vec{i} \le j \le \overleftarrow{i} \}$ where $\text{card}(\hat{T}_i^d) = \overline{p}$.

Pragmatically, the set $\hat{T}_i^d$ stands for the span of the text in the tree *i* of the document *d*, which is considered as far as it is seen by the annotators when segmenting and attributing the nucleus and satellite properties to segments, for the main part that helps readers understand the sense and the content of the tree *i*. For example in figure 2, $\vec{i} = 2$, $\overleftarrow{i} = 8$, $\hat{T}_i^d = \{ s_5^d \}$, $\overline{p} = 1$, the SP is hence the path of the 5^{th} segment.

We call the k-nearest neighborhood of the SP the k satellite segments having the minimum distance with SP. We define the distance d_m as the metric distance which is defined from $T^d * T^d$ to the set of natural numbers where $T^d = \bigcup_{i=1..n} T_i^d$. The metric distance between two segments is given as following;

d_m (s_i^d , s_j^d)= |j-i|. We define also *ds* as the structural distance applied between the nodes of a tree and which has the value of the length of the minimum path between the nodes. The length of a path in the tree is given by the number of nodes it contains between its two extremities.

The process of alignment

The semantic and pragmatic annotation as described previously models the text as a forest of trees. Each tree holds in its structure a set of segments related to each other by means of semantic or pragmatic relations and ordered by their pragmatic importance in the way they hold and convey information. This is modeled by the hierarchical structure and the concept of satellite/nucleus. The first step in the process of alignment is to define a mapping M that establishes the correspondence between the trees composing the two documents. M can be defined according to three approaches:

Approach 1: alignment based on the SP segments

The window of comparison of the texts is limited to certain specific segments, in this case only those that form the set $\hat{T}_i^d$. M is defined as following: M(T_i^d)= $T_j^{d'}$ iif S($\hat{T}_i^d$, $\hat{T}_j^{d'}$) $\leq h$ where S is the similarity measure and h is an empirical threshold. S is calculated basing on the lexical distance between the streams of characters and referring to thesaurus database (see next section).

Approach 2: alignment based on the SP and its k-nearest neighbor segments

The window of comparison of the texts is limited to the segments that form the set $\hat{T}_i^d$ and their k-nearest neighbors. The k-nearest neighbors can be obtained either by considering the M(T_i^d)= $T_j^{d'}$ iif D($\hat{T}_i^d$, $\hat{T}_j^{d'}$) $\leq h$ where D is the similarity distance measure and h is an empirical threshold. The k-nearest neighbors can be obtained applying the metric distance or the structural distance. In both cases we search for the k-nearest satellite to the SP. In the case the structural distance, when the found neighbor is a sub-tree, we investigate its local SP which points to the first neighbor, the next ones will be recursively determined according the first neighbor.

Approach 3: alignment based on the nature of the relations

Relations are classified according to their semantic and pragmatic similarities. We distinguish three main classes, semantic relations, inter-personal and textual relations. Semantic or informational relations are mainly relations used to describe how information is conveyed for instance elaboration, comparison, circumstance, condition and causative. Inter-personal or planning relations are relations that hold a pragmatic intention for example interpretation, evidence, explanation and argumentation. Textual relations are rather relations that have an influence on logical structure of the text for instance list, conclusion, disjunction, conjunction, summary, joint, topic-drift and sequence. Each subclass can itself be more detailed. This approach of alignment of trees is based on a statistical comparison of the frequency and the order of relations.

6. 2. lexical alignment of segments

Despite of the great variation of the conversion from one version to another, there still exists a certain local similarity between words, particularly proper nouns, nouns, adjectives and verbs. Certain words might be subjected to tiny morphological variations (change i/y, s/z, etc). For these reasons, we found the word comparison (Brown et al. 1991) approach more relevant than character comparison. This form of comparison is used only between certain segments of the text hypothesized by the structural alignment (according to the previous approaches) to give a further heuristic of similarity.

Each segment is reduced to a word list where noisy data (article, determiners, pronouns etc) is eliminated. A first approach is to apply a stemming algorithm and then compare the obtained stems. The main drawback of this method is that inaccuracy in rules can map different forms to same stem. A second approach consists in estimating the Levenshtein distance between each couple of words taking into consideration some particular rules when estimating the costs of modification (substitution, omission, and addition). A vector of words' distance is hence obtained. The similarity distance measure is the cost of the minimum path through the previous vector.

Besides the lexical similarity, we observed a kind of semantic similarity between the words used in the different versions of the manuscript, for example mal/mauvais, prouffit/enseignement, etc. A thesaurus that provides synonymy relation between words is in construction. When comparing two segments, the synonymy relation is checked between each couple of words and the vector of words' distance will be modified.

7. Conclusions and future work

When dealing with comparable documents where manual alignment needs an elaborated and deep understanding of the content, automatic alignment becomes a difficult task. Human interaction is then needed to aid in the process structuring and modeling the content. In this framework, we proposed a content-driven alignment whose heuristic is partially based on a manual human annotation of the semantic and pragmatic structure of documents. The main drawback of this annotation is the subjectivity of this task that depends closely on the profile of the annotator. Much of the current work in devoted to define a simplified golden standard and to automate some sub-tasks in order to facilitate the annotator's task.

We have also presented in this article some approaches of content-driven alignment based on the semantic structure of the texts. The alignment process establishes a mapping between spans of texts represented in a tree structure by comparing the significant (from a semantic view) segments. Therefore the semantic structure built upon the text limits the window and drives the order of comparison. The lexical comparison still remains an important heuristic of similarity. Preliminary results seem encouraging and much of the future work will be focused on the evaluation of each approach of alignment proposed (structural, lexical and thesaurus-based alignment) and on the integration of these heuristics. The alignment at the segment level is also one of the interesting points we are investigating, so that the granularity of the correspondence and the comparison could be refined.

References

Ballim A, Coray G, Linden A, Vanoirbeek C 1998 The use of automatic alignment on structured multilingual documents. In *Proceedings of the Seventh International Conference on Electronic Publishing*, Saint Malo, pp 464-475.

Brown P, Della Pietra S, Della Pietrs V, Mercer R 1988 A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary, pp 1-6.

Brown P, Lai J, Mercer R 1991 Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, USA, pp 169-176.

Burstein J, Marcu D 2000 Towards Using Text Summarization for Essay-Based Feedback. *Septième Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000*. Lausanne, Switzerland, pp 51-59.

Catizone R, Russell G, Warwick S 1989 Deriving translation data from bilingual texts. In Zernik U (eds), *Proceedings of the first Lexical Acquisition Workshop* Detroit, Mich, USA.

Chen S 1996 *Building Probabilistic Models for Natural Language*. PhD thesis Harvard University.

Cranias L, Papageorgio H, PiperidisS 1994. A matching technique in example-based machine translation. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, Kyoto, Japan, pp100-104.

Cristea D, Ide N, Romary L 1998 Veins theory: A model of global discourse cohesion and coherence. *In Proceedings of the 17th International Conference on Computational Linguistics and the 36th annual meeting of the Association for Computational Linguistics* (COLING, ACL), Montreal, Canada.

Dagan I 1996 Bilingual word alignment and lexicon construction. Tutorial notes, *34th Annual meeting of the Association for Computational Linguistic*, Santa Cruz, California.

Debili F, Sammouda E 1992 Appariement des phrases de textes bilingues. In *Proceedings of the 12th International Conference on Computational Linguistics*, Nantes, France, pp 517-538.

Gale W, Church K 1991 A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, USA, pp 177-184.

Ghorbel H (forthcoming) 2001 *Guidelines for semantic annotation*. Internal report, Swiss Federal Institute of Technology IN Ecublens, 1015 Lausanne, Switzerland.

Grosz B, Sidner C 1986 Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*. 12(3):175-204.

Hobbs J 1985 *On the coherence and structure of discourse*. Center of the study of language and information, CSLI-85-37, Leland Stanford Junior University.

Kay M, Roescheisen M 1993 Text-translation alignment. *Computational Linguistics*. 19(1):121-142.

Kupiec J 1993 An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, Ohio, 17-22.

Mann W, Thompson S 1988 Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*. 8(3): 243-281.

Marcu D 1997 *The Rhetorical Parsing, Summarization, and Generation of Natural language Texts*. PhD Thesis. Department of Computer Science, University of Toronto, Canada.

Marcu D 1998 Improving summarization through rhetorical parsing tuning. In *Proceedings of The sixth workshop on very large corpora*, Montreal, Canada. pp 206-215.

Marcu D, Carlson L, Watanabe M 2000 The Automatic Translation of Discourse Structures. *The 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2000)*, Seattle, Washington, pp 9-17.

Matsumoto Y, Ishimoto H, Utsuro T 1993 Structural matching of parallel texts. *In 31st Annual Meeting of Computational Linguistics*, Columbus, Ohio, pp23-30.

McEnery A, Oakes P 1995 Cognate extraction in the Crater project. In *Proceedings of the EACL-SIGDAT workshop*, Dublin, pp 77-86.

McEneryA, Wilson, A 1996 *Corpus Linguistic*. Edinburgh, Edinburgh University Press.

Owen C, Makedon F, Steinberg T 1998 Parallel text alignment. *Research and Advanced Technology for Digital Libraries*, Special Issue containing invited ECDL'98 papers, Computer Science Lecture Note Series, Springer Verlag, Christos Nikolaou and Constantine Stephanidis (eds) pp 235-260.

Romary L, BonHomme P 2000 Parallel alignment of structured documents. In Véronis, J. (eds.). *Parallel Text Processing*. Kluwer Academic Publishers, Dordrecht.

Sato, S, Nagao M 1990 Towards memory-based translation. In *Proceedings of the 12th International Conference on Computational Linguistic COLING'90*, Helsinki, pp 247-252.

Simard M, Foster G, Isabelle P 1992 Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Monreal, Canada, pp 67-81.

Sumita E, Hitoshi I, Hideo K 1990 Translating with examples: a new approach to Machine Translation. In the *Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, Austin, Texas, pp. 203-12.

Van der Eijk P 1993 Automating the acquisition of bilingual terminology. *In sixth Conference of the European Chapter of the Association of Computational Linguistics*, Utrecht, The Netherlands, pp 113-119.

Véronis J 2000 Alignement de corpus multilingues. In Pierrel, J-M (eds). *Ingénierie des langues*. Editions Hermès, Paris.

Warwick S, Russel G 1990 Bilingual concordance and bilingual lexicography. In *Proceedings of Eurolax'90*, Malaga, Spain.