SAM: System for Multi-criteria Text Alignment

Hatem Ghorbel, Giovanni Coray, André Linden

Swiss Federal Institute of Technology EPFL
Faculté Informatique et Communications
Computer Science Theory Laboratory LITH
IN Ecublens, 1015 Lausanne, Switzerland
{hatem.ghorbel, giovanni.coray, andre.linden}@epfl.ch

Abstract

The problem of text alignment is to establish the correspondence between subparts of two ore more translations or versions of the same document. Most of the methods used in alignment are based on the statistical analysis of word or character frequencies or of string occurrences. In order to achieve more accurate results, other methods have incorporated some structural properties of the documents as further criteria.

When addressing the problem of alignment to align different versions of medieval texts namely prose and verse versions, we need to consider more efficient methods of content comparison. In this article, we propose an extension to the existing methods of alignment where we consider further linguistic and structural properties of the texts. As a linguistic criterion of alignment, we propose some heuristics to calculate similarities at the lexical, morphological, syntactic and semantic level of the texts. On the other hand, as a structural criterion, we extend the similarity measures to take into account different properties of the rhetorical structure of the texts. The process of alignment is therefore an optimization problem that maximizes linguistic and structural similarities between aligned pairs of parallel 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. The logical structure of documents (e.g. sections, chapters, titles...) was for instance the basis of some research work to estimate an optimal matching of components of texts in structured documents. When addressing the problem of alignment to align different versions of medieval texts namely prose and verse versions, we need to consider more efficient methods of content comparison. In other words, we need to apply techniques of natural language analysis and understanding in order to provide a pool of linguistic and structural criteria that can improve the process of comparison and alignment.

1.1. State of the art

The problem of alignment seems to have first been raised when Brown and his colleagues (1988) tried to build a probabilistic model for automatic translation. Debili et al. (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 et al. (1990) 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 & Church, 1991) and (Simard et al., 1992). Since then several other approaches have been used, both for sentence, word and character alignment (Kay & Roescheisen, 1993; McEnery et al., 1995). All these methods are mainly based on statistics, some based on word frequencies, others on characters occurrences.

There has been some innovative work 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, 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 et 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, 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 and BonHomme (2000) have used the TEI annotation guidelines to calculate the best alignment pairs from the multilingual texts at division, paragraph and sentence level.

1.2. Motivation

In the framework of the MEDIEVAL[1] project (An automatic model for the edition of medieval manuscripts visualized by alignment) we are interested in alignment and in the comparison of French medieval manuscripts, in particular the manuscripts produced between the XIIth and the XVth century. These manuscripts are sets of parallel conversions over time of the same original source texts in prose and verse structure. 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 practical goal of this project, we intend to develop an environment that facilitates for experts and scholars the on-line comparative analysis of ancient texts and the navigation through the various components of the different parallel versions. Based on the particularity of the structural properties of our corpus, we found that it is interesting to investigate current approaches of texts alignment and to enrich this environment with further methods founded on linguistic and structural knowledge.

2. The problem of multiple versions alignment

Early experiments[2] on alignment applied to ancient manuscripts of medieval French have shown the limits of statistical approach 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. This is generally due first to the partial evolution of the language, second to the variation of the text style (verse and prose) and finally to the personal interpretations that could come about when rewriting new versions. In fact, the reproduction of multiple versions of medieval manuscripts is not a task of translation, but a task of transformation or adaptation of the content to a new style, structure and culture.

For example, rewriting a prose version from a verse one requires a transformation from the verse style to the prose one, in addition to other modifications that can affect the linguistic and organizational structure of the text.

When addressing the problem of alignment to this kind of document, we need to consider other methods than those used in the alignment of translations especially deeper methods of linguistic and structural comparison.

In addition to the linguistic similarities (lexical, syntactic and semantic nature) that could be detected on the sentence level of the parallel versions, medieval manuscripts share also a large variety of similarities in their physical structure and textual organization. In fact ancient texts are well organized and structured in order to facilitate their understanding and make stories more pleasant and attractive. Generally, the global organization and meaning of the content are kept invariant when reproducing a further version of a text.

From the perspective of these similarities, the alignment we propose is based on several criteria. First, the linguistic criterion where the comparison is situated on the word or expression level using statistical measures on textual content and their frequencies in the sentence or in the text. These measures are tuned with linguistic knowledge from local morphological or syntactic analysis or from linguistic resources (lexical databases).

The second criterion deals with the structural properties of texts. Texts are annotated according to a hierarchical model that describes their typographic structure on the one hand and their organizational structure on the other hand. Typographical structure is defined according to the layout of the manuscripts, whereas organizational structure is defined by domain experts according to a light text organization model elaborated on the basis of the Rhetorical Structure Theory (RST) (Mann & Thompson, 1988). This tree-like structure describes how text segment (often sentences) are coherently related to each other by means of rhetorical relations. Based on these structural properties, similarities are detected from the perspective of the comparison of the tree nodes nature as well as the characteristics of the relations paths describing the texts spans.

3. Multi-criteria Alignment

The approach of text alignment that we propose is in fact an extension to the existing methods which was limited to texts translations. Our approach is able to detect correspondence not only between translated versions but also between converted or interpreted version. Multi-criteria alignment that we propose in the following section is based on similarity measure from the perspective of two main criteria: linguistic and structural criteria.

3.1. Linguistic similarity

Linguistic similarity between segments is measured by the relative frequency of the similar words they share with respect to the total number of words. Similar words are detected according to the following criteria:

- Heuristic criteria that measure the distribution of shared characters in the word (cognates);
- Philological knowledge which is represented in a set of rules elaborated by domain experts. These rules specify the possible historical variations that could modify the graphics of a word;
- Linguistic knowledge that concerns either the morphosyntactic level or the semantic level. On the morphosyntactic level word similarity is defined by similarity between lemmas of the inflected or derived forms and between some syntactic attributes. On the semantic level, similarity is defined according to linguistic resources such as databases of synonyms and analogical expressions.

3.1.1. Lexical similarities: cognates

A cognate is a word in language A similar in form and meaning to another word in language B. For instance thèse and thesis, respectively in French and English, are cognates. Several heuristic methods have been developed to detect cognates, all being based on the computation of distances and similarities at the graphical level of words. Among the more advanced methods (Simard et al., 1992;

[1] A project in collaboration with the University of Geneva, financed by the Swiss National Fund.

[2] Using Vanilla aligner (implementation of Gale and Church alignment by P. Danielsson and D. Ridings (1997)) and the Talcc aligner (adapted implementation of the Multext aligner developed in ISSCO (Ballim,1996).

Hofland et al., 1998; McEnery et al., 1995; Melamed, 2000), the heuristic method of Dice, which computes the ratio of the percentage of bi-grams shared in two candidate words with the set of their bi-grams, can take into account graphical variations at the word level and is well adapted to medieval texts. At a precision threshold of 0.8 the computation of cognates between the prose and the verse versions of the medieval corpus with Dice's method has been evaluated at an order of 80 %. However, this precision level also makes it possible to account for pairs which have no cognate relation (for instance aoure and autre; beste and estre; rien and bien; ceste and cesse; etc.). In order to solve this precision problem, we have increased the threshold at 0.82. The latter gave a better precision of the order of 87%. In order to palliate to the problem of the loss of weakly correlated cognates (less than 0.82), we have included a few rules of graphical transformations that allow to refine the method. These rules are in general the product of dialectal and orthographic factors and can be applied to simple contexts without the need of complex phonetic constraints.

3.1.2. Morphosyntactic similarity: canonical form

The lexical similarity function based on statistical heuristics and on the application of a few graphic rules does not make it possible to determine those correspondences which are subject to variations of a linguistic order. As a matter of fact, the re-writing of different versions, and especially of a structure into another or of a style into another require some kind of syntactic transformation at the level of the sentence structure, which is at the origin of a variety of lexical and morphosyntactic forms. Such variations cannot be accounted for by the lexical similarity heuristics but rather we must consider morphosyntactic transformations for a more efficient comparison. A morphosyntactic study of medieval French has enabled us to build a lexical database, MEDIEVLEX[3], on the basis of which other similarity of morphological and syntactic nature can be detected using features such us:

Lemmas

In the MEDIEVLEX database, lemmas allow to define a canonical form for lexical entries. This form is the infinitive for verbs, and the masculine singular for nouns, adjectives, pronouns and determiners. Thanks to this lemmatization, it is possible to find the correspondence between tensed forms of verbs (e.g., *croy* and *croire*) and derived forms (e.g., *commencement* and *commençaille*).

Families

The family feature in the MEDIEVLEX database makes it possible to define the etymological origin of lexical entries. This feature can also provide a canonical form for the various derived syntactic categories. This more abstract form of lemmatization makes it possible to find a correspondence between verbs, adjectives and adverbs having the same origin, as, for instance, between the verb *profitter* and the adjective *profitable* or between the noun *humilité* and the adjective *humble*.

[3] A lexical database of medieval French in the XML format. At the present state, it contains about 1500 lexical entries.

3.1.3. Semantic similarity: synonyms

Medieval texts reveal other similarities than those at the lexical and morphosyntactic levels. Words can also be replaced by their synonyms, expressions by equivalent expressions, etc. In order to find the correspondence between these semantically equivalent terms, other methods than those based on lexical and morphosyntactic criteria must be found. For modern languages, Wordnet (Fellbaum, 1998) is an example of a semantic network which represents the lexical relation between words on the one hand and the semantic relation between concepts on the other hand.

As there is a similarity of the semantic structure of the lexicon between medieval French and modern French, in the MEDIEVLEX database this layer is expressed by a link towards modern French. This redirecting link (which is the *sens* (or meaning) feature in the database) indicates the translation of lexical entries in modern French. The comparison of words therefore becomes a comparison of the *synsets* of their translations in the French component of *Wordnet*. Two words in medieval French are synonymous if their translations in modern French belong to the same *synset*.

This concept only concerns words. In order to enrich the semantic knowledge with expressions and groups of words, a thesaurus was developed. This thesaurus is a simple analogical table which explicits the pairs of expressions linked by a relation of synonymy.

3.1.4. The linguistic similarity function

Another important criterion in the comparison process is the order of words. To find a suite of n similar words in two segments is a good indicator for their pairing. We thus propose other characteristic functions which are based on an n-gram model of a suite of words which are linguistically similar. For example, *Quant Dieu ordoneement* is similar to *Quant dieu eut ordonné* if one applies a tri-gram model considering only words having more than 3 characters.

Based on the different methods of comparison on the lexical, morphosyntactic and semantic level of words or expressions so far presented, the linguistic similarity is measured using the Dice ratio that calculates the frequency of similar n-grams of words with respect to the total n-grams of words in the textual data. We used respectively one-gram, bi-gram, and tri-gram Dice ratio to define different similarity functions namely g_1, g_2 and g_3. The final linguistic similarity function f_1 is finally defined as a linear combination where the coefficients evaluate the importance of the nature of similarity functions. f_1 is written as follows:

$$f_1 = \mu_1^1 g_1 + \mu_2^1 g_2 + \mu_3^1 g_3$$

where μ^1_i stands for the weight of each i-gram similarity function.

3.2. Structural similarity

A first structure that we considered as a criterion in the structural alignment is the typographic structure. In fact, the layout of medieval manuscripts is enhanced with graphical elements that reveal a certain knowledge about their logical role in the texts. For instance the use of rubric is coupled with the beginning of an episode in the poem, and so forth. This kind of graphical elements play a role of

triggers and clues to build a coherent structure that describe not only the physical layout, but also the textual organization and the way the author presents his ideas.

Although the typographic structure in ancient texts is very reliable in the understanding of the content, it remains very general and inexact in some manuscripts. That is why we decided to enrich this structure with further precisions about the text organization and about the way the author presents his ideas and arguments. Such a choice is argued with the particularity of the texts we are dealing where the organizational structure of the discourse is often made explicit by the authors for instance with expressions like "Now I am going to talk about ...".

As a paradigm we chose the Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) to propose a model that enables to enrich the typographic structure with rhetorical relations between spans of texts. RST is a discourse theory that models discourse as elementary non-overlapping units of various sizes related to each other by means of rhetorical, cohesive and cohesion relations. RST is an abstract model that defines in a general manner how such an organizational structure of text is formed without imposing any semantic constraint about how units are formed and how relations are fixed. Founded on such a paradigm, we elaborated a light model that describes an abstract structure compatibles with the typographic structure as well as with the linguistic structure.

3.2.1. Rhetorical Structure Theory

Rhetorical Structure Theory 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 independently 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 (1997). It is a binary tree whose terminal nodes represent the elementary units and non-terminal nodes represent the relations holding between spans of texts.

3.2.2. Rhetorical annotation

Recent developments in computational linguistic have created the means for the automatic derivation of rhetorical structures of unrestricted texts. Marcu (1997) 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 the 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, 2002).

The first task of the annotation is the process of segmentation. Unlike previous work where segmentation is basically situated at 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 process of segmentation is achieved using a probabilistic model trained on manual segmented corpora. This issue is out of the scope of the present paper.

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. 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, 1998, 2000; Cristea ,1998) 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.

3.2.3. Structural similarity measures

The second criterion deals with the structural properties of texts. Similarity is defined from the

perspective of the comparison of structural entities (trees) forming the parallel documents. The main distance measure used is the editing distance to compare rhetorical trees and some specific paths in the structure. These measures are described as follows:

Similarity of salient sub-trees

The *salient sub-tree* (SST) of a tree T is the a sub-tree having the same root as T and composed only of *salient paths*. A salient path (SP) (Ghorbel et al., 2001) in a tree is the path linking the root to a leaf node and determined as follows: when we navigate from the root down to the leaves, we choose the nucleus nodes each time coming across a relation node. In a tree there exists only one SP if and only if all the chosen relations are mononuclear otherwise if all chosen relations are multinuclear, the SST of T is equivalent to T. In the general case the number of SP ranges between 1 and the number of leaves (elementary segments) in the tree, and depends on the number of multinuclear relations in the tree. The hypothesis that we adopt in this section is the following :

Two trees are candidates to be aligned they have similar salient sub-tree.

This hypothesis is motivated by the fact that parallel texts have very similar rhetorical structure. In fact this structure, as it was constructed in the annotation task, not only takes into consideration the rhetorical organization of the text spans, but also considers the typographic and the linguistic properties of the documents. When producing parallel translations or versions of a same original document (medieval texts, for instance) authors generally try to generate authentic versions that communicate the same information, but with a different language or style or manner. Even if the rhetorical structure submits some modifications from one version to an other, either effective modifications caused by the process of reproduction, or those due to the process of annotation, there still exists some invariant elements which play an important role in the process of understanding. These invariant elements whenever they exist, happen to be in the SST; this is because the SST stands for the sub-tree holding the principal information which helps in the understanding of the text, at least as it is considered by the annotator in the process of the rhetorical annotation.

Such a structural property is therefore an important criterion in the process of alignment. In fact, it appears instinctively evident that two trees having *similar* SST are more likely to be a candidate pair of alignment. Similarity is calculated using statistical heuristics either on the linguistic level of the textual content of the salient segments (leaves of the SST) or on the level of the relations composing the SP.

For the first similarity measure we apply the linguistic similarity function f_1, on the salient segments trees as follows:

$$h_1(T_1, T_2) = f_1(SS_1, SS_2)$$

where SS1 and SS2 stands for the respective salient segments of T1 and T2.

For the second similarity measure of the SST, we compare the relation paths of the SP using the techniques of approximate string matching (Lecroq, 1995). Each SP is assumed to be a chain of relation names. In the approximate string matching we consider the editing operation of insertion, substitution and deletion, each is associated to a numerical cost. The cost of the substitution is a dynamic value which depends on the relations we are substituting, for instance, it is more costly to substitute a contrast relation with an elaboration relation then with an antithesis relation. The hierarchical classification of the set the relations is elaborated during the annotation phase by the domain experts.

The approximate string matching of two strings x of length m and y of length n is the minimum cost of edit operation needed to make y as a sub-string of x. This is a dynamic programming problem which can be solved in O(mn) time complexity. For example, if we consider the first path x formed by (background(b), elaboration(e), contrast(c), elaboration(e)) and the second path y formed by (elaboration(e), antithesis(a), sequence(s), comment(m)), the minimum cost of finding y in x is determined by the following matching of y in x:

$$\begin{bmatrix} - & e & a & s & m \\ b & e & c & e & - \end{bmatrix}$$

The ratio of similarity is given by the following formula:

$$h_2 = \frac{\min(l(SP_1), l(SP_2)) * d - ASM(SP_1, SP_2)}{\min(l(SP_1), l(SP_2)) * d}$$

where the l function returns the number of relations in each path, the ASM is the minimum approximate string matching of SP_2 in SP_1, and d is the highest cost of substitution.

Similarity of relation paths

For each segment, there exists a unique relation path which links the root of the tree and the leaf which contains the textual segment. The ordering and the type of relation in this path describe the cohesion and the coherence of the structure of texts. The computation of the similarity between texts at the level of the segments can thus have a structural dimension : two segments are homologous in the alignment process if they have a similar structural and rhetorical context. The relation path is therefore an important element which explains in part this structural context and it can be a good comparison indicator. Moreover we apply the same similarity measure between the segments' paths than that used in the salient paths, that is, the function h_2.

Similarity of tree structure

Rhetorical trees are binary trees in the RST form representing the way segments are related to each other by means of coherent relations (rhetorical relations) to form cohesive spans of texts. The comparison of these structures is a tree comparison where nodes are labeled with relation names. We used a framework of dynamic programming to calculate a distance formulated in terms of cost of edit operations to transform one tree into an other with analogy to string edit operations (Wagner and Fischer, 1974). We considered also a predefined taxonomy of relations in order to determine the cost of substituting one with an other. We have adapted the algorithm of Zhang et al. (1994) to calculate the edit tree distance between rhetorical trees. The similarity is hence deduced with the following formula:

$$h_3 = \frac{\max(size(T_1), size(T_2)) * d - edist(T_1, T_2)}{\max(size(T_1), size(T_2)) * d}$$

where the size function returns the total number of relations in each tree and the *edist* returns the editing distance between T1 and T2, and *d* is the highest cost of substitution.

Now, we can define the global structural similarity function as a linear combination of the three similarity measures h_1, h_2, and h_3 as follows:

$$f_2 = \mu_1^2 h_1 + \mu_2^2 h_2 + \mu_3^2 h_3$$

where each stands μ_i^2 for the weight of the similarity according to the criteria i.

3.3. Alignment model

Our approach of alignment is similar to that proposed by Gale and Church (1991), where alignment is viewed as an optimisation problem. The resolution is hence based on a dynamic programming paradigm. However, the criteria of the alignment are different: whereas the basic criteria in Gale and Church work was the sentence length and other lexical information (cognates, ..), in our approach we rather use similarities between document elements (trees and segments) as a basic criteria. Linguistic and structural similarities are calculated according to the previous sections and linearly combined as follows:

$$Similarity = \lambda_1 f_1 + \lambda_2 f_2$$

The weights λ_i are fixed experimentally according to the genre of documents and the pertinence of the criteria in the process of alignment (for instance whether documents are fully or partially annotated, etc…). Alignment is therefore the maximization of the global similarities between trees and segments.

We defined the usual edit operations of insertion, deletion and substitution. The substitution is a general operation that can substitute m segments with n other segments. We usually apply the alignment with a model of (m=3,n=3).

4. Results

From a computational point of view, the algorithms of comparison and similarity measure detailed previously were developed. The alignment algorithm is currently in the phase of test and evaluation. A very first result, applying only the linguistic criteria, shows a precision of about 60% of the aligned pairs of segments (sentences) of prose and verse version with respect to a manual alignment. We intend to validate our method with other corpora of modern texts such as the Systematic Corpus of Federal Laws (French / German / Italian) from the Swiss Federal Chancellery.

5 Conclusion

As a conclusion, we believe that classical approaches of alignment mainly based on language-independent statistical methods are efficient in certain application domains typically in regular translations. However when it comes to interpreted or converted versions where texts undergo important variations on the level of the form and the structure, language-driven approaches are essential to enrich the space of comparison. Therefore, deeper methods of analysis and comparison at the linguistic and structural level of the texts are worth investigating in order to aid in the process alignment at the segment and word level.

5. Acknowledgments

We would like to thank the team of the MELA Department at the Literature Faculty of the University of Geneva, in particular Olivier Collet and Wagih Azzam for their collaboration during the MEDIEVAL project.

6 References

Ballim, A. (1996). *Multext Aligner v.2.0*. Technical Report Institut Dalle Molle pour les Etudes Sémantiques et Cognitives ISSCO. Switzerland.

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

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

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

Burstein, J. and Marcu, D. (2000). Towards Using Text Summarization for Essay-Based Feedback. *In Proceedings of the Septième Conférence Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000* (pp. 51--59). Lausanne, Switzerland.

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

Cristea, D., Ide, N. and 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)* (pp. 281--285). Montreal, Canada.

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

Danielsson, P., and Ridings, D. (1997). *Practical presentation of a Vanilla aligner.* Technical Report 'GU-ISS-97-2'. Swedish Goteborg University. Sweden.

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

Fellbaum, C. (1998). *WordNet, An Electronic Lexical Database.* The MIT Press.

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

Ghorbel, H. (forthcoming) (2002). *Alignement Multicritère des Textes : Critères linguistiques et structurels appliqués aux documents médiévaux*. Thèse

de Doctorat en Informatique, Ecole Polytechnique Fédérale de Lausanne, Suisse.

Ghorbel H., Ballim A. and Coray G. (2001). Rosetta: Rhetorical and semantic environment for text alignment. *In Proceedings of Corpus Linguistics conference* (pp. 224--233). Lancaster.

Hofland, K. and Johansson. (1998). The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In S. Johansson and S. Oksefjell (Eds.) *Corpora and Cross linguistic Research: Theory, Method, and Case Studies* (pp. 87--100). Amsterdam: Rodopi.

Kay, M. and 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* (pp17--22.) Columbus, Ohio.

Lecroq, T. (1995). Experimental results on string-matching algorithms. *Software Practice and Experience* 25(7),727--765 .

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 6th workshop on very large corpora* (pp. 206--215). Montreal, Canada.

Matsumoto, Y., Ishimoto, H. and Utsuro, T. (1993). Structural matching of parallel texts. *In 31st Annual Meeting of Computational Linguistics* (pp. 23--30). Columbus, Ohio.

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

Melamed, D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1),107--130.

Romary, L. and BonHomme, P. (2000). *Parallel alignment of structured documents.* In J. Véronis (Eds.), Parallel Text Processing (pp. 201--217). Kluwer Academic Publishers, Dordrecht.

Simard ,M., Foster, G. and 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* (pp. 67--81). Montreal, Canada.

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

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

Wagner, R.A and Fischer, M.J. (1974). The string-to-string correction problem. *Journal of the ACM* 21(1), 168--173.

Zhang, K., Shasha D., and Wang, J. T. L. (1994) Approximate tree matching in the presence of variable length don't cares. *Journal of Algorithms*, 16(1),33--66.