A language-independent method for the alignement of parallel corpora

Nguyễn Thị Minh Huyền[1] and Mathias Rossignol[2]

[1] Hanoi University of Sciences
334 Nguyễn Trãi, Hà Nội, Vietnam
huyenntm@vnu.edu.vn
[2] International Research Center MICA
HUT - CNRS/UMI 2954 - INPG
1, Đại Cồ Việt, Hà Nội, Vietnam
mathias.rossignol@mica.edu.vn

Abstract. The automatic alignment of parallel corpora is a very rich source of information for automatic translation, multilingual document indexing, information retrieval, *etc.* The rapid growth of the use of "minority" languages in online documents makes it necessary to develop methods that can easily adapt to any language. We present an evolution over previous works, notably by Church and Gale [1], that performs the alignement of parallel texts in any language without any need for information concerning these languages, as is often the case in existing systems. We also introduce a more experimental system, that shows promise for the alignment of degraded translations. The systems have taken part to the ARCADE II evaluation campaign [2], of which we present the results.

1 Introduction

As more and more countries and cultures develop a popular use of the Internet, the amount of distinct languages composing the information found online has grown tremendously in the last five years. Once a mostly English-speaking medium totally based on the latin alphabet, Internet has now become a place where information is represented in radically different ways, in languages that for example may not even share the basic notion of "word".

Being able to find relevant information for a given need in any language, whatever language the need is expressed in, is one of the challenges associated with that evolution. A much explored way to solve it is, of course, automatic translation; the problem of that approach is that each new language to take into account involves a huge amount of human work for the creation of lexicons, grammars, *etc.*, which soon becomes un-tractable as the number of languages increases. But full text translation is not necessary to know that two documents address a same topic: often, a few known keywords or statistical properties of the considered texts are sufficient for that task. The interest of that more "loose" approach is that the information needed can be automatically acquired from training data presented as aligned parallel corpora.

Parallel corpora are instances of a same text written in different languages, and aligning them means exposing the links between portions of text bearing the same meaning. The interest of that information for the automatic acquisition of knowledge concerning semantic equivalences between languages is of course very important. Although parallel corpora are quite frequent– for example, multilingual news outlets –, aligning them still requires work. Doing so manually is extremely tedious and time-consuming, and studies have therefore been conducted in order to do it automatically.

Church and Gale [1] were among the first to propose a system to perform the task of sentence-level parallel text alignment. Given two texts carrying the same message, they make the assumption that the ratio between the lengths of sentences (in numbers of characters) having the same

meaning is roughly constant. Using the DTW (dynamic time warping) algorithm, they then progressively match sentences according to their lengths, while taking into account the fact that a single sentence may be translated as two, or reciprocally, or that a sentence may be deleted. Many works have used this basis, sophisticating it by using, for example, lists of known translations, or looking for corresponding language independent text elements, like numbers. The main difficulty raised by this system is that it is language-dependent, since the sentence length ratio for the considered languages must be provided. The authors claim that simply setting it to 1 has little impact on the system performance: that is true as long as the actual ratio remains relatively close to 1, as is the case for European languages, but performance does degrade in the case of more radically different sentence lengths (typically when the notion of " character " changes, as when aligning English and Chinese texts).

Kay and Röscheisen [3] introduce a different approach that considers globally the repartition of words over text sentences. Words having a similar distribution are assumed to be translations of one another, and the correspondances between their occurrences are used as anchors for sentence alignment. The weakness of that approach, which we call " lexical alignment ", is that it features two potential sources of mistakes: computed word matches may be wrong, and even when they are not, occurrence matches may be incorrect too. Moreover, the performance of that system degrades if the texts to be aligned are in languages having a very different conception of what is a " word " (e.g. flexional *vs* isolating language).

The work we present is inspired by both of those fundational studies, but attempts to create a truly language-independent parallel text alignement system. Contrary to many works carried on in the field on parallel corpus alignment, we do not wish to improve upon the results of Church and Gale so much as to broaden the range of application of their method: our system must be applicable to any language without any prerequisite, for example in situations where dictionary-based approaches such as [4] cannot be used, due to a lack of electronic dictionaries in the considered languages.

We present in Section 2 our modification of the Church-Gale algorithm to make it depend on the studied texts only, and our improvement on previous work by Romary and Bonhomme [5] to take into account the logical structure of documents. We then introduce in Section 3 an attempt at making that system cooperate with another based on lexical alignment. We finally give in Section 4 the results of several evaluations; one is performed using the text of the novel *Le petit prince* in French, English and Vietnamese, and two come from the ARCADE II [2] evaluation campaign: the first between European languages, the second between French and more " exotic " languages such as Russian or Persian.

2 Sentence-level alignment

We present in this section several improvements over the work presented in [1], designed to make it independent on the languages of the texts to align, and to improve its performance through the exploitation of document structure.

We first describe in Section 2.1 the DTW algorithm used by Church and Gale in [1]. We then present in Section 2.2 the alternative sentence matching probability measure we define. In Section 2.3, we introduce an existing work by Romary and Bonhomme [5] to exploit document structure information in order to improve the performance of the alignment, and describe an evolution that we propose to make the algorithm tolerant to discrepancies in document structure encoding. That evolution is notably necessary in the case of documents gathered online, for which we have no *a priori* knowledge of the reliability of their source.

2.1 DTW algorithm

Given two series $a_{i,1\leq i\leq n}$ and $b_{j,1\leq j\leq p}$ of sentence lengths, and given a cost function $c(a_i, b_j)$ reflecting how much a_i and b_j are unlikely to be the lengths of matching sentences, the DTW algorithm computes the sequence of sentence pairings that minimizes the sum of all alignment costs while respecting the ordering of the elements. It does so by building a matrix $M = m_{ij,1\leq i\leq n,1\leq j\leq p}$, where m_ij is the minimum total cost of pairing elements $(1,1)$ through (i,j). Then, naturally, m_{np} is the minimum total cost to align the two series. The m_{ij} are computed recursively by:

$$m_{ij} = \min \begin{cases} m_{i-1,j-1} + c(a_{i-1}, b_{j-1}) & \text{(simple)} \\ m_{i-1,j} + c(a_{i-1}, 0) + penalty_{del} & \text{(delete)} \\ m_{i,j-1} + c(0, b_{j-1}) + penalty_{ins} & \text{(insert)} \\ m_{i-2,j-1} + c(a_{i-1} + a_{i-2}, b_{j-1}) + penalty_{comp} & \text{(compression)} \\ m_{i-1,j-2} + c(a_{i-1}, b_{j-1} + b_{j-2}) + penalty_{exp} & \text{(expansion)} \\ m_{i-2,j-2} + c(a_{i-1} + a_{i-2}, b_{j-1} + b_{j-2}) + penalty_{mix} & \text{(mix)} \end{cases}$$

To each scenario other than a simple match (sentence deletion, compression of two sentences into one, *etc.*) is associated a constant penalty inversely proportional to its probablity in a "typical" aligned corpus. Each cell of the matrix keeps a link to the cell that was chosen to yield the minimal cost, in order to rebuild the optimal path once all values of the matrix have been computed.

The cost function c is central to the functioning of the DTW algorithm. Church and Gale define it using a constant reflecting the typical sentence length ratio between the two considered languages. Since we are interested in developping a system that lets us align texts in any language, we cannot depend on such *a priori* information. That is why we define a different formula to evaluate the sentence matching probability, which we now present.

2.2 Language-independent sentence matching probability

To define our sentence matching probability measure, we observe the repartition of sentence lengths in the two documents or corpora to align. Assuming a roughly gaussian distribution, we compute the means and standard deviations of the two distributions: $\bar{l_1}$ and σ_{l1} for the source text, and $\bar{l_2}$ and σ_{l2} for the destination. To compute the probabilty that a sentence of length l_2 is the translation of a sentence of length l_1, we first center and reduce those values:

$$l'_1 = \frac{l_1 - \bar{l_1}}{\sigma_{l1}} \qquad l'_2 = \frac{l_2 - \bar{l_2}}{\sigma_{l2}}$$

and then compute the area under the normal distribution curve between l_1 and l_2. That area gathers all the values of l that would be a better fit for l_1 than l_2, in other words it represents the probability to find a better value than l_2 to match with l_1. Hence the probability of a l_1 - l_2 match is equal to one minus that value; and we convert it to logarithmic scale to obtain the desired cost function:

$$c(l_1, l_2) = -\log\left(1 - \int_{l'_1}^{l'_2} e^{\frac{-t^2}{2}} dt\right)$$

(assuming that $l'_1 \leq l'_2$). That new measure lets us achieve better results than that of Church and Gale in the case where the languages to align are strongly different (*cf.* Section 4.2 for the result of French-Chinese or French-Arabic alignment). Another way to improve Church and Gale's work is through the refinement of the alignment algorithm itself. For that, we extend existing work taking into account document structure, as we describe in the following subsection.

2.3 Structure-driven document alignment

Romary and Bonhomme present in [5] a system based on the DTW approach of Church and Gale, but that exploits the fact that many documents are encoded with rich structure information. We introduce here the principle of that evolution, and describe the method we have devised to overcome the main limitation of that approach, which is its great sensitivity to structure encoding discrepancies.

The work presented in [5] assumes that, if sentences can be put into correspondance, so can larger text divisions such as paragraphs, sections, *etc.* Moreover, those text segments being longer and in smaller number than sentences, their alignment should be less prone to error. The algorithm proposed by Romary and Bonhomme consists in first performing the alignment of the largest divisions, and then recursively, within those divisions, of smaller and smaller text units.

That system permits important quality improvements over the basic DTW approach, but is not able in its current state to deal with cases where there are differences in structure encoding between the documents to align. For example, structure information may be incomplete for one document, or one may indicate sections and paragraphs when the other only indicates sections. To alleviate that problem, before the alignment of each level in two documents, we first check the consistency of their encoding. If there is an encoding incoherence (mix of paragraphs and divisions, for instance), we ignore the encoding at this level and pass to the next level. Then, if we find an important difference in number of text segments between the two levels to be aligned (ratio > 2 or < 0.5), then we pass to the next level of the document having fewer elements. The process is iterated until we obtain two comparable levels in the two texts.

The evolutions we have brought to the original Church and Gale algorithm thus make for a robust system able to align texts in any languages, and well adapted in particular to online documents, that often feature structural information, but not always encoded in a coherent manner.

In the next section, we present experiments aiming at combining that first system with the lexical alignement approach of [3].

3 Hybrid alignment system

Although it has limitations that have led us to bring our attention in priority to the Church-Gale algorithm, the approach of Kay and Röscheisen [3], which we have called " lexical alignment ", has the advantage of being less sensitive to inaccurate translations. We therefore try to make the two approaches cooperate in a feedback loop, using the lexical alignment as a guide for sentence alignment, and the sentence alignment as an indication for the lexical alignment.

We have not used the original algorithm of [3], but another one of the same family, called DK-Vec. That method is presented in [6] as a good starting point for lexical alignment. We then present the way in which the feedback loop takes place between it and our improved version of the Church-Gale algorithm, and conclude this section by evaluating the practical interest of that development.

3.1 DK-Vec algorithm

The DK-Vec algorithm is presented by Fung and McKeown in [7]: with each word w in the texts to be aligned is associated the vector of successive distances between its consecutive occurrences, $D^w =< d_1^w, \ldots, d_n^w >$, where n is the number of occurrences of word w in the corpus. The similarity between the distance vectors of two words s and t of the source and target corpora, respectively, is computed using the DTW algorithm, this time to match word occurrences. To reduce the computational load, the pairs of s - t candidates are pre-filtered using a simpler statistical criterion ensuring that their distance vectors have similar means and standard deviations. The pairs of

words retained are the ones having the smallest occurrence alignment cost. We now describe how that information can be used in conjunction with the structural alignment algorithm.

3.2 Feedback loop

The reciprocal action of enrichment of each approach by the other is integrated into the hierarchical structure of the algorithm. Texts are first aligned at the largest available text scale (typically, chapters) using the first system. Then, the DK-Vec algorithm is applied, exploiting that first result to scale word positions. The best 25 % of propositions of pairs of words given by DK-Vec are retained, and then used to influence the computation of sentence alignment costs in the first system, applied this time to a smaller textual unit. We thus treat smaller and smaller units until the sentence level is reached. The methods used to compute the influence of each method on the other are as follows:

Influence of lexical alignment on structural alignment: this is done by modifying the sentence alignment cost function. Instead of only taking into account the relative lengths of two sentences, the system also considers if word occurrences are aligned between them. To reflect that, the cost of sentence alignment is divided by the proportion of aligned word occurrences between the considered sentences.

Influence of structural alignment on lexical alignment: in the original DK-Vec algorithm, distance vectors are built from word positions in number of characters from the beginning of the text. We change that measure to take into account our already acquired partial knowledge of the structural alignment. If we assume that the algorithm has functionned properly until the current stage, and text division s_i in the source document is aligned with text division t_j in the target document, then word occurrences in s_i should be aligned with word occurrences in t_j. Moreover, the relative positions of aligned word occurrences within aligned text segments should also be similar.

To reflect that principle, we transform word positions into real numbers defined by:

- an integer part such that the coordinates of all word occurrences of a same text segment have the same integer part, and that if segment s_i in the source document is aligned with segment t_j in the target document, then word occurrences of s_i and t_j have the same integer parts (the order of integer parts is the same as the natural text order);
- a fractional part equal to the relative position of the word occurrence within the enclosing text segment.

3.3 Impact on result quality

Many experiments made to compare that hybrid system with the original one show three kinds of behaviour: in the case of similar languages with parallel corpora of good quality, the results of the two systems are very similar; in the case of very different languages, the hybrid system suffers from the shortcomings of the DK-Vec algorithm in that context, and the first one obtains better results; finally, in the case of parallel texts in similar languages, but with an important degradation of one version, the hybrid system performs significantly better, as could be expected thanks to the superior robustness of the lexical alignment.

Hence, the results we now present are the evaluation of both systems: the hybrid one has been used for European languages, but due to its shortcomings we have employed the basic method in other cases.

Table 1. Alignment performance on the *Le Petit prince* corpus.

	fr - en	fr - vn	en - vn
Precision	96.02%	90.46%	81.42%
Recall	90.96%	87.73%	76.21%
F-measure	93.42%	89.08%	78.73%

4 Results

The system we have presented in this article has been evaluated in many different conditions of application. We first present here the results obtained on a corpus composed of the French, Engligh and Vietnamese versions of the novel *Le Petit Prince*. We then detail the results of the Arcade II evaluation campaign [2], for which we took part to two tasks: alignment between 5 european languages, and alignment between French and 6 languages not based on the latin alphabet.

4.1 *Le petit prince* corpus

The *Le Petit Prince* corpus consists of three texts in French, English and Vietnamese, featuring respectively 18.2, 20.8 and 18.5 thousand words, and 1674, 1660, and 1663 sentences.

The result of the automatic alignment compared to a manually aligned reference is presented in Table 1. The system used for this experiment is the hybrid one, and the results confirm its relative weakness in the case of very different languages (here, two flexional and one isolating). The poor performance of the system for the en-vn pair is also due to the fact that, the original text being in French, the English and Vietnamese version are separated by the distance of two translations.

4.2 ARCADE II evaluation campaign

Following the first ARCADE campaign of automatic text alignment tool evaluation [8], ARCADE II proposed several new challenges, notably the extension of the list of studied languages. Our systems have taken part in two evaluations:

- the JOC corpus contains five versions of excerpts from the Official Journal of the European Community, of about 1 million word each, aligned at the sentence level. The translation is very thorough, designed to make the documents as similar as possible, and most alignment methods perform well;
- the MD corpus contains articles from the newspaper *Le Monde diplomatique* translated into Greek, Russian, Arabic, Persian, Chinese and Japanese. The translation is less litteral than for the JOC corpus. Table 2 details the amount of text available for each language.

Table 3 shows the result on the JOC corpus; given the considered languages, we used the hybrid technique for that task. Table 4 presents the results obtained on the MD corpus, this time using the basic system. The first (“official”) figures correspond to the result published by ARCADE II: the very poor performance of our system here is due to an error in the document structure analysis code, which we have since corrected, obtaining the results labeled by “corrected system”. Those figures show that our approach does permit to reach quite good alignment performance for very different kinds of languages.

Table 2. Composition of the MD corpus.

	# doc.	# Kwords		# Kseg.		# Kalign.
fr - ar	150	517	403	14	11	11
fr - zh	59	197	-	5.2	5.5	4.45
fr - el	50	179	190	4.3	4.4	4.37
fr - ja	52	240	-	5.7	6.1	5.51
fr - fa	53	214	220	5.2	5.3	4.61
fr - ru	50	173	158	4.2	4.2	4

ar: Arabic, el: Greek, fa: Persian, fr: French, ja: Japanese, ru: Russian, zh: Chinese.

Table 3. Alignment performance on the JOC corpus (values in percents).

	Raw corpus				Segmented corpus			
	prec.	recall	f-meas.	rank	prec.	recall	f-meas.	rank
English	95.54	98.18	96.84	2	97.33	98.82	98.06	1
German	95.04	98.15	96.57	1	97.26	98.67	97.96	2
Spanish	95.98	98.64	97.29	1	97.67	98.81	98.23	3
Italian	96.29	98.47	97.32	2	96.92	98.25	97.58	3
Global	**95.69**	**98.37**	**97.01**	**1**	**97.29**	**98.63**	**97.95**	3

5 Conclusion

We have presented a system for the automatic alignment at the sentence level of documents written in any Unicode-encodable language. Two variants are proposed: the first one is mostly an improvement over original work by Church and Gale that makes it applicable to any document by, first, defining the sentence match probability in a way that depends only on the studied document, and second, by devising a mechanism to take into account document structure in a fault-tolerant way. That first system performs relatively well on a wide range of languages, including between very different ones, as the ARCADE II evaluation campaign has shown. The second variant combines the first system with another, based on lexical alignment, in a feedback loop. Being based on lexical alignment, that second system has difficulties with languages where word units are not easily definable, such as Chinese, but in other cases it proves to be more robust for the alignment of strongly deteriorated translations.

The first results of the "hybrid" system show, when it is fully applicable, a significant improvement over the basic system. We therefore plan on making it applicable to all languages by enabling it to match character sequences of any size, using pattern-research algorithms, thus freeing it from the necessity of a lexical pre-treatment of texts.

References

1. Gale, W., Church, K.: A program for aligning sentences in bilingual corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), Berkeley, US (1991)
2. Chiao, Y.C., Kraif, O., Laurent, D., Nguyen, T., Semmar, N., Stuck, F., Véronis, J., Zaghouani, W.: Evaluation of multilingual text alignment systems : the ARCADE II project. In: Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC06), Genoa, Italy (2006)
3. Kay, M., Röscheisen, M.: Text-translation alignment. Computational Linguistics **19**(1) (1993) 121–142

Table 4. Alignment performance on the MD corpus (values in percents).

	Official Result				Corrected System			
	prec.	recall	f-meas.	rank	prec.	recall	f-meas.	rank
Arabic	88.47	93.23	90.79	1	88.47	93.23	90.79	1
Persian	59.20	69.13	63.78	2	**74.75**	**86.04**	**80.00**	**2**
Russian	80.80	83.10	81.94	2	80.80	83.10	81.94	2
Greek	45.61	47.76	46.66	2	**95.54**	**96.40**	**95.97**	**2**
Chinese	77.67	86.35	81.78	1	77.67	86.35	81.78	1
Japanese	3.23	12.25	5.11	2	**87.13**	**91.06**	**89.05**	**1**
Global	**70.18**	**79.10**	**74.37**	**2**	**87.37**	**88.39**	**87.88**	**1**

4. Alexander Gelbukh, Grigori Sidorov: Alignment of Paragraphs in Bilingual Texts using Bilingual Dictionaries and Dynamic Programming. In: Proceedings of the 11th Iberoamerican Congress on Pattern Recognition (CIARP 2006), Cancun, Mexico (2006) to appear.
5. Romary, L., Bonhomme, P.: Parallel alignment of structured documents. In Véronis, J., ed.: Parallel Text Processing. Kluwer Academic Publishers (2000) 201–217
6. Choueka, Y., Conley, E., Dagan, I.: A comprehensive bilingual word alignment system. Application to disparate languages: Hebrew and English. In Véronis, J., ed.: Parallel Text Processing. Kluwer Academic Publishers (2000) 69–96
7. Fung, P., McKeown, K.: A technical word and term translation aid using noisy parallel corpora across language groups. Machine translation **12**(1/2) (1997)
8. Véronis, J., Langlais, P.: Evaluation of parallel text alignment systems: ARCADE. In Véronis, J., ed.: Parallel Text Processing. Kluwer Academic Publishers (2000) 369–388