Lexical-Based Alignment for Reconstruction of Structure in Parallel Texts*

Alexander Gelbukh[1], Grigori Sidorov[1], and Liliana Chanona-Hernandez[2]

[1] Center for Research in Computer Science, National Polytechnic Institute,
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City, Mexico
www.Gelbukh.com, sidorov@cic.ipn.mx
[2] Faculty of Electric and Mechanical Engineering,
National Polytechnic Institute,
Mexico City, Mexico

Abstract. In this paper, we present an optimization algorithm for finding the best text alignment based on the lexical similarity and the results of its evaluation as compared with baseline methods (Gale and Church, relative position). For evaluation, we use fiction texts that represent non-trivial cases of alignment. Also, we present a new method for evaluation of the algorithms of parallel texts alignment, which consists in restoration of the structure of the text in one of the languages using the units of the lower level and the available structure of the text in the other language. For example, in case of paragraph level alignment, the sentences are used to constitute the restored paragraphs. The advantage of this method is that it does not depend on corpus data.

1 Introduction

For a text in two different languages, the parallel text alignment task consists in deciding which element of one text is translation of which one of the other text. Various researchers have tried different approaches to text alignment, usually at sentence level [5], and a number of alignment tools are available. Some methods rely on lexical similarity between two texts [3]. In our previous paper [2], we have suggested an alignment method based on measuring similarity using bilingual dictionaries and presented an approximate heuristic greedy alignment algorithm. We evaluated it on fiction texts that represent difficult cases for alignment. In this paper, our goals are to introduce an optimization algorithm that finds the best solution, instead of the approximate heuristic-based algorithm, using the same measure of lexical similarity as well, and to propose an alternative method of evaluation of alignment algorithms based on reconstruction of the global text structure in one of the languages.

* Work done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (SIP, COFAA).

Z. Kedad et al.(Eds.): NLDB 2007, LNCS 4592, pp. 401–406, 2007.

2 Similarity Measures

For assigning weight to a possible correspondence, we need to calculate the similarity between two sets of paragraphs. We define this function as similarity between two texts that are obtained by concatenation of the corresponding paragraphs.

The first baseline method is relative position of the paragraphs. Common sense suggests that the corresponding pieces of texts are located at approximately the relative same distance from the beginning of the whole text. We define the baseline distance between two pieces of text, T_A in the language *A* and T_B in the language *B*, as follows:

$$\text{Distance}(\mathrm{T}_A, \mathrm{T}_B) = |\text{start}(T_A) - \text{start}(T_B)| + |\text{end}(T_A) - \text{end}(T_B)|, \qquad (1)$$

where $\text{start}(T_X)$ is the relative position of the first word of the text T_X measured in percentage of the total number of words in the text in the corresponding language, and similarly for $\text{end}(T_X)$. We could also use the position of the paragraph instead of word as percentage of the total number of paragraphs, but the measure based on word counts has been reported as better than the one based on paragraph counts, which agrees with our own observations.

We also used the well-known algorithm by Gale and Church [1] as another baseline for comparison.

As far as lexical similarity is concerned, we define the similarity between two texts in different languages as the number of words in both texts that are not mutual translations of each other [5]. Note that it is more correct to call this penalization; we use the term "similarity" just for the sake of uniformity with other approaches. The greater is this value, the less similar are the paragraphs.

For calculating this, we take into account the number of words that are such translations taken from a dictionary. Then we calculate the number of word tokens without translation in both paragraphs, under the hypothesis that these two paragraphs correspond to each other, namely:

$$\text{Distance}(T_A, T_B) = |\,\mathrm{T}_A\,| + |\,T_B\,| - 2 \times \text{translations}. \qquad (2)$$

The cost of an alignment hypothesis is the total number of words in both texts that are left without translation under this hypothesis. Note that under different hypotheses this number is different: here we consider two word tokens to be translations of each other if both of the following conditions hold: (a) they are dictionary translations (as word types) and (b) the paragraphs where they occur are supposed to be aligned. Note that we perform morphological lemmatization and filter out the stop words.

3 Algorithm

To find the exact optimal alignment, we apply a dynamic programming algorithm. It uses a $(N_A + 1) \times (N_B + 1)$ chart, where N_X is the number of paragraphs in the text in the language *X*.

The algorithm works as follows. First, the chart is filled in:

```
$a_{00} := 0$, $a_{i0} := -\infty$, $a_{0j} := -\infty$ for all $i, j > 0$.
for $i$ from 1 to $N_A$ do
    for $j$ from 1 to $N_B$ do
        $a_{ij} := \min (a_{xy} + \text{Distance } (T_A [x + 1 .. i], T_B [y + 1 .. j]))$
```

Here, a_{ij} is the value in the (i,j)-th cell of the chart, $T_X [a .. b]$ is the set of the paragraphs from a-th to b-th inclusive of the text in the language X, and the minimum is calculated over all cells (x,y) in the desired area to the left and above the (i,j)-th cell.

As in any dynamic programming algorithm, the value a_{ij} is the total weight of the optimal alignment of the initial i paragraphs of the text in the language A with the initial j paragraphs of the text in the language B. Specifically, upon termination of the algorithm, the bottom-right cell contains the total weight of the optimal alignment of the whole texts. The alignment itself is printed out by restoring the sequence of the assignments that led to this cell:

```
$(i,j) := (N_A, N_B)$.
while $(i,j) \neq (0, 0)$ do
    $(x,y) := \text{argmin } (a_{xy} + \text{Similarity } (T_A [x + 1 .. i], T_B [y + 1 .. j]))$
    print “paragraphs in $A$ from $x + 1$ to $i$ are aligned with
            paragraphs in $B$ from $y + 1$ to $j$.”
    $(i,j) := (x,y)$
```

Here, again, the minimum is sought over the available area to the left and above the current cell (i,j). Upon termination, this algorithm will print (in the reverse order) all pairs of the sets of paragraphs in the optimal alignment.

4 Experimental Results: Traditional Evaluation

We experimented with a fiction novel *Advances in genetics* by Abdón Ubídia and its original Spanish text *De la genética y sus logros*, downloaded from Internet. The English text consisted of 114 paragraphs and Spanish 107, including the title.[1] The texts were manually aligned at paragraph level to obtain the gold standard.

As often happens with literary texts, the selected text proved to be a difficult case. In one case, two paragraphs were aligned with two: the translator broke down a long Spanish paragraph 3 into two English paragraphs 4 and 5, but joined the translation of a short Spanish paragraph 4 with the English paragraph 5. In another case, the translator completely omitted the Spanish paragraph 21, and so on.

Both texts were preprocessed by lemmatizing and POS-tagging, which allowed for correct dictionary lookup. Stop-words were removed to reduce noise in comparison; leaving the stop-words in place renders our method of comparison of paragraphs completely unusable. Then our algorithm was applied, with both baseline and suggested distance measures.

We evaluate the results in terms of precision and recall of retrieving the hyperarcs (union of several units, or arcs in hypergraph that corresponds to alignment):

[1] We did not experiment with a larger corpus because we are not aware of a gold-standard manually aligned Spanish-English parallel corpus.

precision stands for the share of the pairs in the corresponding alignments; recall stands for the share of the pairs in the gold standard that are also found in the row corresponding to the method. Alternatively, we broke down each hyperarc into pair-wise correspondences, for example, 48–50=47 was broken down into 48 ~ 47, 49 ~ 47, 50 ~ 47, and calculated the precision and recall of our algorithm on retrieving such pairs; see the last two columns of Table 1.

Table 1. Comparison of the similarity measures

Measure	Hyperarcs		Single arcs	
	Precision, %	Recall, %	Precision, %	Recall, %
Proposed	89	85	88	90
Baseline	65	28	43	54
Gale-Church	89	86.5	87.5	91.5

One can see that the proposed distance measure based on the bilingual dictionaries greatly outperforms the pure statistically-based baseline and is practically at the same level as the algorithm of Gale and Church. Still, algorithm of Gale and Church uses certain parameters especially pre-calculated, thus, it cannot be considered an unsupervised algorithm as it is in our case. Also, it relies on the hypothesis of normal distribution, in contrast with our algorithm that does not rely on any distribution.

5 Evaluation Based on Reconstruction of Text Structure

Traditional evaluation schemes usually invoke direct comparison with gold standard, or reference text alignment, see formal definitions of this kind of alignment in [4]. Both precision and recall can be computed, as well as the derived F-measure. It is mentioned in that paper that we can measure these values using different granularity, i.e., for alignment on the sentence level, correctly aligned words or characters can be measured. The authors do not mention the task of paragraph level alignment.

We suggest considering evaluation of an alignment algorithm as the task of global text structure reconstruction. Namely, if we are evaluating the correctness of correspondences at the paragraph level, let us eliminate all paragraph boundaries in one of the texts and allow the algorithm to put back the paragraph marks based on the paragraph structure of the other text and the data of the alignment algorithm itself. Then we evaluate the correctness of the restored paragraph marks using the structure of paragraphs in the other language. We cannot rely on the known paragraph structure for the same language, because the paragraphs can be aligned correctly in different manner (2-1, 3-1, etc.). In practice, this is done by considering all sentences in one of the text as paragraphs, and then paragraph-level alignment is performed.

The restoration of text structure is somehow similar to the evaluation technique based on counting the correspondences on the other level of granularity (say, using sentences for paragraphs, etc.), because it also uses the units of the lower level, but it is essentially the different task. The main difference is that while the algorithm is trying to recreate the text structure using the units of the lower level of granularity, it comes across many possibilities that it never would consider working only with the existing units. It is especially well-seen for alignment at the paragraph level. Usually, the alignment of paragraphs is not considered as an interesting task since in the majority of existing parallel text the paragraphs, even the large ones, have clear correspondences. Meanwhile, if we consider the task of text reconstruction, the paragraph alignment task becomes an interesting problem. Thus, we can evaluate and compare different approaches to paragraph level alignment. This technique can be useful also for automatic search of parallel texts in Internet.

Another consideration is related to corpus structure. As the majority of parallel texts have very similar structure at paragraph level, the problem of alignment at this level is difficult to evaluate, because in any corpus there are few interesting cases of paragraph alignment. Applying the suggested method of evaluation, we resolve the problem of the lack of non-trivial cases of the paragraph level alignment, because now any paragraph of any text is split into sentences and it is a challenge for aligning algorithms.

We conducted experiments using dynamic programming approach described above. Our goal was to compare the performance of the statistical and lexical approaches to similarity calculation using the proposed evaluation method based on reconstruction of the global text structure.

As an example of statistical approach, we used an implementation of Gale and Church algorithm [1], though we had to modify it according to the task. The problem is that this algorithm only takes into account alignment of maximum 2-2 correspondences (i.e., 3-2 is impossible, etc.) and it is penalizing the correspondences that are different from 1-1. We had to remove these penalizations because there can be many more possible correct correspondences, like, for example, 10-1, etc., and these should not be penalized. Obviously, it affects the original algorithm performance. It is the question of further investigations to determine how to modify penalizations in this algorithm or what improvements should be added to achieve the best performance.

For the lexical approach, we used the implementation of our lexical-based alignment algorithm for English-Spanish text pairs (see previous sections). For the moment, we also do not add any penalization for size of fragments, for absolute positions of fragments, or for relative position of lexical units in fragments. We expect that implementation of these parameters will improve the performance of our algorithm.

We made our experiments using the extract of 15 paragraphs from the text mentioned above. Note that it is a difficult case of non-literal translation. We made complete analysis using dynamic programming. The information about Spanish paragraphs was suppressed.

The results of the comparison using both methods are as follows for precision: 84% in lexical approach vs. 26% in statistical approach. We count the correct correspondences using the paragraph structure of the English text. When the algorithm united two paragraphs that were separated both in the Spanish text and in the English text, we counted it as an error for the half of the restored sentences. Still, it is interesting to analyze if it is the same type of error as failing to find the correct correspondence. Note that the information about the paragraph separation in Spanish text was not used.

The problem with the statistical method is that once it makes incorrect alignment, it is difficult for it to return to the correct correspondences.

6 Conclusions

We described a dynamic programming algorithm with lexical similarity for alignment of parallel texts. This is unsupervised algorithm. We conducted the experiments of the traditional evaluation obtaining very similar results with the supervised algorithm of Gale and Church. We used fiction texts that are difficult cases for alignment.

We also presented a new method for evaluation of the algorithms of parallel texts alignment. This method consists in restoration of the structure of the text in one of the languages using the units of the lower level and the structure of the text in the other language. For example, in case of the paragraph level alignment, the sentences are used to constitute the restored paragraphs in one of the languages. The advantage of this method is that it does not depend on corpus data that is random. Another consideration is that in case of paragraphs the corpus data often is trivial. Applying the proposed method, we obtain the basis for comparison of different alignment algorithms that is not trivial at the paragraph level. We conducted experiments on a fragment of English-Spanish text using the restoration method. The text was a fiction text with non-literal translation. Lexical and statistical approaches were tried for calculation of similarity using dynamic programming approach. We obtained much better results for the lexical method, though we expect that the statistical method can be improved for the proposed task.

References

[1] Gale, W.A., Church, K.W.: A program for Aligning Sentences in Bilingual Corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991)

[2] Gelbukh, A., Sidorov, G., Vera-Félix, J.Á.: Paragraph-Level Alignment of an English-Spanish Parallel Corpus of Fiction Texts using Bilingual Dictionaries. In: Sojka, P., Kopeček, I., Pala, K. (eds.) Text, Speech and Dialogue. LNCS (LNAI), vol. 4188, pp. 61–67. Springer, Heidelberg (2006)

[3] Chunyu, K., Webster, J.J., Sin, K.K., Pan, H., Li, H.: Clause alignment for Hong Kong legal texts: A lexical-based approach. International Journal of Corpus Linguistics 9(1), 29–51 (2004)

[4] Langlais, Ph., M. Simard, J. Veronis, Methods and practical issues in evaluation alignment techniques. In: Proceeding of Coling-ACL-98 (1998)

[5] Moore, R.C.: Fast and Accurate Sentence Alignment of Bilingual Corpora. AMTA-2002, pp. 135–144 (2002)