IJCLA VOL. 2, NO. 1–2, JAN-DEC 2011, PP. 257–263
RECEIVED 01/10/11 ACCEPTED 15/11/10 FINAL 16/12/11

A New Combined Lexical and Statistical based Sentence Level Alignment Algorithm for Parallel Texts

GRIGORI SIDOROV,[1] JUAN-PABLO POSADAS-DURÁN,[1]
HECTOR JIMÉNEZ-SALAZAR,[2] AND LILIANA CHANONA-HERNANDEZ[1]

[1] *National Polytechnic Institute (IPN), Mexico*
[2] *Autonomous University of Mexico State (UAEM), Mexico*

ABSTRACT

Parallel texts alignment is an active research area in Natural Language Processing field. In this paper, we propose a method for sentence alignment of parallel texts that is based both on lexical and statistical information. The alignment procedure uses dynamic programming technique. We made our experiments for Spanish and English texts. We use lexical information from bilingual Spanish-English dictionary, as well as the sentence length measured in words and in characters. The proposed method was tested on a corpus of fiction texts, where the frequency of multiple alignments, omissions and insertions is higher than in other types of texts. We obtained better results than the standard Vanilla aligner system that uses a purely statistical approach.

KEY WORDS: *Parallel texts alignment, English, Spanish, dynamic programming, lexical-based alignment, statistical alignment, anchor points, sentence length, fiction writing.*

1 INTRODUCTION

Parallel texts alignment is an active area of research in Natural Language Processing field. Parallel texts are translations of each other

in different languages. The works in this area started at wide scale in 1990s [1, 2, 4, 7, 8]. Some modern ideas of optimization, for example, with genetic algorithms [5], and the task of alignment of more challenging types of texts (like fiction) [6] maintain the interest to this topic.

The usefulness of the aligned parallel texts is related to the fact that they explicitly contain the relationship between the elements in a text in one language and elements of the same text translated into another language, thus, allowing performing of numerous tasks that can exploit the knowledge of the alignment. For example, alignments are useful for machine learning and automatic extraction of information for various purposes, such as statistical machine translation, bilingual word sense disambiguation, etc.

In this work, we propose a method for sentence alignment in parallel texts. We tested it for Spanish and English language pair, though the algorithm is language-independent. It is based on both lexical and statistical information and uses dynamic programming framework [6]. The lexical information is contained in a bilingual dictionary, while statistical information is obtained from the sentence lengths like in [4] measured both in words and in characters.

The proposed method was tested on a corpus of fiction texts. Note that fiction texts are much more difficult case for alignment: the frequency of multiple alignments, omissions and insertions is higher than in texts of other types, like technical or law texts.

We obtained precision about 96%. We compared our results with the results obtained by the Vanilla aligner system [3] for the same corpus. Note that Vanilla uses a purely statistical approach. Vanilla obtained precision about 92%.

The developed method is superior in cases of multiple alignments, omissions and insertions. The results that we obtained show that the use of lexical information contained in a bilingual dictionary combined with statistical information allows developing of the robust method for sentence alignment in fiction texts, which gives better results than purely statistical methods.

The paper is organized as follows. First we describe the alignment algorithm, then we present the corpus used in the experiments and after this present and discuss the obtained results, finally, conclusions are drawn.

2 PROPOSED ALGORITHM

In the task of sentence alignment, we need to find the correspondences of sentences in one text and the sentences of its translation. This correspondence is not trivial because some sentences can be omitted and some sentences can be translated by two or more sentences. The task of sentence level alignment is well-known and intuitively clear, we send the reader for details to the works [4, 6].

We use dynamic programming approach to find the best alignment for all sentences in a pair of texts [4, 6]. Note that this approach implies continuity, i.e., if we found an alignment of a sentence, then no posterior sentence can be aligned before the already existing alignment.

In this approach, we assign scores to every possible pair of aligned sentences. The final score of the alignment is the sum of the scores of each sentence pair that constitutes this alignment.

For weighting the similarity, we use the concept of *significant elements* that are words of open POS classes: nouns, verbs, adjectives, adverbs, and also proper names, abbreviations, signs of admiration, signs of interrogation, and numbers.

We use the following equation for calculating the similarity between sentences. In fact, this function is dissimilarity, i.e., penalization, but penalization is "inverse" function to similarity. Thus, the lower this value, the better is the similarity of the sentences.

$$\begin{aligned} \mathit{Similarity}\,(S_S; S_T) = {} & \mathit{DictionaryDiff}\,(S_S; S_T) + \\ & \mathit{SignificantElementsDiff}\,(S_S; S_T) + \\ & \mathit{CharLengthDiff}\,(S_S; S_T) \end{aligned} \quad (1)$$

where S_S is a source sentence, S_T is a target sentence, the value *SignificantElementsDiff* $(S_S; S_T)$ is the absolute value of the difference of number of significant elements, and *CharLengthDiff* $(S_S; S_T)$ obtains the absolute value of the difference of numbers of characters in two sentences, and *DictionaryDiff* is calculated as follows.

$$\begin{aligned} \mathit{DictionaryDiff}\,(S_S; S_T) = {} & \mathit{SignificantElements}\,(S_S) + \\ & \mathit{SignificantElements}\,(S_T) - \\ & 2 \times \mathit{Translations}(S_S; S_T) \end{aligned} \quad (2)$$

Thus, *DictionaryDiff* $(S_S; S_T)$ is the number of significant elements that are not mutual translations. In our experiments, we used Vox dictionary available on the Web.

```
<Preproccessing>
  <text file="holmes25_eng.txt" num_sentence="27" num_char="2703" language="English">
    <sentence id_sentence="1" num_words="4" mean_words="4" char="19" language="English">
      <token char="Sir" class="normal">
        <type char="sir">
          <translation>señor</translation>
          <translation>sir</translation>
        </type>
      </token>
      <token char="Arthur" class="nombre_propio">
        <type char="arthur">
          <translation>arthur</translation>
        </type>
      </token>
      <token char="Conan" class="nombre_propio">
        <type char="conan">
          <translation>conan</translation>
        </type>
      </token>
      <token char="Doyle" class="nombre_propio">
        <type char="doyle">
          <translation>doyle</translation>
        </type>
      </token>
    </sentence>
    <sentence id_sentence="2" num_words="5" mean_words="3" char="29" language="English">
      <token char="The" class="palabra_auxiliar">
        <type char="the"/>
      </token>
      <token char="adventures" class="normal">
        <type char="adventure">
          <translation>aventura</translation>
```

Fig. 1. Text representation using XML scheme.

In addition, we use anchor points as a constraints in the alignment, i.e., the alignment is possible only between established anchor points.

We used XML scheme of representation of the text, see Fig. 1. Traditional preprocessing with morphological normalization was performed.

3 Corpus Description

We used the following corpus for our experiments: Four chapters of "Adventures of Sherlock Holmes" by A. Conan Doyle and two chapters of "Turn of the screw" by Henry James. The text by A. Conan-Doyle contains 2,558 sentences in English and 2,267 sentences in Spanish. The text by Henry James contains 361 sentences in English and 363 sentences in Spanish. Corpus characteristics related to number of different types of alignments are presented in Table 1.

Manual alignment was performed for these texts, thus, representing the golden standard for the evaluation.

Table 1. Corpus characteristics: numbers of different types of alignments.

Alignment type	A. Conan-Doyle	H. James
1:1	2,061	300
1:2	65	12
2:1	64	8
2:2	4	1
3:1	1	0
1:3	2	0
1:0	11	1
0:1	7	1

4 Obtained Results and Discussion

We compared on the developed corpus our method and Vanilla aligner. Vanilla aligner [3] is based on the work of Gale and Church [4] and takes into consideration the following alignment types between sentences 1:1, 1:0, 0:1, 1:2, 2:1 and 2:2. The system uses the same input as our system, i.e., the text is splitted into sentences in the same way. It is purely statistical method based on lengths of sentences and dynamic programming.

We measure the precision of both methods as the percentage of correctly aligned sentences as compared to the golden standard. We used complete evaluation, i.e., if only a part of a multiple alignment is correct, we do not consider such alignment as the correct one.

The results of comparison of our algorithm with Vanilla aligner for the texts of Henry James and A. Conan-Doyle are presented in Table 2.

Table 2. Comparison of the results.

	Henry James	A. Conan-Doyle
Vanilla aligner	93.01%	90.66%
Proposed algorithm	**96.78**%	**95.62**%

The reasons of errors of our method are related with two major phenomena. The first one is that the word cannot be found in the bilingual dictionary, either due to incompleteness of the morphological analysis system, or due to the incompleteness of the bilingual dictionary itself.

The other reason is related to idiomatic expressions, like "*Oh, dear*" that is translated as "*Válgame Díos*" (lit. "*God help me*)".

5 Conclusions

The aim of this work was to propose an alignment algorithm that is based on combination of lexical information obtained from a bilingual dictionary and traditional statistical information related to sentence lengths.

We tested it for a special type of parallel texts that constitutes more complicated case of alignment, namely, fiction texts. Our hypothesis was that for these texts statistical aligner will make mistakes that can be corrected using lexical information.

We conducted our experiments on a relatively large corpus of fragments of works of A. Conan-Doyle and Henry James in Spanish and in English.

We obtained better results than the base line Vanilla aligner software. Our precision was about 96%, while Vanilla aligner obtained about 92%. The difference is due to better performance of our method for cases of one to many alignments, insertions and omissions.

Acknowledgements**.** Work done under partial support of Mexican Government (CONACYT projects 50206-H and 83270, SNI), Government of Mexico City (project PICCO10-120), European project 269180 WIQ-EI, project "Answer Validation through Textual Entailment" CONACYT-DST (India), and National Polytechnic Institute, Mexico (projects SIP 20111146, 20113295; PIFI, COFAA).

References

1. Brown, P., Lai, J., Mercer, R.: Aligning sentences in parallel corpora. In: Proceedings 29ht Annual Meeting of the ACL, 1991, pp. 169−176.
2. Chen, S.F.: Aligning sentences in bilingual corpora using lexical information. In: Proceedings of ACL-93, 1993, pp. 9−16.
3. Danielsson, P., Riddings, D.: Practical presentation of a Vanilla aligner. In: TELRI Workshop in Alignment and Exploitation of Texts, 1994, pp. 1−2.

4. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991, pp. 177−184.
5. Gautam, M., Sinha, R.: A hybrid approach to sentence alignment using genetic algorithm. In: Proc. of International Conference on Computing: Theory and Applications, 2007, pp. 480−484.
6. Gelbukh, A., Sidorov, G.: Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming. Lecture Notes in Computer Science, vol. 4225, Springer-Verlag, pp. 824−833.
7. Kay, M., Roscheisen, M.: Text-translation alignment. Computational Linguistics, 19(1), 121−142, 1993.
8. Simard, M., Foster, G., Isabelle, P.: Using cognates to align sentences in bilingual corpora. In: Proc. of TMI, 1992, pp. 67−81.
9. Xiong, H., Xu, W., Mi, H., Liu, Y., Liu, Q.: Sub-sentence division for tree-based machine translation. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort'09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 137−140.

GRIGORI SIDOROV
CENTER FOR COMPUTING RESEARCH (CIC),
NATIONAL POLYTECHNIC INSTITUTE (IPN),
COL. NUEVA INDUSTRIAL VALLEJO, CP 07738, DF, MEXICO
E-MAIL: < SIDOROV @ CIC.IPN.MX>

JUAN-PABLO POSADAS-DURÁN
CENTER FOR COMPUTING RESEARCH (CIC),
NATIONAL POLYTECHNIC INSTITUTE (IPN),
COL. NUEVA INDUSTRIAL VALLEJO, CP 07738, DF, MEXICO
E-MAIL: < JPDURAN @ CIC.IPN.MX>

HECTOR JIMÉNEZ-SALAZAR
AUTONOMOUS UNIVERSITY OF MEXICO (UAM),
UNIDAD CUAJIMALPA, DF, MEXICO
E-MAIL: < HGIMENEZS @ GMAIL.COM>

LILIANA CHANONA-HERNANDEZ
ENGINEERING FACULTY (ESIME),
NATIONAL POLYTECHNIC INSTITUTE (IPN),
COL. ZACATENCO, CP 07738, DF, MEXICO
E-MAIL: < LCHANONA @ GMAIL.COM>