Paragraph-Level Alignment of an English-Spanish Parallel Corpus of Fiction Texts Using Bilingual Dictionaries*

Alexander Gelbukh, Grigori Sidorov, and José Ángel Vera-Félix

Natural Language and Text Processing Laboratory
Center for Research in Computer Science
National Polytechnic Institute
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City, Mexico
sidorov@cic.ipn.mx
http://www.Gelbukh.com

Abstract. Aligned parallel corpora are very important linguistic resources useful in many text processing tasks such as machine translation, word sense disambiguation, dictionary compilation, etc. Nevertheless, there are few available linguistic resources of this type, especially for fiction texts, due to the difficulties in collecting the texts and high cost of manual alignment. In this paper, we describe an automatically aligned English-Spanish parallel corpus of fiction texts and evaluate our method of alignment that uses linguistic data-namely, on the usage of existing bilingual dictionaries-to calculate word similarity. The method is based on the simple idea: if a meaningful word is present in the source text then one of its dictionary translations should be present in the target text. Experimental results of alignment at paragraph level are described.

1 Introduction

Current development of corpus linguistics and machine learning methods in text processing leads to increasing importance of text corpora, from raw texts to texts marked up with certain additional linguistic information: phonetic, morphological, syntactic, word senses, semantic roles, etc. The simplest form of such marks is linguistic information on the text itself: e.g., part of speech marks on the words. A more interesting kind of marks relates elements of the text to some external source: some another text, multimedia items [12], multimodal streams [15], pragmatic situation, etc. In spite of great importance of such information, there exist few corpora offering it.

In this paper we are interested in a specific kind of corpora with such external information: aligned parallel texts, i.e., texts that express the same information in two different languages, with the structural parts (units) of these texts that are mutual translations explicitly related to each other by the markup. The procedure of establishing such relation is called alignment. There are various levels of alignment depending on what is considered a unit:

– Document: we just know that two documents express the same information in two different languages (in case of very short texts, such as news messages or paper abstracts, this can be directly useful);

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

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 61–67, 2006.

- Paragraph: it is indicated which paragraphs are mutual translations;
- Sentence: similarly, but the information is given for individual sentences;
- Phrase (such as noun phrase) or word.

A unit in a parallel text can have one, several, or no counterparts: for example, a sentence can be translated by several ones; some words can be omitted, etc. This makes alignment of parallel texts quite a difficult task. Such situation is especially frequent in fiction texts, which we discuss in this paper.

An obvious source of parallel texts is Internet. Unfortunately, texts presented in Internet are difficult to process: they may contain pictures, special formatting, HTML tags, etc. Often the texts are in PDF format and during their conversion into plain text format the information about paragraph boundaries is lost. Rather extended preprocessing, often manual, is necessary for these texts. We will not discuss here this process, as well as the problem of automatic search of parallel texts in Internet.

The importance of aligned parallel corpora is closely related to the presence of structural differences between languages. On the one hand the differences make the alignment task very difficult, but on the other hand, they can be exploited for automatic extraction of information on various linguistic phenomena-for example, for word sense disambiguation. An obvious application of parallel corpora is machine translation [1], from dictionary compilation to example-based machine translation.[1] Examples of other applications are automatic extraction of data for machine learning methods, bilingual lexicography [8,13], and language teaching.

There are two major classes of alignment methods: those based on statistical data and those using additional linguistic knowledge.[2] This distinction is not related to methods of processing but to types of information involved.

Statistical methods usually exploit the expected correlation of length of text units (paragraphs or sentences) in different languages [5,9] and try to establish the correspondence between the units of the predicted size. The size can be measured in words or characters. Linguistic-based methods, on the other hand, use linguistic data for establishing the correspondence between structural units.

Statistical methods work well for texts when very literal translation is necessarily, like texts of laws or technical manuals. For fiction texts, where the structure of the original and the translation can vary significantly, it is more desirable to use linguistic methods, though they require more thorough processing. Linguistic methods can also be used for alignment at word level, though they require not only dictionaries, like the method described in this paper, but also some additional syntactic information or syntactic heuristics. For example, translation of an adjective should not be too far from the translation of the noun it modifies, etc.

The idea of application of dictionaries to alignment problem is not new [2,10,11,13]; of recent works the paper [4] can be mentioned. Still, it is not as popular as statistical methods-probably due to the fact that the dictionaries are not so easily available. Usually, as we have mentioned above, the experiments on parallel texts are conducted using specialized texts, like texts of Canadian or European parliament, legal texts, or programming help files. The

[1] Perhaps the most famous example of application of parallel texts is deciphering of Egyptian hieroglyphs based on the parallel texts of Rosetta stone.

[2] Similarly to, say, word sense disambiguation [6] and many other language processing tasks.

problem we are dealing with in this paper is how a method based on a bilingual dictionary method performs for fiction texts.

The main idea behind this method is that if a meaningful word is present in one text then one of its dictionary translations should be present in the other text. So, even if some words are omitted, the presence of the translation of other words allows for alignment at least at the level of paragraph or sentence. The situation is more complicated at the word level, because, as we mentioned, it requires not only lexical information from the dictionaries, but also syntactic information.

The paper is organized as follows. First we describe the English-Spanish parallel corpus compiled for fiction texts of significant size. Then we present a method of alignment based on the use of bilingual dictionaries. Finally, we discuss the experiments on evaluation of automatic alignment conducted at paragraph level.

2 Preparation of the Parallel Corpus

We began with preparation of a bilingual corpus, i.e., with compilation and preprocessing of parallel texts. Generally speaking, a corpus can contain texts of different genres, like fiction, newspaper articles, technical manuals, etc. In our case, we chose fiction genre because it is the most non-trivial case of translation. Sometimes, fiction texts present the situations that are difficult for automatic processing: for example, if one of the texts contains a joke, it is possible that the other text will contain a joke that has nothing in common with the original. I.e., the fiction translation is not literal.

On the other hand, there are many fiction texts in Internet, while variety of possible parallel texts of other genres is lower.

We included the following titles in our corpus: *Alice's adventures in wonderland, Through the looking-glass, The adventures of Sherlock Holmes, The turn of the screw, The jungle book, Frankenstein, Dracula, Advances in genetics*,[3] *Five weeks in a balloon, From the earth to the moon, Michael Strogoff, Twenty thousand leagues under the sea* by Lewis Carroll, Arthur Conan Doyle, Henry James, Rudyard Kipling, Mary Shelley, Bram Stoker, Ubídia, Abdón, and Jules Verne, correspondingly, with their Spanish equivalents. The texts were originally in PDF format and were preprocessed manually for elimination of special format elements, Internet links, and for restoration of paragraph boundaries. The size of corpus is more than 11.5 MB. The corpus size might seem too small, but this is a parallel corpus, for which the data are very difficult to obtain. Our corpus is freely available upon request for research purposes.

3 Alignment Based on Bilingual Dictionaries

Our task is alignment of structural units (parts) of texts, such as paragraphs and sentences. We do not consider words for the moment because of the lack of syntactic information necessary for this type of alignment. Note that the correspondence of paragraphs and sentences in the source and target texts is not necessarily one-to-one; see an example in Table 1.

We use lemmatization of words. For Spanish, we apply the morphological analyzer AGME [14] that allows for analysis of grammar forms of 26,000 lexemes, i.e., more than a million grammar forms; for English we use a similar morphological analyzer [6] based

[3] This is a fiction text, not a scientific text.

Table 1. Example of alignment of paragraphs with pattern 2-to-1

Spanish	Literal English translation	Real English text
Luego, al percatarse de mi gesto estupefacto, corrigió:	*After this, when she noticed my surprised gesture, she corrected herself:*	*When she saw my baffled look, she corrected herself: "My grandmother."*
— No. Es mi abuela.	*"No, this is my grandmother."*	

on the morphological dictionary of WordNet with about 60,000 lexemes. Thus, all words in the pair of texts are normalized before we begin any further processing. We do not resolve morphological homonymy; instead, we consider the presence of homonymic lemmas as a source of possible translations. It is justified by the fact that often the only difference between morphological homonyms is their part of speech, i.e., semantically they are similar (*work* vs. *to work*). Also, often a word can be translated into the other language using a different part of speech. It may be even useful in future work to search the translations of all related words of different parts of speech—at least in some cases. If the morphological analyzer does not succeed with analysis of a word then we treat it as a literal string and search for exact matching in the other text. This is useful, for example, for proper names.

We filter out all auxiliary words, i.e., if at least one morphological homonym is an auxiliary word (preposition, article, conjunction, or pronoun) then we do not consider it for further comparison. This corresponds to the use of a stop word list, but in our case we rely on morphological analyzers to determine whether the word should be considered or not. It is justified by the very high frequency of this type of words, so that they can be present practically in any sentence or paragraph.

3.1 Similarity Measure

Dictionary-based methods use some similarity measure. Sometimes it is a global optimization, as in [10], and sometimes local, as in [4]. We used the simplest local measure. The main idea of the implemented method is that if a meaningful word appears in a sentence or paragraph of the source text, then one of its possible translations given in a bilingual dictionary should appear in the corresponding part of the target text. It is not always so, because a word can be translated by several words, omitted, or substituted by a far synonym that does not appear as a translation variant in the dictionary.

In the future it is interesting to analyze the cases of absence of translation. For example, we expect that many such cases occur due to the presence of idiomatic expressions. This may allow for automatic extraction of such expressions. In case of translation of a word using non-idiomatic word combination, it is natural to expect that at least one word in this word combination would be its dictionary translation.

These considerations allow for introducing a measure of similarity between two structural units for a pair of parallel texts. We use as the measure of similarity a coefficient similar to the well-known Dice coefficient:

$$\mathit{Similarity} = \frac{2x}{y+z} \times \begin{cases} \frac{1}{K} & \text{if } K \geq 1, \\ K & \text{if } K < 1, \end{cases}$$

where y is the size in words of the source text unit, z is the size of the target text unit, x is the number of intersections of the two units counted as the presence of any dictionary translation of a source text word in the target text unit; $K = y \times k/z$, k is an expected coefficient of correspondence of the number of words in source and target texts: namely, $k = Z/Y$, where Z and Y are the total number of words in the target and the source text, correspondingly. This measure differs from the traditional Dice coefficient in the parameter K, which penalizes the units with too different sizes.

For the current version of the algorithm we used Spanish-English dictionary with about 30,000 entries. In the future, we plan to use English-Spanish dictionary as well, because it is not guaranteed that the dictionaries are symmetric. We will use the average value of similarity calculated with both dictionaries.

3.2 Algorithm

Currently we use an *ad hoc* alignment algorithm that passes through the units of the source text and analyzes three immediately available units from the target text, i.e., we search only the alignment patterns 1-to-1, 1-to-2, 1-to-3, 2-to-1, and 3-to-1 paragraphs. It is not a complete scheme, but it serves for evaluation of quality of the results of our dictionary-based method of alignment. Other patterns are extremely rare and in fact did not occur in our experiments with paragraphs. In the future, we plan to use a genetic algorithm for searching for the global alignment optimum, as in [7] for word sense disambiguation. Other possibilities are to use dynamic programming [5] or simulated annealing [3].

For improvement of the performance of the algorithm we also implemented an anchor points technique. Anchor points are short units with very high similarity. They serve for limiting the effect of cascaded errors during alignment, since the algorithm restarts after each anchor point. With this, if an error occurs, it will not affect alignment of the whole text. The algorithm performs two passes. First, it searches for anchor points, and then it processes the units between the anchor points.

The algorithm implements the following processing: it takes a unit from the source text (Spanish in our case), calculates the similarity of the patterns 1-to-1, 1-to-2, 1-to-3, 2-to-1, and 3-to-1 taking into account two subsequent units from the source text and three current units from the target text (English). Then it selects the pattern with the best similarity score and continues from the next available units in the source and target texts.

4 Experimental Results

The results of an experiment for 50 patterns of paragraphs of the text *Dracula* are presented in Table 2. The precision of the method in this experiment was 94%. Here we dealt with non-literal translations of the fiction texts.

An example of an error of our method is presented in Table 3. The paragraph is rather small that makes the correct solution less probable. Usually, the larger is the unit, the more probable it is to obtain correct similarity. It is so because it is more probable to find a word and its translation.

The paragraph in Table 3 would present difficulties for statistical methods as well, because the size of the English unit is 21 words, while the size of the Spanish one is only 11 words. There are only three words that have translations given in the dictionary or direct

Table 2. Alignment results for 50 patterns of paragraphs

Patterns found	Correct	Incorrect
1–1	27	0
1–2	8	2
1–3	6	0
2–1	7	0
3–1	2	1

Table 3. Alignment results for 50 patterns of paragraphs

English text	Spanish text	Literal translation
"Suppose that there should turn out to be no such person as Dr. Fergusson?" exclaimed another voice, with a malicious twang.	*>Y si el doctor Fergusson no existiera? –preguntó una voz maliciosa.*	*"And if Doctor Fergusson does not exist?" asked a malicious voice.*

string matching: *voice*, *malicious*, *Fergusson*. There are 9 meaningful words in the English paragraph and 6 in the Spanish paragraph. The words that are presented as translations but do not have correspondence in the dictionary are: *ask* and *exclaim*. In order to detect that it is something related, we should use, for example, some additional dictionary with marked-up hyponymic relations, such as WordNet. The other words are really different, though they carry the same meaning: *turn out to be no such person* = *not exist*. A grammatical way of expression of conditional mode is used in Spanish, while in English a special construction appears: *suppose that there should exist* vs. Spanish *existiera*.

This is a representative case of the most difficult situation consisting in non-literal translation of a short unit. The situation is worse if there are several subsequent units of this type because the possibility of wrong alignment increases. We expect that adding more dictionary information (synonyms, hyponyms), syntactic information, and genetic algorithm as optimization technique will alleviate this problem.

5 Conclusions and Future Work

We have presented an English-Spanish parallel corpus of fiction texts of considerable size freely available for researchers and discussed a dictionary-based method of alignment as applied to fiction texts. The experiment conducted at the paragraph alignment level shows that the dictionary-based method has high precision (94%) for non-literal translations.

In the future, we plan to implement a better alignment algorithm based on genetic algorithm with global optimization. Another direction of improvement of the method is the use of other types of dictionaries with synonymic and homonymic relations, such as WordNet. Also, the method can beneficiate from weighting the distance between a word and its possible translation, especially in case of the large units, because some words can occur in a unit as translation of the other word and not the one we are looking for.

It is also interesting to analyze the influence of the special treatment of morphological homonyms. As to alignment at the word level, we plan to try some additional syntactic heuristics. As a possible application, we plan to analyze the cases of absence of translations, as discussed above.

References

1. Brown, P. F., Lai, J. C., and Mercer, R. L. 1991. Aligning Sentences in Parallel Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, pp. 169–176.
2. Chen, S. 1993. Aligning sentences in bilingual corpora using lexical information. In: *Proceeding of ACL '93*, pp. 9–16.
3. Cowie, J., J. A. Guthrie, L. Guthrie. Lexical disambiguation using simulated annealing. In *Proc. of the International Conference on Computational Linguistics*, 1992, 359–365.
4. Kit, Chunyu, Jonathan J. Webster, King Kui Sin, Haihua Pan, Heng Li. 2004. Clause alignment for Hong Kong legal texts: A lexical-based approach. *International Journal of Corpus Linguistics* 9:1. pp. 29–51.
5. Gale, W. A., and Church, K. W. 1991. A program for Aligning Sentences in Bilingual Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California.
6. Gelbukh, Alexander, and Grigori Sidorov. 2003. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. *Lecture Notes in Computer Science*, N 2588, Springer-Verlag, pp. 215–220.
7. Gelbukh, Alexander, Grigori Sidorov, SangYong Han. 2005. On Some Optimization Heuristics for Lesk-Like WSD Algorithms. *Lecture Notes in Computer Science*, Vol. 3513, Springer-Verlag, pp. 402–405.
8. McEnery, A. M., and Oakes, M. P. 1996. Sentence and word alignment in the CRATER project. In: J. Thomas and M. Short (eds), *Using Corpora for Language Research*, London, pp. 211–231.
9. Mikhailov, M. 2001. Two Approaches to Automated Text Aligning of Parallel Fiction Texts. *Across Languages and Cultures*, 2:1, pp. 87–96.
10. Kay, Martin and Martin Roscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.
11. Langlais, Ph., M. Simard, J. Veronis. 1998. Methods and practical issues in evaluation alignment techniques. In: *Proceeding of Coling-ACL-98.*
12. Li, Wei, Maosong Sun. Automatic Image Annotation based on WordNet and Hierarchical Ensembles. In: A. Gelbukh (Ed.) Computational Linguistics and Intelligent Text Processing. Proc. of CICLing 2006. *Lecture Notes in Computer Science* N 3878, Springer, pp. 551–563.
13. Meyers, Adam, Michiko Kosaka, and Ralph Grishman. 1998. A Multilingual Procedure for Dictionary-Based Sentence Alignment. In: *Proceedings of AMTA '98: Machine Translation and the Information Soup*, pages 187–198.
14. Velásquez, F., Gelbukh, A. and Sidorov, G. 2002. AGME: un sistema de análisis y generación de la morfología del español. In: Proc. *Of Workshop Multilingual information access and natural language processing of IBERAMIA 2002 (8th Iberoamerican conference on Artificial Intelligence)*, Sevilla, España, November, 12, pp. 1–6.
15. Villaseñor Pineda, L., J. A. Massé Márquez, L. A. Pineda Cortés. Towards a Multimodal Dialogue Coding Scheme. In: A. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing*. Proc. of CICLing 2000. IPN, Mexico, pp. 551–563.