A Bilingual Corpus of Novels Aligned at Paragraph Level*

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
www.Gelbukh.com

Abstract. The paper presents a bilingual English-Spanish parallel corpus aligned at the paragraph level. The corpus consists of twelve large novels found in Internet and converted into text format with manual correction of formatting problems and errors. We used a dictionary-based algorithm for automatic alignment of the corpus. Evaluation of the results of alignment is given. There are very few available resources as far as parallel fiction texts are concerned, while they are non-trivial case of alignment of a considerable size. Usually, approaches for automatic alignment that are based on linguistic data are applied for texts in the restricted areas, like laws, manuals, etc. It is not obvious that these methods are applicable for fiction texts because these texts have much more cases of non-literal translation than the texts in the restricted areas. We show that the results of alignment for fiction texts using dictionary based method are good, namely, produce state of art precision value.

1 Introduction

There are many sources of linguistic data. Nowadays, Internet is one of the most important sources of texts of various kinds that are used for investigations in the field of computational linguistics. Also, advances of corpus linguistics give more and more possibilities of accessing of various types of corpora – raw texts as well as texts marked with certain additional linguistic information: phonetic, morphological, syntactic, information about word senses, semantic roles, etc. One of the important types of the linguistic information is the "relative" information that is not specific to the text itself, but is related to some other text or pragmatic situation, in contrast with the "absolute" information specific to the text itself.

One of the clear examples of the relative information is the case of parallel texts, i.e., the texts that are translations of each other, or, maybe, translations of some other text. Sometimes it is interesting to compare two different translation of the same text. The relative information is represented by the relation between different structural

* The work was done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (CGPI, COFAA, PIFI).

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 16 – 23, 2006.
© Springer-Verlag Berlin Heidelberg 2006

parts (units) of these texts. The procedure of establishing these relations is called alignment, and the resulting parallel corpus is called aligned. Obviously, there are various levels of alignment: text, i.e., we just know that the texts are parallel (it may be useful in case of very short texts, like news messages or paper abstracts, for example); paragraphs; sentences; words and phraseological units.

It is important to emphasize that each unit in a parallel text can have one, several or zero correspondences in the other text, for example, one sentence can be translated with various, some words can be omitted, etc. Thus, the alignment of parallel texts is not a trivial task. This situation is especially frequent in fiction texts that we discuss in the present paper.

One of the most accessible sources of parallel texts is Internet. Unfortunately, the texts presented in Internet are very "dirty", i.e., they may have pictures, special formatting, special HTML symbols, etc. Often, the texts are in the PDF format and during their conversion into the plain text format the information about the ends of paragraphs is lost. Thus, rather extended preprocessing, sometimes inevitably manual, is necessary.

The importance of the aligned parallel corpora is related with the fact that there are structural differences between languages. These differences can be exploited for automatic extraction of various linguistic phenomena. The other obvious application of these corpora is machine translation [1], especially, machine translation based on examples. Another application is automatic extraction of data for machine learning methods. Also, these resources are useful in bilingual lexicography [7], [11]. Another natural application is language teaching. The famous example of the application of parallel texts is deciphering of Egyptian hieroglyphs based on the parallel texts of Rosetta stone.

Generally speaking, there are two very large classes of methods in computational linguistics. One class is based on statistical data, while the other one applies additional linguistic knowledge. Note that the basis of this distinction is related with the kind of data being processed independently of the methods of processing. This is typical situation, for example, the same happens in word sense disambiguation [6].

As far as alignment methods are concerned, the classic statistical methods exploit the expected correlation of length of text units (paragraphs or sentences) in different languages [4], [8] and try to establish the correspondence between the units of the expected size. The size can be measured in number of words or characters.

On the other hand, the linguistic methods, one of which was used for obtaining the results presented in this paper, use linguistic data (usually, dictionaries) for establishing the correspondence between structural units. Application of dictionary data for text alignment was used, for example, in [2], [9], [10], [11]. Among more recent works, let us also mention the paper [3]. The experiments described in this paper were conducted using texts of laws. This is typical for parallel texts because there are many translations of specialized texts, like technical manuals, parliament debates (European or Canadian), law texts, etc. Still, fiction texts are different from these types of technical texts because translation of fiction is much less literal than translation of specialized documents.

The motivation of our paper is presentation of a bilingual parallel corpus of novels and evaluation of how a dictionary-based method performs for fiction texts.

2 Corpus Description

We present the English-Spanish parallel corpus of fiction texts aligned at the paragraph level. The first step is preparation of a corpus, i.e., compilation and preprocessing of the parallel texts. In our case, we chose novels because it is one of the most non-trivial cases of translation of the data of considerable size, for example, advertising is even more complicated, but the corresponding texts are very short. The titles that we included in our corpus are presented in Table 1, as well as the number of paragraphs of each text.

Table 1. Texts included in the corpus with correspoding number of paragraphs

Author	English title	Par.	Spanish title	Par.
Carroll, Lewis	*Alice's adventures in wonderland*	905	*Alicia en el país de las maravillas*	1,148
Carroll, Lewis	*Through the looking-glass*	1,190	*Alicia a través del espejo*	1,230
Conan Doyle, Arthur	*The adventures of Sherlock Holmes*	2,260	*Las aventuras de Sherlock Holmes*	2,550
James, Henry	*The turn of the screw*	820	*Otra vuelta de tuerca*	1,141
Kipling, Rudyard	*The jungle book*	1,219	*El libro de la selva*	1,428
Shelley, Mary	*Frankenstein*	787	*Frankenstein*	835
Stoker, Bram	*Dracula*	2,276	*Drácula*	2,430
Ubídia, Abdón	*Advances in genetics*[2]	116	*De la genética y sus logros*	109
Verne, Jules	*Five weeks in a balloon*	2,068	*Cinco semanas en globo*	2,860
Verne, Jules	*From the earth to the moon*	894	*De la tierra a la luna*	1,235
Verne, Jules	*Michael Strogoff*	2464	*Miguel Strogoff*	3,059
Verne, Jules	*Twenty thousand leagues under the sea*[3]	3,702	*Veinte mil leguas de viaje submarino*	3,515

[2] This is a fiction text, not a scientific text.
[3] There are two English translations of this novel available.

The texts had originally the PDF format. They were converted into plain text and preprocessed manually. Special formatting elements were eliminated and the paragraph structure was restored. The total size of corpus is more than 11.5 MB. The corpus size might seem too small, but let us remind that it is a parallel corpus, where the data is not so easy to obtain. The corpus is freely available on request for research purposes.

Table 2. Some corpus parameters

Corpus parameter	English part	Spanish part
Number of words	848,040	844,156
Tokens (wordforms)	25,877	43,176
Paragraphs	18,701	21,540
Most frequent words	52,597 (*the*) 25,159 (*and*) 25,147 (*of*) 22,041 (*to*) 18,225 (*I*) 17,280 (*a*) 13,473 (*in*)...	43,451 (*de*) 28,714 (*que*) 26,768 (*la*) 24,498 (*y*) 21,871 (*el*) 20,043 (*a*) 18,182 (*en*)...

The difference between numbers of tokens is explained by the presence of morphological variants in Spanish as compared with English.

3 Method Used for Corpus Alignment

Let us remind that the correspondence of paragraphs in source and target texts is not necessarily one-to-one. One of such examples is presented in Table 3.

Table 3. Example of alignment of paragraphs with pattern "2-1"

... Antes de que yo dijese una palabra, María se apresuró a decirme, azorada: *-Es mi madre.*	*Before I could say a word, Maria, disturbed, said hastily, "It's my mother."*
(Lit.: *...Before I could say a word, Maria hurried to say hastily:* "*It is my mother.*")	

This often happens in English-Spanish text pairs, when the direct speech constitutes a separate paragraph in Spanish, while it is part of the previous paragraph in English.

According to the used method, the texts are compared using bilingual dictionaries. Dictionaries have their entries as normalized words. So, it is necessary to implement morphological normalization of tokens (wordforms) converting them into types (lemmas). We performed normalization of Spanish texts using our morphological analyzer AGME [12]. The number of entries of the morphological dictionary is about 26,000 that is equivalent to more than 1,000,000 wordforms.

We have similar morphological analyzer [5] for English language. It is based on WordNet dictionary. The English morphological dictionary contains about 60,000 entries.

Another necessary feature in text alignment is filtering of the auxiliary words. These words should be ignored because their presence is arbitrary and can carry false information about paragraph matching.

Our alignment was based on Spanish-English dictionary that contains about 30,000 entries.

For the moment, we developed a heuristic algorithm that performs the alignment. The algorithm takes into account possible patterns of three paragraphs in each text, starting from the beginning of the texts, and tries to find the best match calculating the similarity for possible patterns of three paragraphs: 1 to 1, 1 to 2, 1 to 3, 2 to 1, 3 to 1. The best possible correspondence is taken. Then the algorithm proceeds to the next three available paragraphs. This algorithm implies the usage of the local optimization as in [3]. It cannot use the global optimization like in [9]. In future, we plan to try the global optimization strategies as well, for example, it is possible to apply genetic algorithm as we already did for word sense disambiguation [6] or, say, dynamic programming [4].

We also implemented the anchor point technique. It implies that we search the small paragraphs, where we are very sure of the alignment (=anchor points). It happens when the paragraphs contain dates or numbers or proper names or some metadata, like chapters. Further, the main algorithm works only between anchor points. It allows avoiding the completely wrong alignment that could be produced due to only one error influencing the rest of alignment.

For calculation of the similarity measure used in the algorithm, we used Dice coefficient with the only modification that we penalize paragraphs with too different sizes. Dice coefficient for two sets – in our case, the set of words and the set of their possible translations– is equal to the intersection (multiplied by two) of these sets, normalized by dividing to the total size of both sets. The penalization is made by multiplication to the number that is the difference of the expected correlation of lengths of sets and the actual correlation.

4 Example of Non-literal Translation

Let us consider an example of non-literal translation of paragraphs. This case is presented in Table 4. The example is taken from the text 8 of the corpus.

The English paragraph has only 85 words, while the Spanish one has 157 words (nearly double size). Note that alignment of these paragraphs is difficult for statistical methods because of the difference in sizes. Obviously, the final correct or incorrect alignment depends on the structure of the context paragraphs.

Table 4. Example of non-literal translation of a paragraph

<table>
<tr><td colspan="2">(Original) Le hice ver que no había dado importancia al asunto. Pero (y hablo de esa vez), a duras penas pude controlar la emoción que me vapuleó de arriba a abajo. "Qué puedes temer del tiempo, si tiempo es lo que más tienes", le dije desde mi interior, pensando en que su turbación se debía al súbito sufrimiento que le ocasionaba el solo pensar que, de todas maneras, en aquella foto, en aquel rostro desdibujado, estaba escrito ya el arribo inevitable, el futuro corrupto y degradado que mis ojos inquisidores (y mis teorías acerca de las herencias físicas) podían prefigurar para ella; algo como una vergüenza impuesta por un pecado aún no cometido, algo como una culpa asumida sin razón; en el fondo, la réplica infantil de una conciencia demasiado tierna. "Qué puedes temer del tiempo, chiquilla", le repetí desde mí mismo, mientras me alejaba de ella para darle lugar a recomponerse, a retomar su serenidad de siempre.</td></tr>
<tr><td>(Translation) I shrugged and smiled and nodded. But I could barely control the emotion that shook me from head to toe. I figured that she must have been upset by the thought that in that blurry face in the photo was written the inevitable, corrupted, degraded future of her own old age. "What do you have to fear from time, little girl, if you have so much of it?" I wondered, as I withdrew to give her space to compose herself and regain her customary serenity.</td><td>(Literal translation) I made her see that I did not give importance to the situation. But (and I speak about this time) I could barely control the emotion that whipped me from head to toe. “Why should you be afraid of time, if you have a lot of time?” I told her in my inside, thinking that her abashment is due to the abrupt anguish that was caused by the mere idea that, anyway, at the photograph, in that blurry face, there was written something inevitable, the corrupted and degraded future, which my inquisitional eyes (and my theories about physical inheritance) could foresee for her. Something like a shame of a sin not committed yet or a fault assumed without any reason, deep down, the infantile copy of a too immature conscience. “Why are you afraid of time, little girl” I repeated inside, while I was withdrawing to allow her to compose herself and regain her customary serenity.</td></tr>
</table>

If we calculate the words that these paragraphs have in common according to their translations in dictionaries, then we will get the value that usually would not appear in relatively big paragraphs, namely, 20 words, i.e., 23% for English and 12% for Spanish. Still, this value keeps being rather solid for their alignment using the dictionary-based method.

5 Alignment Evaluation

We made the experiment for 50 patterns of paragraphs of the text *Dracula*. The results presented in Table 5 were obtained.

Note that we deal with translations of the fiction texts that are not literate; still the precision of the method in this experiment is 94%. The result is the state of art value that shows that the dictionary-based method can be applied to the alignment of fiction texts.

Table 5. 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

The errors of the alignment methods based on dictionaries happened for paragraphs that have small sizes, because they do not have enough significant words for using them in alignment. Note that from the point of view of statistical methods, the small-size paragraphs are also unreliable.

For a dictionary-based method, a possible solution can be the following: if there are very few o none significant words, some kinds of auxiliary words that have reasonable translations (say, prepositions) can be used in comparison, i.e., treated like significant words.

We expect that adding more dictionary information (synonyms, hyponyms), syntactic information will allow improvements in resolving this problem.

6 Conclusions and Future Work

The paper describes the English-Spanish parallel corpus of novels of a considerable size. Also, we present the evaluation of the performance of the dictionary-based method of alignment at the paragraphs level applied to this corpus. Note that the corpus contains fiction texts that usually do not have very literal translation. The experiment conducted on a small sample and verified manually shows that the dictionary based method has high precision (94%) for this type of non-literal translations. So, we expect that this kind of methods is applicable for fiction texts.

The corpus is freely available for research purposes.

In future, we plan to implement better algorithm of alignment instead of the described heuristic-based algorithm. For example, we plan to use genetic algorithm with global optimization and dynamic programming.

Another direction of improvement of the method is usage of other types of the dictionaries with synonymic and homonymic relations, like WordNet. Also, the method can beneficiate from weighting of the distance between a word and its

possible translation, especially in case of the large paragraphs, because some words can occur in a paragraph as a translation of the other word, and not the one that we are searching.

References

[1] Brown, P. F., Lai, J. C. & 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] 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.

[4] Gale, W. A. & 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.

[5] 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.

[6] Gelbukh, Alexander, Grigori Sidorov, SangYong Han. 2005. On Some Optimization Heuristics for Lesk-Like WSD Algorithms. *Lecture Notes in Computer Science, N 3513*, Springer-Verlag, pp. 402–405.

[7] McEnery, A. M. & Oakes, M. P. 1996. Sentence and word alignment in the CRATER project. In: J. Thomas & M. Short (eds), *Using Corpora for Language Research*, London, pp. 211 – 231.

[8] Mikhailov, M. 2001. Two Approaches to Automated Text Aligning of Parallel Fiction Texts. *Across Languages and Cultures, 2:1*, pp. 87 – 96.

[9] Kay, Martin and Martin Roscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121-142.

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

[11] 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.

[12] Velásquez, F., Gelbukh, A. & Sidorov, G. 2002. AGME: un sistema de análisis y generación de la morfología del español. In: *Proc. Of Workshop Multilingual information access & natural language processing of IBERAMIA 2002 (8th Iberoamerican conference on Artificial Intelligence)*, Sevilla, España, November, 12, pp 1-6.