Building a Spanish-Portuguese Parallel Corpus for Statistical Machine Translation

Wilker F. Aziz, Thiago A. S. Pardo
NILC, ICMC, Universidade de São Paulo
Av.Trabalhador São-Carlense, 400
São Carlos, Brazil
wilker.aziz@usp.br, taspardo@icmc.usp.br

Ivandré Paraboni
EACH, Universidade de São Paulo
Av. Arlindo Bettio, 1000
São Paulo, Brazil
ivandre@usp.br

ABSTRACT
Parallel corpora have long been recognised as valuable resources for building MT applications, but their usefulness have often been limited to the translation between language pairs that include English. In this work we describe our efforts to build a parallel corpus for the Brazilian Portuguese and European Spanish languages. The corpus has been aligned at sentence and word levels and manually inspected for correctness, representing a first step towards the development of translation models for this language pair.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]: *machine translation*.

General Terms
Statistical Machine Translation.

Keywords
Parallel corpus, statistical machine translation.

1. INTRODUCTION
Machine Translation (MT) systems and other text-generating applications often make use of large amounts of translation parallel corpora – collections of text translated into two or more languages – for training purposes. Parallel corpora are most useful when texts are *aligned* at some level, i.e., when each text element in one language (e.g., paragraphs, sentences or words) is mapped onto its translation in a second language. While resources of this kind have been built on a large scale for language pairs involving English, some of the world's most widely-spoken languages still lack behind. This may be especially true of so-called 'closely related' language pairs [2] such as Romance languages, whose translation from one to another has been shown to be nevertheless nontrivial.

In this paper we describe our efforts to build an aligned parallel corpus for the Brazilian Portuguese and Spanish languages. The present work is part of an underlying MT project intended to implement and evaluate a number of statistical translation models for Romance languages.

2. DATA COLLECTION AND PRE-PROCESSING
We collected 645 Portuguese-Spanish text pairs from the Environment, Science, Humanities, Politics and Technology supplements of the on-line edition of the "Revista Pesquisa FAPESP"[1], a Brazilian journal on scientific news. The corpus consists of 17,681 sentence pairs comprising 908,533 words in total (being 65,050 distinct). The Portuguese version consists of 430,383 words (being 32,324 distinct) and the Spanish version consists of 478,150 words (being 32,726 distinct).

Although Portuguese and Spanish may be regarded as 'closely related' languages [2], we are aware that our data set is considerably smaller than standard training data used in statistical MT, for example. This limitation not withstanding, we decided to leave the issue of whether to expand the corpus to a later stage of our investigation, when we expect to have gained a better understanding of what the translation between these two languages actually entails.

Portuguese text segmentation was performed using SENTER [4], a simple rule-based segmentation tool. The tool was also employed in the segmentation of the Spanish texts with a number of changes to handle Spanish abbreviations. Despite the similarities between the two languages, it is immediate to observe that word-to-word translation is not feasible: besides the differences in word order, there are subtle changes in meaning (e.g., "espantosa" vs. "impresionante", analogous to "amazing" vs. "impressive"), additional words (e.g. "ubicada") and others, as illustrated in Table 1.

Table 1. Portuguese and Spanish text fragments

Portuguese	Spanish
Ao desencadear uma cascata de eventos físico-químicos poucos quilômetros acima da floresta, a espantosa concentração de aerossóis na Amazônia no auge da estação (...)	Esa impresionante concentración de aerosoles en la Amazonia, al desencadenar una cascada de eventos fisicoquímicos ubicada a algunos kilómetros arriba del bosque, en el auge de la estación (...)

[1] http://www.revistapesquisa.fapesp.br/

3. SENTENCE ALIGNMENT

A *sentence alignment* a is taken to be an ordered set of p(a) sentences in our Portuguese corpus and an ordered set of s(a) related sentences in the Spanish corpus. Values of p(a) and s(a) can vary from zero to an arbitrary large number. For example, a Portuguese sentence may correspond to exactly one sentence in the Spanish translation, and such 1-to-1 relation is called a *replacement* alignment. If a Portuguese sentence is simply omitted from the Spanish translation then we have a 1-to-0 alignment or *deletion*. In this work we focus on replacement alignments only, investigating how n-to-m alignments become 1-to-1 without compromising the quality of the alignment. By doing so we expect not only to reduce the computational complexity of our next task – word alignment – but also to provide the required input format for MT tools such as GIZA++ [3].

For the sentence alignment task, we used an implementation of the *Translation Corpus Aligner* (TCA) method called TCAalign [1]. The choice was based on the high precision rates reported for Portuguese-English (97.10%) and Portuguese-Spanish (93.01%) language pairs. The set of alignments produced by TCAalign consists of m-to-n relations marked with XML tags. As our goal is to produce an aligned corpus as accurate as possible, the data were inspected semi-automatically for potential misalignments, which were in turn collapsed as follows:

1. An alignment a is *incorrect* if it is not a replacement, i.e., if p(a) <> s(a);
2. An alignment a is *correct iff* it is of the replacement type and if its m surrounding alignments (i.e., above and below a) are replacements as well, in which m is the number of sentences in each version of the text; for example, a replacement r containing three sentences is considered to be correct *iff* the three alignments previous to r and the three alignments following r are all replacements.
3. All other replacements are considered *unsafe* and marked for manual revision.

Following the above, 10% of the alignments were classified as unsafe, and their manual inspection revealed that 1,668 instances (9.43%) were indeed incorrect. A large number of misalignment were due to segmentation errors, which caused two or more sentences to be regarded as a single unit:

Table 2. Misalignment due to incorrect segmentation

Portuguese	Spanish
<s snum=14> "Partículas provenientes da Amazônia já foram encontradas nos Andes e em São Paulo." Isso não quer dizer que, em razão do resfriamento e da estiagem associados à ação dos aerossóis, a venda de malhas tenha disparado. </s>	**<s snum=13>** "Las partículas provenientes de la Amazonia han sido encontradas en los Andes y en São Paulo."</s> **<s snum=14>** Esto no quiere decir que, en razón del enfriamiento y la sequedad asociados a la acción de los aerosoles, la venta de ropa se haya disparado.</s>

These were adjusted manually so that the resulting corpus contained a set of Portuguese sentences and their Spanish counterparts in 1-to-1 relationships. Cases in which *n* Portuguese sentences were (correctly) aligned to *n* Spanish sentences in a different order were split into individual 1-to-1 alignments.

Example 3. A 2-to-2 alignment to be split into 2 1-to-1

Portuguese	Spanish
<s snum=28> Até agora, não há registro de nenhum grande acidente na retirada ou transporte do óleo na região do Urucu.**</s>** **<s snum=28>** Embora pequenos vazamentos de petróleo já afetem, de modo ainda pouco conhecido, a diversidade de peixes de ecossistemas tropicais.**</s>**	**<s snum=28>** Pequeños derrames afectan, aunque de una manera aún poco conocida, la diversidad de peces en los ecosistemas tropicales.**</s>** **<s snum=28>** Pero hasta ahora no existen registros de ningún accidente de magnitud en el retiro o el transporte de petróleo en la región del Urucú.**</s>**

Other kinds of misalignment were introduced by the alignment tool itself:

Table 4. A 1-to-2 alignment followed by a 2-to-1 alignment

Portuguese	Spanish
<s snum=3> Após analisar amostras do parasita vindas da África, do Oriente Médio e da América do Sul, pesquisadores da Universidade de São Paulo (USP) encontraram mutações genéticas que causam a troca de um único nucleotídeo - molécula formada por uma das quatro bases nitrogenadas que formam o DNA, A (adenina), T (timina), C (citosina) e G (guanina) - e alteram as proteínas indicadas pela OMS como alvos para a criação de vacinas.**</s>** **<s snum=4>** Em conseqüência, variações nos genes dessas proteínas poderiam diminuir o efeito da vacina.**</s>** **<s snum=4>**"Para realmente funcionar, uma vacina deveria conter todas as variações encontradas nessas proteínas", diz Emmanuel Dias Neto, o coordenador do estudo, feito em parceria com Sérgio Verjovski-Almeida.**</s>**	**<s snum=3>** Tras analizar muestras del parásito provenientes de África, Oriente Medio y América del Sur, investigadores de la Universidad de São Paulo (USP) hallaron mutaciones genéticas que causan el cambio de un solo nucleótido - una molécula formada por una de las cuatro bases nitrogenadas que forman el ADN: A (adenina), T (timina), C (citosina) y G (guanina)- y alteran las proteínas indicadas por la OMS como blancos para la creación de vacunas.**</s>** **<s snum=3>** Como consecuencia de ello, las variaciones en los genes de esas proteínas podrían reducir el efecto de la vacuna.**</s>** **<s snum=4>**"Para funcionar realmente, una vacuna debería contener todas las variaciones halladas en esas proteínas", dice Emmanuel Dias Neto, coordinador del estudio, realizado en conjunto con Sérgio Verjovski-Almeida.**</s>**

Finally, different choices in translation may lead to correct n-to-m alignments as in the following Table 5:

Table 5. A 2-to-1 alignment

Portuguese	Spanish
<s snum=11> Camargo disse não.**</s>** **<s snum=11>** Preferiu voltar, decidido a criar um centro de proteínas no Brasil.**</s>**	**<s snum=11>** Pero Camargo dijo que no, prefirió volver, decidido a crear un centro de proteínas en Brasil.</**s>**

Since punctuation will be removed in the generation of our translation models, in cases as the above it was possible - when there was no change in meaning - to manually split the Spanish sentence and create two individual 1-to-1 alignments as in the following Table 6:

Table 6. A manually created 1-to-1 alignment

Portuguese	Spanish
<s snum=11>Camargo disse não.**</s>** **<s snum=12>**Preferiu voltar, decidido a criar um centro de proteínas no Brasil.**</s>**	**<s snum=11>**Pero Camargo dijo que no.</s> **<s snum=12>**Prefirió volver, decidido a crear un centro de proteínas en Brasil.</**s>**

4. WORD ALIGNMENT

Two versions of the corpus have been produced: one represents the aligned corpus in its original format, with capital letters, punctuation marks and alignment tags; the other represents the aligned corpus in GIZA++ [3] format, in which the entire text was converted to lower case, punctuation marks and tags were removed, and the correspondence between sentences is given simply by their relative position within each text file.

Using GIZA++, the second version was aligned at word level to produce a basic translation model (namely, model 4 in GIZA++) of the Portuguese-Spanish language pair. The tool produced 489,594 word alignments, about 82% of which were of the word-to-word type. Moreover, very few mappings (701 cases or 0.14% in total) involved more than three words in the target language (i.e., alignments 1-4 to 1-9.) These results may suggest a strong similarity between Portuguese and Spanish, as argued in [2], although we presently do not seek to prove this claim.

5. FINAL REMARKS

We have described the preliminary stages of development of a Portuguese-Spanish parallel corpus for statistical machine translation purposes. The corpus has been automatically aligned at sentence and word levels and semi-automatically revised at sentence level for correctness.

We are now in the process of building additional resources as required for the translation task proper, including the evaluation of the existing lexical alignment and translation model, which will be used as part of an MT system under development.

A first experiment using the corpus described in this paper to translate from Portuguese to Spanish (and vice-versa) is described in [5], in which the results of a statistical MT system for this language pair are compared to those obtained by the rule-based MT approach in [2].

6. ACKNOWLEDGMENTS

This work has been supported by FAPESP and CNPq / Brazil.

7. REFERENCES

[1] Caseli, H. M. 2007. Indução de léxicos bilíngües e regras para a tradução automática. Doctoral thesis, University of São Paulo.

[2] Corbí-Bellot, et. al. 2005. An open-source shallow-transfer machine translation engine for the romance languages of Spain. 10th Annual Conference of the European Association for Machine Translation, pp. 79-86.

[3] Och, F.J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, vol. 29, n.1, pp. 19-51.

[4] Pardo, T. A. S. SENTER: Um Segmentador Sentencial Automático para o Português do Brasil. 2006. NILC Technical Reports Series NILC-TR-06-01. University of São Paulo.

[5] Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo and Ivandré Paraboni. 2008. An Experiment in Portuguese-Spanish Statistical Machine Translation. 19th Brazilian Symposium on Artificial Intelligence (SBIA-2008) Salvador. Springer LNAI vol. 5249, pp. 248-257.