Fully Automatic Compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora

Wilker Aziz[1] **, Lucia Specia**[1]

[1] Research Group in Computational Linguistics, University of Wolverhampton, UK

{W.Aziz,L.Specia}@wlv.ac.uk

Abstract. *This paper reports the fully automatic compilation of parallel corpora for Brazilian Portuguese. Scientific news texts available in Brazilian Portuguese, English and Spanish are automatically crawled from a multilingual Brazilian magazine. The texts are then automatically aligned at document- and sentence-level. The resulting corpora contain about 2,700 parallel documents totaling over 150,000 aligned sentences each. The quality of the corpora and their usefulness are tested in an experiment with machine translation.*

1. Introduction

Parallel corpora are collections of texts that are mutual translations. Machine Translation (MT) systems have recently become very popular due to the latest advances in the field of Statistical Machine Translation (SMT). The most advanced SMT approaches, such as phrase-based [Koehn et al. 2003] and hierarchical SMT [Chiang 2005], learn how to perform translation from parallel corpora.

A popular way of building a parallel corpus is collecting large amounts of data from the web. The alignment between documents, sentences and words allows the development of resources such as dictionaries and SMT systems.

[Esplà-Gomis and Forcada 2010] present *Bitextor*, a language-independent tool to harvest parallel documents from multilingual web sites. It uses features from the HTML structure of the documents and their URLs to perform document alignment. [Aziz et al. 2008] describe the creation of a Brazilian Portuguese-Spanish (`pt-es`) parallel corpus of scientific texts. The corpus is manually crawled from the Brazilian magazine *Pesquisa FAPESP Online*[1] and the document alignment is manually performed, followed by automatic sentence alignment using the Translation Corpus Aligner (TCA) method from [Hofland 1996]. The same magazine and methodology were used to create a Brazilian Portuguese-English (`pt-en`) corpus. Both `pt-es` and `pt-en` corpora are freely available[2] and will be referred to as `Fapesp-v1`. Despite small ($\sim$ 18K sentences, the `Fapesp-v1` corpora have been used in interesting applications: rule-based MT [Caseli et al. 2006], SMT [Aziz et al. 2009] and coreference resolution [Souza and Orasan 2011].

In this paper we present a sentence-aligned parallel corpora (`pt-en` and `pt-es`) using all the currently available issues of the magazine *Pesquisa FAPESP Online*. Unlike in [Aziz et al. 2008], our compilation is fully automatic from the data crawling to the sentence alignment. Unlike *Bitextor*, data crawling is fit to our type of text, resulting in very good recall and precision at document-level alignment with surface-based features.

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

[2] `www.nilc.icmc.usp.br/lacioweb/english/corpora.htm`

2. Methodology

The compilation of the corpora is divided in three steps: data crawling, document alignment and sentence alignment, as described in the following sections.

2.1. Data Crawling

The source for our corpora is the Brazilian scientific magazine *Pesquisa FAPESP Online*, which contains parallel electronic documents in three languages: Brazilian Portuguese (original release), English and Spanish (human translations). The magazine currently contains approximately 2,700 parallel articles and it is monthly updated.

We use *GNU wget*[3] and a URL template to download HTML pages from the magazine's website. For a given issue we first download its index, a single page containing links to articles, and then parse those links and download the actual articles. Issues are identified by a sequential number that is consistent across the different versions of the website (original, English and Spanish). Therefore the alignment of issues is straightforward. However, there is no obvious correspondence between articles' identifiers, and therefore content-based document alignment techniques need to be applied. The crawling took about 1 hour and resulted in $3,658$ original articles ($233,490$ sentences), $2,740$ English translations ($184,587$ sentences) and $2,750$ Spanish translations ($180,525$ sentences). Hereafter we will refer to this compilation as `Fapesp-v2`.

2.2. Document Alignment

Assuming the existence of a bilingual dictionary (take `pt-en` as example), we use a very simple alignment technique:

1. For every issue we gather its `pt` and `en` articles, lowercase and tokenize them;
2. For every `pt` document we replace the original words by their *n*-best translations according to our bilingual dictionary;
3. We represent documents as distributional profiles (DP), that is, a vector that contains a word w and its frequency f in the document: $\text{profile} : w \mapsto f$;
4. For every possible document pair we compute the cosine similarity of their DPs.
5. We align each `en` document to its best `pt` candidate if the similarity score is above a given threshold.

We compiled bilingual dictionaries using the small corpus `Fapesp-v1` running 5 iterations of IBM Model 1 [Brown et al. 1993]. IBM Model 1 estimates a word-based translation probability distribution from a sentence aligned corpus. We set the number of best translations from the dictionary to 3 and the similarity metric threshold to 0.15 after a short round of evaluation described in Section 3. The same procedure was used to align `pt-es` articles. We obtained $2,675$ `pt-en` and $2,668$ `pt-es` parallel texts.

2.3. Sentence Alignment

We use an implementation of TCA specially written for aligning Brazilian Portuguese to a foreign text [Caseli and Nunes 2003]. While the most desired alignment type is the substitution (*1-1*), i.e., one source sentence aligned to one target sentence, other alignment types occur as consequence of different translation decisions. 85% of the alignments produced

[3] `www.gnu.org/software/wget/`

are substitutions. For the remaining cases, the following heuristics were designed (based on manual inspection) to convert as many as possible cases into valid *1-1* substitutions. Figure 1 presents examples of the application of these heuristics.

H1: A deletion followed by an insertion, or vice-versa, many times result from a wrong decision of the automatic aligner, in reality the two alignment sets should be a single *1-1* alignment.

H2: Alignments of the type *1-2* or *2-1* usually result from incorrect sentence splitting or different translation decisions. For instance, a text separated by a colon in the source text is separated by a period in the target text. If two chunks of one sentence, separated by a punctuation mark, independently align to each of the two sentences in the other language, the original alignment is replaced by two *1-1* alignments. We align the chunks using the same metric used to align documents.

H1	[Uma hiptese na contramo]$_1$	[A hypothesis on the wrong side of the road]$_1$
H2	[So peixes que, como os linguados, dos quais so aparentados, tm os dois olhos e o colorido do corpo de um lado s - e assim conseguem se camuflar:]$_1$ ["Vivem enterrados na areia e os predadores no os enxergam", relata Menezes.]$_2$	[They are fish that, like flounders, to which they are related, have their two eyes and the body colored on one side only – and thus succeed in camouflaging themselves.]$_1$ ["They live buried in the sand, and their predators do not see them", Menezes reports.]$_2$

Figure 1. Examples of the application of our alignment heuristics

In our experiments with `pt-en`, for example, $\sim$ 10K sentence pairs were generated using the conversion heuristics. After the application of the heuristics, all the remaining non-substitution alignments sets (3% of the total) were removed from the corpus, since these are likely to contain inconsistencies. The resulting `Fapesp-v2` corpora contain $156,712$ `pt-en` and $152,238$ `pt-es` sentence pairs.

3. Experiments

To assess the quality of the corpus, we performed intrinsic (`pt-en`) and extrinsic evaluations (`pt-en` and `pt-es`). Starting with the document alignment strategy, we manually checked 310 pairs of `pt-en` texts. A cosine threshold of 0.15 guaranteed that 97.66% of the `en` documents were aligned to `pt` ones with 99.35% accuracy.

For the `pt-en` sentence alignment, we assessed 900 substitution (*1-1*) alignments produced by the TCA aligner and 300 resulting from the conversion heuristics. Results showed that 98% of the original substitution alignments were correct and that the heuristics H1 and H2 achieved an accuracy of 95.91% and 82%, respectively.

For both document and sentence alignment, we expect the results for the `pt-es` corpus to be similar or even more positive, given the proximity of the language pair.

3.1. Phrase-based SMT

To assess the utility of the corpora in an external task, we used them to build Phrase-based SMT (PBSMT) systems and evaluated their quality using automatic metrics. For each language, we split the corpus in one training, two developments and two test sets as shown in Table 1 (the last column points to the HTML documents in the website).

Table 1. Splitting of the `Fapesp-v2` for training and test purposes

Set	Sentences		Issues	HTML pointer
	pt-en	pt-es		
training	150,968	146,755	April, 2005 - March, 2011	19 - 945
dev	1,375	1,302	June, 2003	118
dev-test	1,608	1,601	June, 2010	934
test-a	1,314	1,201	July, 2003	119
test-b	1,447	1,379	July, 2010	935

We trained a PBSMT model using the Moses toolkit [Koehn et al. 2007] following its "baseline" settings and truecased data [Koehn et al. 2008]. For comparison purposes, we trained a model using the old version of the training set ($18,232$ sentence pairs): `pt-en Fapesp-v1` (UoW-fapesp-v1), and another using the new version: `pt-en Fapesp-v2` (UoW-fapesp-v2). In both experiments we used the `Fapesp-v2 dev` for tuning the features of the PBSMT model. The systems were tested using 3 test sets: i) `Fapesp-v1 test` (667 sentences), ii) `Fapesp-v2 test-a`, and iii) `Fapesp-v2 test-b`. We assessed the system's performances in terms of BLEU [Papineni et al. 2002] and its case-sensitive version BLEU-c. We also compared the translations against those produced by Google Translate, an off-the-shelf, out-of-domain PBSMT system. Table 2 shows that the model trained using our automatically compiled corpus significantly outperforms both Google Translate and a model trained on the previous, smaller version of the corpus, UoW-fapesp-v1, in all test sets.

Table 2. Performance on the test sets

PBSMT system	UoW-fapesp-v1		UoW-fapesp-v2		Google	
Test set	BLEU	BLEU-c	BLEU	BLEU-c	BLEU	BLEU-c
fapesp-v1	39.99	37.94	**57.09**	**54.25**	37.62	36.97
fapesp-v2-a	43.24	40.98	**57.69**	**54.87**	40.4	39.86
fapesp-v2-b	30.5	28.8	**44.69**	**42.15**	38.35	37.71

For `pt-es`, a PBSMT system trained using `Fapesp-v2` achieved BLEU scores of 70.42 (`Fapesp-v1` test), 70.26 and 71.49 (`Fapesp-v2` test-a and test-b).

4. Conclusions

We have presented an approach to automatically build parallel corpora involving Brazilian Portuguese. The corpora are aligned at the sentence level and contain over 150K `pt-en` and `pt-es` sentence pairs. This is comparable in size to the News Commentaries parallel corpora used for standard MT evaluations by the Workshop on Machine Translation[4]. We showed that the proposed strategies for document and sentence alignment achieve good performance and demonstrated the usefulness of our corpora for building SMT systems.

Both `pt-en` and `pt-es` corpora are available[5], along with the tools for the complete pipeline of corpus creation. We also intend to keep updating the corpora as the magazine releases additional issues. We expect these resources will be useful for people working with multilingual applications that require parallel corpora involving Brazilian Portuguese.

[4] `http://www.statmt.org/wmt11/`

[5] `http://pers-www.wlv.ac.uk/~in1676/resources`

References

Aziz, W., Pardo, T., and Paraboni, I. (2008). Building a Spanish-Portuguese parallel corpus for statistical machine translation. In *VI Workshop on Information and Human Language Technology*, Vila Velha, Brazil.

Aziz, W., Pardo, T., and Paraboni, I. (2009). Statistical phrase-based machine translation: Experiments with Brazilian Portuguese. In *XXIX CSBC VII Encontro Nacional de Inteligência Artificial*, pages 769–778, Bento Gonçalves, Brazil.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computacional Linguistics*, 19:263–311.

Caseli, H. M. and Nunes, M. G. V. (2003). Sentence alignment of Brazilian Portuguese and English parallel texts. In *Proceedings of the Argentine Symposium on Artificial Intelligence (ASAI)*, pages 1–11.

Caseli, H. M., Nunes, M. G. V., and Forcada, M. L. (2006). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20:227–245.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270.

Esplà-Gomis, M. and Forcada, M. L. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. In *Research in Humanities Computing*, pages 165–178. Oxford University Press.

Koehn, P., Arun, A., and Hoang, H. (2008). Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio.

Koehn, P., Hoang, H., Birch, A., Burch, C. C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54, Edmonton, Canada.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Souza, J. G. C. and Orasan, C. (2011). Coreference resolution for portuguese using parallel corpora word alignment. In *III Knowledge Engineering: Principles and Techniques Conference*, Cluj-Napoca, Romania.