BABYLON Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages

Michael Mohler, Rada Mihalcea

Department of Computer Science
University of North Texas
mgm0038@unt.edu, rada@cs.unt.edu

Abstract

This paper describes BABYLON, a system that attempts to overcome the shortage of parallel texts in low-density languages by supplementing existing parallel texts with texts gathered automatically from the Web. In addition to the identification of entire Web pages, we also propose a new feature specifically designed to find parallel text chunks within a single document. Experiments carried out on the Quechua-Spanish language pair show that the system is successful in automatically identifying a significant amount of parallel texts on the Web. Evaluations of a machine translation system trained on this corpus indicate that the Web-gathered parallel texts can supplement manually compiled parallel texts and perform significantly better than the manually compiled texts when tested on other Web-gathered data.

1. Introduction

The majority of the world's languages are poorly represented in informational media like radio, television, newspapers, and the Internet. Native speakers of these languages (called low-density languages) are ill-equipped to access and utilize the informational resources of our increasingly technology- and information-driven world. Translation into and out of these languages may offer a way for speakers of these languages to interact with the wider world. For instance, automatically translating the content of a news site into low-density languages like Quechua or Nahuatl could allow speakers of those languages to access world news in a way that is not dependent upon local monolingual news agencies which are often run by local governments, are focused on a narrow range of interests, or simply have insufficient resources to cover news from the wider world.

Parallel texts are crucial resources for automatic cross-language communication tools. [1] Current statistical machine translation models are only effective when large corpora of high-quality sentence-aligned parallel texts are available. Other language processing tasks such as cross-language information retrieval (CLIR), electronic resource translation, and annotation projection across parallel texts also depend upon the availability of parallel corpora.

Most of the work to date in the area of gathering parallel texts has focused on high-density language pairs, i.e. English-French (Resnik and Smith, 2003; Chen and Nie, 2000), English-Spanish (Resnik, 1998; Tomás et al., 2005), English-German (Ma and Liberman, 1999) or English-Chinese (Chen, Chau and Yeh, 2004; Resnik and Smith, 2003; Chen and Nie, 2000; Shi et al., 2006). These techniques typically assume the existence of a large number of Web sites containing parallel texts, which is a fair assumption for major languages like English, Spanish or Chinese, but it is not applicable to less widely spoken languages such as Quechua or Nahuatl. In our work, we attempt to harvest the best of these techniques and apply them to the more difficult task of finding parallel texts for languages with scarce resources. We also supplement these techniques with additional features specifically designed for low-density languages. We are interested specifically in discovering how much parallel text exists on the Web for low-density languages, and the extent to which the text found is useful in a variety of natural language processing tasks – especially machine translation.

2. Related Work

Systems that attempt to gather parallel texts from the Web can typically be reduced to two steps. First, a set of pages must be found that are likely to have some parallel content. Then, the parallel content must be discovered among the pages found, extracted, and aligned.

Several of the highest quality parallel text discovery systems find initial sets of potentially parallel pages using existing search engines by querying for language-specific content (Tomás et al., 2005) or for markers that indicate the presence of multilingual content (Resnik, 1998; Resnik and Smith, 2003; Chen and Nie, 2000). For example, the text *Spanish version* found on a Web page indicates that a translation is likely to be found on a linked page. Other systems (Ma and Liberman, 1999; Tomás et al., 2005) analyze all Web sites in a top level domain (e.g. the .de domain).

Most parallel text discovery systems make use of the fact that administrators of multilingual Web sites often name translated pages similarly with language-identifying markers in the URL. In most cases, this is accomplished by first building a set of language-pair specific tags (e.g., *en*, *_spanish*, or *big5* for Chinese) and either performing a substitution from one language to the other (Chen, Chau and Yeh, 2004; Resnik and Smith, 2003), deleting all language markers from the URL (Resnik and Smith, 2003), or adding language markers (Chen and Nie, 2000) and looking for a match.

The seminal system for finding structure-based parallelism is the STRAND system (Resnik, 1998) in which a Web page is converted into a set of tags describing the structure of a page (the HTML tags) and the length of the text between the tags. STRAND then aligns the pair of pages according to

[1] Parallel texts are documents that contain the same information in two or more languages, typically aligned on a sentence by sentence basis.

the set of tags and calculates a score based upon the alignment that indicates whether the pair are or are not parallel. Other systems concerned with discovering structure-based parallelism (Tomás et al., 2005; Shi et al., 2006; Resnik and Smith, 2003) generally follow Resnik's example.

However, in many cases (for instance, when the Web page contains text without markup), page structure is not a useful feature. In these cases, parallelism can only be found by analyzing the content of the page. The BITS system (Ma and Liberman, 1999) and the PTI system (Chen, Chau and Yeh, 2004) make use of a bilingual lexicon to perform a word-by-word substitution from one language to the other. They then analyze the similarity between the two texts to determine if the documents are parallel. The WebMining system (Tomás et al., 2005) attempts to align the documents (without a lexicon) and uses the results of each alignment to classify the documents as parallel or not parallel and to retrain the alignment system.

3. Building a Parallel Text

Our work attempts to combine existing techniques for Web-based gathering of parallel texts with a new on-page translation detection component. We apply this combination to the task of gathering parallel texts for a low-density language paired with a higher-density language. For our experiments, we have selected Quechua (spoken by around 10 million people in Peru, Bolivia, and the surrounding region) and Spanish (spoken natively by over 300 million people worldwide). However, our method was not designed with any language pair in mind and can easily be applied to other language pairs.

Figure 1 describes the overall system flow of the BABYLON Parallel Text Builder. Given a short monolingual text in Quechua as a starting point, we produce a parallel text composed of sentence-aligned Spanish-Quechua text. Using the monolingual Quechua text, we randomly select up to 1,000 words which are then used to query Google using the Perl SOAP API. The Web pages that are returned by Google are used as starting points from which to run the Web crawler. The crawler, starting at these seed URLs, searches for pages in Quechua by performing a modified breadth-first search on the underlying hyperlink graph in which links from Quechua pages are preferred to links from pages with no Quechua content.[2] The crawler is stopped after one million pages have been scanned.

The pages are then categorized in one of three ways based upon the language content of text chunks between high-level HTML tags in a document. Either they have a large proportion of Quechua text (and are labeled as *strong* pages), they have a roughly even proportion of Spanish and Quechua text in which case it is likely that an on-page translation exists (*weak* pages), or the page contains too little Quechua and is irrelevant to the search. Once a set of Quechua pages has been found, a second crawler searches for Spanish-language counterpart pages by performing a breadth-first search starting from the Quechua page and returning all Spanish pages within a few links of the original Quechua page. Hence, for each *strong* Quechua page found, there are zero or more Spanish pages which may or may not be translations of the Quechua.

Next, each Spanish-Quechua pair is passed through a set of four filters to remove pairs that are unlikely to be translations of one another. We chose to prefer high recall to high precision and so a pair must only pass one filter to be considered a candidate in the final alignment stage. The first filter accepts pairs that have a high similarity score between their two URLs based upon the edit distance. For example, the two URLs ”en.wikipedia.org” and ”de.wikipedia.org” differ by two characters and so have a high similarity score. The second filter accepts pairs whose page structure have a high similarity score based upon a modified edit-distance[3]. Finally, the third and fourth filters compare the content of the two texts (once with a lexicon and once without). If a bilingual lexicon is available, the Spanish text is converted word-by-word into Quechua and the tokens in both texts are mapped to features in a term frequency vector. If the lexicon is not available, it is still possible to find any cognates, punctuation, numerals and proper nouns that co-occur in both documents. Following the assumption that the exact co-occurrence of a token is a good indication of an alignment (Ma, 2006), all tokens are also mapped to themselves. The system then finds the Jaccard Coefficient and takes it as a similarity score for the two vectors. This score is obtained again after the lexicon and both documents are naively stemmed by truncating each token to the first four characters in hopes of normalizing terms to their morphological form found in the (now stemmed) lexicon. The results of the stemmed and non-stemmed calculations are combined to produce the final score for a pair of documents.

After the set of document pairs has been filtered, the resulting set of *strong* pages, together with the Spanish and Quechua text chunk pairs identified on the *weak* pages, are aligned using the CHAMPOLLION sentence alignment tool (Ma, 2006).[4] This tool produces an alignment which our system uses to generate a score for the pair based upon the number of one-one alignments and the one-zero alignments where one-one alignments are strongly favored and one-zero alignments are strongly disfavored. For each Quechua page, the Spanish page with the highest score (above a threshold) is selected for the final parallel text. The size of the data at each stage in the pipeline is shown in Table 1.

Stage	Size
Seed Quechua from Google (pages)	5,249
Strong Quechua (pages)	1,433
Weak Quechua (pages)	507
Strong Quechua-Spanish (page pairs)	4,927
Filtered *strong* Quechua-Spanish (page pairs)	2,720
Final alignment (page pairs)	364
Final alignment (sentences)	5,485

Table 1: Experiment size for the BABYLON stages

[2] Language identification is done using a modified version of TEXTCAT (http://lit.csci.unt.edu/˜babylon/tools/text_cat_new).

[3] The page is first converted into the STRAND tag/chunk format.

[4] http://champollion.sourceforge.net

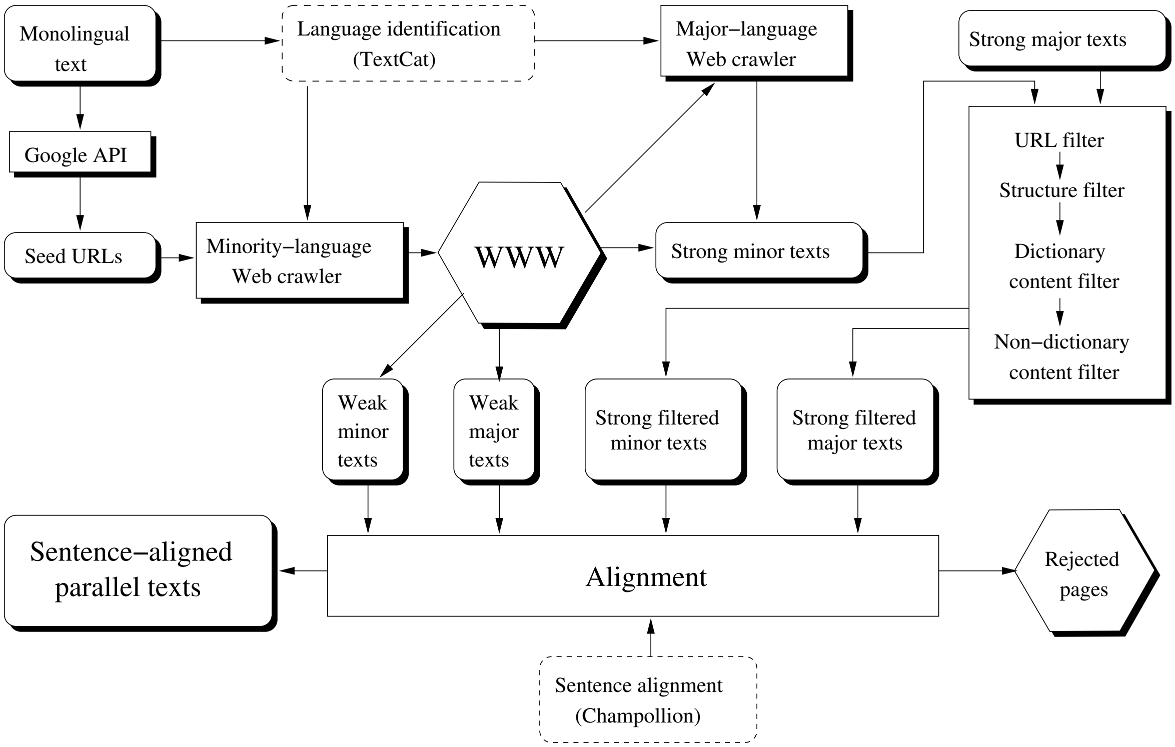
Monolingual text
Language identification (TextCat)
Major−language Web crawler
Strong major texts
Google API
URL filter
Structure filter
Dictionary content filter
Non−dictionary content filter
Seed URLs
Minority−language Web crawler
WWW
Strong minor texts
Weak minor texts
Weak major texts
Strong filtered minor texts
Strong filtered major texts
Sentence−aligned parallel texts
Alignment
Rejected pages
Sentence alignment (Champollion)

Figure 1: Flow of the BABYLON parallel text builder

4. Machine Translation Evaluation

Since a sufficiently sized pool of bilingual Quechua-Spanish speakers was not readily available to judge the quality of the produced texts, we instead use the produced corpus to train a statistical machine translation system. We use the Moses translation system (Koehn, 2007), and we evaluate the quality of the automatically produced translations by using the BLEU evaluation tool (Papineni et al., 2001). BLEU takes as input two sentences in the same language – one the output of a machine translation system and the other a gold standard translation produced by an expert. It measures the N-gram overlap of the two sentences and produces a score $B(N) = \frac{N{-}grams\ that\ overlap}{total\ N{-}grams}$. The scores in Tables 3 and 4 are produced using one to four-grams which are combined into a single score with the following equation:

$$score \leftarrow k * e^{log(B(1))+log(B(2))+log(B(3))+log(B(4)))/4}$$

where *k* is a function of the length of the translations which penalizes sentences that are too short.

The translations obtained when training on the Web-based parallel texts are compared to training on a parallel corpus consisting of an electronic version of the Bible in Quechua, aligned to one or four Spanish Bible translations. The sizes of the parallel texts used to train the machine translation system are shown in Table 2.

Translation models were built for five different training data sets: the crawled parallel texts alone, one translation of the Bible alone, the Bible translation plus the crawled texts, all four Bible translations alone, and all four Bible translations plus the crawled texts. In order to observe the usefulness of Web-gathered parallel texts, we evaluated each of the five translation models on both a subset of the Bible and a subset of the crawled text. The results of translating from Spanish to Quechua are shown in Table 3. The results of translating from Quechua to Spanish are shown in Table 4. The baseline score indicated below is the result of proposing that the input Spanish sentence is actually a Quechua sentence and comparing it to the reference Quechua translation and vice-versa. This baseline indicates how much of the N-gram overlap is actually due to cognates, numerals, proper nouns, and other non-translatable tokens that would be the same in both languages.

	Bible	Crawled
Lines	31,095	5,485
Quechua Words	484,638	87,398
Spanish Words	747,448	99,618
Quechua Size	4.6MB	550KB
Spanish Size	4.2MB	540KB

Table 2: Parallel text size

Note that the upper-bound on these data sets when using the BLEU evaluation metric is about 25-35, which corresponds to the BLEU N-gram overlap measured between two expert-produced Spanish translations.

As a final test, to validate the portability of our system, we replicated the entire experiment for the Spanish-Nahuatl language pair (Nahuatl is a low-density language spoken by about 1 million people in Mexico). The evaluation scores for a machine translation system trained on this data strongly correlated with those obtained for the Quechua-Spanish experiment, which demonstrates the portability of our system to new language pairs.

	Test Set	
Training Set	Bible	Crawled
Baseline	0.39	3.80
Crawled	0.62	6.42
Bible	2.89	2.65
Bible + Crawled	3.32	5.16
4 Bibles	4.70	2.66
4 Bibles + Crawled	4.55	5.70

Table 3: Spanish to Quechua Translation Results

	Test Set	
Training Set	Bible	Crawled
Baseline	0.38	3.81
Crawled	0.70	7.17
Bible	4.82	3.56
Bible + Crawled	4.79	6.26
4 Bibles	7.99	3.32
4 Bibles + Crawled	8.02	6.46

Table 4: Quechua to Spanish Translation Results

5. Conclusions

While the result of adding the crawled texts to the existing biblical texts yield little if any improvement over training on the biblical texts alone, it is apparent that the crawled texts alone produce a positive scoring translation even when the effects of including non-translatable tokens are ignored. This suggests that the texts that were created exhibit some usable quantity of parallelism.

While it has been suggested that corpus size contributes the most to the overall translation quality, our results seem to indicate otherwise. Adding three additional Bibles to the training set yielded negligible improvement on the crawled test set, but training on the much smaller Web-domain crawled set produced a higher score than with any biblical training set. In addition, the score found by evaluating the crawled training data against the crawled evaluation set is higher than with any other combination of training data. This score suggests that the texts contain enough parallelism to be able to predict how other crawled text might be translated.

Acknowledgments

This research has been supported in part by a research grant on “Hispanic and Global Studies” from the University of North Texas.

6. References

Jiang Chen and Jian-Yun Nie. 2000. *Parallel Web Text Mining for Cross-Language IR*. Proceedings of RIAO-2000: Content-Based Multimedia Information Access.

Jisong Chen and Rowena Chau and Chung-Hsing Yeh. 2004. *Discovering Parallel Text from the World Wide Web*. ACSW Frontiers ’04: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalization.

Philipp Koehn et al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. Proceedings of the Association for Computational Linguistics (ACL 2007), demonstration session.

Xiaoyi Ma. 2006. *Champollion: A Robust Parallel Text Sentence Aligner*. LREC 2006: Fifth International Conference on Language Resources and Evaluation.

Xiaoyi Ma and Mark Y. Liberman. 2006. *BITS: A Method for Bilingual Text Search over the Web*.

Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation*. ACL ’02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Philip Resnik. 1998. *Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text*. AMTA ’98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup.

Philip Resnik and Noah A. Smith. 2003. *The Web as a Parallel Corpus*. Computational Linguistics 29(3):349–380.

J. Tomás and E. Sánchez-Villamil and L. Lloret and F. Casacuberta. 2005. *WebMining: An Unsupervised Parallel Corpora Web Retrieval System*. Proceedings from the Corpus Linguistics Conference.

Lei Shi and Cheng Niu and Ming Zhou and Jianfeng Gao. 2006. *A DOM Tree Alignment Model for Mining Parallel Data from the Web*. ACL ’06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL.