An Unsupervised Method for Generating Parallel Corpora for Medium and High Density Languages

Max Kaufmann

***Abstract*—Parallel corpora are an extremely useful tool in many natural language processing tasks. However, it is often difficult, impossible, or expensive to obtain parallel corpora for certain language pairs. We propose an unsupervised method that is capable of automatically creating high quality parallel corpora, as well as bilingual lexicons. The outlined method is also capable of automatically augmenting the size of the corpora. The resulting parallel corpora are made freely available for download.**

Index Terms: **Wikipedia, Parallel Corpora, DBpedia**

I. Introduction

The Wikipedia project has resulted in an unprecedented amount of collaboration to create an astounding number of authoritative documents. It is a one of the largest free collections of humans knowledge in existence, and because of this it has received much media attention. But there is another aspect of Wikipedia that receives much less attention. Wikipedia is a large, if not the largest, multilingual repository. Articles in Wikipedia, especially for the more popular topics, are often written in a multitude of languages. However, these articles are not direct translations of each other, so Wikipedia is not a truly parallel corpora. Previous literature has referred to it as a document-aligned corpora or noisy parallel corpora [1] . While the articles are usually semantically similar, they can vary widely in their structure as well as syntax. This makes the task of creating parallel corpora from these articles fairly difficult. However, several projects such as DBpedia and OmegaWiki have organized the knowledge available on Wikipedia in a format that makes it easier to extract aligned sentences. By utilizing these resources, we can build a collection of parallel corpora for a large number of language pairs. In this paper, we will attempt to extract all information that would aid in NLP tasks involving two languages, such as machine translation or relationship extraction. The end result is a set of parallel corpora as well as a bilingual lexicon for each language pair.

II. Motivation

Parallel corpora are used in all types of multilingual research, especially in Machine Translation (MT) tasks. However, creating these corpora is an arduous task for several reasons. Creating parallel corpora requires humans who are proficient in both the source and target language(s). These translators usually are compensated in some form, which means that parallel corpora generation is both expensive and time consuming. Given that the strength of any MT system, particularly statistical MT systems, hinges largely on the quality and size of parallel corpora, we believe that automatic parallel corpora generation could greatly aid the task of machine translation by reducing the dependency on human generated corpora. The costs of creating parallel corpora are further exacerbated by the fact that as time goes on, new corpora have to be generated to account for changes in both languages. A parallel corpora generated 100 years ago would have far less utility in a machine translation task than a corpora made in the past 5 years.

The corpora generated by our approach attempt suffer much less from the previously mentioned flaws. The approach which we will discuss creates parallel corpora and lexicons by extracting and organizing information from two sources: DBpedia and OmegaWiki. Both of these projects are currently active, which suggests that as time goes on they will continue to grow in size and accuracy. The only costs associated with creating parallel corpora from these sources are the initial cost of creating the programs that create the corpora, and the computational cost of running them. These costs are minor compared with the expense invested in paying human translators to generate the parallel corpora.

Our research is further motivated by the fact that, despite the importance of parallel corpora, they are somewhat difficult to obtain. Despite the problems previously discussed, there are many parallel corpora available on the web. However, there are several problems with these corpora. First of all, they are not always free. The Linguistic Data Consortium[1] is one the largest unified providers of parallel corpora. However, membership is required to download these corpora. The cheapest membership is $1,000, while the most expensive is $24,000. The second issue is that the majority of these parallel corpora only exist for more popular language pairs. When attempting to translate from two very popular languages, such as English-German, finding parallel corpora is not difficult. But finding parallel corpora for languages pairs such as Thai-Greek is considerably more difficult. The parallel corpora resulting from this research will include corpora for small languages, and potentially language pairs for which parallel corpora do not exist. The size of the corpora may be small, depending on the size of the Wikipedias of the respective languages, but due to the self-modifying nature of these parallel corpora, they will be capable of automatically growing as their Wikipedias grow.

III. Previous Work

As far as we are aware, there have been no public projects which attempt to extract parallel corpora from this particular combination of resources. However, several projects have tackled the issue of multilingual retrieval from Wikipedia to build parallel corpora. One such example is [2]. They tested two approaches for aligning sentences from Dutch-English Wikipedia. Their first approach used an already existing SMT system to translate the Dutch article into English. They then compared the similarity of the sentences between the English article and the Dutch article (which was translated into English). The second approach create a bilingual lexicon by leveraging the fact that articles across languages translated the titles into English. They then used this lexicon and compared the sentences by computing their lexical similarity. While computing lexical similarity is an extremely effective way to measure multilingual sentence similarity, as evidence by [3], using an already made machine translation system appears rather impractical. The purpose of parallel corpora generation is to create a statistical machine translation, and if one already exists, and is capable of doing a decent job of translating a Wikipedia article, then there seems little utility in generating parallel corpora. This approach also lacks extensibility, because one cannot always assume that a MT system will already exist to aid in the translation. In their study, they found that the bilingual lexicon approach was, unsurprisingly, less accurate then the MT approach. This is potentially due to the fact that lexical similarity, especially when computed with an incomplete lexicon, is not powerful enough to truly capture sentence similarity. One of the main problems with their approach is that too many possible sentences candidates were generated. They compared every sentence in a Wikipedia article to every sentence in the L2 article, resulting in 80 million candidate pairs [2].

[3] used the lexicon methodology outlined in [2], but also attempted to utilize character lengths to compute sentence similarity to align Persian-English Wikipedias. They used the statistical model created by [4] to figure out the approximate correlation between the length of an English sentence and the length of a Persian sentence. Doing this, they were able to discard translation pairs that were too long or too short to actually be translations. This approach shortens the number of possible candidates significantly, which solves the problem of overmatching presented in [2].

Both of the previous approaches have only dealt with aligning one language pair, and it has always been from L1-English or English-L1. This phenomenon is not unique to these papers, the large majority of Wikipedia alignment work has focused on aligning Wikipedia in one language to English Wikipedia. Even projects that have used more than 2 languages, such as [1] have always had English as one of the languages. [1] attempted to align Spanish-English, German-English and Bulgarian-English Wikipedias by building a Maximum-Entropy classifier to determine if sentences from aligned articles were parallel or not. However, their classifier was trained on seed parallel data. Seed data may not always available, especially for small language pairs, and so their approach is not as general as the one we will outline here. We will use techniques similar to theirs for extracting parallel sentences from Wikipedia, but with seed

[1] http://www.ldc.upenn.edu/

data that we have generated ourselves.

IV. Resources

We aim to produce high quality parallel corpora for a wide variety of language pairs, and to achieve that we leveraged two sources of comporable corpora: DBpedia[2] and OmegaWiki[3]. These resources are described in more detail below:

A. OmegaWiki

OmegaWiki describes itself as a c̈ollaborative project to produce a free, multilingual dictionary in every language, with lexicological, terminological and thesaurus information.̈ It has entries for many sense-disambiguated words (each sense-disambiguated words is called an ëxpression)̈. Each entry includes a translation of the expression and its definition in several languages. An entry may also include information for certain grammatical information about an expression such as its POS tag, gender, corresponding Wikipedia article, and hypernms/hyponyms. A database including all of this information is freely downloadable. We have downloaded this database, and have created a program that, given a language pair, can extract all of the expressions in both languages, as well as the corresponding grammatical information available. We also extract the definitions of each expression and if they are available in L1 and L2, add them to our parallel corpora.

B. DBpedia

The DBpedia project is an attempt to take the data in Wikipedia and turn it into a structured representation of knowledge. We exploit the fact that during their attempts to turn Wikipedia into an ontology, DBPedia generates automatically generates short and long summaries, called abstracts for each entry in the ontology. While these abstracts are not entirely parallel, they may contain parallel data. Our approach leverages these abstracts, and attempts to find parallel sentences using the method escribed in the next section. DBpedia contains both long abstracts, which may be 5-10 sentences in length, as well as short abstracts, which are generally 1-2 sentences. The DBpedia ontology also has a l̈abelf̈ields from which we can extract additional vocabulary of L1 and L2. The l̈abelf̈ield which contains the Wikipedia title of the article for every language it is available in. The Wikipedia article titles are direct translations.

V. Parallel Corpora Generation

There are several differences between the two resources we are exploiting (DBpedia and OmegaWiki). The first is that DBpedia contains a large amount of noisy data, while OmegaWiki contains a much smaller amount of clean data. This is due to the fact that DBpedia data is generated automatically, while OmegaWiki data is created by human contributors. Therefore, our approach will first extract the higher quality data from OmegaWiki, and then use that data to aid in gathering data from DBpedia. As previously stated, the two fields that contain helpful information in OmegaWiki are the word translations, and definitions. Therefore, we first extract all parallel words and definitions for a given language pair. We then leverage the parallel words extracted from OmegaWiki to build a bilingual lexicon. However, for smaller languages, this method does not often yield a large lexicon.

Since OmegaWiki does not provide enough data for multilingual experiments, we turn to DBpedia as an additional source of parallel data. DBpedia itself does not contain any parallel data, however there is comparable data that we can exploit to gain parallel data. We posit that since the a pair of abstracts for a given entry in DBpedia are semantically similar, we might find syntactically parallel sentences in these abstracts to add to our corpora. However, parallel data will not necessarily be found in corresponding sentences, so extraction is not a trivial task. If we are attempting to extract English and Chinese parallel sentences, the first sentence in the English abstract may be parallel to the third sentence in the Chinese abstract. To solve this issue, we align each sentence in an L1 DBpedia abstract to every sentence in the corresponding L2 DBpedia abstract. This guarantees that all potential parallel sentence pairs are considered. However, this alignment suffers from two problems. First, it requires that the abstracts, which are in paragraph form, are split into sentences. To solve this problem, we use the Lingua sentence splitter[4], which has support for

[2]http://wiki.dbpedia.org/

[3]http://www.omegawiki.org

[4]http://code.google.com/p/corpus-tools/

splitting Catalan, Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish sentences. The second issue that arises from aligning sentences between abstracts is that it creates a significant amount of non parallel data. If we are aligning two abstracts, one with 4 sentences, and one with 7, we generate 28 sentence pairs, even though there can only be 4 parallel sentences. To solve this, we use HunAlign, a multi lingual open source parallel sentence aligner.

A. HunAlign

HunAlign [5] was originally created to detect whether sentences between Hungarian and English are parallel. However, the algorithm has been extended to work for any language pair. HunAlign primarily uses Gale-Church sentence alignment algorithm, which detects parallel sentences by comparing their length. However, if a bilingual dictionary exists, it can be used to compute the lexical similarity of sentences. We construct a bilingual lexicon by first getting all of the parallel words for a given language pair from OmegaWiki. We then concatenate the ẗabelf̈ield from DBpedia, which represents the titles of the article. We use this bilingual lexicon as the dictionary for HunAlign. We then run HunAlign on all of the sentence pairs generated in the previous step. HunAlign then generates a probability, p, between 0 and 1 which represents the likelihood that the sentences are parallel pairs. To ensure that our translations are truly parallel, we only accept values of $p > .99$. The end result is a set of parallel sentences between a given language pair.

VI. Results

We tested our method on several language pairs. Our goal was to create to a system which was capable of creating parallel corpora for uncommon language pairs. Since our system did not require any linguistic knowledge of the language pairs, excluding the ability to determine where sentences ended, we expected that it would perform equally well on extracting parallel data from similar and distinct languages. For the purposes of this experiment, we define languages as similar languages as languages from the same family (e.g., romance, germanic, hellenic) and distinct languages as languages from different families. Below are tables showing the data generated for several language pairs from 3 language families. The language pairs are described using their ISO 639-1 codes.

Language Pair	Parallel Words	Parallel Sentences	Family
DE-AF	2,151	5,559	Germanic
DE-SV	31,918	17,482	Germanic
EN-DE	69,210	13,213	Germanic
EN-ES	343,416	51,421	Mixed
EN-HI	17,234	1,550	Mixed
PL-UK	99,758	7,150	Slavic
RU-IT	51,496	11,251	Mixed

VII. Evaluation

From Table 1, we can conclude several things about the effectiveness of our method. Overall, more parallel sentences tend to be produced when the two languages are from similar families. The exception to this is EN-ES, which is probably due to the fact that the English and Spanish are the first and third largest Wikipedia's, respectively. This doesn't reflect on the effectiveness of our method, and is probably just due to the initial sizes of the corpora. It is also interesting to note that RU-IT performed significantly well, despite them being languages from different families. This is probably due to the size of the corpora.

A. Comparison to Other Corpora

It is difficult to evaluate the quality of a parallel corpora by any other metric other than looking at its size. Therefore, we will compare our corpora with two existing free parallel corpora.

There are several parallel corpora which are freely available. These corpora were all generated automatically. The smallest of these is Europarl, which contains aligned sentences between 11 European languages and English [6]. This corpora suffers from the same flaw as [2] and [3], which is that it only contains data aligned between English and another language.

The second is JRC-ACQUIS [7]. [7] contains parallel corpora for 22 languages, which were generated by running HunAlign on a series of EU documents. Below is a table comparing the size of the parallel data generated using our algorithm to the size of the two previous parallel corpora.

Despite the fact that we offer smaller parallel corpora than the two previous sources, there are several features that our corpora has that the previous

corpora lack. First, our corpora are self-updating, meaning that its size grows automatically. Both DBpedia and OmegaWiki are constantly increasing the size of their databases, and making that data publicly available. Our algorithm can use this updated data to generate bigger parallel corpora. Secondly, both of these corpora were generated from European Union legal documents. This means that, while they offer substantial amounts of data, it does not cover a wide variety of domains. Our corpora are based on Wikipedia, which covers many domains.

VIII. Future Work

The method used to generate parallel corpora in this paper, while only applied to several language pairs, is highly extensible. DBPedia has 97 languages, which means that there are a massive number of parallel corpora that could be generated using this algorithm. The approach we outlined here performs fairly well in extracting data from similar language pairs, but future work could improve its robustness on languages from distinct languages. One way to do that involves translating through three languages to create artificial parallel data. For example, if we have a large volume of English-Hindi sentences, and a large volume of English-Bengali parallel sentences, it would be possible to create new Bengali-Hindi aligned sentences by translating through English as an intermediary language.

Another way to improve this system would be to actually exploit Wiktionary to add grammatical data to the parallel corpora. Wiktionary contains translations, which could augment the size of the generated corpora, but it also contains grammatical information, such as case, part of speech, and conjugations for its entries. This grammatical information could be incorporated into HunAlign to help decide whether sentences are parallel or not, and it could be incorporated into the published parallel corpora to make them more useful to researchers.

IX. Contributions

In this experiment, we have outlined an algorithm for extracting parallel sentence data from DBPedia and OmegaWiki. This algorithm is capable of extracting parallel data from comparable documents with almost no previous linguistic knowledge. We have produced large parallel corpora and lexicons for common language pairs, and medium sized corpora and lexicons for distinct language pairs. We have created several parallel corpora for languages which previously did not have parallel corpora.Furthermore, these corpora are freely available and can be used in many multilingual NLP tasks. This algorithm extracts data from resources which are continually growing, so as these resource grow, so will the size of the parallel corpora and lexicons generated.

References

[1] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting parallel sentences from comparable corpora using document level alignment," Stroudsburg, PA, USA, pp. 403–411, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?id=1857999.1858062

[2] S. F. Adafre and M. de Rijke, "Finding similar sentences across multiple languages in wikipedia," in *11 Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[3] M. Mohammadi and N. GhasemAghaee, "Building bilingual parallel corpora based on wikipedia," in *Proceedings of the 2010 Second International Conference on Computer Engineering and Applications - Volume 02*, ser. ICCEA '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 264–268. [Online]. Available: http://dx.doi.org/10.1109/ICCEA.2010.203

[4] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ser. ACL '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1991, pp. 177–184. [Online]. Available: http://dx.doi.org/10.3115/981344.981367

[5] K. Tóth, R. Farkas, and A. Kocsor, "Sentence alignment of hungarian-english parallel corpora using a hybrid algorithm," Szeged, Hungary, Hungary, pp. 463–478, January 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1466514.1466522

[6] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," University of Souther california, Tech. Rep., 2002.

[7] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufiş, "The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages," pp. 2142–2147, 2006.