Parallel Document Identification using Zipf's Law

Mehdi Mohammadi
Department of Computer Science
Western Michigan University, MI, USA
mehdi.mohammadi@wmich.edu

Abstract

Parallel texts are an essential resource in many NLP tasks. One main issue to take advantage of these resources is to distinguish parallel or comparable documents that may have parallel fragments of texts from those that have no corresponding text. In this paper we propose a simple and efficient method to identify parallel documents based on Zipfian frequency distribution of available parallel corpora. In our method, we introduce a score called $CumulativeFrequencyLog$ by which we can measure the similarity of two documents that fit into a simple linear regression model. The regression model is generated based on the word ranks and frequencies of an available parallel corpus. The evaluation of the proposed approach over three language pairs achieve accuracy up to 0.86.

Keywords: Parallel corpora, Comparable Corpora, Parallel document identification, Zipf's Law, Wikipedia.

1. Introduction

Statistical NLP approaches, such as Statistical Machine Translation (SMT), are highly attractive and yield satisfactory results. However, a prerequisite for such methods is a parallel corpus containing a large amount of correct translation pairs i.e. sentences in the source language aligned with their translations in the target language. Constructing parallel corpora for scarce resource languages is an expensive job, since it requires translators who are fluent in both source and target languages. It also takes a lot of time to collect such examples. Therefore, researchers have paid attention to some other online sources like bilingual web sites to create parallel corpora.

Zipf's law is a statistical formulation devised empirically by G. K. Zipf that says in a corpus of natural language tokens, the frequencies of words associate inversely with their rank. This implies that rank-frequency distribution of words falls into an inverse relation. Two parallel corpora have this characteristic in common, so the frequency distribution of the words in one corpus would estimate the frequency of the words in the other side. In other words, the rank and frequency distribution of the terms in both documents are very close to each other.

In this paper we propose a method to identify parallel documents using a heuristic method based on Zipf's law. The essence of the filter is based on Zipfian frequency distribution of two parallel corpora combined with a linear regression model. The linear regression model is obtained from frequency analysis of tokens in the parallel corpora. Zipf's filter determines if two documents should be considered parallel or not using the error of prediction of linear regression function.

The motivation behind this work is to prepare fast and easy-to-build parallel corpora for limited-resource languages like Maori (the native language of New Zealand) to be used in NLP-related tasks. Beyond Statistical Machine Translation, such parallel corpora can be used in dialect identification (Malmasi et al., 2015) or lexicon construction. The proposed approach can also be extended to other NLP applications that deal with parallel corpus such as cross-language plagiarism detection in which a suspicious document is highly correlated to the original document in terms of words frequency distribution.

A primary application of this method is to find parallel documents among a set of comparable documents. Another interesting use case would be identifying comparable articles in Wikipedia and extracting parallel fragments of text from those comparable articles. Wikipedia is a source of multilingual texts that can be used to extract bilingual phrases or sentences automatically. Extracted parallel texts have been used as a complementary resource to Statistical Machine Translation systems in order to improve the performance of translation (Pal et al., 2014). Each article in Wikipedia may have a link to other languages. So, Wikipedia articles are aligned at document level. But they are not necessarily translations of each other. Although the articles with the same title in different languages are not exact translations of each other, it is possible to extract chunks of texts that have corresponding translations.

The rest of this paper is organized as follows. Section 2. presents an overview of the current approaches in this field. Section 3. presents details to undertake Zipf's filter for parallel documents identification. In section 4. we show our experimental results and evaluations. Finally we conclude the paper in section 5.

2. Related Work

There are many attempts to align parallel texts at document level. Among the existing approaches, heuristic methods have been shown to be attractive and efficient for identifying comparable and parallel documents. The main advantage of these methods is that they are usually easy to implement as well as easy to understand.

The work in (Paramita et al., 2013) reports implementing two simple filters to detect comparable documents in Wikipedia articles. These filters are document's minimum size and length's difference. Using these filters they rule out over 80% of the initial document pairs.

Zafarian et al. (2015) use different characteristics of German-English documents in four modules to identify their similarity. These modules perform reducing the size of target space, Name Entity recognition, building topic

Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, pages 21–25,
LREC 2016, Portorož, Slovenia, May 23, 2016.

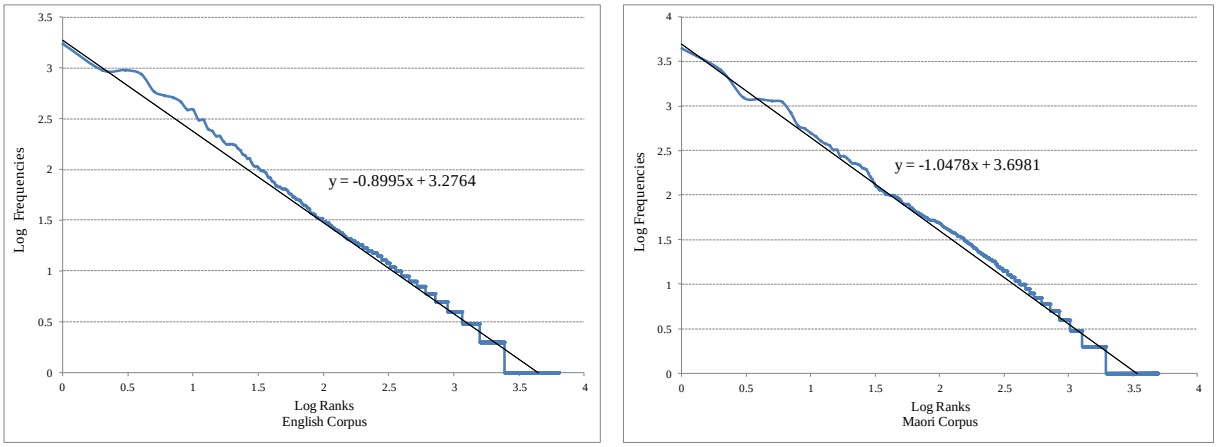

Figure 1: Zipf's curve for words in (a) English side, and (b) Maori side of a parallel corpus.

Parameters	English	Maori
Number of sentences	1695	1695
Number of words	30130	39488
Number of unique words	6380	4939

Table 1: The statistics of a small-size parallel corpora to analyze Zipf's law characteristics.

models and SMT. Their approach uses content of documents without links, tags or meta-data. Their results show that their approach can achieve recall of 45% for the first match.

In a system called LINA (Morin et al., 2015), authors use counts of hapax words to identify comparable documents. In their approach, only words that have appeared once in the document are considered for comparability measurement. Two documents that share the largest number of these hapax words are identified as parallel. Their results indicate that the system finds comparable documents with a precision of about 60%.

3. Zipfian-based Filter

Based on Zipf's law, the frequency of words in a large corpus is inversely proportional to their rank (Deane, 2005). The empirical law for single word frequency distribution says that if the words in a corpus are ranked by their frequency, for a given word with rank r, the function $f(r)$ gives the frequency of the word such that

$$f(r) = \frac{C}{r^{\alpha}} \quad (1)$$

where C is a normalizing constant for the corpus and α is a free parameter for specifying the degree of skew. For single word frequency distribution, α is close to 1. The study by Ha et al. (2002) shows that beyond one token, a list of n-gram tokens also follow the law very well. Putting the logarithm of frequencies versus the logarithm of the ranks in a graph, a straight-like curve is obtained with slope of -1. For large corpora with about one million tokens, it has been observed that the highest ranked words may have frequencies that deviated slightly from the straight line. However, it is asserted that the law is valid for small corpora (Ha et al., 2002).

The main task of the filter is to distinguish parallel document candidates from those that might have no parallel texts. In order to find out if Zipf's law is applicable to parallel documents, we analyzed the frequency distribution of a small parallel corpus. Table 1 shows the statistics of these data. We observed that our tiny-size corpus almost conform to the Zipf's law for the relationship of the rank and frequency of words in a corpus. Both the source and target languages show largely the same shape of relationship for the logarithm of rank and frequency. By analogy of the whole parallel corpus, we reached two linear functions for both languages with a slope close to -1. Figure 1 shows this observation.

The small size of corpora with this observation leads us to infer that this relationship should be held for two parallel documents as well. In two bilingual parallel documents, the rank and frequency of constituting words probably would be close to each other in two languages (The corresponding words in both sides should have largely the same rank and frequency). If two articles in two languages show the same pattern of relationship (a curve with the same slope) between the words ranks and frequencies, then we can infer that the two articles may have some degree of parallelism. In such cases, if a document in the source language consists of the words that have the ranks between 1 to r_s then the corresponding comparable document in the target language includes words ranks from 1 to r_t. Based on Zipf's law, r_s and r_t have a high probability to be close to each other. Intuitively, the area beneath the two functions as an indicator of parallelism of two documents would be close to each other. Figure 2 illustrates the idea where two candidate documents have some degree of parallelism versus two documents that are not related at all. We compute the area beneath the curve as *cumulative frequency log* for a document D as follows.

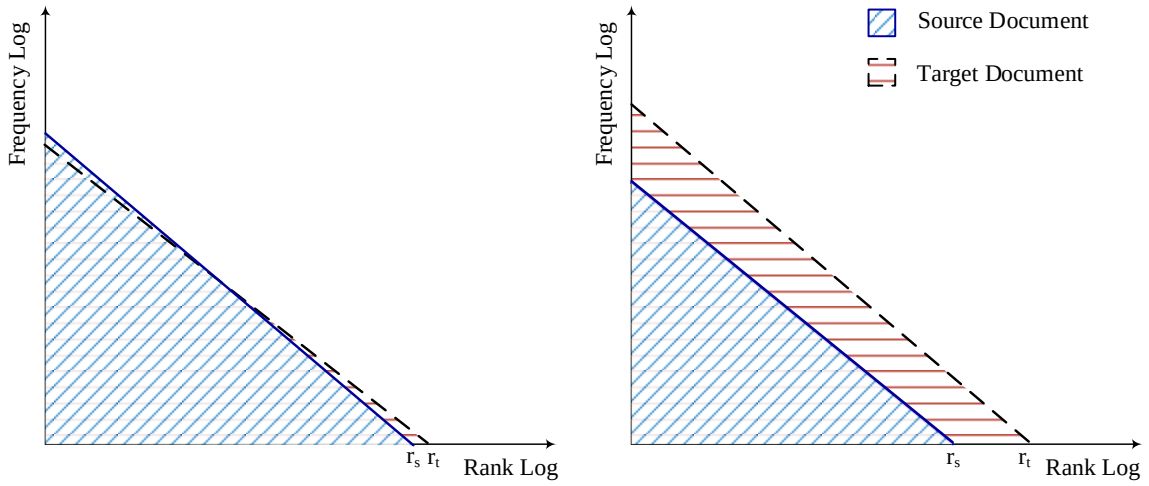

Figure 2: (a) two comparable documents that share parallel texts; (b) two documents that do not contain parallel texts.

$$Score(D) = \sum_{r=1}^{rmax} log(f(r)) \quad (2)$$

where r is the rank of the words in D, $rmax$ is the last rank in the document, and $f(r)$ is the frequency associated to the rank r.

Analyzing the cumulative frequency log of parallel documents reveals that for a given language, this score is linearly related to its counterpart in the other language. Figure 3 depicts this relationship for 40 Spanish-English parallel documents that are generated from Spanish part of Europarl corpora (Koehn, 2005). In this set, the lengths of document pairs are considered different.

Therefore, having the *Cumulative Frequency Log* of source documents will estimate the *Cumulative Frequency Log* of the target documents. In the training process with a set of n parallel documents, we use a Linear Regression Model to predict the response to n data points (x_1, y_1),(x_2, y_2), ...,(x_n,y_n) where x_i and y_i are the cumulative frequency log of ith parallel document pair in the source and target language, respectively. The linear regression model is given by

$$y = a_0 + a_1 x \quad (3)$$

where a_0 and a_1 are the constants of the regression model. A measure of best-fitting line, i.e, how well $a_0 + a_1 x$ predicts the cumulative frequency log of y is the magnitude of the error of predictions (ϵ_i) at each of the n data points.

$$\epsilon_i = y_i - (a_0 + a_1 x_i) \quad (4)$$

The regression parameters can be obtained by minimizing these errors of predictions by Least Square methods.

In the core of the filter, with two given documents in the source and target languages, namely D_s and D_t, the cumulative frequency log of two documents are computed as $x = Score(D_s)$ and $y = Score(D_t)$. Then x is put to the regression model to obtain the predicted cumulative

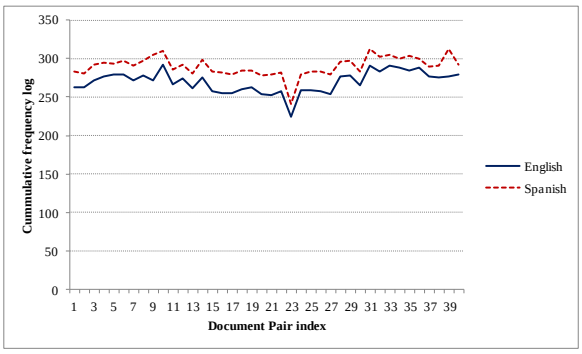

Figure 3: Cumulative frequency log of parallel documents

frequency log of target document. By computing the absolute value of error of prediction (ϵ), we determine the parallelism of two documents if ϵ is smaller than or equal to a threshold called δ.

$$Par(D_s, D_t) = \begin{cases} 1, & \text{if } |\epsilon| \leq \delta \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The best result for Eq. 5 is obtained when $\epsilon = 0$ which means the predicted value coincides with the actual value. However, we need to allow some degree of deviation from the regression model using δ. We can find the best value for δ that maximizes the precision and recall of the filter at the same time. Our experiments in the next section find different best δ for different language pairs.

4. Experiment and Results

We have used the English-Spanish (en-es), English-Dutch (en-nl), and English-Swedish (en-sv) parallel corpora in the Europarl dataset (Koehn, 2005) to evaluate our proposed method. In this regard, we split each parallel corpus to 77 parallel document pairs with different sizes. The range of size of these documents is from a couple of lines to about

Language pair	#test doc pairs	#parallel test docs	training data (MB)	
			source	target
English-Spanish (en-es)	314	27	182	201
English-Dutch (en-nl)	336	12	184	203
English-Swedish (en-sv)	290	26	170	177

Table 2: Statistical information of test and training dataset.

δ	English-Spanish			English-Dutch			English-Swedish		
	precision	**recall**	**F1**	**precision**	**recall**	**F1**	**precision**	**recall**	**F1**
1	0.57	0.15	0.24	0.57	0.33	0.42	0.86	0.23	0.36
2	0.60	0.22	0.32	0.47	0.58	0.52	0.69	0.35	0.46
3	0.62	0.30	0.40	0.47	0.75	0.58	0.74	0.65	0.69
4	0.55	0.41	0.47	0.41	0.75	0.53	0.71	0.77	0.74
5	0.52	0.44	0.48	0.41	0.92	0.56	0.65	0.77	0.70
6	0.50	0.67	0.57	0.37	0.92	0.52	0.62	0.81	0.70

Table 3: Evaluation results of the proposed method applied on three language pairs.

Length ratio threshold (β)	English-Spanish	English-Dutch	English-Swedish
0.1	0.74	0.44	0.57
0.2	0.47	0.31	0.47
0.3	0.36	0.21	0.37

Table 4: Accuracy of length-based filter to identify parallel documents

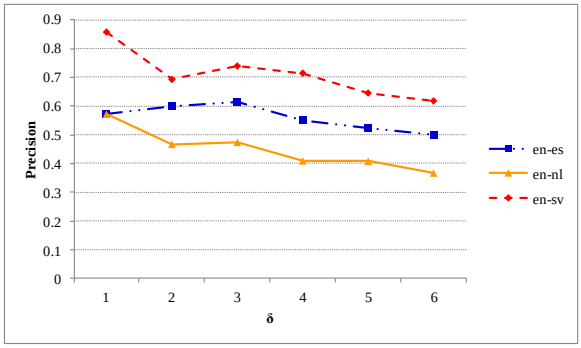

Figure 4: Precision trend versus delta for three language pairs.

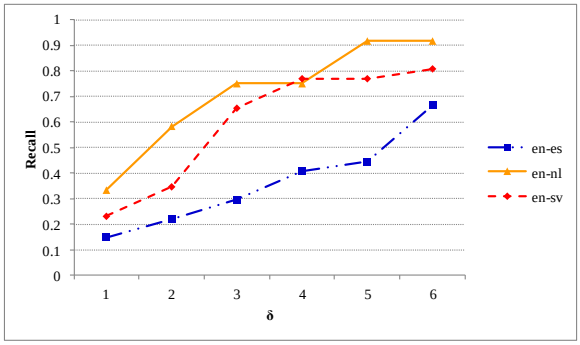

Figure 5: Recall of the proposed method with different delta for three language pairs.

100K lines in which each line represents a sentence. For each language pair, we use 50 document pairs for training the model and use the remaining document pairs to create test data. The test data are generated using randomly picking one document from the source language and one from the target language. Actual parallel documents are identified by a same name in the source and target languages. Table 2 shows some statistical information about the training and test data.

In the experiment, we perform several runs with different threshold (δ) from 1 to 6. We go through interval of 1 for δ since we can see bigger changes in the precision and recall. Table 3 summarizes the precision, recall and F measure obtained by the proposed approach for three language pairs. Figure 4 illustrates the precision results for three given language pairs with varying δ. Figure 5 also shows the recalls with the same settings.

Our results show that using a low threshold yields higher precision and lower recall compared to using a high threshold that leads to lower precision and higher recall. We can rely on F-measure to find out the best setting for threshold. From the results in Table 3, the thresholds that maximize the F-measure for Spanish-English, Dutch-English, and Swedish-English are 6, 3, and 4, respectively. With these best configurations in the language pairs of the study, the filter achieves a precision between 0.47 to 0.71, recall between 0.67 to 0.77, and F-measure between 0.57 to 0.74.

Compared to the related works like (Zafarian et al., 2015) and (Morin et al., 2015) in which the precision is reported as 0.46 and 0.57, respectively, our approach achieves competitive results, in particular when the parameter δ is fine-tuned.

We also run another experiment over test data using a length-based filter to identify parallel documents and benchmark against the proposed Zipfian-based filter. We compute the length ratio of each two documents i and j ($length\_ratio_{ij}$) based on their word counts and decide over their parallelism if $|length\_ratio_{ij} - 1| \leq \beta$, where β is a predefined threshold. Table 4 presents the precision

results obtained by this method using different threshold values. The results show that the length based filter performs relatively well for English-Spanish documents, but its performance for English-Dutch and English-Swedish is not very good. In contrast, our Zipfian-based filter outperforms the length based filter for English-Dutch and English-Swedish documents.

5. Conclusion and Future Works

Parallel texts are an essential source of NLP and machine translation tasks while they are hardly available for under-resource languages. In this paper we proposed to identify parallel documents from a set of comparable articles using a filter based on Zipfian characteristic of parallel documents. We performed experiments over three language pairs to evaluate the proposed approach. Based on our results, the approach achieves promising results in terms of precision and recall of the identified parallel documents. The proposed method is language independent and does not rely on any linguistic knowledge.

Potential pathways for future works include extensive evaluation of the proposed method on larger experiment test cases that covers more language families. Another pathway would be to apply the proposed approach to some well-known existing methods for parallel text identification to improve the phase of document-level alignment in these approaches. In particular, applying the proposed method on linked Wikipedia articles to extract parallel articles from Wikipedia resources would be beneficial for low-resource languages.

6. References

Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 605–613. Association for Computational Linguistics.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., and Smith, F. J. (2002). Extension of zipf's law to words and phrases. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–6. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Malmasi, S., Refaee, E., and Dras, M. (2015). Arabic dialect identification using a parallel multidialectal corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015), Bali, Indonesia*, pages 209–217.

Morin, E., Hazem, A., Boudin, F., and Clouet, E. L. (2015). Lina: Identifying comparable documents from wikipedia. In *Eighth Workshop on Building and Using Comparable Corpora*.

Pal, S., Pakray, P., and Naskar, S. K. (2014). Automatic building and using parallel resources for smt from comparable corpora. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, pages 48–57.

Paramita, M. L., Guthrie, D., Kanoulas, E., Gaizauskas, R., Clough, P., and Sanderson, M. (2013). Methods for collection and evaluation of comparable documents. In *Building and Using Comparable Corpora*, pages 93–112. Springer.

Zafarian, A., Aghasadeghi, A., Azadi, F., Ghiasifard, S., Alipanahloo, Z., Bakhshaei, S., and Ziabary, S. M. M. (2015). Aut document alignment framework for bucc workshop shared task. *ACL-IJCNLP 2015*, page 79.