Building & Using Comparable Corpora for Machine Translation

A report submitted for the Summer Internship Project (2018)

By

Rahul Ranjan

Bachelor of Technology, IV Semester

Under the guidance of

Dr. Anil Kumar Singh

Department of Computer Science and Engineering

Indian Institute of Technology (BHU), Varanasi

July, 2018

Abstract

The size and quality of the parallel corpus used for training, greatly impacts the quality of translation of an SMT system. But, there are very few sources of parallel corpora for many language pairs. This is a major hurdle in the development of good SMT systems. To alleviate this problem, comparable or non-parallel corpora, which are largely available, can be exploited to extract parallel data. We study the recent work done in this area, and explore various approaches for extraction of parallel sentences, parallel fragments of sentences from comparable corpora.

Keywords - Comparable Corpus, Parallel Corpus, Parallel Sentence Extraction

Declaration

I declare that this submission represents my idea in my own words and where others' idea or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/sources in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from proper permission has not been taken when needed.

(Signature)

(Name of the student)

Date:

Department of Computer Science & Engineering
Indian Institute of Technology (BHU), Varanasi

Dr. Anil Kumar Singh
Associate Professor
Email: aksingh.cse@iitbhu.ac.in

To Whom It May Concern

This is to certify that the report entitled **"Building & Using Comparable Corpora for Machine Translation"** submitted by "Rahul Ranjan", has been carried out under my supervision and that this work has not been submitted elsewhere for a degree,diploma or a course.

Signature of Supervisor

(Dr. Anil Kumar Singh)

Acknowledgement

I express my sincere and heartfelt gratitude towards my guide, Dr Anil Kumar Singh*, for all the valuable guidance and support he has given me in this* Project*. I also thank the entire* NLP Lab Members(Phd Researchers *and* Staffs) *at* CSE *dept.,* IIT BHU *for the insightful discussions and their valuable suggestions in the weekly meetings.*

- Rahul Ranjan

Contents

List of Figures

List of abbreviations

A

ASCII	American Standard Code for Information Interchange
ASPEC	ASEAN Patent Examination Co-operation

C

CFG	Context Free Grammar
CRF	Conditional Random Fields

H

HTML	Hyper Text Mark-up Language
HMM	Hidden Markov Model

M

MLP	Multilayer Perceptron
MST	Maximum Spanning Tree parser
MT	Machine Translation
SMT	Statistical Machine Translation

N

NN	Neural Network
NMT	Neural Machine Translations

S

SVM	Support Vector Machine
SVC	Support Vector Classification

X

XML	Xtended Mark-up Language
TMX	Translation Memory Exchange

Chapter 1

Introduction

A machine translation system that is based on the probabilistic translation models (Brown et al., 1993), are generally trained using parallel corpora. A parallel corpus is a sentence aligned pair of documents in which each pair of aligned sentences are trans- lations of each other. During training, the translation model of the SMT system learns the statistics over the training data provided to it and estimates its parameters ac- cordingly. The translations that are produced are very much dependent on the parallel corpus used for training. Larger the size of the training corpus, better is the parameter estimation and thus, better is the translation quality. But, one of the major challenges faced by SMT is that of scarce availability of parallel corpora. There are some language pairs like English-French or English-Spanish, for which huge set of parallel corpora are readily available. But for many other language pairs, there is scarcity of parallel corpora. Creation of a large parallel corpus manually, can be very costly in terms of efforts and man hours. So, we need some ways to automatically and efficiently create parallel corpora. Comparable corpora and non-parallel corpora are largely available for all language pairs. The next section introduces comparable corpora and its various sources.

1.1 Comparable Corpora

A noisy parallel or comparable corpus consists of bilingual documents that are not sentence aligned, but are rough translations of each other. In fact, in a noisy paral- lel corpus, the documents can possibly contain many parallel sentences in a roughly same order (Smith et al., 2010). But, in a comparable corpus, the sentences are not really translations of each other but convey almost the same information, and hence, may contain some parallel sentences. Such comparable corpora can be exploited to find parallel sentences or parallel phrases. Examples of such comparable corpora are multilingual news feeds provided by news agencies like Agence France Presse, Xinhua News, Reuters, CNN, BBC, etc (Munteanu and Marcu, 2005). Also, comparable cor- pora could contain documents that are not even rough translations of each other but are only topic-aligned.

1.1.1 Wikipedia

Wikipedia is a huge collection of articles on a large variety of topics and various lan- guages. So, it is rich in information from various domains and that too, in many different languages.

1. **Interwiki-links**

   Articles on the same topic in different languages in Wikipedia are connected through interwiki links. These links are annotated by users. Thus, document alignment for mul- tilingual documents on similar topics is already provided in Wikipedia. These aligned documents, can be directly given to the sentence extraction step.

2. **Markup**

   Wikipedia's markup can be very useful in providing numerous cues for parallel sentence extraction. For example, text in a typical Wikipedia article contains hyperlinks that point to other articles. If the hyperlinks in a bilingual sentence pair match, then that sentence pair can be said to be parallel. Hyperlinks match if the articles they point to, are connected by interwiki links.

1.2 Outline of Report

- Chapter 2 - Gives a general architecture of a parallel corpora extraction system. Discusses approaches for document alignment.
- *Chapter 3* -In this Chapter, Parallel system analysis, discussion ,results are taken into point. The existing work in extracting parallel sentences from non-parallel or comparable corpora using various approaches has been discussed.
- *Chapter 4* - Basically Conclusions, results and future scopes to be taken is discussed.

Chapter 2

Parallel Sentence Extraction System

The overview of our parallel sentence extraction system is presented in Figure 1. We first align articles on the same topic in Wikipedia via the interlanguage links ((1) in Fig- ure 1). Then we generate all possible sentence pairs by the Cartesian product from the aligned articles, and discard the pairs that do not fulfill the conditions of a filter to reduce the candidates keeping more reliable sentences ((2) in Fig- ure 1). Next, we use a classifier trained on a small number of parallel sentences from a seed parallel corpus to identify the parallel sentences from the candidates ((3) in Figure 1).The strategy of the filter and the features used for the clas- sifier will be described in Section 2.1. and Section 2.2. in detail.

2.1 Parallel Sentence Identification by Binary Classification

As the quality of the extracted sentences is determined by the accuracy of the classifier, the classifier becomes the core component of the extraction system. In this section, we first describe the training and testing process, and then introduce the features we use for the classifier.

2.2 Training & Testing

We use a support vector machine (SVM) classifier. Training and testing instances for the classifier are created follow- ing the method of (Munteanu and Marcu, 2005). We use a small number of parallel sentences from a seed parallel cor- pus as positive instances. Negative instances are generated by the Cartesian product of the positive instances exclud- ing the original positive instances, and they are filtered by the same filtering method used in Section 2.1.. Moreover, we randomly discard some negative instances for training when necessary, 1 to guarantee that the ratio of negative to positive instances is less than five for the performance of the classifier.

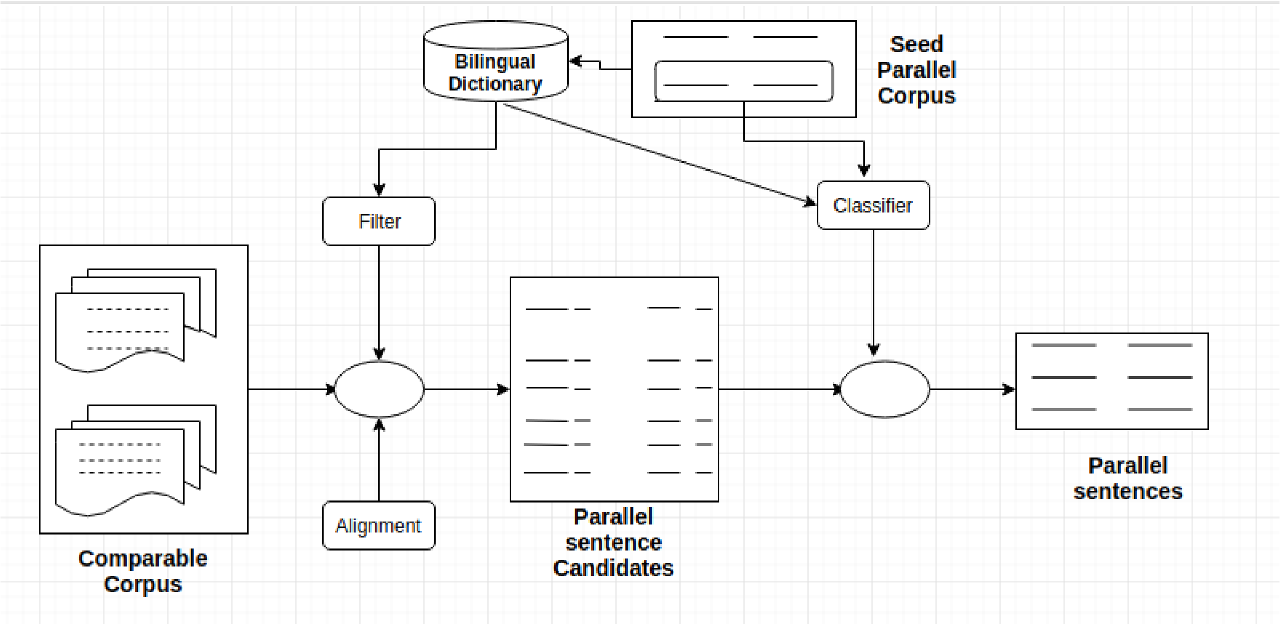

Figure 2.1: Parallel Sentence Extraction System

2.2.1 Alignment Technique

Sentence alignment was a very active field of research in the early days of statistical machine translation. An influential early method is based on sentence length, measured in words (Brown et al., 1991; Gale and Church, 1991; Gale and Church, 1993) or characters (Church, 1993).Here, Gale & Church is used for the alignment purpose as it is very much efficient appropriate.Aligning sentences is just a first step toward constructing a probabilistic dictionary (Table 3) for use in aligning words in machine translation (Brown et al. 1990), or for constructing a bilingual concordance (Table 4) for use in lexicography (Klavans and Tzoukermann 1990). Although there has been some previous work on the sentence alignment (e.g., Brown, Lai, and Mercer 1991 [at IBM], Kay and R6scheisen [this issue; at Xerox], and Catizone, Russell, and Warwick, in press [at ISSCO], the alignment task remains a significant obstacle preventing many potential users from reaping many of the benefits of bilingual corpora, because the proposed solutions are often unavailable, unreliable, and/or computationally prohibitive. Most of the previous work on sentence alignment has yet to be published. Kay's draft (Kay and R6scheisen; this issue), for example, was written more than two years ago and is still unpublished.
The sentence alignment program is a two-step process. First paragraphs are aligned, and then sentences within a paragraph are aligned. It is fairly easy to align paragraphs in our trilingual corpus of Swiss banking reports since the boundaries are usually clearly marked. However, there are some short headings and signatures that can be confused with paragraphs. Moreover, these short ”pseudo-paragraphs” are not always translated into all languages. On a corpus this small the paragraphs could have been aligned by hand. It turns out that ”pseudo-paragraphs” usually have fewer than 50 characters and that real paragraphs usually have more than 100 characters. We used this fact to align the paragraphs automatically, checking the result by hand. The procedure correctly aligned all of the English and German paragraphs. However, one of the French documents was badly translated and could not be aligned because of the omission of one long paragraph and the duplication of a short one. This document was excluded for the purposes of the remainder of this experiment.

2.2.2 Bilingual Dictionary

A bilingual dictionary or translation dictionary is a specialized dictionary used to translate words or phrases from one language to another. Bilingual dictionaries can be unidirectional, meaning that they list the meanings of words of one language in another, or can be bidirectional, allowing translation to and from both languages. Bidirectional bilingual dictionaries usually consist of two sections, each listing words and phrases of one language alphabetically along with their translation. In addition to the translation, a bilingual dictionary usually indicates the part of speech, gender, verb type, declension model and other grammatical clues to help a non-native speaker use the word. Other features sometimes present in bilingual dictionaries are lists of phrases, usage and style guides, verb tables, maps and grammar references. In contrast to the bilingual dictionary, a monolingual dictionary defines words and phrases instead of translating them.

Here, a bilingual dictionary (Hungarian-English) is used. The Bilingual Dictionary is also used in filtering the comparable corpus and used in classifier to train the corpus model to achieve better accuracy.

2.2.3 Seed Parallel Corpus

The Seed Parallel Corpus used here is a parallel corpus of English(en)-Hungarian(hu) Language used for testing the training data as it will provide better accuracy.The seed parallel corpus is obtained from OPUS (opus.nlpl.eu) and contains 1.3k source tokens (en)and 1.2k target tokens (hu)and 0.6k sentence alignment as the file is in XML/TMX format.The amount of parallelism among sentences is quite normal not highly parallel sentences.To extract text from XML/TMX file there are certain codes/programs which can be implemented.The seed parallel corpus is used in classifier as testing data to classify parallel sentences and improve the accuracy of classifier.

Chapter 3

System Analysis & Experiments

3.1 Introduction

We evaluated classification accuracy, and conducted extrac- tion of data.As the quality of the extracted sentences is determined by the accuracy of the classifier, the classifier becomes the core component of the extraction system.The classifier helped in achieving the result by classifying the sentences which are parallel.

3.1.1 Data

The seed parallel corpus we used is the English-Hungarian section of the OPUS, containing 0.6k sentences pairs (1.3K English and 1.2 Hungarian tokens, respectively). Also, we downloaded English (latest) and Hungarian (latest) Wikipedia database dumps. We used an open-source Python script to extract and clean the text from the dumps. We aligned the articles on the same topic using Gale & Church Algorithm for the alignment.

3.1.2 Classification Accuracy Evaluation

We evaluated classification accuracy using two distinct sets of 0.6k parallel sentences from the seed parallel corpus for testing and Parallel sentence candidates corpus (containing 54,866 tokens) for training purpose for the classifie.

3.1.2.1 Settings

- Word-Alignment : Gale & Church Algorithm.
- Dictionary : Bilingual Dictionary of English-Hungarian Languages.
- Classifier : For the classification we have used Support Vector Machine (SVM) with 5-fold cross-validation and radial basis function (Linear) kernel,Pandas library,Keras library.
- Sentence length ratio threshold: 2
- Word overlap threshold: 0.25
- Classifier probability threshold: 0.9

3.1.2.2 Evaluation

We evaluated the performance of classification by comput- ing precision, recall and F-measure, defined as:

$$\textit{precision} = (100 * classified\_well)/classified\_parallel$$

$$\textit{\textbf{recall}} = (100 * classified\_well)/true\_parallel$$
$$\textit{\textbf{F-measure}} = (100 * precision * recall)/precision + recall$$

Where classified_well is the number of pairs that the clas- sifier correctly identified as parallel, classified_parallel is the number of pairs that the classifier identified as parallel, true_parallel is the number of truly_parallel pairs in the test set.

3.1.2.3 Results

We have obtained following result from applied system in fig 2.1 that uses different features:

- Baseline : The Parallel Sentence Extraction System

Results are shown in Table 1.To understand why our proposed method contributed to the recall but not the precision, we analyzed the classification results.Together with the other features, Baseline judges this sentence pair as non- parallel.

System	Accuracy	Precision	Recall	f-measure
Baseline (SVM)	0.6437	0.5639	0.6178	0.5896

We looked at methods based on classification,- ranking, similarity, dis- criminative modeling, etc. for finding parallel sentences from comparable Corpora. Most of these techniques make use of word alignment based features to distinguish parallel sentences from non-parallel or Comparable corpora.

3.2 Related work

Several Studies and Projects have been done regarding extraction of parallel sentences from comparable corpora, some are (Munteanu and Marcu, 2005), (Sutskever et al., 2014),(Chu et al., 2014), Dabre et al., 2015,(Bahdanau et al., 2014),etc.
(Cho et al., 2014) scored the phrase pairs of a SMT system with a neu- ral translation model,

and used the scores as additional NN features for decoding. (Dabre et al., 2015) used the NN features for a pivot-based SMT system for dictionary con- struction. In contrast, we score the sentence pairs of with a neural translation model, and use the scores as NN features for parallel sentence extraction from comparable corpora.

Chapter 4

Conclusion & Future Direction

Lack of parallel corpora is a major bottleneck in development of SMT systems for many language pairs. There are large amounts of comparable and non-parallel corpora available for many language pairs. These can be exploited to extract parallel data from them.

From any non-parallel set of documents, document alignment gives a definite region in the entire search space to look for parallel sentence pairs. The sentence extraction step must be robust in order to filter any noise and give accurate, parallel sentence pairs. Good use of feature functions for ME classifiers and ranking models can be used to serve the purpose.

The amount of data extracted is often adversely affected by a poor dictionary cov- erage. This can easily be resolved by using a few boostrapping iterations or other approaches as discussed which will be implemented in future for better performance of project. In this paper, we have not incorporated the NN features for parallel sentence extraction from comparable corpora for the but that will surely improve the accuracy what achieved till now is not sufficient, it is just an start will improve alot in future.In future neural network and deep learning will be applied to improve its accuracy more. As future work, we plan to address the domain problem of "+NN-ASPEC" by a NN based sentence selection method. Namely, we train NN models on the sentences extracted by the baseline system, and then use these models to select sentences from ASPEC (or other corpora) that are similar to the sentences in Wikipedia. Hopefully, it could further improve the performance of our system.

Bibliography

[1] Chenhui Chu , Raj Dabre , Sadao Kurohashi *Parallel Sentence Extraction from Comparable Corpora with Neural Network Features* (2014)

[2] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical ma- chine translation: Parameter estimation. Computational Linguistics, 19(2):263–312

[3] Sanjika Hewavitharana and Stephan Vogel *Extracting Parallel Phrases from Comparable Data*

[4] William A. Gale and Kenneth W. Church , *A Program for Aligning Sentences in Bilingual Corpora* 1991.

[5] Quoc Hung Ngo, Werner Winiwarter *Building an English-Vietnamese Bilingual Corpus for Machine Translation*

[6] Franz Josef Och, Hermann Ney (2003). *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics, 29:19–51

[7] Ker Sue J. and Jason S. Chang (1997). *A class-based approach to word alignment.* Computational Linguistics, 23(2):313–343

[8] Quoc Hung-Ngo, Werner Winiwarter (2012).*A Visualizing Annotation Tool for Semi-Automatically Building a Bilingual Corpus*, In Proceedings of the 5th Workshop on Building and Using Comparable Corpora, LREC2012 Workshop, pp. 67-74.

[9] Van Bac Dang, Bao Quoc Ho (2007). *Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining. Research*, Innovation and Vision for the Future (RIVF), IEEE International Conference. pp. 261-266.

[10] Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi,*Constructing a Chinese–Japanese Parallel Corpus from Wikipedia* 2009

[11] Sadaf Abdul-Rauf and Holger Schwenk. 2011. *Parallel sentence generation from comparable corpora for im- proved smt.* Machine Translation, 25(4):341–375.

[12] Rucha C. Kulkarni, under the guidance of Prof. Pushpak Bhattacharyya (IIT Bombay & IIT Patna) Extraction of Parallel Corpora from Comparable Corpora (Survey Report)

[13] Anil Kumar Singh,Harshit Surana, *Can Corpus Based Measures be Used for Comparative Study of Languages?*

[14] Francis Grégoire and Philippe Langlais *A Deep Neural Network Approach To Parallel Sentence Extraction*

[15] Philipp Koehn ,*Europarl: A Parallel Corpus for Statistical Machine Translation*

[16] Jason R. Smith , Chris Quirk and Kristina Toutanova, *Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment*

[17] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy,László Németh , Viktor Trón, *Parallel corpora for medium density languages*

[18] Dragos Stefan Munteanu, Daniel Marcu, *Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora*

[19] Percy Cheung and Pascale Fung. 2004. Sen- tence alignment in parallel, comparable, and quasi- comparable corpora. In LREC2004 Workshop.