Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 43 (2016) 56–69

www.elsevier.com/locate/csl

Parallel fragments : Measuring their impact on translation performance

Sadaf Abdul-Rauf[a,b,*], Holger Schwenk[b], Mohammad Nawaz[c]

[a] Fatima Jinnah Women University, Rawalpindi, Pakistan
[b] LIUM, LUNAM University, University of Le Mans, France
[c] BUITEMS, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan

Received 29 April 2016; received in revised form 28 July 2016; accepted 2 December 2016
Available online 21 December 2016

Abstract

Lack of parallel corpora have diverted the direction of research towards exploring other arenas to fill in the dearth. Comparable corpora have proved to be a valuable resource in this regard. Interestingly other than the parallel sentences extracted from comparable corpora, parallel phrase fragments have also proved to be beneficial for statistical machine translation. We present a novel approach based on an efficient framework for parallel fragment extraction from comparable corpora. Using the fragments as additional corpus for translation, we are able to obtain an improvement of 0.88 and 0.89 BLEU points on test data for Arabic−English and French−English systems respectively. We have also conducted a detailed analysis of impact of fragments extracted from related vs non-related corpus. A comparison of impact of parallel fragments vs. parallel sentences is also presented highlighting the significance of parallel segments for statistical machine translation. The article concludes with a crude comparative analysis of our approach with an existing fragment extraction technique at various stages of the fragment extraction pipeline.

Keywords: Parallel fragments; Statistical machine translation; Comparable corpus

1. Introduction

In recent decades, construction and research on bilingual corpora has become an area of immense importance and interest. Due to its emerging importance, comparable corpora have become a significant object of study by researchers. These have proved to be beneficial in a variety of tasks such as improving SMT performance using extracted parallel sentences (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2011), extracting phrasal alignments (Kumano et al., 2007), word sense disambiguation (Kaji, 2003), acquiring synonyms (Shimohata and Sumita, 2005), parallel fragment extraction (Munteanu and Marcu, 2006; Cettolo et al., 2010), extracting lay paraphrases of specialized expressions (Deléger and Zweigenbaum, 2009) and language and translation model adaptation (Snover et al., 2008; Abdul Rauf et al., 2016) etc. They have specifically proved to be valuable for languages and domains which lack parallel corpora.

[*] Corresponding author at: Fatima Jinnah Women University, Rawalpindi, Pakistan.
E-mail address: sadaf.abdulrauf@gmail.com (S. Abdul-Rauf).

http://dx.doi.org/10.1016/j.csl.2016.12.002

The world has become a global village and translations play a vital role in bridging communication gaps all over the world. It is not affordable for human beings to translate everything manually so the demand of machine translation (MT) is growing rapidly all over the world. In Statistical MT, translation information is automatically obtained from parallel corpora; parallel corpus is a corpus that contains sentence aligned source texts and their translations. Parallel corpora can be bilingual or multilingual, once a parallel corpus is available, then the rapid development of SMT systems for different language pairs is possible; by examining many samples of human-produced translation, SMT algorithms automatically learn how to translate (Brown et al., 1993).

Because of the high dependence on parallel corpora, the quality and quantity of parallel corpora are crucial for SMT, the insufficiency of parallel data is an issue of concern for SMT system development. Lack of parallel corpus and linguistic resources for many languages and domains is one of the major obstacles for diverse and good SMT systems. Reasons for the scarceness of these resources are the inherent richness of languages, domain diversity etc. Moreover, languages evolve over time, the SMT training corpora also needs to be updated accordingly. Again, this is difficult in the case of parallel corpora.

There are many language pairs which do not have enough parallel corpora. Building such corpora can take much time as corpus building is a slow process for less spoken languages. Germann (2001) report that it takes 140 translation hours to create a 1300 sentence (24,000 tokens) Tamil−English parallel corpus at an average translation rate of 170 words per hour. At this rate it would require 4−5 full time translators to translate 100,000 words in a month. This of course forces to explore other scenarios for parallel corpus creation. For many language pairs, most of the times, comparable corpora do exist. A comparable corpus is a collection of texts in two or more languages which has similar contents in each language but do not have the exact translation of each language pair. We can gather and compile comparable corpora from multilingual newspapers, Wikipedia and different websites which contain articles on same topic in different languages.

Extraction of parallel sentences and segments from comparable corpora is a challenging task. The usual sentence alignment techniques applicable for parallel corpora rely on equivalent sentences and paragraphs, which have same order in the two parts of the bitext. Due to this the search space in sentence alignment is significantly reduced. This is not the case for comparable corpora, finding matching sentences and phrases remains a challenging task. Typically, comparable corpora don't have any information regarding document pair similarity. Generally, there exist many documents in one language which don't have any corresponding document in the other language. Also, when the correspondence information among the documents is available, the documents in question are not literal translations of each other. Thus, extracting parallel data from such corpora requires special algorithms designed for the corpora in question.

Depending upon the comparability of the comparable corpus, it is not sure that there will always be parallel sentences in comparable corpora rather there might be or might be not, but there could be parallel fragments in comparable sentences abundantly. Parallel fragments have also proved to be helpful for improving SMT performance (Munteanu and Marcu, 2006; Fu et al., 2013; Rahimi et al., 2014). An interesting aspect of using parallel fragments is that domain dependency is a bit alleviated, the fragments that are found are often named entities and everyday use phrases, thus adding out-domain sentence fragments to in-domain parallel data also helps improve MT performance as shown by Gupta et al. (2013) for English−Bengali language pair. Other than improving MT performance, fragments have also been helpful in other NLP domains requiring parallel data, Belz and Kow (2010) report parallel fragment extraction for improving Natural Language Generation(NLG) systems.

In this article, we present an efficient fragment extraction algorithm making use of information retrieval (IR) and SMT itself to retrieve parallel fragments. Firstly, the foreign language side of the comparable corpus is translated into English and potential matching sentences are retrieved from the English side of the comparable corpus (as described in Sections 5 and 6). Parallel phrase fragments are then identified using Levenshtein distance and the phrase alignment information. Our proposed scheme of fragment extraction is comparable in efficiency and results to all the previous works. We also present a detailed analysis of impact of fragments extracted from related vs nonrelated corpus. Other than the design efficiency of our approach, we also present a comparison of SMT improvement using fragments versus paralllel sentences extracted by Abdul-Rauf and Schwenk (2009a) from the same corpus thus highlighting the utility of fragments by comparative analysis.

We start by giving a brief overview of comparable corpora and the related work in this field followed by a high-level overview of our parallel sentence extraction system. Section 3 describes our fragment extraction framework followed by description of SMT and IR frameworks used in Section 5 and 6, respectively. Sections 7.2 and 7.1 report