ABSTRACT

Word alignment is the task of aligning bilingual words in a corpus of parallel sentences, and determining the probabilities for these aligned bilingual word pairs. It is the most important factor affecting the quality of any Statistical Machine Translation (SMT) systems. The IBM word alignment models are most well-known in the SMT research community. These models are pure statistical models and therefore they are not good for some language pairs which have differences in linguistic aspects (e.g. grammatical structures). This paper aims to improve the IBM models by using syntactic information. The authors first propose a new type of constraint based on bilingual syntactic patterns, and then integrate it into the IBM models. Finally, they show how to estimate the models' parameters using this new type of constraint. The experiments are conducted on the English-Vietnamese language pair for evaluation.

Keywords: EM Algorithm, IBM Models, Segment Constraint, Statistical Machine Translation, Syntactic Pattern, Word Alignment

1. INTRODUCTION

Word alignment, which is defined as an object for indicating the corresponding words in a parallel text (Liu et al., 2010), is the first step of most current approaches to SMT (Moore, 2005). Various methods have been proposed for finding word alignments between parallel sentences. These methods can be classified into two main approaches. The first one relies on discriminative training of a set of features, and thus it is more flexible to integrate new features, as shown in Moore (2005), Liu et al. (2005), and Ma et al. (2008). However, this approach has also faced some disadvantages as pointed out in Liu et al. (2010): (1) the need for annotated training data which is very expensive for the alignment task; (2) it is also difficult to select such a representative training corpus to ensure that the model will work well on unseen data, especially when the bilingual corpus to be aligned consists of parallel texts.
from different domains. The second approach is generative, and the typical models are IBM models 1-5 developed by Brown et al. (1993) which are widely used to align large bilingual corpora. This approach is pure statistical and therefore can cause wrong alignments because of the co-occurrence of words by chance. Some studies have used external knowledge sources such as lexical or syntactic information to alleviate this problem.

There have been several methods previously proposed for integrating external knowledge sources into alignment models. Och and Ney (2003) used a bilingual dictionary as an additional knowledge source for extending the training corpus. The authors assigned each entry in the dictionary, which really co-occur in the training corpus, a high weight and assigned the remaining entries a very low weight. Talbot (2005) proposed a method that uses auxiliary information sources such as cognate relations, bilingual dictionary, and numeric pattern matching to constrain the procedure directly by restricting the set of alignments explored during parameter estimation. Baobao et al. (2002) has extracted translation units from Chinese-English parallel corpora by using a set of predefined syntactic patterns. Holmqvist (2010) used alignment patterns (which uses part-of-speech tags and word forms) from the training data to align new bilingual texts. However, this method didn’t work well if the alignment patterns don’t appear in them.

Different from previous studies, our work focuses on using bilingual syntactic patterns to constrain the alignment candidates in the IBM models. It is worth noting that there are five models named IBM models 1-5, and each model in the series is extended from its predecessor. Therefore, the IBM model 1 is the essential component for all IBM models 1-5. In this paper, we first propose a new type of constraint based on bilingual syntactic patterns, called segment constraint, and then integrate it into the IBM model 1. After that, we show how to estimate the alignment parameters with the segment constraint. These improved IBM models will be then applied in SMT systems for the language pair of English and Vietnamese.

The rest of this paper is organized as follows. Section 2 presents the necessary backgrounds. Section 3 describes our proposed method to word alignment for SMT. Experimental results are shown in Section 4. Finally, conclusions are derived in Section 5.

2. BACKGROUND

2.1. Word Alignment

Given a source language sentence $f$ consisting of $J$ words $f_1,...,f_J$ and a target language sentence $e$ consisting of $I$ words $e_1,...,e_I$, the alignment $a$ between $e$ and $f$ is defined as a subset of the Cartesian product of the word positions:

$$a \subseteq \{(j,i) : j = 1,...,J; i = 0,...,I\}$$

The alignment $a$ connects words in the target language sentence to words in the source language sentence. The set of alignments $a$ is defined as the set of all possible connections between each word at position $i$ in the target language sentence to one word at position $j$ in the source language sentence. Figure 1 illustrates a word alignment between an English-Vietnamese sentence pair. The Vietnamese word $tôi$ is aligned to the English word $me$ because they are translations of one another. Similarly, the Vietnamese word $vượt$ is aligned to the English word $passed$, etc.

2.2. Statistical Machine Translation

The task of a SMT system is to model the translation probability $Pr(e|f)$ where the source sentence $f$ is translated into the target sentence $e$. Brown et al. (1993) use Bayes’ rule to formulate the translation probability for translating a source sentence $f$ into target sentence $e$ as follows:
Knowledge-Scientific Evaluation of a Social Service System
Fei Meng, Yoshiteru Nakamori and Van-Nam Huynh (2016). *International Journal of Knowledge and Systems Science* (pp. 60-77).
www.igi-global.com/article/knowledge-scientific-evaluation-of-a-social-service-system/163050?camid=4v1a

Evolution of IC Science and Beyond
www.igi-global.com/chapter/evolution-science-beyond/68208?camid=4v1a