Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 38 (2016) 13–27

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Phrase-boundary model for statistical machine translation

Shahram Salami[a,*], Mehrnoush Shamsfard[a], Shahram Khadivi[b]

[a] *Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran*
[b] *Human Language Technology Lab, Amirkabir University of Technology, Tehran, Iran*

Received 17 February 2015; received in revised form 6 September 2015; accepted 12 November 2015
Available online 12 December 2015

Abstract

This paper proposes a new probabilistic synchronous context-free grammar model for statistical machine translation. The model labels nonterminals with classes of boundary words on the target side of aligned phrase pairs. Labeling of the rules is performed with coarse grained and fine grained nonterminals using POS tags and word clusters trained on the target language corpus. Considering the large size of the proposed model due to the diversity of nonterminals, we have also proposed a novel approach for filtered rule extraction based on the alignment pattern of phrase pairs. Using limited patterns of rules, the extraction of hierarchical rules gets restricted from phrase pairs that are decomposable to two aligned subphrases. The proposed filtered rule extraction decreases the model size and the decoding time considerably with no significant impact on the translation quality. Using BLEU as a metric in our experiments, the proposed model achieved a notable improvement rate over the state-of-the-art hierarchical phrase-based model in the translation from Persian, French and Spanish to English language. This is applicable for all languages, even under-resourced ones having no linguistic tools.

Keywords: Statistical machine translation; Hierarchical models; Rules filtering

1. Introduction

The phrase-based model (Zens et al., 2002; Koehn et al., 2003) improved the previous state-of-the-art statistical machine translation approaches using phrases instead of words. The hierarchical phrase-based model (Chiang, 2005) started with the phrase-based model and supported the translation of longer phrases using hierarchical phrases. In that model, hierarchical phrases are induced from the parallel corpus by substituting subphrases with one generic nonterminal. The model suffers from high ambiguity and a very large number of grammar rules. Some solutions have been proposed to fix these problems. One set of solutions decreases the ambiguity of decoding using the context of input or syntactic knowledge which may reduce the model scope to languages that have linguistic tools. Another set of solutions filters grammar rules to reduce the model size without significantly affecting the translation quality.

This paper proposes the phrase-boundary model as a hierarchical model in which both grammar rules and nonterminals are induced from the parallel corpus without using syntactic trees. This model has a higher precision than the hierarchical phrase-based model with one generic nonterminal. The nonterminals of phrase-boundary grammar

[*] Corresponding author. Tel.: +98 21 29904111.
E-mail addresses: sh_salami@sbu.ac.ir (S. Salami), m-shams@sbu.ac.ir (M. Shamsfard), khadivi@aut.ac.ir (S. Khadivi).

http://dx.doi.org/10.1016/j.csl.2015.11.005

are defined by boundary word classes on the target side of aligned phrase pairs. In other words, each phrase pair is known by the first and last word of its target side. For example, the French-English phrase pair < idée possible, feasible idea > is known as *JJ-NN* by the concatenation of the boundary POS tags on the English side. Thus, the following rule exists in the phrase-boundary model:

$$JJ\text{-}NN \rightarrow < \text{ idé e possible, feasible idea } > \tag{1}$$

Instead of *JJ-NN* label, the left hand side of this rule is labeled with the generic nonterminal *X* in the hierarchical phrase-based model. Additionally, word classes can be determined by clustering words on the target training corpus. In this case, the words of the target corpus are clustered in an arbitrary number of word classes. Automatic clustering of the words generalizes the usability of the model to the language pairs for which no linguistic tool may be available. Although defining nonterminals based on source word classes is straightforward, the use of target side word classes is proposed. This is because the translation output can be structured based on the target syntax in the rules, while decoding is directed by the input. Target side syntax has also been used in other models such as SAMT (Zollmann and Venugopal, 2006). Although the use of both side word classes is possible, it may lead to model sparseness.

In the absence of syntactic categories, the hierarchical phrase-based model uses one generic nonterminal, while the granularity of nonterminals in the proposed phrase-boundary model depends on the number of word classes used. Nonterminal diversity increases the precision of the phrase-boundary model. On the other hand, phrases with the same syntactic category may have different nonterminals in the phrase-boundary grammar, which may result in an extra number of non-overlapped hierarchical rules in the model. Larger models consume more time and memory in the training stage and decoding process. According to the fact that hierarchical rules represent the decomposition pattern of phrase pairs, this paper proposes a novel approach to filter out grammar rules based on the alignment pattern of the phrase pairs. This filtering limits the patterns of the hierarchical rules extracted from phrase pairs that are decomposable to two aligned subphrases.

This study examined filtered and non-filtered phrase-boundary models with POS tags and automatic word clustering in language pairs with similar and different word ordering. The experiments showed that the proposed filter reduces model size by more than 70% without significantly impacting translation quality. Furthermore, experimental results demonstrated that the phrase-boundary model is a considerable improvement over the hierarchical phrase-based model using BLEU metric for evaluation. The design and implementation method of the proposed model is presented in this paper. Related work is referenced in Section 2. Section 3 explains the proposed phrase-boundary model. Section 4 introduces the methods for filtered rule extraction. Section 5 shows the results of experiments. Finally, the paper is concluded in Section 6.

2. Related work

Unsupervised grammar induction based on phrase alignment was introduced by Chiang (2005) as hierarchical phrase-based model. This model labeled all nonterminals with one generic label. For better rule selection in the decoding process of the hierarchical phrase-based model, the context of input was used in forms such as POS tags (He et al., 2008) and CCG (Combinatory Categorial Grammar) tags (Haque et al., 2010). Discontinuous generation of target words limits pruning of the decoding space with the target language model. Watanabe et al. (2006) generated a target sentence in left-to-right order using hierarchical rules, the target sides of which were in the Greibach Normal Form class. Another work limited decoding space of the hierarchical phrase-based model by avoiding the recursion of hierarchical rules (Huck et al., 2012). It used two different nonterminals on the left and right hand sides of hierarchical rules. Zhou et al. (2008) scored derivations during translation decoding using syntactic knowledge. The hierarchical phrase-based model was augmented with syntactic categories (Zollmann and Venugopal, 2006) and CCG tags (Almaghout et al., 2010) to increase model precision. The precision of the proposed hierarchical model was increased with various nonterminals and without using syntactic categories.

There is a long history of using word classes in statistical machine translation. The alignment template model (Och and Ney, 2004) uses word classes to explicitly define word reordering. In some recent work, the classes of the boundary words in the aligned phrases are used to improve reordering in the hierarchical phrase-based model (Huck et al., 2012) and the phrase-based model (Cherry, 2013). Vilar et al. (2010) induced grammar nonterminals from the training corpus by clustering aligned phrases based on the source and target word classes. They used all words in the phrases, while