Expert Systems With Applications 84 (2017) 186–199

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Syntax- and semantic-based reordering in hierarchical phrase-based statistical machine translation

Arefeh Kazemi[a], Antonio Toral[b], Andy Way[c], Amirhassan Monadjemi[a,*], Mohammadali Nematbakhsh[a]

[a] Department of Computer Engineering, University of Isfahan, Iran
[b] University of Groningen, The Netherlands
[c] ADAPT Centre, School of Computing, Dublin City University, Ireland

ARTICLE INFO

*Article history:*
Received 28 December 2016
Revised 1 May 2017
Accepted 2 May 2017
Available online 5 May 2017

*Keywords:*
Statistical machine translation
Reordering
Dependency structure
HPB-SMT
Syntax
Semantic

ABSTRACT

We present a syntax-based reordering model (RM) for hierarchical phrase-based statistical machine translation (HPB-SMT) enriched with semantic features. Our model brings a number of novel contributions: (i) while the previous dependency-based RM is limited to the reordering of head and dependant constituent pairs, we also model the reordering of pairs of dependants; (ii) Our model is enriched with semantic features (Wordnet synsets) in order to allow the reordering model to generalize to pairs not seen in training but with equivalent meaning. (iii) We evaluate our model on two language directions: English-to-Farsi and English-to-Turkish. These language pairs are particularly challenging due to the free word order, rich morphology and lack of resources of the target languages.

We evaluate our RM both intrinsically (accuracy of the RM classifier) and extrinsically (MT). Our best configuration outperforms the baseline classifier by 5–29% on pairs of dependants and by 12–30% on head and dependant pairs while the improvement on MT ranges between 1.6% and 5.5% relative in terms of BLEU depending on language pair and domain. We also analyze the value of the feature weights to obtain further insights on the impact of the reordering-related features in the HPB-SMT model. We observe that the features of our RM are assigned significant weights and that our features are complementary to the reordering feature included by default in the HPB-SMT model.

## 1. Introduction

Statistical machine translation (SMT) is a data-driven approach for translating from one natural language into another. It can be used to attain fully automatic high quality translation in some specific domains, gisting, translation on hand-held devices, assisting human translators, etc (Koehn, 2012). Apart from the usefulness of SMT systems as standalone applications, they can also play a relevant role in broader applications including a wide range of intelligent systems such as multilingual customer service chat bots and intelligent tourist guide systems, cross-lingual recommendation systems, etc. SMT uses statistical methods to examine collections of human-produced translations and automatically learn how to produce optimum translations. Natural languages vary greatly not only in their vocabularies, but also in the manner that they arrange words in a sentence. Accordingly, the SMT task consists of two interrelated problems: finding the appropriate words in the translation and predicting their order in the target-language sentence (“reordering”).

Reordering is a fundamental problem in SMT that significantly affects the quality of the final translation. The amount of reordering needed varies for different language pairs. For translation between language pairs with similar or close word order, such as French–English, reordering can be restricted to short, local movements. In contrast, language pairs with different syntactic structure and major differences in word order, such as Farsi–English, Turkish–English and Japanese–English, are more challenging because they require long-distance movements. The basic word order of the first language in these pairs (Farsi, Turkish and Japanese) is primarily Subject-Object-Verb, while the word order of the second (English) is Subject-Verb-Object. Moving the verb from the end of the sentence to the position just after the subject results in a movement over a potentially large number of words for mid-length and long sentences.

* Corresponding author.
*E-mail addresses:* kazemi@eng.ui.ac.ir (A. Kazemi), a.toral.ruiz@rug.nl (A. Toral), andy.way@adaptcentre.ie (A. Way), monadjemi@eng.ui.ac.ir, ui.paper.submission@gmail.com (A. Monadjemi), nematbakhsh@eng.ui.ac.ir (M. Nematbakhsh).

http://dx.doi.org/10.1016/j.eswa.2017.05.001
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

Due to the high computational complexity of exploring all possible word orders, a standard phrase-based SMT system (PB-SMT) (Koehn et al., 2003b) only allows local and short movements and penalizes long-distance reordering. This means that a standard PB-SMT system is able to perform the required reordering for similar language pairs relatively well. However, it does not have any principled fashion to deal with long-distance reordering, which is needed for language pairs with differing word orders. There are two possible ways to overcome this limitation: (i) incorporating a reordering model (RM) into the PB-SMT system; and (ii) using a tree-based SMT system. A RM tries to capture differences in word order in a probabilistic framework, by assigning a probability to each possible order of words. It allows the SMT system to consider not only local movements, but also those long-distance movements which have a high probability in the RM framework. Different proposed RMs for PB-SMT can be generally categorized into three main groups: distance-based models (Koehn, Och, & Marcu, 2003a), lexical phrase-based models (Koehn et al., 2005; Tillmann, 2004) and hierarchical phrase-based models (Galley & Manning, 2008). Tree-based SMT systems (e.g. HPB-SMT Chiang, 2005) merge the translation model and reordering model in a synchronous grammar. Although all of these approaches have successfully improved SMT systems, a reordering approach may suffer from the following flaws:

- *Inability to perform long-distance reordering*. In PB-SMT, reordering can be generally handled by distance-based, lexical or hierarchical models. Distance-based RMs generally prefer monotone decoding (Koehn et al., 2003a), so long-distance reordering would be penalized heavily by these models. Lexical phrase-based models assign reordering probability to adjacent words or phrases. As a consequence, they are able to perform local reordering between neighboring units successfully, but they fail to capture long-distance reordering. Hierarchical RMs merge neighboring phrases and try to predict the ordering of the new phrases. This model outperforms lexical and distance-based RMs as it is able to capture longer-distance reorderings. However, it cannot capture all of the required long-distance reorderings due to data sparseness.

  The weakness of PB-SMT systems in handling long-distance reordering motivates research on tree-based SMT, such as HPB-SMT (Chiang, 2005). Despite the good performance offered by the HPB-SMT in medium-range reordering, it has been shown that it still performs weakly on long-distance reordering (Birch, Blunsom, & Osborne, 2009).

  In order to capture long-distance reorderings, modern approaches try to consider reordering not only between adjacent words, but also between words with longer distance. These words can be simply all the word pairs in the source sentence (Hayashi, Tsukada, Sudoh, Duh, & Yamamoto, 2010) or the words that are in syntactic relation to each other (Gao et al., 2011; Huang et al., 2013)
- *Inability to generalize*. Some of the approaches can perform reordering of common words or phrases relatively well, but they have no ability to generalize to unseen words with the same linguistic structure. For example, if in the source language the object follows the verb and in the target language it precedes the verb, these models still need to see particular instances of verbs and objects in the training data to be able to perform reordering between them; the fact that this may be a regular pattern over a class of lexical items cannot be captured in a global operation (unlike might be the case in rule-based MT). In order to overcome this limitation, some reordering approaches use extracted features from phrases and use them in the model, instead of the phrase itself (Xiong, Liu, & Lin, 2006; Zens & Ney, 2006).
- *Context insensitivity*. Lexical and hierarchical models try to predict the ordering of the phrases based solely on the surface form of phrases. Distance-based models focus on the position of phrases in the source and target sentences to determine the ordering between them. However, reordering will not regularly occur in the same way for the same phrases or same phrase positions over different contexts. It is, therefore, essential to include context information in order to accurately capture the reordering behavior.

This paper presents a syntax-based RM for HPB-SMT enriched with semantic features. It extends our previous works (Kazemi, Toral, & Way, 2016; Kazemi, Toral, Way, Monadjemi, & Nematbakhsh, 2015) with major new additions, including: (i) switching the classifier from Naive Bayes to Maximum Entropy, (ii) improving the experimental set-up and obtaining more stable results by averaging over three tuning runs, (iii) evaluating the proposed method on an additional language pair (English–Turkish) and also on an additional corpus from a different domain for English–Farsi, (iv) comparing the proposed model to two state-of-the-art RMs, (v) analyzing the feature weights, and (vi) adding translation examples.

The proposed method: (i) is able to perform long-distance reordering as it is based on dependency information; (ii) can generalize to words unseen in the training data that hold the same syntactic and semantic structures, by using features based on syntactic and semantic information, respectively; (iii) it is sensitive to context as it uses the dependency relations of the words being reordered; and (iv) due to the fact that the proposed model uses only source-side information, the model can be precomputed offline, which leads to low added complexity during decoding.

The main contributions are as follows:

- We introduce a novel syntax-based RM based on the dependency structure of the source sentence. While the state-of-the-art dependency-based RM (Gao, Koehn, & Birch, 2011) is limited to the reordering of head and dependant constituent pairs, we also model the reordering of pairs of dependants.
- Our model is enriched with semantic features in order to allow the RM to generalize to pairs not seen in training but with equivalent meaning. While semantic structures such as Predicate-Argument-Structure (PAS) and Semantic-Role-Labeling (SRL) have been previously used for MT reordering, e.g. Liu and Gildea (2010), Xiong, Zhang, and Li (2012), Li, Resnik, and III (2013), in this work we include semantic features jointly with lexical and syntactic features in the framework of a syntax-based RM.

  Our model is evaluated on the English-to-Farsi and English-to-Turkish language pairs. Designing a reordering model for these pairs is particularly challenging because the target languages (Farsi and Turkish) are free word order, rich in morphology and comparatively under-resourced.

The remainder of this paper is organized as follows. Section 2 reviews the related work and puts our work in its proper context. Section 3 presents in detail our RM, including its conceptualization, the features used, its training regime, and its integration in HPB-SMT. Section 4 contains the experiments carried out to evaluate our RM. This is followed by more in-depth analyzes in Section 5. Finally, we outline conclusions and lines of future work in Section 7.