University of Groningen

Syntax-and Semantic-based Reordering in Hierarchical Phrase-based Statistical Machine Translation

Kazemi, Arefeh; Toral Ruiz, Antonio; Way, Andy; Monadjemi, Amirhassan; Nematbakhsh, Mohammadali

Published in:
Expert systems with applications

DOI:
[10.1016/j.eswa.2017.05.001](https://doi.org/10.1016/j.eswa.2017.05.001)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Kazemi, A., Toral Ruiz, A., Way, A., Monadjemi, A., & Nematbakhsh, M. (2017). Syntax-and Semantic-based Reordering in Hierarchical Phrase-based Statistical Machine Translation. *Expert systems with applications*, 84, 186-199. <https://doi.org/10.1016/j.eswa.2017.05.001>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

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.

© 2017 Elsevier Ltd. All rights reserved.

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).

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 pre-computed 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 Ill (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.

2. Related work

2.1. Reordering in PB-SMT systems

PB-SMT systems translate phrases as entire units, so they can perform a considerable amount of local (short-distance) reordering occurring within phrases. However, non-local (medium- and long-distance) reordering between phrases cannot be captured by standard PB-SMT systems. Previous work on reordering in PB-SMT can be generally categorized into two groups.

2.1.1. Reordering as pre-processing

Considering the complexity of solving reordering during the decoding process, approaches in the first group separate reordering from the translation process and perform it before decoding, as a pre-processing step. These approaches apply some reordering rules to the source sentence to render its order more similar to the order of the target language, so that the translation process can be then performed monotonically. The reordering rules can be manually written linguistic rules (Collins, Koehn, & Kučerová, 2005) or automatically learned from data (Genzel, 2010; Xia & McCord, 2004). Although all of these approaches result in significant improvement when added to a PB-SMT system, there is a fundamental problem with separating the reordering task from the translation process (Koehn, 2012). The reordering rules might be unreliable, so applying them to the source sentences can lead to making faulty reordering decisions in the pre-processing step. Such erroneous decisions will be passed along as hard decisions which cannot be reversed by the SMT system. This also violates one of the main principles underpinning the probabilistic modeling of SMT, namely the avoidance of hard decisions and maintaining the possibility of reversing previous faulty decisions.

2.1.2. Reordering during decoding

Approaches in the second group address reordering during the decoding step, as part of the translation process. They handle reordering with a probabilistic RM which assigns probabilities to each possible ordering of the target sentence. This probability is then used in combination with the other probabilistic models in SMT to find the most likely output translation. Approaches in this line range from simple distance-based models to lexical and hierarchical phrase orientation models. Distance-based models (Koehn et al., 2003a) generally prefer monotone translation and penalize any deviation from this norm. Although monotone translation may work for similar language pairs, it is not a realistic framework for language pairs with different syntactic structure which require long-distance reordering. On top of this limitation, these models are not sensitive either to the specific word being translated nor to its context and thus they do not generalize well.

Given the fact that some phrases are reordered more frequently than others, lexical phrase orientation models take content into consideration and condition orientation on actual phrases (Koehn et al., 2005; Tillmann, 2004). These models try to predict the orientation of a source phrase with respect to its neighboring phrases, by learning patterns from training data. Despite the satisfactory improvement of lexical models over pure PB-SMT systems and distance-based RMs, they continue to suffer from two major limitations. First, these models have no ability to be generalized to unseen phrases, or phrases with low frequency in the training data. In order to overcome this limitation, some approaches extract features from phrases and use them in the model, instead of the phrase itself (Xiong et al., 2006; Zens & Ney, 2006). Second, these models cannot capture long- and even medium-distance reordering, since they predict orientation solely between adjacent phrases. To handle longer distance reorderings, some approaches move beyond short and adjacent phrases and attempt to predict the ori-

entation between non-adjacent words (Bisazza & Federico, 2013) or longer phrases (Galley & Manning, 2008). Word-after-word reordering model tries to predict the orientation of a source word in a given position with respect to another, based on the words at those positions and their context (Bisazza & Federico, 2013). Hierarchical phrase orientation models merge adjacent phrases together and form longer phrases and then try to predict orientation between these new phrases (Galley & Manning, 2008). Since these models predict orientation for longer phrases, they are able to capture longer distance reordering compared with lexical models. However, these models are still weak on long-distance reordering as they are still restricted to phrases with no gaps and adjacent phrases. This model is also limited by a maximum reordering distance often used in decoding which allows only a limited number of adjacent phrases to be merged.

2.2. Reordering in HPB-SMT systems

One of the reasons why HPB-SMT was proposed was to overcome the weakness of PB-SMT systems in handling long-distance reordering (Chiang, 2005). Rather than using only continuous source words as phrases, HPB-SMT uses hierarchical phrases which contain gaps, so it should lead to better reordering than PB-SMT. However, it has been shown that despite the superiority of HPB-SMT over PB-SMT in medium-distance reordering, it is equally weak in long-distance reordering (Birch et al., 2009). A number of approaches have been proposed to improve reordering in HPB-SMT. Common approaches include handling reordering as a pre-processing step (Xu, Kang, Ringgaard, & Och, 2009) and labeling non-terminals in the translation rules (Zollmann & Venugopal, 2006). The first approach suffers from the same issues presented in Section 2.1.1 for performing reordering as a pre-processing step in PB-SMT. The second causes additional complexity to the system and increases data sparsity (Hanneman & Lavie, 2013).

Other work tries to integrate a RM into HPB-SMT as an external feature function. Huck, Wuebker, Rietig, and Ney (2013) and Nguyen and Vogel (2013) successfully incorporated lexical phrase orientation model and distortion models – initially proposed for PB-SMT – into HPB-SMT. Hayashi et al. (2010) integrated the word reordering model of Tromble and Eisner (2009) into HPB-SMT. This model tries to predict ordering of any two words in the source sentence. He, Meng, and Yu (2010) added a swap glue grammar rule ($X \rightarrow X_1X_2|X_2X_1$) to the hierarchical model, which allows phrases to be swapped. They use a maximum entropy classifier to predict the probability of swapping for each neighboring phrase pairs. Rather than conditioning on the actual phrases, they use some extracted features from them, such as the first and last word of the source and target phrases.

Our work falls into a recent line of research which has focused on using extracted information from the syntactic parse tree of the source sentence to predict the orientation of source elements. Work in this line differs in two respects: (i) the type of the parser that they use to extract syntactic information; and (ii) the type of the source elements from which they try to predict orientations. Gao et al. (2011) use the dependency parse tree of the source sentence and try to predict the monotone or swap orientation of each dependant with respect to its head. They used a maximum entropy classifier which uses the surface form of the head and dependants and the dependency relation as features. Rather than using a dependency parse tree, Huang, Devlin, and Zbib (2013) use constituency parses and train a maximum entropy classifier to predict the orientation between each pair of words. It is worth noting that to date, all these approaches have been applied solely to one language pair, namely Chinese-to-English.

Our RM can be compared to that of Huck et al. (2013), Hayashi et al. (2010) and Gao et al. (2011). All these approaches try to pre-

Fig. 1. An example dependency tree for an English source sentence, its translation in Farsi and the word alignments.

dict the orientation for two source elements. However, they differ in how they select the type of these elements. Huck et al. (2013) predict orientation for adjacent hierarchical phrases. Despite the satisfactory performance of this model on handling short- and medium-distance reorderings, it cannot capture long distance reorderings. In contrast, Hayashi et al. (2010) predict orientation for all word pairs in the source sentence. This model has a high computational complexity since it considers any two words in the source sentence for reordering. That is, for a sentence of length n , this model predicts orientation for $O(n^2)$ pairs. In addition to this, as we will show later in Section 4, considering more word pairs for reordering does not always lead to better results. Accordingly, it is essential to strike the right balance between these approaches by selecting the right word pairs for reordering (Zhang, Utiyama, Sumita, Zhao, & Neub, 2016). Gao et al. (2011) use dependency structure and predict the orientation of each dependant with respect to its head. In the proposed method, we predict the orientation between each pair of dependants.

2.3. Reordering using semantic information

A recent line of research has focused on incorporating semantic information into the RM. To the best of our knowledge, all research in this strand has used PAS and SRL as the basis of the model. Some work uses PAS to perform reordering as a pre-processing or post-processing step. Komachi, Matsumoto, and Nagata (2006) and Wu, Sudoh, Duh, Tsukada, and Nagata (2011) use some reordering rules based on the PAS of the source sentence in order to reorder source arguments in the pre-processing step. The pre-reordering rules are either written manually (Komachi et al., 2006) or extracted automatically (Wu et al., 2011). Zhai, Zhang, Zhou, and Zong (2012) use PAS transformation rules to convert the PAS of the source sentences to target-like PAS. Wu et al. (2011) perform reordering as a post-processing step. They use the PAS of the output translation and reorder the arguments to match the semantic role of the output to that of the source sentence. Rather than using semantics in pre-processing or post-processing steps, some other work integrates semantic-based features into the decoder. Liu and Gildea (2010) design a semantic-role reordering feature which tries to model the reordering of the source-side semantic roles in the target side and integrate them into a tree-to-string SMT system. Xiong et al. (2012) and Li et al. (2013) perform semantic role labeling on the source sentence and try to predict the translation order of two arguments (Xiong et al., 2012) or an argument and its predicate (Li et al., 2013).

Our work differs from existing work in the following respects:

1. We propose a dependency-based RM which predicts the orientation not only between a dependant and its head (Gao et al., 2011), but also between pairs of dependants.
2. Rather than using a *semantic structure* (e.g. PAS) as the basis of the RM and trying to predict orientation between semantic constituents, we use the syntactic structure and add semantics to our model by using *semantic features*. This is motivated by two issues presented by semantic-based RMs that use PAS; compared to syntax-based models, these models have fewer opportunities to influence reordering, and reordering predictions are worse as the linguistic phenomena involved are harder to classify (Li, Marton, Resnik, & Daumé III, 2014). Specifically, we use WordNet (Miller, 1995) to incorporate semantics into our RM.
3. We apply dependency-based RMs to English-to-Farsi and English-to-Turkish. Our target languages are (i) free word order, (ii) rich in morphology and (iii) comparatively under-resourced.

3. Dependency-based reordering model

This section covers our RM in detail. First, we introduce the idea and concepts behind our model (Section 3.1). Then we describe the implementation of our model as a classifier, its features and training (Section 3.2). Finally, we show how the model has been integrated into the HPB-SMT model (Section 3.3).

3.1. Rationale

The proposed RM is based on the dependency parse tree of the source sentence, an example of which is shown, for English, in Fig. 1. The dependency tree of a sentence shows the grammatical relations between pairs of head and dependants of that sentence. For example, for the sentence in Fig. 1, the arrow with label “obj” from “tent” to “bought” indicates that the dependant “tent” is the object of the head word “bought”. Based on the assumption that constituents move as a whole in translation (Quirk, Menezes, & Cherry, 2005), our model takes the dependency tree of a source sentence and tries to find the ordering of its counter-part constituents. To be more precise, our model aims to predict the orientation of a dependant with respect to its head (*head-dep*), and also with respect to other dependants of that head (*dep-dep*). We consider two types of orientation: *monotone* (M) and *swap* (S). If the order of two words S_1 and S_2 in the source sentence is the same as the order of their translations T_1 and T_2 in the target sentence, the orientation is monotone and otherwise it is swap. For example, for the sentence in Fig. 1, our model tries to predict the orientations (ori) between the (*head-dep*) and (*dep-dep*) pairs that are shown in Table 1.

Table 1
head–dep and dep–dep pairs for the sentence in Fig. 1 and their corresponding orientations when translating into Farsi.

Head	Dependant	ori	dep1	dep2	ori
bought	he	M	he	camels	M
bought	camels	S	he	tent	M
bought	tent	S	he	wife	M
bought	wife	S	he	wandered	M
bought	wandered	M	camels	tent	M
wife	a	M	camels	wife	M
tent	a	M	camels	wandered	M
tribes	wandered	S	tent	wife	M
			tent	wandered	M
			wife	wandered	M

Table 2
Features for the (head–dep) pairs.

Feature	Description
<i>Lex(head)</i>	The surface form of the head word
<i>Lex(dep)</i>	The surface form of the dependant
<i>depRel(dep)</i>	The dependency relation of the dependant
<i>syn(head)</i>	The synset of the head word
<i>syn(dep)</i>	The synset of the dependant

Our motivation for using the dependency structure of the source sentence as the basis of our RM comes from the assumption that, if it is the case that a reordering pattern has been applied for translating one English sentence with a specific structure into Farsi, this pattern will also work for translating another English sentence with the same dependency structure. For example, in translating from English into Farsi, all of the following English sentences follow the same word order in their translation in Farsi: “he puts the desk on the ground”, “I put the book on the table”, and “he puts his hand on her shoulder”. In general, almost all sentences with the dependency structure “subject” put “object” “preposition-on” have the same order in their Farsi translation: “subject” “object” “preposition-on” “translation of verb put”.

3.2. Classifier

Since the proposed RM aims to classify the (head–dep) and (dep–dep) pairs into their correct orientation class (monotone or swap), its task can be seen as a binary classification task. In order to build such a classifier, we used a Maximum Entropy (henceforth maxEnt) classifier. Instead of using maxEnt to perform a hard classification, we use it to estimate the probability distribution over two orientation classes. The maxEnt classifier estimates the probability of the orientation type *ori* given the constituent pair *pair* as shown in Eq. (1) (Manning & Klein, 2003), where $h_n : n = 1 \dots N$ are N binary features whose values depend on the constituent pair, *pair* and orientation type *ori* and λ_n are the weights of these features.

$$P(ori|pair) = \frac{\exp \sum_{n=0}^N \lambda_n h_n(ori, pair)}{\sum_{ori' \in \{monotone, swap\}} \exp \sum_{n=0}^N \lambda_n h_n(ori', pair)} \quad (1)$$

Various features can be used to characterize constituent pairs in the maxEnt model. We used only source-side features for two reasons. First, this enabled us to extract the features offline, as a pre-processing step, so as not to slow down the decoding phase. Second, it has been shown that source-side features largely outperform target-side features (Zens & Ney, 2006). Tables 2 and 3 show the features that we used for (head–dep) and (dep–dep) pairs, respectively. Given a word *w*, *synset(w)* shows the synonym set of *w* as found in Wordnet (Fellbaum, 1998), a lexical database of English

Table 3
Features for the (dep–dep) pairs.

Feature	Description
<i>Lex(head)</i>	The surface form of the mutual head word
<i>Lex(dep1)</i>	The surface form of the first dependant
<i>Lex(dep2)</i>	The surface form of the second dependant
<i>depRel(dep1)</i>	The dependency relation of the first dependant
<i>depRel(dep2)</i>	The dependency relation of the second dependant
<i>syn(head)</i>	The synset of the mutual head word
<i>syn(dep1)</i>	The synset of the first dependant
<i>syn(dep2)</i>	The synset of the second dependant

Table 4
Features and their values for the (dep–dep) pair (“camels”, “tent”) in Fig. 1.

Features	Values
<i>lex(head)</i> , <i>lex(dep1)</i> , <i>lex(dep2)</i>	bought, camels, tent
<i>syn(head)</i> , <i>syn(dep1)</i> , <i>syn(dep2)</i>	02186766-V, 02414816-N, 04359185-N
<i>depRel(dep1)</i> , <i>depRel(dep2)</i>	doj, doj

which groups words into sets of cognitive synonyms called *synsets*. In other words, in WordNet a set of synonym words belong to the same synset. For example, the words “story”, “narrative” and “tale” are in the same synset.

As an example, Table 4 shows the features that we use for the (dep–dep) pair “camels” and “tent” in our example in Fig. 1. Synsets are represented by their unique identifiers in WordNet.

To the best of our knowledge, previous work on syntax-based reordering have considered lexical and syntactic features, e.g. surface forms of the words, part-of-speech tags and syntactic labels (cf. related work in Section 2). Our model uses lexical, syntactic and also semantic features. The use of syntactic features, such as dependency relations, enables us to capture the general syntactic reordering patterns that are not available to pure lexical models. For instance, our model can learn that when translating between SVO and SOV language pairs, the object and the verb should be swapped. In addition to this syntactic generalization, our model can be generalized from words seen in the training data to any of their synonyms by using synsets as additional features.

In order to generate training examples for the maxEnt model, we use the dependency parse tree of the source sentence which is obtained automatically with a dependency parser and the word alignments between the source and target words in the parallel corpus. We extract all possible (head–dep) and (dep–dep) pairs for each sentence and determine the orientation type for each pair. With the alignment points (P_{S1} , P_{T1}) and (P_{S2} , P_{T2}) for two source words ($S1$, $S2$) and their aligned target words ($T1$, $T2$), we define the orientation type (*ori*) as shown in Eq. (2). If a source word is aligned to more than one target word¹, we only consider the last aligned target word in determining the orientation type. Using the position of the first or last aligned target word is a heuristic way that has been used previously (Gao et al., 2011). In order to choose between the first or the last aligned word, we conducted a small experiment and extracted 200 English words from the parallel corpus which were aligned to more than one Farsi word. We observed that in 67% of them, the correct alignment is the last aligned one.

$$ori = \begin{cases} monotone, & \text{if } (p_{S1} - p_{S2}) \times (p_{T1} - p_{T2}) > 0 \\ swap, & \text{otherwise.} \end{cases} \quad (2)$$

For example, for the pair (*bought*, *camels*) in Fig. 1 with alignment points (1,7) and (2,2) for the head word *bought* and the dependant

¹ In the parallel corpora used in our experiments only 9% of the source words are aligned to more than one target word.

camels respectively, we consider swap orientation between *bought* and *camels* based on Eq. (2).

Once we have obtained the orientation type for each constituent pair in the training part of our parallel corpus, we train the maxEnt classifier to estimate the probability of a source dependent word having monotone or swap order with respect to its head and the other dependants of that head.

3.3. Integration into HPB-SMT

During translation, the HPB-SMT decoder searches for the most probable translation hypothesis based on the combination of several feature functions. To do so, it estimates the probability of a translation hypothesis H given the source sentence S through a log-linear combination of several feature functions, as shown in Eq. (3) where a is the latent word alignment between H and S , $f f_i$ are N arbitrary feature functions and w_i are the weights of these features. w_i shows the importance of $f f_i$ in the log-linear model.

$$\begin{aligned} \log P(H|S) \\ = \sum_{i \in \{\text{monotone}, \text{swap}, \text{dependency-coherence}, \text{unaligned-pairs}\}} \log(w_i f f_i(H, S, a)) \end{aligned} \quad (3)$$

We implemented our RM as four feature functions: RM0: $f f_{\text{monotone}}$, RM1: $f f_{\text{swap}}$, RM2: $f f_{\text{dependency-coherence}}$ and RM3: $f f_{\text{unaligned-pairs}}$. $f f_{\text{monotone}}$ and $f f_{\text{swap}}$ compute the sum of orientation probability of those constituent pairs which are translated in monotone or swap order, respectively. $f f_{\text{dependency-coherence}}$ shows the number of constituent pairs which are translated in the hypothesis and encourage concurrent translation of constituents, based on the assumption that constituents move together in the translation process (Quirk et al., 2005). These three feature functions can be computed for a constituent pair only if both of the words in the pair are aligned to at least one target word. The feature function $f f_{\text{unaligned-pairs}}$ counts the number of pairs with at least one unaligned word and penalizes those hypotheses on which the orientation features could not be applied. After computing the four feature functions for the translation hypothesis, we combine them with the other feature functions in the HPB-SMT model, as shown in Eq. (3).

In order to compute our feature functions for hypothesis H which has been applied to source sentence S with constituent pairs $\text{Pairs}(S)$, we perform three steps:

1. Determining the constituent pairs of the source sentence S that have been covered by hypothesis H . We call this set $\text{Covered}(H, \text{Pairs}(S))$. A constituent pair is covered by hypothesis H , if H covers both words in the pair. We compute the dependency-coherence feature as the number of covered constituent pairs by hypothesis H , as in Eq. (4).

$$f f_{\text{dependency-coherence}}(H, S) = |\text{Covered}(H, \text{Pairs}(S))| \quad (4)$$

2. Determining the aligned covered pairs based on the word alignment within the phrases. A constituent pair is aligned if both words in the pair are aligned to at least one target word. We call the set of aligned covered pairs $\text{Aligned}(H, \text{Pairs}(S), a)$, where a is the word alignment between S and H . It might be the case that a word in a constituent pair is not aligned to any target word. In this case, we cannot determine the orientation and compute swap or monotone features for this pair. As this may lead to a search error (Gao et al., 2011), we apply a penalty by means of an unaligned-pairs feature $f f_{\text{unaligned-pairs}}$ computed as the number of covered constituent pairs with at

least one unaligned word, as in Eq. (5).

$$f f_{\text{unaligned-pairs}}(H, S, a) = |\text{Covered}(H, \text{Pairs}(S))| - |\text{Aligned}(H, \text{Pairs}(S), a)| \quad (5)$$

3. Determining the orientation class and its probability for each aligned covered pair. We determine the orientation class and its probability based on Eqs. (2) and 1, respectively. Based on the orientation class for a pair, we consider its score for calculating monotone or swap feature functions. We compute monotone and swap features as in Eq. (6).

$$\begin{aligned} f f_{\text{monotone}}(H, S, a) = \sum_{\text{pair} \in \text{Aligned}(H, \text{Pairs}(S), a), \text{ori}(\text{pair}) = \text{monotone}} P(\text{ori}|\text{pair}) \\ f f_{\text{swap}}(H, S, a) = \sum_{\text{pair} \in \text{Aligned}(H, \text{Pairs}(S), a), \text{ori}(\text{pair}) = \text{swap}} P(\text{ori}|\text{pair}) \end{aligned} \quad (6)$$

It is worth mentioning that although we have integrated the proposed RM into HPB-SMT, integrating it into other tree-based MT paradigms is straightforward and does not require any crucial modification. In order to integrate the proposed model into PB-SMT, the distortion-limit should be set to a large number, to enable the system to produce hypotheses with long-distance reorderings, which can be then scored by the proposed RM.

4. Experiments

4.1. Set-up

We used an English–Turkish parallel corpus and two English–Farsi parallel corpora with different sizes and from different domains to evaluate the proposed method. The English–Turkish corpus (En–Tr)² (Oflazer, 2008) is composed of documents about international relations and legal documents from sources such as the Turkish Ministry of Foreign Affairs, the European Union, etc. It consists of approximately 100,000 sentences. The Mizan corpus³ (Supreme Council of Information & Communication Technology, 2013) is an English–Farsi corpus, extracted from novels and their translations and contains about 1 million sentence pairs. Tep++ (Passban, Way, & Liu, 2015) is another English–Farsi corpus, gathered from film subtitles and consisting of about 600,000 sentences. For each experiment, training and test sets were selected randomly from the corpus used in that experiment, with the remaining part of the corpus used for training. Table 5 gives the statistics of training, tuning and test sets for the parallel corpora.

Knowing the required reordering types for a specific language pair is an important step in designing a suitable RM for them (Birch, 2011). Therefore, we analyze the parallel corpora by extracting the size and number of reorderings detected in them (further details can be found in Birch (2011)). Fig. 2 shows the distribution of reorderings in the parallel corpora broken down by the width of the reorderings. As shown in the figure, Tep++ has many more short distance reorderings, but it has fewer reorderings over medium and long distances.⁴ There are two reasons for this. First, the average sentence length for Tep++ is shorter than that of Mizan and En–Tr (8.08 for Tep++ vs. 13.7 for Mizan and 12.02 for En–Tr). Second, Tep++ contains film subtitle translations in spoken Farsi, while Mizan and En–Tr contain translations of novels and legal documents, respectively. As human translations of the film subtitles are required to follow the structure and rhythm of the English

² <http://www.denizyuret.com/2006/11/turkish-resources.html>

³ <http://dadegan.ir/catalog/mizan>

⁴ As in Birch (2011), we define short, medium or long distance reordering as a reordering of width 1–4 words, 5–8 and more than 8 words, respectively.

Table 5
Parallel corpora statistics.

	Corpus	Train		Tune		Test	
		Sentences	Words	Sentences	Words	Sentences	Words
Mizan	English	1,016,758	13,919,071	3000	40,831	1000	13,165
	Farsi	1,016,758	14,043,499	3000	41,670	1000	13,444
Tep++	English	575,208	4,652,389	2000	16,152	1000	8136
	Farsi	575,208	4,421,994	2000	15,388	1000	7850
En-Tr	English	100,957	1,084,233	647	13,302	644	12,371
	Turkish	100,957	1,148,388	647	13,969	644	13,048

Table 6
Reordering type distribution over the training data for the parallel corpora.

Corpus	Mizan		Tep++		En-Tr	
	head-dep	dep-dep	head-dep	dep-dep	head-dep	dep-dep
Instances	6,391,956	5,247,526	2,693,793	1,964,448	535,892	480,726
Monotone (%)	57.54	67.91	63.04	71.92	55.70	60.93
Swap (%)	42.46	32.09	36.96	28.08	44.30	39.07

Fig. 2. Percentage of reorderings with different widths in the parallel corpora.

source sentences, a lower amount of reordering is expected than in formal translation of English written texts (Cettolo, Niehues, Stker, Bentivogli, & Federico, 2014).

We parsed the English sides of our parallel corpora with the Stanford dependency parser (Chen & Manning, 2014) and used the “collapsed representation” of the parser output to obtain the direct dependencies between the words in the English sentence. The accuracy of the Stanford parser on the English Penn Treebank is 91.8%⁵ (Chen & Manning, 2014). We aligned the words in the parallel corpora using GIZA++ (Och & Ney, 2003). Having obtained both the dependency structure and the word alignment, we extracted (*head-dep*) and (*dep-dep*) pairs from the training data sets and determined the orientation for each pair based on Eq. (2). Table 6 shows the number of extracted constituent pairs and the reordering type distribution over the training set. To perform the classification task, we used the Stanford maxEnt classifier (Manning & Klein, 2003) with default settings.

Our baseline SMT system is the Moses implementation (Hoang, Koehn, & Lopez, 2009) of the HPB-SMT model, with standard settings.⁶ It uses a log-linear model with eight built-in feature func-

Table 7
Features.

Surface form of head	<i>Lex(head)</i>
Surface form of 1st dependant	<i>Lex(dep1)</i>
Surface form of 2nd dependant	<i>Lex(dep2)</i>
Synset head	<i>Syn(head)</i>
Synset of 1st dependant	<i>Syn(dep1)</i>
Synset of 2nd dependant	<i>Syn(dep2)</i>
Dep. relation of 1st dependant	<i>depRel(dep1)</i>
Dep. relation of 2nd dependant	<i>depRel(dep2)</i>

tions: *n*-gram language model (LM), translation model (four features: inverse phrase translation probability (TM0), inverse lexical weighting (TM1), direct phrase translation probability (TM2) and direct lexical weighting (TM3)), glue grammar rule application model (GG), word penalty (WP) and phrase penalty (PP). We integrated our reordering feature functions as four additional features as described in Section 3.3. In all experiments the weights of our reordering feature functions and the other built-in feature functions were tuned by MIRA (Cherry & Foster, 2012). We used a 5-gram LM trained on the Farsi side of our training corpus. As mentioned earlier, we used Wordnet synsets of the English words in our RM. Since some of the words have more than one sense, and thus are assigned to multiple synsets in Wordnet, we consider only the most common sense of these words to extract their synset.

4.2. Translation results and analysis

We used different feature sets in our maxEnt classifier in order to determine the impact of each feature set on the quality of the MT system. Table 7 shows the features that we examine in this paper. As mentioned earlier, all these features are entirely based on the source sentence.

We build six MT systems according to the constituent pairs and feature sets examined. Three different feature sets were examined in this paper, including information from (i) surface forms (*surface*), (ii) synsets (*synset*) and (iii) both surface forms and synsets (*both*). The features used in each MT system are then as follows:

- *hd-surface* (Gao et al., 2011): *Lex(head)*, *Lex(dep)*, *depRel(dep)*

⁵ The accuracy of the parser and the word-aligner affect our reordering model. However, since we do not have access to gold-standard parse trees and word alignments for our parallel corpus, we cannot perform an in-depth analysis of the error propagation.

⁶ Max-chart-span parameter in Moses shows the maximum number of words that can be covered by a hierarchical rule. The standard value of this parameter

is 10. However, to be able to produce longer distance reorderings, we change its value to 20 in all experiments.

Table 8

MT scores for all systems on English–Farsi corpora. Metric scores are averages over three runs. The best result according to each metric (highest for BLEU and lowest for TER) is shown in bold.

Corpus	System	BLEU $\uparrow$				TER $\downarrow$			
		Avg	$\bar{s}_{\text{sel}}$	s_{Test}	p -value	Avg	$\bar{s}_{\text{sel}}$	s_{Test}	p -value
Mizan	Baseline	10.9	0.6	0.0	–	80.3	0.8	0.0	–
	word-based RM (Hayashi et al., 2010)	10.4	0.6	0.1	0.00	79.8	0.8	0.2	0.00
	hd-surface (Gao et al., 2011)	11.1	0.6	0.1	0.08	80.9	0.8	0.3	0.01
	hd-syn	11.3	0.6	0.1	0.00	80.5	0.8	0.2	0.40
	hd-both	11.1	0.6	0.1	0.06	81.1	0.8	0.3	0.00
	dd-surface	11.4	0.7	0.1	0.00	79.7	0.8	0.2	0.01
	dd-syn	11.3	0.6	0.2	0.01	79.8	0.8	0.2	0.05
	dd-both	11.5	0.7	0.2	0.00	79.8	0.8	0.5	0.02
	all-surface	11.2	0.6	0.1	0.01	80.9	0.8	0.1	0.01
	all-syn	11.3	0.6	0.1	0.00	81.0	0.8	0.3	0.00
	all-both	11.3	0.6	0.0	0.01	81.1	0.8	0.0	0.00
Tep++	Baseline	22.8	1.1	0.1	–	66.2	1.1	0.2	–
	word-based RM (Hayashi et al., 2010)	23.1	1.1	0.1	0.31	66.1	1.1	0.2	0.07
	hd-surface (Gao et al., 2011)	23.1	1.1	0.0	0.22	66.8	1.2	0.1	0.02
	hd-syn	23.2	1.1	0.1	0.08	66.7	1.1	0.3	0.06
	hd-both	23.2	1.1	0.1	0.12	66.6	1.1	0.1	0.10
	dd-surface	23.4	1.1	0.1	0.00	65.7	1.1	0.1	0.03
	dd-syn	23.4	1.1	0.1	0.00	65.5	1.1	0.1	0.00
	dd-both	23.3	1.1	0.1	0.02	65.8	1.1	0.1	0.05
	all-surface	23.3	1.1	0.3	0.07	66.3	1.1	0.2	0.75
	all-syn	23.2	1.1	0.1	0.09	66.1	1.1	0.0	0.61
	all-both	23.2	1.1	0.1	0.13	66.1	1.1	0.1	0.78

- hd-synset: *Syn(head)*, *Syn(dep)*, *depRel(dep)*
- hd-both: *Lex(head)*, *Lex(dep)*, *Syn(head)*, *Syn(dep)*, *depRel(dep)*
- dd-surface: *Lex(head)*, *Lex(dep1)*, *Lex(dep2)*, *depRel(dep1)*, *depRel(dep2)*
- dd-synset: *Syn(head)*, *Syn(dep1)*, *Syn(dep2)*, *depRel(dep1)*, *depRel(dep2)*
- dd-both: *Lex(head)*, *Lex(dep1)*, *Lex(dep2)*, *Syn(head)*, *Syn(dep1)*, *Syn(dep2)*, *depRel(dep1)*, *depRel(dep2)*

We also integrate both the hd and dd systems into one system and build three additional MT systems, which have the capacity to reorder both (*head–dep*) and (*dep–dep*) pairs:

- all-surface: hd-surface + dd-surface
- all-syn: hd-syn + dd-syn
- all-both: hd-both + dd-both

The hd-surface system was proposed by Gao et al. (2011). We also implemented the word-based RM proposed by Hayashi et al. (2010) which considers any two words in the source sentence for reordering. As in our own model, we use a maximum entropy classifier for their model. As for features, we use surface forms and POS tags of the words.⁷

We compare the proposed RM to three MT systems: (i) a standard HPB-SMT system as the baseline, (ii) the hd-surface model proposed by Gao et al. (2011), and (iii) the word-based model proposed by Hayashi et al. (2010)

We evaluate each of the MT systems according to two widely used automatic evaluation metrics (BLEU, (Papineni, Roukos, Ward, and Zhu, 2002); TER, (Snover, Dorr, Schwartz, Micciulla, & Weischedel, 2006)). Each system is tuned three times and we report the average scores (Avg), $\bar{s}_{\text{sel}}$, s_{Test} and p -value obtained with multeval⁸ (Clark, Dyer, Lavie, & Smith, 2011) on the MT outputs (without further processing such as detokenisation or recasing). In Tables 8 and 9, s_{Test} indicates the standard deviation due to optimizer instability and $\bar{s}_{\text{sel}}$ indicates the standard deviation due to

test set selection and optimizer instability, quantified jointly (Clark et al., 2011). The p -values are relative to the baseline and indicate whether a difference of this magnitude (between the baseline and the system on that line) is likely to be generated again by some random process (a randomized optimizer). In order to calculate the p -values, multeval applies the approximate randomization (AR) test (Noreen, 1989). AR is a nonparametric test that approximates a paired permutation test by sampling permutations (Clark et al., 2011). In the case of estimating the p -value for SMT systems, the test statistic in the AR test is the absolute value of the difference in metric score (i.e. BLEU or TER) produced by two systems on the same test set (Riezler & Maxwell, 2005). AR estimates the probability (p -value) that a test statistic occurs by chance by randomly exchanging output translations between the two SMT systems. If there is no significant difference between two SMT systems (i.e. the null hypothesis is true) and the difference in metric score is produced due to the optimizer instability, any output translation produced by one of the systems could have been produced by the other system. Therefore, shuffling output translations between the two systems and recomputing the test statistic, creates an approximate distribution of the test statistic under the null hypothesis. The p -value is then computed as the percentage of runs where the test statistic computed on the shuffled data is greater than or equal to the test statistic computed on the test data. Since the AR test assumes that the samples being used are representative of all extraneous variables that could affect the output of SMT systems, multeval includes multiple optimizer replications in the shuffling⁹ (Clark et al., 2011).

We derive the following observations from the results:

- Almost all of the systems that model the orderings between pairs of constituents outperform the baseline HPB-SMT system according to BLEU. Most of these differences are significant (considering statistical significance as $p \leq 0.05$).
- The use of our novel semantic feature (synset) leads to better scores than using surface features alone, except for the dd system for Mizan corpus, with a slight but insignificant increase in terms of TER (79.8 vs. 79.7). While the semantic features

⁷ We do not have access to the dependency relation feature for Hayashi's model. Instead, as in their original work, we use the POS tags of the words instead.

⁸ <https://github.com/jhclark/multeval>

⁹ The standard deviation due to optimizer instability is shown with s_{Test} .

Table 9

MT scores for all systems on English–Turkish corpus. Metric scores are averages over three runs. The best result according to each metric (highest for BLEU and lowest for TER) is shown in bold.

Corpus	System	BLEU $\uparrow$				TER $\downarrow$			
		Avg	$\bar{s}_{\text{sel}}$	s_{Test}	p -value	Avg	$\bar{s}_{\text{sel}}$	s_{Test}	p -value
En–Tr	Baseline	25.0	0.9	0.1	–	65.9	1.1	0.3	–
	word-based RM (Hayashi et al., 2010)	23.7	0.9	0.4	0.00	66.0	0.9	2.6	0.80
	hd-surface (Gao et al., 2011)	25.2	0.9	0.1	0.47	66.2	1.0	0.0	0.25
	hd-syn	25.2	0.9	0.1	0.30	66.0	1.0	0.4	0.80
	hd-both	25.2	0.9	0.1	0.31	66.4	1.1	0.3	0.05
	dd-surface	25.0	0.9	0.0	0.24	65.8	1.0	0.3	0.59
	dd-syn	25.4	0.9	0.1	0.05	65.7	1.0	0.0	0.35
	dd-both	25.1	0.9	0.1	0.83	65.7	1.0	0.1	0.44
	all-surface	25.2	0.9	0.1	0.55	65.8	1.0	0.1	0.51
	all-syn	25.1	0.9	0.1	0.67	66.0	1.0	0.2	0.56
	all-both	24.9	0.9	0.0	0.23	66.2	1.0	0.2	0.30

Fig. 3. An example English sentence with its partial dependency tree, its reference translation in Farsi and the word alignments.

have only a small impact on the final results, this small effect is consistently positive; in the majority of the cases in our experimental setup the best scores are obtained by using semantic features

- We observe that the systems that consider reordering for all constituent pairs by combining both hd and dd systems into one system (systems “all-”) show lower performance than that of the dd systems. Similarly, the word-based system that considers reordering for any two words in the sentence has lower performance than that of all- systems. In other words, in most cases, by considering more word pairs for reordering, the performance of the system decreases. We hypothesize that this is because there are redundant and probably conflicting predictions among the systems (Zhang et al., 2016). That is, a reordering between two words in the word-based system can be performed implicitly by a sequence of reorderings between some constituent pairs in all- systems. For example, in Fig. 1, if the all- system predicts the orientation for two word pairs (bought, wondered) and (wondered, tribes) as monotone, the orientation for the word pair (bought, tribes) can be simply determined to be monotone. As a consequence, predicting orientations for more word pairs does not essentially improve the MT results. In addition to this, inconsistent predictions for the orientations decrease the overall performance of the system. For example, if the word-based model incorrectly predicts a swap orientation for (bought, tribes), which is inconsistent with the implicit orientation for this pair, it has a negative affect on the quality of the final translation.
- While the systems that model the orderings between head and dependant pairs (systems hd, first reported in Gao et al. (2011)) are capable of handling non-local reorderings, the scores obtained by systems that implement our novel reordering between pairs of dependants (systems dd) achieve better results than both the baseline and hd systems, across both evaluation metrics. We hypothesize that this can be attributed, at least in part, to the fact that the classifier learns reordering for *dep–dep* pairs better than for *head–dep* pairs (cf. Table 12).

Table 10

Significance test results over *hd – surface* system (Gao et al., 2011), using multeval (Clark et al., 2011) w.r.t. BLEU and TER. The symbol $\gg$ represents a significant difference at the $p < 0.05$ level and – means not significantly different at $p = 0.05$.

Corpus	Mizan		Tep++		En–Tr	
	BLEU	TER	BLEU	TER	BLEU	TER
dd-surface	$\gg$	$\gg$	$\gg$	$\gg$	–	–
dd-syn	–	$\gg$	$\gg$	$\gg$	–	$\gg$
dd-both	$\gg$	$\gg$	–	$\gg$	–	$\gg$

As mentioned earlier, using dependency information in the RM for HPB-SMT system was first proposed by Gao et al. (2011). While they try to predict the orientation of each dependant with respect to its head, we predict the orientation for pairs of dependants. Table 8 shows that in English–Farsi parallel corpora, *hd – surface* (proposed by Gao et al., 2011) improves the baseline in terms of BLEU. However, it is significantly worse than the baseline in terms of TER. The proposed systems which predict the orientation between dependants (systems dd) significantly improve upon the baseline both in terms of both BLEU and TER. Table 9 shows that for the English–Turkish corpus *hd – surface* (Gao et al., 2011) improves upon the baseline in terms of BLEU but degrades in terms of TER. The proposed systems dd improve upon the baseline both in terms of BLEU and TER (except for dd-surface). For system *dd – syn* this improvement is significant.

We also compare the proposed systems dd and *hd – surface* (which is proposed by Gao et al., 2011) directly. Table 10 shows the significance test results between dd systems and *hd – surface*. As shown in Table 10, for Mizan and Tep++, systems dd are significantly better than *hd – surface* both in terms of BLEU and TER (except for *dd – syn* on Mizan in terms of BLEU, and *dd – both* on Tep++ in terms of BLEU). For the English–Turkish corpus, both *dd – syn* and *dd – both* are significantly better than *hd – surface* in terms of TER.

Table 11

Example translations from Mizan corpus, output by the baseline and proposed MT systems.

Input sentence	It took two hours to select the equipment that jamie would take.
Reference	
Translation	do saat tool keshid ta lavazemi ke jamie bayad mibord ra entekhab konad.
Order	Two hours took the equipment that jamie would take to select .
Baseline	
Translation	do saat be entekhab lavazemi ke jamie ra.
Order	Two hours to select the equipment that jamie would take.
hd-surface (Gao et al., 2011)	
Translation	do saat be lavazem ra entekhab konad ke jamie.
Order	Two hours the equipment would take to select that jamie.
dd-surface	
Translation	do saat tool keshid ta lavazemi ke jimi ra entekhab.
Order	Two hours took the equipment that jamie would take to select .

Table 12Classification accuracy (Acc.) on the test data set, using different features. Column *diff* show the relative improvement over the baseline. The best result for each classifier is shown in bold.

Corpus		Mizan		Tep++		En-Tr	
Constituent	Classifier	Acc.	diff	Acc.	diff	Acc.	diff
<i>dep-dep</i>	Baseline	65.69	–	74.46	–	57	–
	dd-surf	73.14	11.34%	78.64	5.61%	73.58	29.08%
	dd-syn	72.47	10.32%	78.46	5.37%	73.85	29.56%
	dd-both	72.59	10.50%	78.81	5.84%	73.61	29.14%
<i>head-dep</i>	Baseline	55.07	–	63.48	–	56.11	–
	hd-surface	69.01	25.31%	71.16	12.09%	73.09	30.26%
	hd-syn	69.02	25.33%	71.75	13.02%	72.83	29.79%
	hd-both	68.84	25.00%	71.35	12.39%	72.97	30.04%

4.3. Translation examples

Fig. 3 shows an example English sentence, its partial dependency tree, its reference translation in Farsi and the word alignment. For simplicity, it only shows the dependants of the verbs “select” and “take”. Table 11 shows the translation output and the order of translation by the baseline, hd-surface and dd-surface systems for the sentence in Fig. 3, demonstrating how the proposed RM that predicts the ordering between pairs of dependants outperforms the RMs that predict the ordering between head–dependant pairs (Gao et al., 2011).

As shown in the reference order, the translation of the main verb *select* should be placed at the end of the sentence, but the baseline system fails to perform any reordering and puts it at the beginning of the sentence. The hd-surface model (Gao et al., 2011) tries to predict the ordering between the head words “select” and “take” and their dependants, and encourages concurrent translation of the head and its dependants. It correctly predicts the ordering of “equipment” and “select” as swap, so it places the translation of “select” after that of “equipment” and hence moves the main verb closer to its correct position at the end of the sentence. However, due to a faulty prediction about the ordering of “select” and “take”, it translates them monotonically. The dd-surface system tries to predict the ordering of dependant pairs and encourages concurrent translation of such pairs. It predicts the ordering of all of the dependant pairs correctly as monotone. Since it prefers concurrent translation of dependants, it translates “equipment” and “took” (as dependants of the head word “select”) together. In the same way, it translates “jaime” and “would” (as dependants of the head word “take”) concurrently and hence moves the translation of the main verb “select” to the end of the sentence.

5. Analyses

5.1. Analysis of classification accuracy

In Section 4.2 we presented the impact of different feature sets on the quality of a range of MT systems, which can be seen as an extrinsic evaluation of the maxEnt classifier. In this section, we investigate the impact of those feature sets on the accuracy of the maxEnt classifier, i.e. intrinsic evaluation. We train the classifiers separately for (*head-dep*) and (*dep-dep*) pairs on the training data and then we evaluate them on the test data. We compare the accuracy of the classifiers with that of the baseline classifier, which considers monotone orientation (the majority class) for all of the constituent pairs in the data set. Table 12 shows the performance of the classifiers using different feature sets, for (*head-dep*) and (*dep-dep*) constituent pairs.

Here we observe:

- All classifiers outperform the baseline, confirming the trends shown in translation results, cf. Table 8.
 - The best classification accuracy for (*head-dep*) on both English–Farsi corpora is obtained by using synsets, which corroborates the best MT result according to both BLEU and TER for this constituent pair. The best classification accuracy for (*dep-dep*) on Mizan is obtained using solely surface forms, which is corroborated by the MT results in terms of TER but not in terms of BLEU, for which the best result was obtained using both surface forms and synsets. This shows that higher classification accuracy does not always lead to higher MT scores, but the general trend follows this direction. Note that the mismatch between classification performance and MT performance has been reported previously (Gao, 2010).
- For English–Farsi parallel corpora, classification of orientation between pairs of (*head-dep*) is a more difficult problem than the classification of (*dep-dep*) as shown by the difference in

Fig. 4. Weights and standard deviations (average over 3 tunes) for each feature function and system for Mizan.

scores obtained by the baseline, over 10 points higher for the latter. Note, however, that the relative improvement in classification accuracy for *head-dep* vs. *dep-dep* pairs is much higher, about 15% on Mizan and 7% on Tep++.

5.2. Analysis of feature weights

In order to obtain insights regarding the impact of the RM in the SMT systems, we study the values assigned by the tuning algorithm to each of the 4 reordering feature functions (RM0, RM1, RM2 and RM3) and to the feature function used for reordering in the baseline (GG). In principle, the higher the absolute value of the weight assigned to a feature function, the higher its impact. Figs. 4–6 show, for the Mizan, Tep++ and En-Tr corpora, respectively, the weights assigned to each of the 5 reordering feature functions¹⁰ for each MT system. The error bars in Figs. 4–6 show the standard deviations.

While it is not possible or desirable to extract definite conclusions from analyzing the weights (the tuning process is after all non-deterministic), we think it worthwhile analyzing them in detail, as they might provide indicative global trends. In order to be able to derive more solid conclusions, we tune each system three times, and for each weight and system consider the average value obtained across all the tuning executions. We also calculate for each weight and system the standard deviation across tuning executions. The standard deviation for reordering-related features stays below 0.025.

From the analysis of the weights related to reordering (RM0, RM1, RM2, RM3 and GG), we derive the following conclusions:

- The RM feature functions RM0, RM1 and RM2 have a considerable impact (in all cases their weights are higher than 0.018). Conversely, RM3 seems to be considerably less relevant for English–Farsi corpora as attested by its lower absolute val-

ues. RM3 is the only reordering feature function for which there are instances of negative values (hd systems on Mizan and all the systems on Tep++).

- In all systems with an RM (except for dd-syn and dd-both systems on En-Tr corpus), the weight for RM0 (monotone) is higher than for RM1 (swap). We hypothesize that this is because the classifier's performance for the monotone class are higher than for the swap class. As we have more monotone instances in training data sets, the decoder learns to trust monotone scores more than swap scores.
- The glue grammar rule (the feature function that deals with reordering in the HPB-SMT model) and the RM proposed seem to be complementary. Compared to the baseline, most of the systems using the RM have a considerably lower score to the GG but still significant (values between 0.03 and 0.1 compared to 0.25 for Mizan, and between 0.01 and 0.03 compared to 0.24 for Tep++).

6. Some advantages and limitations of the proposed RM

We have proposed a syntax-based RM enriched with semantic features. Rather than performing reordering in a preprocessing step (Collins et al., 2005; Genzel, 2010; Xia & McCord, 2004; Xu et al., 2009), we address reordering during decoding, i.e. as part of the translation process (Hayashi et al., 2010; Huck et al., 2013; Nguyen & Vogel, 2013). In this way, the proposed RM avoids making hard decisions and maintains the possibility of reversing previous faulty decisions (Koehn, 2012). By using only source-side information in the proposed RM, it can be precomputed offline, which leads to low added complexity during decoding. The proposed RM uses dependency information (Gao et al., 2011) and tries to predict the orientation between pairs of dependants in the dependency tree of the source sentence. Therefore, it is able to perform long-distance reordering. We enrich the proposed RM with semantic features based on WordNet to allow it to generalize to pairs not seen in the training data but with equivalent meaning.

While the proposed RM offers these advantages, it has two limitations, in terms of language resources required.

¹⁰ Features not related to reordering (i.e. translation model and penalty features) are not included in the figures as their values were not substantially different for the baseline and our proposed systems.

Fig. 5. Weights and standard deviations (average over 3 tunes) for each feature function and system for Tep++.

Fig. 6. Weights and standard deviations (average over 3 tunes) for each feature function and system for En-Tr.

- Since the proposed RM uses the dependency information of the source sentence, it cannot be applied for translating from languages for which automatic dependency parsers are not available.
- In order to enrich the proposed RM with semantic features, we should have access to WordNet for the source language.

With respect to the first limitation, the contemporary development of Universal Dependencies¹¹ is lowering the barrier to have parsers available for new languages. As for the second, for lan-

guages where WordNet is not available, one could obtain synonyms in an unsupervised way, e.g. by building word embeddings.

As we have shown in Section 4.2 the proposed RM outperforms the baseline HPB-SMT system both in terms of BLEU and TER. However, there is a trade-off between the SMT performance and its speed. Since our RM requires the dependency information of the source sentence, the source sentence should be parsed before translation. This adds extra time complexity to the SMT system. By using only source-side information which can be precomputed offline, we try to keep the complexity as low as possible.

¹¹ <http://universaldependencies.org/>

7. Conclusions and future work

This paper has presented a novel syntax-based RM for HPB-SMT enriched with semantic features to address some of the deficiencies of previous RMs: (i) it is able to perform long-distance reordering as it predicts the orientation between pairs of dependants; (ii) it generalizes 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 our model uses only source-side information, the model can be pre-computed offline, which leads to low added complexity during decoding.

Our model builds upon the state-of-the-art RMs for HPB-SMT that use the source-side dependency tree, to which it contributes across multiple dimensions: (i) while the previous dependency-based RM (Gao et al., 2011) 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 RM to generalize to pairs not seen in training but with equivalent meaning; (iii) we evaluate our model on the English-to-Farsi and English-to-Turkish language pairs, challenging due to their rich morphology and under-resourced status. We evaluate our RM both intrinsically (accuracy of the reordering classifier) and extrinsically (impact on MT). In both cases, the RM outperforms the baseline (majority class for the intrinsic evaluation and HPB-SMT for MT). The 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 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 in the HPB-SMT model.

As for future work, we consider several avenues of research worth investigating. (i) Perform word sense disambiguation. While on this paper we have considered the synset that correspond to the most common sense of each word, an alternative would be to perform disambiguation to find the right sense. While this should lead to using more accurate synsets, which might have an impact on the final MT results, one should factor in the error rate of disambiguation tools. (ii) Experiment with other language pairs to further assess the language-independent nature of our methodology. (iii) Conduct a human evaluation to assess the impact of our RM for human uses of MT, be it gisting, post-editing, etc.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This research is supported by University of Isfahan and by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (<http://www.adaptcentre.ie>) at Dublin City University, the European Union Seventh Framework Programme FP7/2007–2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran). The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Birch, A. (2011). *Reordering metrics for statistical machine translation*. University of Edinburgh Ph.D. thesis.

- Birch, A., Blunsom, P., & Osborne, M. (2009). A quantitative analysis of reordering phenomena. In *Proceedings of the fourth workshop on statistical machine translation* (pp. 197–205). Athens, Greece.
- Bisazza, A., & Federico, M. (2013). Dynamically shaping the reordering search space of phrase-based statistical machine translation. *Transactions of the Association for Computational Linguistics*, 1, 327–340.
- Cettolo, M., Niehues, J., Stker, S., Bentivogli, L., & Federico, M. (2014). Report on the 11th IWSLT Evaluation Campaign. *Proceedings of the eleventh workshop on spoken language translation* (pp. 2–16).
- Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Empirical methods in natural language processing (EMNLP)*.
- Cherry, C., & Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 427–436).
- Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the forty-third annual meeting on association for computational linguistics* (pp. 263–270). doi:10.3115/1219840.1219873.
- Clark, J. H., Dyer, C., Lavie, A., & Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the forty-ninth annual meeting of the association for computational linguistics: Human language technologies, HLT '11*: 2 (pp. 176–181). Stroudsburg, PA, USA: Association for Computational Linguistics. Short papers.
- Collins, M., Koehn, P., & Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the forty-third annual meeting on association for computational linguistics* (pp. 531–540). doi:10.3115/1219840.1219906.
- Fellbaum, C. (1998). *WordNet*. Wiley Online Library.
- Galley, M., & Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 848–856).
- Gao, Y. (2010). Discriminative reordering model for machine translation.
- Gao, Y., Koehn, P., & Birch, A. (2011). Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 857–868).
- Genzel, D. (2010). Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the twenty-third international conference on computational linguistics* (pp. 376–384).
- Hanneman, G., & Lavie, A. (2013). Improving syntax-augmented machine translation by coarsening the label set. In *Proceedings of the North American chapter of the association of computational linguistics* (pp. 288–297).
- Hayashi, K., Tsukada, H., Sudoh, K., Duh, K., & Yamamoto, S. (2010). Hierarchical phrase-based machine translation with word-based reordering model. In *Proceedings of the twenty-third international conference on computational linguistics (COLING 2010)* (pp. 439–446).
- He, Z., Meng, Y., & Yu, H. (2010). Extending the hierarchical phrase based model with maximum entropy based BTG. In *Proceedings of the 2010 Conference of the association for machine translation in the Americas (AMTA)*.
- Hoang, H., Koehn, P., & Lopez, A. (2009). A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the 2009 international workshop on spoken language translation (IWSLT)* (pp. 152–159).
- Huang, Z., Devlin, J., & Zbib, R. (2013). Factored soft source syntactic constraints for hierarchical machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 556–566).
- Huck, M., Wuebker, J., Rietig, F., & Ney, H. (2013). A phrase orientation model for hierarchical machine translation. In *Proceedings of the eighth workshop on statistical machine translation* (pp. 452–463).
- Kazemi, A., Toral, A., & Way, A. (2016). Using WordNet to improve reordering in hierarchical phrase-based statistical machine translation. In *Proceedings of the eighth meeting of the global WordNet conference*. Bucharest, Romania.
- Kazemi, A., Toral, A., Way, A., Monadjemi, A., & Nematbakhsh, M. (2015). Dependency-based reordering model for constituent pairs in hierarchical SMT. In *Proceedings of the eighteenth annual conference of the European association for machine translation* (pp. 43–50). Antalya, Turkey.
- Koehn, P. (2012). *Statistical machine translation*. Cambridge University Press.
- Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., & Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the international workshop on spoken language translation* (pp. 68–75).
- Koehn, P., Och, F., & Marcu, D. (2003a). Statistical phrase-based translation. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 127–133).
- Koehn, P., Och, F. J., & Marcu, D. (2003b). Statistical phrase-based translation. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology – volume 1* (pp. 48–54). doi:10.3115/1073445.1073462.
- Komachi, M., Matsumoto, Y., & Nagata, M. (2006). Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proceedings of the 2006 international workshop on spoken language translation (IWSLT)* (pp. 77–82).
- Li, J., Marton, Y., Resnik, P., & Daumé III, H. (2014). A unified model for soft linguistic reordering constraints in statistical machine translation. In *Proceedings of the fifty-second annual meeting of the association for computational linguistics: 1* (pp. 1123–1133). Baltimore, Maryland: Association for Computational Linguistics.

- Li, J., Resnik, P., & III, H. D. (2013). Modeling syntactic and semantic structures in hierarchical phrase-based translation. In *Proceedings of the annual conference of the North American chapter of the association for computational linguistics human language technologies* (pp. 540–549).
- Liu, D., & Gildea, D. (2010). Semantic role features for machine translation. In *Proceedings of the twenty-third international conference on computational linguistics* (pp. 716–724).
- Manning, C., & Klein, D. (2003). Optimization, Maxent models, and conditional estimation without magic. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology: Tutorials* (p. 8).
- Miller, G. A. (1995). Wordnet: A Lexical database for english. *Communications of the ACM*, 38(11), 39–41.
- Nguyen, T., & Vogel, S. (2013). Integrating phrase-based reordering features into a chart-based decoder for machine translation. In *Proceedings of the fifty-first annual meeting of the association for computational linguistics* (pp. 1587–1596).
- Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses: An introduction*. Wiley.
- Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51.
- Ofllazer, K. (2008). Statistical machine translation into a morphologically complex language. In *Proceedings of the conference on intelligent text processing and computational linguistics*.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the fortieth annual meeting on association for computational linguistics* (pp. 311–318). doi:10.3115/1073083.1073135.
- Passban, P., Way, A., & Liu, Q. (2015). Benchmarking SMT performance for farsi using the TEP++ corpus. In *Proceedings of the eighteenth annual conference of the european association for machine translation*. Antalya, Turkey.
- Quirk, C., Menezes, A., & Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the forty-third annual meeting on association for computational linguistics* (pp. 271–279).
- Riezler, S., & Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation methods for machine translation and summarization* (pp. 57–64). Ann Arbor, Michigan.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Weischedel, R. (2006). A study of translation error rate with targeted human annotation. In *Proceedings of the association for machine translation in the Americas*.
- Supreme Council of Information and Communication Technology (2013). *Mizan English–Persian Parallel Corpus*. Tehran, I.R. Iran.
- Tillmann, C. (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the 2004 HLT-NAACL* (pp. 101–104). Short papers.
- Tromble, R., & Eisner, J. (2009). Learning linear ordering problems for better translation. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 1007–1016).
- Wu, X., Sudoh, K., Duh, K., Tsukada, H., & Nagata, M. (2011). Extracting pre-ordering rules from predicate-argument structures. In *Proceedings of the fifth international joint conference on natural language processing* (pp. 29–37).
- Xia, F., & McCord, M. (2004). Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the twentieth international conference on computational linguistics*. doi:10.3115/1220355.1220428.
- Xiong, D., Liu, Q., & Lin, S. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the twenty-first international conference on computational linguistics and the forty-fourth annual meeting of the association for computational linguistics* (pp. 521–528). doi:10.3115/1220175.1220241.
- Xiong, D., Zhang, M., & Li, H. (2012). Modeling the translation of predicate-argument structure for SMT. In *Proceedings of the fiftieth annual meeting of the association for computational linguistics: 1* (pp. 902–911). Long papers.
- Xu, P., Kang, J., Ringgaard, M., & Och, F. (2009). Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of the 2009 annual conference of the North American chapter of the association for computational linguistics on human language technologies* (pp. 245–253).
- Zens, R., & Ney, H. (2006). Discriminative reordering models for statistical machine translation. In *Proceedings of the workshop on statistical machine translation (StatMT '06)* (pp. 55–63).
- Zhai, F., Zhang, J., Zhou, Y., & Zong, C. (2012). Machine translation by modeling predicate argument structure transformation. In *Proceedings of the International Conference on Computational Linguistics (COLING 2012)* (pp. 3019–3036).
- Zhang, J., Utiyama, I., Sumita, E., Zhao, H., & Neub, G. (2016). Learning local word reorderings for hierarchical phrase-based statistical machine translation. *Machine Translation Journal*, 30(1), 1–18.
- Zollmann, A., & Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings of the workshop on statistical machine translation (StatMT '06)* (pp. 138–141).