Using Wordnet to Improve Reordering in Hierarchical Phrase-Based Statistical Machine Translation

Arefeh Kazemi[†], **Antonio Toral**[⋆], **Andy Way**[⋆]
[†] Department of Computer Engineering, University of Isfahan, Isfahan, Iran
{kazemi}@eng.ui.ac.ir
[⋆] ADAPT Centre, School of Computing, Dublin City University, Ireland
{atoral,away}@computing.dcu.ie

Abstract

We propose the use of WordNet synsets in a syntax-based reordering model for hierarchical statistical machine translation (HPB-SMT) to enable the model to generalize to phrases not seen in the training data but that have equivalent meaning. We detail our methodology to incorporate synsets' knowledge in the reordering model and evaluate the resulting WordNet-enhanced SMT systems on the English-to-Farsi language direction. The inclusion of synsets leads to the best BLEU score, outperforming the baseline (standard HPB-SMT) by 0.6 points absolute.

1 Introduction

Statistical Machine Translation (SMT) is a data driven approach for translating from one natural language into another. Natural languages vary in their vocabularies and also in the manner that they arrange words in the sentence. Accordingly, SMT systems should address two interrelated problems: finding the appropriate words in the translation ("lexical choice") and predicting their order in the sentence ("reordering"). Reordering is one of the hardest problems in SMT and has a significant impact on the quality of the translation, especially between languages with major differences in word order. Although SMT systems deliver state-of-the-art performance in machine translation nowadays, they perform relatively weakly at addressing the reordering problem.

Phrased-based SMT (PB-SMT) is arguably the most widely used approach to SMT to date. In this model, the translation operates on *phrases*, i.e. sequences of words whose length is between 1 and a maximum upper limit. In PB-SMT, reordering is generally captured by distance-based models (Koehn et al., 2003) and lexical phrase-based models (Tillmann, 2004; Koehn et al., 2005), which are able to perform local reordering but they cannot capture non-local (long-distance) reordering. The weakness of PB-SMT systems on handling long-distance reordering led to proposing the Hierarchical Phrase-based SMT (HPB-SMT) model(Chiang, 2005), in which the translation operates on tree structures (either derived from a syntactic parser or unsupervised). Despite the relatively good performance offered by HPB-SMT in medium-range reordering, they are still weak on long-distance reordering (Birch et al., 2009).

A great deal of work has been carried out to address the reordering problem by incorporating reordering models (RM) into SMT systems. A RM tries to capture the differences in word order in a probabilistic framework and assigns a probability to each possible order of words in the target sentence. Most of the reordering models can perform reordering of common words or phrases relatively well, but they can not be generalized to unseen words or phrases with the same meaning ("*semantic generalization*") or the same syntactic structure ("*syntactic generalization*"). 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 required reordering between them. Likewise, if two words in the source language follow a specific reordering pattern in the target language, these models can not generalize to unseen words with equivalent meaning in the same context.

In order to improve syntactic and semantic generalization of the RM, it is necessary to incorporate syntactic and semantic features into the model. While there has been some encouraging work on integrating syntactic features into the RM, to the best of our knowledge, there has been no previous work on integrating semantic

Reordering Model	Features Types	Features
Zens and Ney (2006)	lexical	surface forms of the source and target words
		unsupervised class of the source and target words
Cherry (2013)	lexical	surface forms of frequent source and target words
		unsupervised class of rare source and target words
Green *et al.* (2010)	lexical	surface forms of the source words, POS tags of the
	syntactic	source words, relative position of the source words
		sentence length
Bisazza and Federico (2013)	lexical	surface forms and POS tags of the source words
and Goto *et al.* (2013)	syntactic	surface forms and POS tags of the source context words
Gao *et al.* (2011) and	lexical	surface forms of the source words
Kazemi *et al.* (2015)	syntactic	dependency relation
The proposed method	lexical	surface forms of the source words
	syntactic	dependency relation
	semantic	synset of the source words

Table 1: An overview of the used features in the SOTA reordering models

features. In this paper we enrich a recently proposed syntax-based reordering model for HPB-SMT system (Kazemi et al., 2015) with semantic features. To be more precise, we use WordNet[1] (Fellbaum, 1998) to incorporate semantics into our RM. We report experimental results on a large-scale English-to-Farsi translation task.

The rest of the paper is organized as follows. Section 2 reviews the related work and puts our work in its proper context. Section 3 introduces our RM, which is then evaluated in Section 4.2. Finally, Section 5 summarizes the paper and discusses avenues of future work.

2 Related Work

Many different approaches have been proposed to capture long-distance reordering by incorporating a RM into PB-SMT or HPB-SMT systems. A RM should be able to perform the required reorderings not only for common words or phrases, but also for phrases unseen in the training data that hold the same syntactic and semantic structure. In other words, a RM should be able to make syntactic and semantic generalizations. To this end, rather than conditioning on actual phrases, state-of-the-art RMs generally make use of features extracted from the phrases of the training data. One useful way to categorize previous RMs is by the features that they use to generalize. These features can be divided into three groups: (i) lexical features (ii) syntactic features and (iii) semantic features. Table 1 shows a representative selection of state-of-the-art RMs along with the features that they use for generalization.

Zens and Ney (2006) proposed a maximum-entropy RM for PB-SMT that tries to predict the orientation between adjacent phrases based on various combinations of some features: surface forms of the source words, surface form of the target words, unsupervised class of the source words and unsupervised class of the target words. They show that unsupervised word-class based features perform almost as well as word-based features, and combining them results in small gains. This motivates us to consider incorporating supervised semantic-based word-classes into our model.

Cherry (2013) integrates sparse phrase orientation features directly into a PB-SMT decoder. As features, he used the surface forms of the frequent words, and the unsupervised cluster of uncommon words. Green *et al.* (2010) introduced a discriminative RM that scores different jumps in the translation depending on the source words, their Part-Of-Speech (POS) tags, their relative position in the source sentence, and also the sentence length. This RM fails to capture the rare long-distance reorderings, since it typically over-penalizes long jumps that occur much more rarely than short jumps (Bisazza and Federico, 2015). Bisazza and Federico (2013) and Goto *et al.* (2013) estimate for each pair of input positions x and y, the probability of translating y right after x based on the surface forms and the POS tags of the source words, and the surface forms and the POS tags of the source context words.

[1] http://wordnet.princeton.edu/

Gao *et al.* (2011) and Kazemi *et al.* (2015) proposed a dependency-based RM for HPB-SMT which uses a maximum-entropy classifier to predict the orientation between pairs of constituents. They examined two types of features, the surface forms of the constituents and the dependency relation between them. Our approach is closely related to the latter two works, as we are interested to predict the orientation between pairs of constituents. Similarly to (Gao et al., 2011; Kazemi et al., 2015), we train a classifier based on some extracted features from the constituent pairs, but on top of lexical and syntactic features, we use semantic features (WordNet synsets) in our RM. In this way, our model can be generalized to unseen phrases that follow the same semantic structure.

3 Method

Following Kazemi *et al.* (2015) we implement a syntax-based RM for HPB-SMT based on the dependency tree of the source sentence. The dependency tree of a sentence shows the grammatical relation between pairs of head and dependent words in the sentence. As an example, Figure 1 shows the dependency tree of an English sentence. In this figure, the arrow with label "nsubj" from "fox" to "jumped" indicates that the dependent word "fox" is the subject of the head word "jumped". Given the assumption that constituents move as a whole during translation (Quirk et al., 2005), we take the dependency tree of the source sentence and try to find the ordering of each dependent word with respect to its head ($head$-dep) and also with respect to the other dependants of that head (dep-dep). For example, for the English sentence in Figure 1, we try to predict the orientation between ($head$-dep) and (dep-dep) pairs as shown in Table 2.

We consider two orientation types between the constituents: $monotone$ and $swap$. If the order of two constituents in the source sentence is the same as the order of their translation in the target sentence, the orientation is monotone and otherwise it is swap. To be more formal, for two source words (S_1,S_2) and their aligned target words (T_1,T_2), with the alignment points (P_{S1},P_{S2}) and (P_{T1},P_{T2}), we find the orientation type between S_1 and S_2 as shown in Equation 1 (Kazemi et al., 2015).

$$ori = \begin{cases} \text{if } (p_{S1} - p_{S2}) \times (p_{T1} - p_{T2}) > 0 \\ \quad monotone \\ \text{else} \\ \quad swap \end{cases} \tag{1}$$

For example, for the sentence in Figure 1, the orientation between the source words "brown" and "quick" is monotone, while the orientation between "brown" and "fox" is swap.

We use a classifier to predict the probability of the orientation between each pair of constituents to be monotone or swap. This probability is used as one feature in the log-linear framework of the HPB-SMT model. Using a classifier enables us to incorporate fine-grained information in the form of features into our RM. Table 3 and Table 4 show the features that we use to characterize ($head$-dep) and (dep-dep) pairs respectively.

As Table 3 and Table 4 show, we use three types of features: lexical, syntactic and semantic. While semantic structures have been previously used for MT reordering, e.g. (Liu and Gilda, 2010), to the best of our knowledge, this is the first work that includes semantic features jointly with lexical and syntactic features in the framework of a syntax-based RM. Using syntactic features, such as dependency relations, enables the RM to make syntactic generalizations. For instance, the RM can learn that in translating between subject-verb-object (SVO) and subject-object-verb (SOV) languages, the object and the verb should be swapped.

On top of this syntactic generalization, the RM should be able to make semantic generalizations. To this end, we use WordNet synsets as an additional feature in our RM. WordNet is a lexical database of English which groups words into sets of cognitive synonyms. In other words, in WordNet a set of synonym words belong to the same synset. For example, the words "baby", "babe" and "infant" are in the same synset in WordNet. The use of synsets enables our RM to be generalized from words seen in the training data to any of their synonyms present in WordNet.

4 Experiments

4.1 Data and Setup

We used the Mizan English–Farsi parallel corpus [2] (Supreme Council of Information and Communication Technology, 2013), which contains

[2] http://dadegan.ir/catalog/mizan

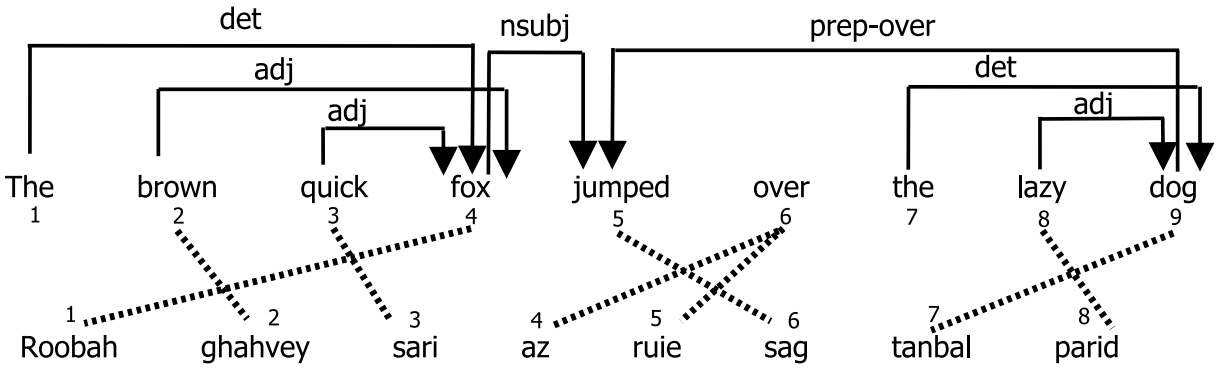

Figure 1: An example dependency tree for an English source sentence, its translation in Farsi and the word alignments

head	jumped	jumped	fox	fox	fox	dog	dog
dependant	fox	dog	the	brown	quick	the	lazy
dependant 1	fox	brown	the	the	the		
dependant 2	dog	quick	brown	quick	lazy		

Table 2: head-dependant and dependant-dependant pairs for the sentence in Figure 1.

around one million sentences extracted from English novel books and their translation in Farsi. We randomly held out 3,000 and 1,000 sentence pairs for tuning and testing, respectively, and used the remaining sentence pairs for training. Table 5 shows statistics (number of words and sentences) of the data sets used for training, tuning and testing.

	Unit	**English**	**Farsi**
Train	sentences	1,016,758	1,016,758
	words	13,919,071	14,043,499
Tune	sentences	3,000	3,000
	words	40,831	41,670
Test	sentences	1,000	1,000
	words	13,165	13,444

Table 5: Mizan parallel corpus statistics

We used GIZA++ (Och and Ney, 2003) to align the words in the English and Farsi sentences. We parsed the English sentences of our parallel corpus with the Stanford dependency parser (Chen and Manning, 2014) and used the "collapsed representation" of its output which shows the direct dependencies between the words in the English sentence. Having obtained both dependency trees and the word alignments, we extracted 6,391,956 (*head-dep*) and 5,247,526 (*dep-dep*) pairs from our training data set and determined the orientation for each pair based on Equation 1. We then trained a Maximum Entropy classifier (Manning and Klein, 2003) (henceforth MaxEnt) on the extracted constituent pairs from the training data set and use it to predict the orientation probability of each pair of constituents in the tune and test data sets. As mentioned earlier, we used WordNet in order to determine the synset of the English words in the data set.

Our baseline SMT system is the Moses implementation of the HPB-SMT model with default settings (Hoang et al., 2009). We used a 5-gram language model and trained it on the Farsi side of the training data set. All experiments used MIRA for tuning the weights of the features used in the HPB model (Cherry and Foster, 2012).

The semantic features (synsets) are extracted from WordNet 3.0. For each word, we take the synset that corresponds to its first sense, i.e. the most common one. An alternative would be to apply a word sense disambiguation algorithm. However, these have been shown to perform worse than the first-sense heuristic when WordNet is the inventory of word senses, e.g. (Pedersen and Kolhatkar, 2009; Snyder and Palmer, 2004).

4.2 Evaluation: MT Results

We selected different feature sets for (*head-dep*) and (*dep-dep*) pairs from Table 3 and Table 4 respectively, then we used them in our MaxEnt classifier to determine the impact of our novel se-

Features	Type	Description
$lex(head)$,$lex(dep)$	lexical	surface forms of the head and dependent word
$depRel(dep)$	syntactic	dependency relation of the dependent word
$syn(head)$,$syn(dep)$	semantic	synsets of the head and dependent word

Table 3: Features for ($head$-dep) constituent pairs

Features	Type	Description
$lex(head)$,$lex(dep1)$,$lex(dep2)$	lexical	surface forms of the mutual head and dependent words
$depRel(dep1)$,$depRel(dep2)$	syntactic	dependency relation of the dependent words
$syn(head)$,$syn(dep1)$,$syn(dep2)$	semantic	synsets of the head and dependent words

Table 4: Features for (dep-dep) constituent pairs

mantic features (WordNet synsets) on the quality of the MT system. 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*). We build six MT systems, as shown in Table 6, according to the constituent pairs and feature sets examined.

We compared our MT systems to the standard HPB-SMT system. Each MT system is tuned three times and we report the average scores obtained with multeval[3] (Clark et al., 2011) on the MT outputs. The results obtained by each of the MT systems according to two widely used automatic evaluation metrics (BLEU (Papineni et al., 2002), and TER (Snover et al., 2006)) are shown in Table 7. The relative improvement of each evaluation metric over the baseline HPB is shown in columns $diff$.

Compared to the use of surface features, our novel semantic features based on WordNet synsets lead to better scores for both (head- dep) and (dep-dep) constituent pairs according to both evaluation metrics, BLEU and TER (except for the dd system in terms of TER, where there is a slight but insignificant increase (79.8 vs. 79.7)).

5 Conclusions and Future Work

In this paper we have extended a syntax-based RM for HPB-SMT with semantic features (WordNet synsets), in order to enable the model to generalize to phrases not seen in the training data but that have equivalent meaning. The inclusion of synsets has led to the best BLEU score in our experiments, outperforming the baseline (standard HPB-SMT) by 0.6 points absolute.

[3] https://github.com/jhclark/multeval

As for future work, we propose to work mainly along the following two directions. First, an investigation of the extent to which using a WordNet-informed approach to classify the words into semantic classes (as proposed in this work) outperforms an unsupervised approach via word clustering. Second, an in-depth human evaluation to gain further insights of the exact contribution of WordNet to the translation output.

Acknowledgments

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (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) and by the University of Isfahan.

References

Alexandra Birch, Phil Blunsom, and Miles Osborne. 2009. A Quantitative Analysis of Reordering Phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Athens, Greece.

Arianna Bisazza and Marcello Federico. 2013. Dynamically shaping the reordering search space of phrase-based statistical machine translation. *Transactions of the ACL*, (1):327–340.

Arianna Bisazza and Marcello Federico. 2015. A survey of word reordering in statistical machine translation: Computational models and language phenomena. In *arXiv preprint arXiv:1502.04938*.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

MT System	Features
hd-surface	$Lex(head)$, $Lex(dep)$, $depRel(dep)$
hd-synset	$depRel(dep)$,$Syn(head)$, $Syn(dep)$
hd-both	$Lex(head)$,$Lex(dep)$,$depRel(dep)$,$Syn(dep)$,$Syn(head)$
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	$Lex(head)$,$Lex(dep1)$,$Lex(dep2)$,$Syn(head)$,$Syn(dep1)$,$Syn(dep2)$, $depRel(dep1)$,$depRel(dep2)$

Table 6: Examined features for MT systems

System	BLEU ↑					TER ↓				
	Avg	diff	$\overline{s}_{\text{sel}}$	s_{Test}	p-value	Avg	diff	$\overline{s}_{\text{sel}}$	s_{Test}	p-value
baseline	10.9	-	0.6	0.0	-	80.3	-	0.8	0.0	-
dd-surface	11.4	4.58%	0.7	0.1	0.00	**79.7**	-0.74%	0.8	0.2	0.01
dd-syn	11.3	3.66%	0.6	0.2	0.01	79.8	-0.62%	0.8	0.2	0.05
dd-both	**11.5**	5.50%	0.7	0.2	0.00	79.8	-0.62%	0.8	0.5	0.02
hd-surface	11.1	2.18%	0.6	0.1	0.08	80.9	0.74%	0.8	0.3	0.01
hd-syn	11.3	3.66%	0.6	0.1	0.00	80.5	0.24%	0.8	0.2	0.40
hd-both	11.1	2.18%	0.6	0.1	0.06	81.1	0.99%	0.8	0.3	0.00

Table 7: MT scores for all systems. 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). Metric scores are averages over three runs. s_{sel} indicates the variance due to test set selection and has nothing to do with optimizer instability. The best result according to each metric (highest for BLEU and lowest for TER) is shown in bold.

Colin Cherry and George Foster. 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*, pages 427–436.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Yang Gao, Philipp Koehn, and Alexandra Birch. 2011. Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 857–868.

Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. 2013. Distortion model considering rich context for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 155–165.

Spence Green, Michel Galley, and Christopher D. Manning. 2010. Improved models of distortion cost for statistical machine translation. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, page 867875.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT*, pages 152–159.

Arefeh Kazemi, Antonio Toral, Andy Way, Amirhassan Monadjemi, and Mohammadali Nematbakhsh. 2015. Dependency-based reordering model for constituent pairs in hierarchical smt. In *Proceedings of*

the 18th Annual Conference of the European Association for Machine Translation, pages 43–50, Antalya, Turkey, May.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–133.

Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 68–75.

Ding Liu and Daniel Gilda. 2010. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 716–724.

Christopher Manning and Dan Klein. 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*, pages 8–8.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Ted Pedersen and Varada Kolhatkar. 2009. Wordnet::senserelate::allwords: A broad coverage word sense tagger that maximizes semantic relatedness. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session*, NAACL-Demonstrations '09, pages 17–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Translation in the Americas*.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Supreme Council of Information and Communication Technology. 2013. Mizan English-Persian Parallel Corpus. Tehran, I.R. Iran.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *StatMT '06 Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63.