THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Investigating the Usefulness of Generalized Word Representations in SMT

Citation for published version:
Durrani, N, Koehn, P, Schmid, H & Fraser, AM 2014, Investigating the Usefulness of Generalized Word Representations in SMT. in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland.* pp. 421-432. <http://aclweb.org/anthology/C/C14/C14-1041.pdf>

Link:
Link to publication record in Edinburgh Research Explorer

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

Published In:
COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 09. Dec. 2024

Investigating the Usefulness of Generalized Word Representations in SMT

Nadir Durrani
University of Edinburgh
dnadir@inf.ed.ac.uk

Philipp Koehn
University of Edinburgh
pkoehn@inf.ed.ac.uk

Helmut Schmid **Alexander Fraser**
Ludwig Maximilian University Munich
fraser,schmid@cis.uni-muenchen.de

Abstract

We investigate the use of generalized representations (POS, morphological analysis and word clusters) in phrase-based models and the N-gram-based *Operation Sequence Model (OSM)*. Our integration enables these models to learn richer lexical and reordering patterns, consider wider contextual information and generalize better in sparse data conditions. When interpolating generalized OSM models on the standard IWSLT and WMT tasks we observed improvements of up to +1.35 on the English-to-German task and +0.63 for the German-to-English task. Using automatically generated word classes in standard phrase-based models and the OSM models yields an average improvement of +0.80 across 8 language pairs on the IWSLT shared task.

1 Introduction

The increasing availability of digital text has galvanized the use of empirical methods in many fields including Machine Translation. Given bilingual text, it is now possible to automatically learn translation rules that required years of effort previously. Bilingual data, however, is abundantly available for only a handful of language pairs. The problem of reliably estimating statistical models for translation becomes more of a challenge under sparse data conditions especially when translating into morphologically rich or syntactically divergent languages. The former becomes challenging due to lexical sparsity and the latter suffers from sparsity in learning underlying reordering patterns. The last decade of research in Statistical Machine Translation has witnessed many attempts to integrate linguistic analysis into SMT models, to address the challenges of (i) translating into morphologically rich language languages, (ii) modeling syntactic divergence across languages for better generalization in sparse data conditions.

The integration of the *Operation Sequence Model* into phrase-based paradigm (Durrani et al., 2013a; Durrani et al., 2013b) improved the reordering capability and addressed the problem of the phrasal independence assumption in the phrase-based models. The OSM model integrates translation and reordering into a single generative story. By jointly considering translation and reordering context across phrasal boundaries, the OSM model considers much richer conditioning than phrasal translation and lexicalized reordering models. However, due to data sparsity the model often falls back to very small context sizes. We address this problem by learning operation sequences over generalized representations such as POS and Morph tags. This enables us to learn richer translation and reordering patterns that can generalize better in sparse data conditions. The model benefits from wider contextual information as we show empirically in our results.

We investigate two methods to combine generalized OSM models with the lexically driven OSM model and experimented on German-English translation tasks. Our best system that uses a linear combination of different OSM models gives significant improvements over a competitive baseline system. An improvement of up to +1.35 was observed on the English-to-German and up to +0.63 BLEU points on the German-to-English task over a factored augmented baseline system (Koehn and Hoang, 2007).

POS taggers and morphological analyzers, however, are not available for many resource poor languages. In the second half of the paper we investigate whether annotating the data with automatic word

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 421–432, Dublin, Ireland, August 23-29 2014.

clusters helps improve the performance. Word clustering is similar to POS-tagging/Morphological annotation except that it also captures interesting syntactic and lexical semantics, for example countries and languages are grouped in separate clusters, animate objects are differentiated from inanimate objects, colors are grouped in a separate cluster etc. Word clusters, however, deterministically map each word type to a unique[1] cluster, unlike POS/Morph tagging, and therefore might be less useful for disambiguation. We use the `mkcls` utility in GIZA (Och and Ney, 2003) to cluster source and target vocabularies into classes and will therefore refer to automatic classes as Och clusters/classes in this paper.

We first use Och classes as an additional factor in phrase-based translation model, along with a target LM model over cluster-ids to improve the baseline system. We then additionally use the OSM model over cluster-ids. Our experiments include translation from English to Dutch, French, Italian, Polish, Portuguese, Russian, Spanish, Slovenian and Turkish on IWSLT shared task data. Our results show an average improvement of +0.80, ranging from +0.41 to +2.02. Compared to the improved baseline system obtained by using Och classes as a factor in phrase-based translation models, adding an OSM model over cluster-ids improved performance in four (French, Spanish, Dutch and Slovenian) out of eight cases. In other cases performance stayed constant or dropped slightly. We also used POS annotations for three tasks, namely translating from English into French, Spanish and Dutch to compare the performance of the two different kinds of generalizations. Surprisingly, using Och classes always performed better than using POS annotations. The rest of the paper is organized as follows. Section 2 gives an account on related work. Section 3 discusses the factor-based OSM model. Section 4 presents the experimental setup and the results. Section 5 concludes the paper.

2 Related Work

Previous work on integrating linguistic knowledge into SMT models can be broken into two groups. The first group focuses on using linguistic knowledge to improve reordering between syntactically different languages. A second group focuses on translating into morphologically rich languages.

Initial efforts to use linguistic annotation focused on rearranging source sentences to be in the target order. Xia and McCord (2004) proposed a method to automatically learn rewrite rules to preorder source sentences. Collins et al. (2005) and Popović and Ney (2006) proposed methods for reordering the source using a small set of handcrafted rules. Crego and Mariño (2007) use syntactic trees to derive rewrite rules. Hoang and Koehn (2009) used POS tags to create templates for surface word translation to create longer phrase translation. A whole new paradigm of using syntactic annotation to address long range reorderings has emerged following Galley et al. (2006), Zollmann and Venugopal (2006), Chiang (2007) etc. Crego and Yvon (2010) and Niehues et al. (2011) used a Tuple Sequence Model (TSM) over POS tags in an N-gram-based search to improve mid-range reorderings. Our work is similar to them except that OSM model is substantially different from the TSM model as it integrates both the translation and reordering mechanisms into a combined model. Therefore both translation and reordering decisions can benefit from richer generalized representations.

A second group of work addresses the problem of translating into morphologically richer languages. The idea of translating to stems and then inflecting the stems in a separate step has been studied by Toutanova et al. (2008), de Gispert and Mariño (2008), Fraser et al. (2012), Chahuneau et al. (2013) and others. Koehn and Hoang (2007) proposed to integrate different levels of linguistic information as factors into the phrase-based translation model. Yeniterzi and Oflazer (2010) used source syntactic structures as additional complex tag factors for English-to-Turkish phrase-based machine translation. Green and DeNero (2012) proposed a target-side, class-based agreement model to handle morpho-syntactic agreement errors when translating from English-to-Arabic. El Kholy and Habash (2012) tested three models to find out which features are best handled by modeling them as a part of translation, and which ones are better predicted through generation, also in the English-to-Arabic task. Several researchers attempted to use word lattices to handle generalized representation (Dyer et al., 2008; Hardmeier et al., 2010; Wuebker and Ney, 2012). Automatically clustering the training data into word classes in order to obtain smoother

[1]We are referring to hard clustering here. Soft clustering is intractable as it requires a marginalization over all possible classes when calculating the n-gram probabilities.

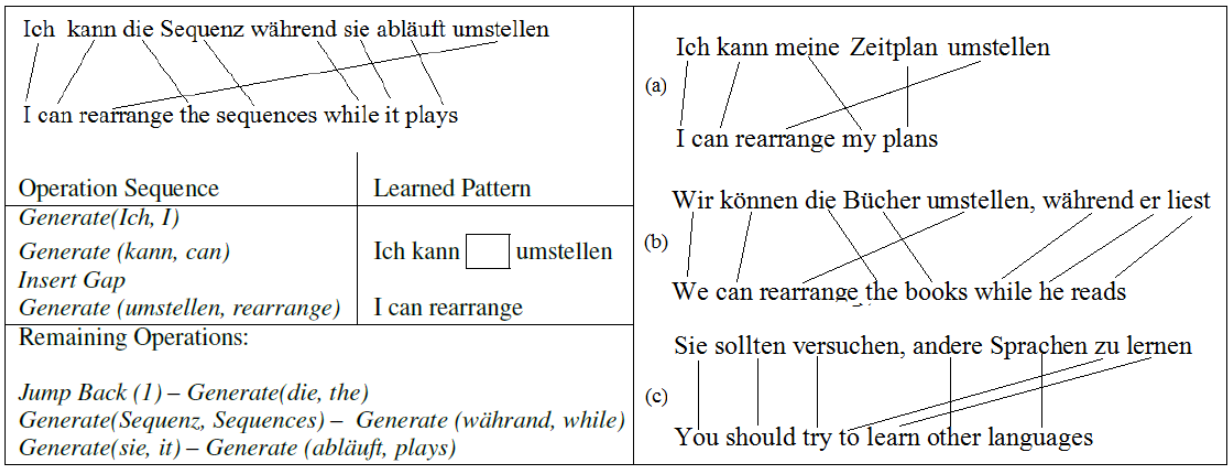


Figure 1: Operation Sequence Model – Training Sentence with Generation and Test Sentences

distributions and better generalizations has been a widely known and applied technique in natural language processing. Training based on word classes has been previously explored by various researchers. Cherry (2013) addressed data sparsity in lexicalized reordering models by using sparse features based on word classes. Other parallel attempts on using word-class models include Wuebker et al. (2013), Chahuneau et al. (2013) and Bisazza and Monz (2014).

More recent research has started to set apart from the conventional maximum likelihood estimates toward neural network-based models that use continuous space representation (Schwenk, 2012; Le et al., 2012; Hu et al., 2014; Gao et al., 2014). Although these methods have achieved impressive improvements, traditional models continue to dominate the field due to their simplicity and low computational complexity. How much of the improvement will be retained when scaling these models to all available data instead of a limited amount will be interesting.

3 Operation Sequence Model

The Operation Sequence Model (Durrani et al., 2011) is an instance of the N-gram based SMT framework (Casacuberta and Vidal, 2004; Mariño et al., 2006). It represents the translation process through a sequence of operations. An operation can be to simultaneously generate source or target words or to perform reordering. Reordering is carried out through jump and gap operations. The model is different from its ancestors in that it strongly integrates translation and reordering into a single generative story in which translation decisions can influence and get impacted by the reordering decisions and vice versa. Given a bilingual sentence pair $< F, E >$ and its alignment A, a sequence of operations $o_1, o_2 \ldots, o_J$ is generated deterministically through a conversion algorithm. The model is learned by learning Markov chains over these sequences and is formally defined as:

$$p_{osm}(F, E, A) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, ..., o_{j-1})$$

Figure 1 shows an example of an aligned bilingual sentence pair and the corresponding operation sequence used to generate it. There is a 1-1 correspondence between a sentence pair and its operation sequence. We thus get a unique sequence for every bilingual sentence pair given the alignment.

3.1 Motivation

Due to data sparsity it is impossible to observe all possible reordering patterns with all possible lexical choices in translation operations. The lexically driven OSM model therefore often backs off to very small context sizes. Coming back to the training example in Figure 1. The useful reordering pattern

learned through this example is:

Ich kann □ umstellen → I can rearrange

which is memorized through the operation sequence:

Generate(Ich, I) – Generate(kann, can) – Insert Gap – Generate(umstellen, rearrange)

It can generalize to the test sentence shown in Figure 1(a). However, it fails to generalize to the sentences in Figure 1(b) and (c) although the underlying reordering pattern is the same. The second part of the German verb complex usually appears at the end of a clause or a sentence and needs to be moved in order to produce the correct English word order. However, due to data sparsity such a combination of lexical decisions and reordering decisions may not be observed during training. The model would therefore fail to generalize in such circumstances. This problem can be addressed by learning a generalized form of the same reordering rule. By annotating the corpus with word classes such as POS tags, we obtain the reordering pattern:

PPER VMFIN □ VVINF → PP MD VB

memorized through the operation sequence:

Generate (PPER,PP) – Generate (VMFIN,MD) – Insert Gap – Generate (VVINF,VB)

This rule generalizes to all the test sentences in Figure 1. Since the OSM model strongly couples translation and reordering, the probability of each translation or reordering operation depends on the n previous translation/reordering decisions. The generalization of the model by replacing words with POS tags allows the model to consider a wider syntactic context, thus improving lexical decisions and the reordering capability of the model. Using different kinds of word classes, we can also control the type of abstraction. Using lemmas for example, we can map different forms of the verb "können – can" (kann, kannst, konnte) to a single class. Och clusters can provide different levels of granularity.

3.2 Models

Given that we can learn OSM models over different word representations, the question then is how to combine the lexically driven OSM model with an OSM model based on a generalized word representation. The simplest approach is to treat each OSM model as a separate feature in the log-linear framework, thus summing up the weighted log probabilities. The effect of this is similar to an *And* operation. A translation is considered good if both, the word-based OSM and the POS-based OSM models indicate that it is a good translation. However, an *Or* operation might be more desirable in some scenarios. The operation *Generate (trotz, in spite of)* should be ranked high although the POS-based operation *Generate(APPR, IN IN IN)* is improbable. Similarly, the generalized operation sequence:

Insert Gap – Generate (ADJ, JJ) – Jump Back – Generate (NOM, NN)

that captures the swapping of noun and adjective in French-English, should be ranked higher even though *noir* (black) never appeared after *cheval* (horse) during training and the sequence:

Insert Gap – Generate (noir, black) – Jump Back – Generate (cheval , horse)

is never observed. Instead of using both the models, a single model that could switch between different generalized OSMs during translation and choose the one which gives the best prediction in each situation, can be used. In order to achieve this effect, we formulated a second model that interpolates the lexically driven OSM model with its generalized variants. However, we can only

interpolate two models that predict the same representation. The lexically driven OSM predicts the surface forms whereas the POS-based OSM predicts POS translations. To make the two comparable, we multiply the POS-based OSM probability with the probability of the lexical operation given the POS operation. More specifically the probability of the generalized model gm can be defined as:

$$p_{gm}(o_j|o_{j-n+1}^{j-1}) = p_{osm_{pos}}(o'_j|o'^{j-1}_{j-n+1})\ p(o_j|o'_j) \tag{1}$$

where $p_{osm_{pos}}$ is the operation sequence model learned over POS tags and $p(o_j|o'_j)$ is the probability of the lexical operation given the POS-based operation. It is 1 for all reordering operations. We assume here that for each lexical operation o_j a corresponding POS-based operation o'_j is uniquely determined. With $p_{osm_{sur}} = p_{osm_{sur}}(o_j|o_{j-n+1}^{j-1})$ (lexically driven OSM model) and $p_{gm} = p_{gm}(o_j|o_{j-n+1}^{j-1})$ (generalized OSM model as described above), the overall probability of the new model p_{osm} is defined as:

$$p_{osm} = \alpha p_{osm_{sur}} + (1-\alpha)p_{gm} \tag{2}$$

Such an interpolation is expensive in the discriminative training. It would require a sub-tuning routine inside of tuning, a main loop to train all the features including the OSM model and an inner loop to distribute the weight assigned to OSM model among lexically driven and POS-based OSM models. We therefore just take the larger one of the two model values and add a POS-based translation penalty ϕ. The value of this penalty is the number of times that the POS-based operation was chosen when translating a sentence. This penalty acts similarly as the prior α above. Using this formulation, the model could therefore be redefined as:

$$p_{osm} = \begin{cases} p_{osm_{sur}} & \text{if } p_{osm_{sur}} \geq e^{\lambda} p_{gm} \\ e^{\lambda} p_{gm} & \text{otherwise} \end{cases} \tag{3}$$

where λ is the weight for the POS driven translation penalty ϕ. This allows the optimizer to control whether it prefers the lexically driven or the POS-driven OSM model. By setting a very low weight λ the optimizer can force the translator to always choose lexically driven OSM. This formulation can be extended to multiple generalized OSM models based on e.g. POS tags, morphological tags, or word clusters. Equation 2 can be rewritten as follows:

$$p_{osm} = \alpha_1 p_{osm_{sur}} + \sum_{i=2}^{n} \alpha_i p_{gm_i} \tag{4}$$

with $\sum_{i=1}^{n} \alpha_i = 1$ and p_{gm_i} defined analogous to Equation 1.

Setting $p_{gm_1} = p_{osm_{sur}}$ and $\lambda_1 = 0$, we can again simplify Equation 4 by taking the maximum to:

$$p_{osm} = \max_{i=1}^{n} e^{\lambda_i} p_{gm_i} \tag{5}$$

We use a translation penalty ϕ_i for each generalized model and tune its weight λ_i along with the weights of other features. We will refer to this model as $\mathbf{Model_{or}}$ in this paper and the commonly used log-linear interpolation of the features as $\mathbf{Model_{and}}$. The intuition behind $\mathbf{Model_{or}}$ is that we back-off to generalized representations only when the lexically driven model doesn't provide enough contextual evidence. The downside of this approach, however, is that unlike $\mathbf{Model_{and}}$, it cannot distribute weights over multiple features and solely relies on a single model.

4 Evaluation

Data: We ran experiments with data made available for the translation task of the IWSLT-13 (Cettolo et al., 2013): International Workshop on Spoken Language Translation[2] and WMT-13 (Bojar et al., 2013): Eighth Workshop on Statistical Machine Translation.[3] The sizes of bitext used for the estimation of translation and monolingual language models are reported in Table 1.

We used LoPar (Schmid, 2000) to obtain morphological analysis and POS annotation of German and MXPOST (Ratnaparkhi, 1998), a maximum entropy model for English POS tags. For other language pairs we used TreeTagger (Schmid, 1994).

[2] http://www.iwslt2013.org/

[3] http://www.statmt.org/wmt13/

Pair	Parallel	Monolingual	Pair	Parallel	Monolingual	Pair	Parallel	Monolingual
de–en	≈4.6 M	≈287.3 M	en–de	≈4.6 M	≈59.5 M	en–fr	≈5.5 M	≈69 M
en–es	≈4.1 M	≈59.6 M	en–nl	≈2.1 M	≈21.7 M	en–ru	≈1.15 M	≈21 M
en–pt	≈1.0 M	≈2.3 M	en–pl	≈0.77 M	≈0.8 M	en–sl	≈0.63 M	≈0.65 M
en–tr	≈0.13 M	≈0.14 M						

Table 1: Number of Sentences (in Millions) used for Training

Model	$\mathbf{iwslt_{10}}$	$\mathbf{wmt_{13}}$	$\mathbf{iwslt_{10}}$	$\mathbf{wmt_{13}}$
	English-to-German		German-to-English	
Baseline	23.56	20.38	31.46	27.27
$\mathbf{M_{and_{(pos,pos)}}}$	**23.93** Δ+0.37	20.61 Δ+0.23	**31.91** Δ+0.45	27.55 Δ+0.28
$\mathbf{M_{and_{(pos,morph)}}}$	**24.62** Δ+1.06	**20.88** Δ+0.50	**32.09** Δ+0.63	**27.62** Δ+0.35
$\mathbf{M_{and_{(all)}}}$	**24.91** Δ+1.35	**20.93** Δ+0.55	**32.00** Δ+0.54	**27.71** Δ+0.44
$\mathbf{M_{or_{(pos,pos)}}}$	23.61 Δ+0.05	20.24 Δ-0.14	31.55 Δ+0.09	27.32 Δ+0.05
$\mathbf{M_{or_{(pos,morph)}}}$	23.83 Δ+0.27	20.44 Δ+0.08	31.58 Δ+0.12	27.20 Δ-0.07
$\mathbf{M_{or_{(all)}}}$	23.88 Δ+0.32	20.55 Δ+0.17	31.40 Δ-0.06	27.15 Δ-0.12

Table 2: Evaluating Generalized OSM Models for German-English pairs – Bold: Statistically Significant (Koehn, 2004) w.r.t Baseline

Baseline System: We trained a Moses system (Koehn et al., 2007), replicating the settings described in (Birch et al., 2013) developed for the 2013 Workshop on Spoken Language Translation. The features included: a maximum sentence length of 80, grow-diag-final-and symmetrization of GIZA++ alignments, an interpolated Kneser-Ney smoothed 5-gram language model with KenLM (Heafield, 2011) used at runtime, a lexically-driven 5-gram operation sequence model (Durrani et al., 2013b) with 4 additional supportive features: 2 gap-based penalties, 1 distance-based feature and 1 deletion penalty, lexicalized reordering, sparse lexical and domain features (Hasler et al., 2012), a distortion limit of 6, 100-best translation options, Minimum Bayes Risk decoding (Kumar and Byrne, 2004), Cube Pruning (Huang and Chiang, 2007), with a stack-size of 1000 during tuning and 5000 during test and the no-reordering-over-punctuation heuristic. We used the compact phrase table representation by Junczys-Dowmunt (2012). For our German-to-English experiments, we used compound splitting (Koehn and Knight, 2003). German-to-English and English-to-German baseline systems also used POS and morphological target sequence models built on the in-domain subset of the parallel corpus using Kneser-Ney smoothed 7-gram models and as additional factors in phrase translation models (Koehn and Hoang, 2007). We used an unsupervised transliteration model (Durrani et al., 2014) to transliterate OOV words when translating into Russian.

Tuning and Test: The systems were tuned on the dev2010 dataset and evaluated on the test2010-2013 datasets made available for the IWSLT-13 workshop. We performed a secondary set of experiments for German-English pairs using tuning and test sets made available for the WMT-13 workshop. We concatenated the news-test sets 2008 and 2009 to obtain a large dev-set of 4576 sentences. Evaluation was performed on the news-test set 2013 which contains 3000 sentences. Tuning was performed using the k-best batch MIRA algorithm (Cherry and Foster, 2012) with at most 25 iterations. We use BLEU (Papineni et al., 2002) as a metric to evaluate our results.

Results I – Using Linguistic Annotation: We trained 5-gram OSM models over different representations and added these to the baseline system. First we evaluated $\mathbf{Model_{and}}$ ($\mathbf{M_{and}}$) which uses a MIRA tuned linear combination of different OSM models versus $\mathbf{Model_{or}}$ ($\mathbf{M_{or}}$) which computes only one OSM model but allows the generator to switch between different OSM models built on various generalized forms. Table 2 shows results from running experiments on German-English pairs. We found that the simpler model $\mathbf{Model_{and}}$ outperforms $\mathbf{Model_{or}}$ in all the experiments. $\mathbf{Model_{or}}$ does not give significant improvements over the baseline system and shows an occasional drop. This result is contrary to the expectation formulated in Section 3.2. We speculate that the optimizer faces problems to train this kind of model, because it cannot take into account that the selected OSM model can change when the weight parameter is modified. It assumes that the feature stays constant. In our formulation the same

derivation can occur with different feature scores in different decoding runs and the optimizer is unable to handle this. Our speculation is based on the observation of λ_ϕ, the weight of feature ϕ which allows the translator to switch between different OSM models. The value of λ_ϕ was not stable across different iterations and different experiments.

$\mathbf{Model_{and}}$ consistently improves the baseline. Adding an OSM model over [pos, morph] (source:pos, target:morph) combination gave the best results, giving a statistically significant gain of +1.06 on the $\mathbf{iwslt_{10}}$ test-set and +0.50 on the $\mathbf{wmt_{13}}$ test-set. Using an OSM model over a [pos,pos] combination also showed improvements, however, not as much as using morphological tags. Morphological tags provide richer information for disambiguation when translating into German. Note that the baseline system also used a target sequence model over morphological tags. Nevertheless using an OSM [pos,morph] model still gives significant improvements which shows that learning a joint model over source and target units is more fruitful than only considering target-side information. Using both the models together gave best results for English-to-German giving a further improvement of +0.29 on the $\mathbf{iwslt_{10}}$ task but no real gain on the $\mathbf{wmt_{13}}$ task. Using morphological tags also produced the best results for the German-to-English pair, giving a statistically significant gain of +0.63 on $\mathbf{iwslt_{10}}$ and +0.35 on $\mathbf{wmt_{13}}$. Using both the models together did not give any further significant improvements. The results changed by +0.10 and -0.09 on the $\mathbf{wmt_{13}}$ and $\mathbf{iwslt_{10}}$ test-sets respectively.

Results-II – Using Och Classes: In our secondary experiments we tested the effect of using Och clusters. The overall goal was to study whether using unsupervised word classes can serve the same purpose as POS tags and to compare the two methods of annotating the data. We obtained Och clusters using the `mkcls` utility (Och, 1999) in GIZA++ (Och and Ney, 2003). This is generally run during the alignment process where data is divided into 50 classes to estimate IBM Model-4. Chahuneau et al. (2013) found mapping data to 600 Och clusters useful, so we used this as well. We additionally experimented with using 200 and 1000 classes. We integrated Och clusters as additional factors[4] when training the phrase-translation models and used a monolingual n-gram model over cluster-ids built on the target-side of the in-domain corpus. Then we added a 5-gram OSM model over cluster-ids. We replace surface forms with their cluster-ids in source and target corpus and convert it to operation sequences, that jointly generate source and target cluster-ids. We only used $\mathbf{Model_{and}}$ for these experiments when adding an OSM model over cluster-ids.

	$\mathbf{B_0}$	**50**	**200**	**600**	**1000**	**POS**	**50**	**200**	**600**	**1000**	**POS**
		Target Sequence Model over Word Clusters					**Operation Sequence Model over Word Clusters**				
en – fr	33.17	33.30	33.40	33.05	33.05	33.14	**33.76**	33.74	33.58	33.75	33.03
en – es	34.14	34.33	34.58	34.46	33.96	33.91	**34.73**	34.62	34.60	34.55	34.35
en – nl	26.51	26.67	26.15	26.31	26.47	26.55	**26.91**	26.52	26.61	26.49	26.62
en – ru	13.12	13.34	13.51	13.53	**13.97**	–	13.61	13.66	13.80	13.63	–
en – sl	17.98	18.67	18.55	17.67	17.97	–	18.64	**18.91**	18.17	17.98	–
en – pt	30.80	31.62	32.21	32.40	32.44	–	31.77	32.44	32.34	31.90	–
en – pl	9.74	9.90	10.11	10.05	10.43	–	10.06	10.19	10.24	10.14	–
en – tr	7.18	7.43	7.45	7.50	7.50	–	7.26	7.28	7.51	7.54	–

Table 3: Evaluating Phrase-based and N-gram-based Translation Models over Och Clusters

Table 3 shows results from using models based on cluster-ids. The left side of the table evaluate the use of adding a target sequence model over cluster-ids using a factored-based translation model. Results improved consistently in all resource poor languages (pt, pl, tr) giving significant improvements in most of the cases. Mixed results were obtained for the pairs with a reasonable amount of parallel data (fr, es, nl), showing an occasional drop in performance. However, improvements can be found for all the language pairs.

[4]Note that adding cluster-ids in factored models alone has no impact in this scenario, as we are using hard clustering (each word deterministically maps onto a unique cluster-id). In a joint source-target factored model which is what we are using, it will result in an identical distribution as the baseline system.

In the right half of the table we tested whether additionally using an OSM model built over cluster-ids, on top of a phrase-based system that uses cluster-ids as factor and target language model, improves the performance any further. Consistent improvements were seen in Spanish and French. Better systems were produced in the case of French, Spanish, Dutch and Slovenian. No improvements were observed for Turkish and Portuguese whereas the performance got worse in Polish and Russian.

Using 50 classes consistently improved the baseline. Different numbers of clusters provide different levels of abstraction and granularity. We also tried using OSM models over different numbers of clusters simultaneously for English-to-Spanish, English-to-French and English-to-Dutch pairs in an effort to explore whether using different numbers of clusters to classify data provides different information. A slight gain was observed for EN-ES as the best system improved from 34.73 to 34.95. No further gains were observed for the other two pairs.

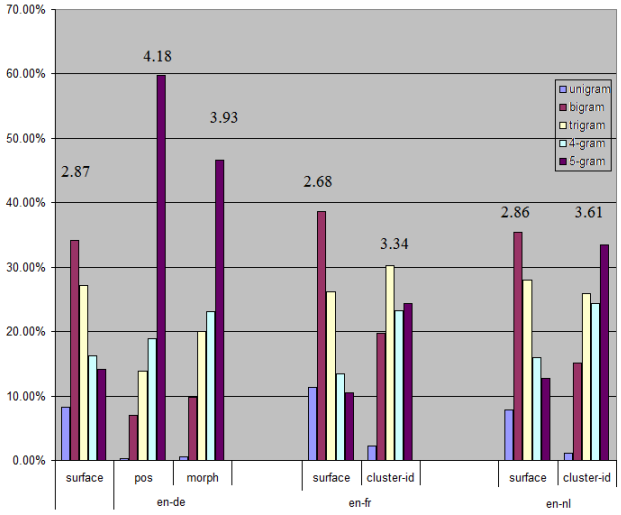
We also used POS annotation as a factor instead of Och clusters in French, Spanish and Dutch. See the POS columns of Table 3. Using POS as an additional factor, did not improve over the baseline performance. A significant drop was seen in the case of English-to-Spanish. Using a POS-based OSM on top of the POS-based phrase-model did not help either except for Spanish where results got improved by +0.44 over its phrase-based variant that used a POS factor. However, using Och clusters produced better results in all three cases. We speculate that the reason for this result is that Och clusters are more evenly distributed as compared to POS tags where the distribution is biased toward noun class and secondly Och clusters are optimized for language modeling. Also each word is deterministically mapped to a single class but can have multiple POS tags. The latter thus causes a sparser translation model. Finally Table 4 shows the comparison of results on $\mathbf{iwslt_{11-13}}$ by running baseline $\mathbf{B_0}$ and best systems $\mathbf{B_x}$ in Tables 3.

	$\mathbf{iwslt_{11}}$		$\mathbf{iwslt_{12}}$		$\mathbf{iwslt_{13}}$		**Avg**		
	$\mathbf{B_0}$	$\mathbf{B_x}$	$\mathbf{B_0}$	$\mathbf{B_x}$	$\mathbf{B_0}$	$\mathbf{B_x}$	$\mathbf{B_0}$	$\mathbf{B_x}$	$\mathbf{\Delta}$
en – fr	39.84	40.63	40.50	41.24	–	–	40.24	40.94	+0.70
en – es	32.89	33.24	26.45	26.81	34.01	34.73	31.12	31.60	+0.48
en – nl	30.01	30.31	26.40	26.72	24.96	25.57	27.12	27.53	+0.41
en – ru	14.93	15.91	13.01	13.53	15.65	16.4	14.53	15.28	+0.75
en – sl	–	–	11.34	12.40	12.85	13.73	12.09	13.10	+1.01
en – pt	31.61	33.62	33.24	34.91	30.83	33.24	31.89	33.92	+2.02
en – pl	12.73	13.13	9.52	10.50	11.30	11.54	11.18	11.72	+0.53
en – tr	7.01	7.42	6.99	7.43	6.21	6.84	6.74	7.23	+0.49
Avg	24.15	24.89	20.93	21.69	19.40	20.29	21.49	22.29	+0.80

Table 4: Evaluating on Test Sets $\mathbf{iwslt_{11-13}}$ – $\mathbf{B_0}$ = Baseline System, $\mathbf{B_x}$ = Best Systems in Tables 2

Analysis: In a post-evaluation analysis we confirmed whether using generalized OSM models actually consider a wider contextual window than its lexically driven variant. The graph shown in Figure 2 shows average context size considered (on top of each set of bars) and percentages of 1-5 gram matches by different OSM models. The results show that the probability of an operation is conditioned on less than a trigram in the OSM model over surface forms. In comparison OSM models over POS, morph or cluster-ids consider a window of roughly 4 previous operations thus considering more contextual information. The percentage of 5-gram matches increases from 15.5% to 59.2% using POS-based OSM model and up to 45.6% in morph-based OSM model, the number of unigram matches are decreased from 8.30% to less than 1% in both the models. Similar observation is made for the OSM models over clusters where 5-gram matches improve from 12% to 30% on average, showing the ability of the generalized models to use richer conditioning thus improving the translation quality.

We also analyzed what kind of words are clustered together using Och classes and found that clusters capture both syntax and lexical semantics. Figure 2 (b) shows several useful clusters to exhibit this. We also saw negative examples where words from different classes are clustered together. “Boy”, “Girl” and “Man” for example were clustered into a single class but “Woman” in another. Similarly “Grey” and “Orange” were grouped together with animated objects.



Colors: Red, Black, Green, Blue …
Animals: Spider, Elephant, Cat, Rabbit …
Cities: London, Paris, Chicago …
Languages: English, Arabic, Farsi …
Royal Family: Queen, King, Prince …
Food: Sugar, Salt, Pepper, Milk …
Stars: Sagittarius, Aquarius, Virgo …
Verbs: Improve, Accelerate, Improvise …
Adjectives: Stupid, Ugly, Funny, Rusty …

Colors: Rouge, Noir, Vert, Bleu…
Animals: Chat, Lapin, Serpent, Crocodile …
Cities: Londres, Paris, Lahore …
Languages: Urdu, Hindi, Tamil …
Royal Family: Roi, Prince, Reine …
Food: Sucre, Sel, Poivre, Lait…
Stars: Sagittarius, Aquarius, Capricorn …
Verbs: Améliorer, Accélérer, Improviser …
Adjectives: Stupide, Drôle …

Figure 2: (a) Average Size of N-grams Used in Different OSM Models and Percentages of 1-5 Gram Matches in Three Language Pairs (b) Different Word Clusters using 50 Classes

5 Conclusion

In this paper we investigated the usefulness of integrating word classes in phrase-based models and Operation Sequence N-gram models. We explored two models of interpolating generalized OSM models and tested variations on the standard IWSLT and WMT tasks. Our results showed that the simpler more commonly used method of integrating the models in the log-linear framework worked best. We showed that by learning OSM models over generalized POS and morphological representations, we were able to build richer models that outperformed state-of-the-art baseline systems. Statistically significant gains of up to +1.35 and +0.63 were observed in English-to-German and German-to-English tasks. We also made use of Och classes as additional factors in phrase translation and language models. These were tested translating from English to 8 different languages which includes a mixture of morphologically rich (French, Spanish and Russian, Dutch, and Turkish) and sparse data (Portuguese, Polish, Slovenian and Turkish) languages. Our results show that using clusters was helpful in all of the cases. Using the OSM model over word-clusters additionally improved the performance further. Our results show an average improvement of +0.80, ranging from +0.41 to +2.02. Our EN-FR systems were ranked third (on tst2013) and second (on tst2011-tst2012) in IWSLT-13 translation task following EU-Bridge (Freitag et al., 2013) which used our output for system combination. The code to train class-based models has been made available to the research community via the Moses toolkit. See Advanced Features[5] in the Moses Decoder for details.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback and suggestions. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 287658 (EU-Bridge) and n° 287688 (MateCat). Alexander Fraser was funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation. Helmut Schmid was supported by Deutsche Forschungsgemeinschaft grant SFB 732. This publication only reflects the authors' views.

References

Alexandra Birch, Nadir Durrani, and Philipp Koehn. 2013. Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation. In *Proceedings of the 10th International Workshop on Spoken Language Translation*,

[5] http://www.statmt.org/moses/?n=Moses.AdvancedFeatures

pages 40–48, Heidelberg, Germany, December.

Arianna Bisazza and Christof Monz. 2014. Class-Based Language Modeling for Translating into Morphologically Rich Languages. In *Proceedings of the 25th Annual Conference on Computational Linguistics (COLING)*, Dublin, Ireland, August.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation*, WMT-2013, pages 1–44, Sofia, Bulgaria.

Francisco Casacuberta and Enrique Vidal. 2004. Machine Translation with Inferred Stochastic Finite-State Transducers. *Computational Linguistics*, 30:205–225.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. *Proceedings of the International Workshop on Spoken Language Translation, Heidelberg, Germany*.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into Morphologically Rich Languages with Synthetic Phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

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, Montréal, Canada, June. Association for Computational Linguistics.

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, Atlanta, Georgia, June. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, MI.

Josep M. Crego and José B. Mariño. 2007. Syntax-Enhanced N-gram-Based SMT. In *Proceedings of the 11th Machine Translation Summit*, MT Summit XI, pages 111–118.

Josep M. Crego and François Yvon. 2010. Improving Reordering with Linguistically Informed Bilingual N-Grams. In *Coling 2010: Posters*, pages 197–205, Beijing, China, August. Coling 2010 Organizing Committee.

Adrià de Gispert and José B. Mariño. 2008. On the Impact of Morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12):1034–1046.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA, June.

Nadir Durrani, Alexander Fraser, and Helmut Schmid. 2013a. Model With Minimal Translation Units, But Decode With Phrases. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013b. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2014)*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing Word Lattice Translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1012–1020, Columbus, OH, USA. The Association for Computer Linguistics.

Ahmed El Kholy and Nizar Habash. 2012. Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation. volume 12.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674, Avignon, France, April. Association for Computational Linguistics.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Nadir Durrani, Matthias Huck, Philipp Koehn, Thanh-Le Ha, Jan Niehues, Mohammed Mediani, Teresa Herrmann, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *International Workshop on Spoken Language Translation*, Heidelberg, Germany, December.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proceedings of COLING-ACL*, pages 961–968, Sydney, Australia. Association for Computational Linguistics.

Jianfeng Gao, Xiaodong He, Scott Wen-tau Yih, and Li Deng. 2014. Learning Continuous Phrase Representations for Translation Modeling. In *Proceedings of the Association for Computational Linguistics*, Baltimore, MD, USA, June.

Spence Green and John DeNero. 2012. A Class-Based Agreement Model for Generating Accurately Inflected Translations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–155, Jeju Island, Korea, July. Association for Computational Linguistics.

Christian Hardmeier, Arianna Bisazza, and Marcello Federico. 2010. FBK at WMT 2010: Word Lattices for Morphological Reduction and Chunk-Based Reordering. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 88–92, Uppsala, Sweden, July. Association for Computational Linguistics.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2012. Sparse Lexicalised features and Topic Adaptation for SMT. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 268–275.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Hieu Hoang and Philipp Koehn. 2009. Improving Mid-Range Re-Ordering Using Templates of Factors. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 372–379, Athens, Greece, March. Association for Computational Linguistics.

Yuening Hu, Michael Auli, Qin Gao, and Jianfeng Gao. 2014. Minimum Translation Modeling with Recurrent Neural Networks. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 20–29, Gothenburg, Sweden, April. Association for Computational Linguistics.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic, June. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2012. Phrasal Rank-Encoding: Exploiting Phrase Redundancy and Translational Relations for Phrase Table Compression. *The Prague Bulletin of Mathematical Linguistics*, 98:63–74.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 187–193, Morristown, NJ.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007 Demonstrations*, Prague, Czech Republic.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July.

Shankar Kumar and William J. Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *HLT-NAACL*, pages 169–176.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous Space Translation Models with Neural Networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–48, Montréal, Canada, June. Association for Computational Linguistics.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-Based Machine Translation. *Computational Linguistics*, 32(4):527–549.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 198–206, Edinburgh, Scotland, July. Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz J. Och. 1999. An Efficient Method for Determining Bilingual Word Classes. In *Processings of EACL*, pages 71–76, Bergen, Norway.

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*, ACL '02, pages 311–318, Morristown, NJ, USA.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283, Genoa, Italy.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 2000. Lopar: Design and implementation. Bericht des sonderforschungsbereiches "sprachtheoretische grundlagen fr die computerlinguistik", Institute for Computational Linguistics, University of Stuttgart.

Holger Schwenk. 2012. Continuous Space Translation Models for Phrase-Based Statistical Machine Translation. In *Proceedings of COLING 2012: Posters*, pages 1071–1080, Mumbai, India, December. The COLING 2012 Organizing Committee.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June. Association for Computational Linguistics.

Joern Wuebker and Hermann Ney. 2012. Phrase Model Training for Statistical Machine Translation with Word Lattices of Preprocessing Alternatives. In *NAACL 2012 Seventh Workshop on Statistical Machine Translation*, pages 450–459, Montreal, Canada, June. Association for Computational Linguistics.

Joern Wuebker, Stephan Peitz, Felix Rietig, and Hermann Ney. 2013. Improving Statistical Machine Translation with Word Class Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, Washington, USA, October. Association for Computational Linguistics.

Fei Xia and Michael McCord. 2004. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of Coling 2004*, pages 508–514, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden, July. Association for Computational Linguistics.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June. Association for Computational Linguistics.