Machine Translation with Source-Predicted Target Morphology

Daiber, J.; Sima'an, K.

Publication date
2015
Document Version
Final published version
Published in
Proceedings of MT Summit XV. - Vol. 1

Machine Translation with Source-Predicted Target Morphology

Joachim Daiber J.Daiber@uva.nl
Khalil Sima'an K.Simaan@uva.nl
Institute for Logic, Language and Computation, University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands

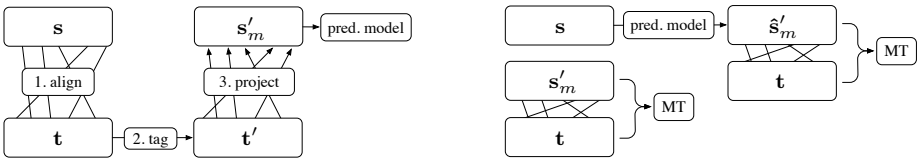
Abstract

We propose a novel pipeline for translation into morphologically rich languages which consists of two steps: initially, the source string is enriched with target morphological features and then fed into a translation model which takes care of reordering and lexical choice that matches the provided morphological features. As a proof of concept we first show improved translation performance for a phrase-based model translating source strings enriched with morphological features projected through the word alignments from target words to source words. Given this potential, we present a model for predicting target morphological features on the source string and its predicate-argument structure, and tackle two major technical challenges: (1) How to fit the morphological feature set to training data? and (2) How to integrate the morphology into the back-end phrase-based model such that it can also be trained on *projected* (rather than predicted) features for a more efficient pipeline? For the first challenge we present a latent variable model, and show that it learns a feature set with quality comparable to a manually selected set for German. And for the second challenge we present results showing that it is possible to bridge the gap between a model trained on a predicted and another model trained on a projected morphologically enriched parallel corpus. Finally we exhibit final translation results showing promising improvement over the baseline phrase-based system.

1 Introduction

Translation into a morphologically rich language poses a challenge for statistical machine translation systems. Rich morphology usually goes together with relatively freer word order of the target language, which makes it difficult to predict morphology and word order in tandem. Technically speaking this difficulty could be due to data sparsity, but possibly also due to morphological agreement between words over long distances. In this paper we explore the idea of combating sparsity by conducting translation in a probabilistic pipeline (chain rule), whereby morphological choice may precede lexical choice and reordering.

Whenever the predicate-argument structures of the source and target strings are similar, we expect that the linguistic information required for determining the morphological inflection of a plausible translation resides in the source sentence and its syntactic dependency structure. Consequently, we explore target morphology as a source-side prediction task which aims at enriching the source sentence with useful target morphological information. Practically (see Figure 1), after word aligning the sentence pairs, we project a subset of the target morphological attributes to the source side via the word alignments, and then train a model to predict these attributes on predicate-argument aspects of source dependency trees (i.e., without the source word order).



(a) Morphology projection and pred. model training. (b) Machine translation system training.

Figure 1: Overview of the training setup and morphology projection.

Our approach differs from other approaches to predict target morphology (e.g. Chahuneau et al. (2013)) mainly in that we predict on the source side only. A related intuition underlies source-side reordering schemes, which have seen a surge of successful applications recently (e.g. Collins et al. (2005) or Lerner and Petrov (2013)). While syntax-driven source-side reordering assumes that source and target syntax are similar, here we make a weaker assumption, namely that the predicate-argument structures are similar.

We explore the prediction of target morphology on the source side because we see several benefits that could potentially be exploited for further improving machine translation into morphologically rich languages. Source-side prediction models can capitalize on the much reduced complexity of having to represent and process only the input source sentence instead of a large lattice of target hypotheses. Hence, morphological agreement can be enforced over long distances by morphological predictions for the full source sentence. Furthermore, while not pursued in the present work, we hypothesize that the morphological information predicted by our model can be exploited in the word alignment process.

Our contributions in this paper are three. Firstly, we report experiments to support the hypothesis that projecting morphology to the source side could be beneficial for translation, and then present a model for learning to predict target morphology on the source side (Section 4). Secondly, we address how to automatically learn the set of morphological attributes that fit with the parallel training data (Section 5). Finally, we introduce methods for integrating this new information into a machine translation system and evaluate on a translation task (Section 6).

2 Related work

Various approaches have been proposed to the problem of translating between languages of varying morphological complexity. Avramidis and Koehn (2008) enrich the morphologically impoverished source side with syntactic information and translate via a factored machine translation model. In spirit, this paper is closely related to the present work; however, while their decorations are source-side syntactic information (e.g. the noun is the subject), we directly predict target morphology and learn to select the most relevant properties. A similar approach, in which source syntax is reduced to part-of-speech tags is used successfully for translation into Turkish (Yeniterzi and Oflazer, 2010). Following the tradition of two step machine translation (Bojar and Kos, 2010), Fraser et al. (2012) translate morphologically underspecified tokens and add inflections on the target side based on the predictions of discriminative classifiers.

Carpuat and Wu (2007), Jeong et al. (2010), Toutanova et al. (2008) and Chahuneau et al. (2013) propose discriminative lexicon models that are able to take into account the larger context of the source sentence when making lexical choices on the target side. These proposals differ mostly in the way that the additional morphological information is integrated into the machine translation process. Jeong et al. (2010) integrate their lexical selection model via features in the underlying treelet translation system (Quirk et al., 2005). Toutanova et al. (2008) survey two basic methods of integration. In the first method, the inflection prediction model is

		Translation		Word order	Lexical choice
Training and test decor.	Tags	MTR	BLEU	Kendall's τ	BLEU-1
None (baseline)	-	35.74	15.12	45.26	49.86
Projected manual set	77	36.34	15.86	45.79	51.30
Projected automatic set	225	36.50	15.73	46.45	51.24
Projected full set	846	36.67	15.96	46.27	51.52

All translation results statistically significant against baseline at $p < 0.01$

Table 1: Translation with various subsets of projected morphology.

allowed to change the inflections produced by the underlying MT system. The second method is a two step method, where the MT system translates into target-language stems, which are then inflected by the inflection model. Chahuneau et al. (2013) create *synthetic phrases*, i.e. phrases with inflections that have not been observed directly in the training corpus but have been created by an inflection model. These synthetic phrases are then added to the training data of the MT system and marked as such. This enables the MT system to learn how much to trust them.

Finally, Williams and Koehn (2011) add unification-based constraints to the target side of a string-to-tree model. The constraints are extracted heuristically from a treebank and violations are then penalized during decoding.

3 Morphology projection hypothesis

A *morphological attribute* is a morphological property of a word. Each morphological attribute can assume any of a predetermined set of values, such as {`nom, acc, dat, gen`} for the morphological attribute `case` in the German language. Further, the morphological attributes are refined based on a set of 9 atomic parts of speech, yielding a set of morphological attributes of the form `noun:case`, `adj:case`, `verb:tense`, etc.

In this paper, we are interested in the question whether target morphology can be addressed directly on the source. We hypothesize that projecting target morphological attributes and learning to predict them on source side trees can be beneficial to machine translation. To test this hypothesis we initially perform translation experiments with a standard phrase-based MT setup with and without projected morphological information.[1] These experiments provide an indication for the potential of such an approach. They do, however, not answer the question to what extent target morphology can realistically be predicted on the source side. This question will be addressed in the next sections. We perform translation experiments with translation systems decorated with *projected* morphological attributes. In these translation systems, the target side of the test set was processed with a morphological tagger[1] and subsets of the resulting morphological attributes were projected to the source side via alignments. These experiments provide a conservative indication of the potential of this approach. They are not oracle translation experiments, but simulate an optimal target morphology prediction model. The three systems listed in Table 1 differ only in the subset of morphological attributes they use.

The experiment is documented in Table 1. We evaluate translation quality with METEOR and BLEU (Denkowski and Lavie, 2011; Papineni et al., 2002), word order with Kendall's Tau (Kendall, 1938) and lexical choice with unigram BLEU. Statistical significance tests are performed for the translation scores (METEOR and BLEU) using the bootstrap resampling method (Koehn, 2004). The results show that projecting target morphological attributes improves trans-

[1] Details of the experimental setup are provided in Section 6.3.

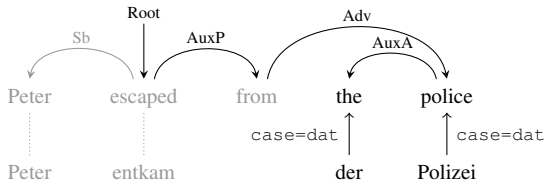


Figure 2: Morphology projection and a source dependency chain.

lation. Improvements result both from better lexical choice and sometimes also better word order. Using the full set of attributes gives the best METEOR and BLEU scores, but it also contributes significantly to data sparsity. Surprisingly, including only a small, manually selected subset of attributes gives comparable improvement while significantly decreasing the number of tags. This *manual* subset is the set of attributes selected for prediction by Fraser et al. (2012), who found that it is beneficial to make some morphological attributes part of the translated word stem instead of predicting them on the target side. The *automatic* selection is a selection of features that an automatic learning procedure determined to be the most beneficial for representing the language pair. This selection performed equally well in our experiments.

Hence, while better translation performance is achievable by including all attributes, the prediction task also becomes significantly harder; comparable translation performance can be achieved with a small, well-chosen set of attributes. The good performance of the *manual* set shows that linguistic intuition can be a good starting point for selecting this set; however, a more empirically beneficial set may be selected by enriching the source side only with attributes which help in selecting the correct target words. The fact that the *automatic* set produces a better METEOR score than the *manual* set further supports this intuition.[2] We highlight the METEOR scores here, since for the language pair English-to-German, METEOR has higher correlation with human judgments than BLEU (Machacek and Bojar, 2014). Now that we established the potential of projecting target morphology on the source side, in the sequel we aim at capitalizing on this potential. In the next section, we present our model for predicting target morphology on source trees based on source side dependency chains.

4 Modeling target-side morphology

Since the word order of the source and target language may differ significantly, predicting morphology in a sequential, word-by-word fashion could be inadequate. We think that source syntax and the source predicate argument structure should be informative for predicting target morphology. Hence, we propose a source-side dependency chain model $P(\mathbf{s}'_m \mid \tau, \mathbf{s})$ to predict the morphologically enriched source string $\mathbf{s}'_m$ given a lexical dependency tree τ of $\mathbf{s}$.

4.1 Source-side dependency chains

A source-side dependency chain is any path from the root of the source dependency tree to any of its leaf nodes, such as `escaped`→`from`→`police`→`the` in Figure 2. Every node with a 1-to-1 alignment to a target node is decorated with the target node's morphological attributes. A standard morphological tagger, such as the n-th order linear chain CRF model (e.g. Mueller et al., 2013), would predict the attribute–value vector for each word left-to-right with a history of $n-1$ tags. Modeling with source-side dependency chains instead, gives various advantages: Besides providing access to the morphological tags assigned to the dependency tree parent and grandparent nodes, it implicitly encourages morphological agreement between a node and its $n-1$ ancestor nodes. The model benefits from access to the node's syntactic role, for example

[2]The difference is statistically significant at $p < 0.05$.

		Manual			Automatic			All
		5	6	7	5	6	7	5
Strict	50k	68.50	70.13	68.86	70.84	69.73	70.97	58.33
	100k	67.08	67.38	67.01	69.33	71.15	69.52	58.65
	200k	67.40	67.40	68.55	69.58	69.82	70.06	57.99
Relaxed	50k	72.67	70.36	72.86	**74.67**	71.42	71.83	62.16
	100k	70.01	71.89	69.82	72.63	72.04	72.61	62.18
	200k	69.40	69.46	69.99	71.44	70.80	69.83	60.86

Best overall F_1 score highlighted in bold.

Table 2: Impact of attribute selection and model parameters on prediction quality (F_1 score).

to predict grammatical case. Finally, training data sparsity is alleviated because the dependency chain formulation allows the extraction of chains from only partially aligned sentences.

4.2 Model estimation

We estimate the source dependency chain model using the general CRF framework. In a linear-chain CRF model, the probability of a tag sequence **y** given a sentence **x** is:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{\exp \sum_{t,i} \lambda_i \cdot \phi_i(\mathbf{y}, \mathbf{x}, t)}{\sum_{\mathbf{y}} \exp \sum_{t,i} \lambda_i \cdot \phi_i(\mathbf{y}, \mathbf{x}, t)}$$

where t is the index of a token, i is the index of a feature and λ_i is the weight corresponding to the binary feature $\phi_i(\mathbf{y}, \mathbf{x}, t)$. To improve training and inference time, we use a coarse-to-fine pruned CRF tagger (Mueller et al., 2013). The training procedure is identical to the linear-chain case, except that we use dependency chains instead of left-to-right chains as training examples. The dependency chain model's feature set is based on the set used in the linear chain CRF for morphological tagging (Mueller et al., 2013). Additionally to the features used by Mueller et al. (2013), we add the following feature templates: the dependency label of the current token, the dependency label of the parent token, the number of children of the current token, the source-side part-of-speech tag of the token, and the current token's child tokens if they are a determiner (*AuxA*), auxiliary verb (*AuxV*), subject (*Sb*) or a preposition (*AuxP*).

4.3 Intrinsic evaluation

To evaluate the quality of the source dependency chain predictions, we perform experiments on a heldout dataset. Models are trained on a subset of the parallel Europarl data. Evaluation is performed using the F_1 score of the predictions compared to the *projected* morphological attributes obtained by automatic alignment of the source and target side of the evaluation set.

Impact of model parameters Table 2 shows prediction performance of the dependency chain model in relation to a selection of model parameters. For each morphological attribute set, we train models of order 5, 6 and 7. All models are trained on sets of 50k, 100k and 200k dependency chains, which are randomly sampled from the training data. In *strict* training mode, we require that target words and source words connected by alignment links agree in their coarse part of speech tags. This restriction enforces a weak form of isomorphism between the source and the target sentence and hence limits the training set to training instances of potentially higher quality. In the *relaxed* setup, no such agreement is enforced.

Up to a certain point, higher order models perform better than models with shorter dependency histories; however, these models are also prone to the issues of data sparsity and overfit-

ting. The results show that *strict* training performs worse than the *relaxed* training regime. The *strict* training regime could possibly produce cleaner training examples; however, since it also enforces a potentially unrealistic isomorphism between the two sentences, those examples may also be less helpful for the final prediction.

Impact of morphological attribute selection As shown in Section 3, it is possible to reduce the set of morphological attributes without major losses in translation quality. For the dependency chain model, smaller attribute sets are preferable since they lead to less complex models and faster training times. Individual attributes may be difficult to predict; hence, the exact selection of attributes is also important for prediction quality.

	Manual	Automatic	All
Training time, 50k	36m	45m	77m
Training time, 100k	58m	82m	2h51m
Training time, 200k	1h54m	3h5m	6h44m
Tags	77	225	846
Best F_1	72.86	74.67	62.18

Table 3: Training times and best scores for the three attribute sets.

Table 3 summarizes training times and prediction performance of the three morphological attribute sets. Larger attribute sets and more training examples lead to longer training times. Overall, the automatic set produces more accurate results than the manual selection. Our analysis shows that this is largely due to difficult to predict verb attributes, which are included in the manual selection but are not part of the automatically learnt set. The finding that these attributes are hard to predict is in line with Fraser et al. (2012), who equally dropped the prediction of verb attributes in later work.

5 Learning salient morphological attributes

Decorating the source language with all morphological properties of the target language will lead to data sparsity and will complicate the prediction task. Therefore, it is necessary to reduce this set to only morphological attributes which are helpful for a given language pair. We consider a morphological attribute to be salient if it enables the machine translation system to perform better lexical selection. It is computationally infeasible to test all possible combinations of morphological attributes in a full machine translation system; hence, we approximate the machine translation system's ability to perform lexical selection with a word-based translation system given by IBM model 1 (Brown et al., 1993). Based on this simplified translation model, the set of salient features which improve the translation performance can be chosen by a clustering procedure.

5.1 Learning procedure

Let $(\mathbf{s}, \mathbf{t})$ be a pair of parallel sentences in source and target language. IBM model 1 provides an iterative method for estimating the translation model $P(\mathbf{t} \mid \mathbf{s})$ from a set of parallel sentences. We add the morphological decoration $\mathbf{s}'_m$ to this model. The translation model now takes the following form:

$$P(\mathbf{t} \mid \mathbf{s}) = \sum_{\mathbf{s}'_m \in \Theta_m(\mathbf{s})} P(\mathbf{s}'_m \mid \mathbf{s}) P(\mathbf{t} \mid \mathbf{s}'_m)$$

where $P(\mathbf{t} \mid \mathbf{s}'_m)$ is the standard IBM model 1 formulation applied to morphologically decorated source tokens. In this simple machine translation model, the morphological attributes are directly attached to the source words. For example, if the English token `police` is decorated with grammatical case, gender and number, it would be replaced by the string `police/case=dat+gender=female+number=singular`. We define the log-likelihood of a set of parallel sentences $\mathbf{X}$ to be:

$$\mathcal{L}(\mathbf{X}) \equiv \log \prod_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} P(\mathbf{t} \mid \mathbf{s})P(\mathbf{s}) = \sum_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} \log P(\mathbf{t} \mid \mathbf{s}) + \log P(\mathbf{s})$$

Let M_0 be the initial set of all morphological attributes observed in the training corpus. Our goal is to find the set $M_n \subseteq M_0$ which maximizes the likelihood of a heldout dataset. By ${\mathbf{s}'_m}^{(i)}$ we denote the decorated source sentence containing only the morphological attributes in M_i. We formulate the search for the set M_n as follows:

$$\begin{aligned}
M_n &= \arg\max_{M_i\subset M_0} \sum_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} \log P(\mathbf{t} \mid \mathbf{s}) + \log P(\mathbf{s}) \\
&= \arg\max_{M_i\subset M_0} \sum_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} \log P(\mathbf{t} \mid \mathbf{s}) \\
&= \arg\max_{M_i\subset M_0} \sum_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} \log\Bigg(\sum_{\mathbf{s}'_m\in\Theta_m(\mathbf{s})} P(\mathbf{s}'_m \mid \mathbf{s}) P(\mathbf{t} \mid {\mathbf{s}'_m}^{(i)}) \Bigg)
\end{aligned}$$

We found the estimates for $P(\mathbf{s}'_m \mid \mathbf{s})$ using the full set of attributes M_0 to be reasonable, with sufficient probability mass assigned to the most likely path. Therefore, we approximate this model by only using the first-best (Viterbi) assignment $\mathbf{s}''_m$. The final, simplified search objective is therefore:

$$\begin{aligned}
M_n &= \arg\max_{M_i\subset M_0} \sum_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} \log\Big(P(\mathbf{s}''_m \mid \mathbf{s}) P(\mathbf{t} \mid {\mathbf{s}''_m}^{(i)}) \Big) \\
&= \arg\max_{M_i\subset M_0} \sum_{(\mathbf{s},\mathbf{t})\in\mathbf{X}} \log P(\mathbf{t} \mid {\mathbf{s}''_m}^{(i)})
\end{aligned}$$

The optimal set of attributes can now be determined with a clustering procedure starting from the full set of morphological attributes M_0. This procedure is reminiscent of Petrov et al. (2006) since as in their work, we can simulate the removal of a morphological attribute by merging the statistics of each of its occurrences.[3]

1. Initialization:
   - Estimate the source dependency chain model $P({\mathbf{s}'_m}^{(0)} \mid \mathbf{s})$, apply it to decorate the training and heldout set, producing $\mathbf{T}_0$ and $\mathbf{H}_0$ (datasets $\mathbf{T}$ and $\mathbf{H}$ decorated with M_0).
   - Estimate $P(\mathbf{t} \mid {\mathbf{s}''_m}^{(0)})$: perform 5 iterations of IBM Model 1 training on $\mathbf{T}_0$.
2. Start with $i = 0$.
3. Calculate $P(\mathbf{t} \mid {\mathbf{s}''_m}^{(i)})$ for each sentence pair in the heldout set $\mathbf{H}_i$.

[3] For example, to simulate the removal of the attribute `gender`, we would merge the statistics of every occurrence of the attribute (either `gender=male` or `gender=female`). The two tags `case=nom+gender=female` and `case=nom+gender=male` would therefore be merged into one tag `case=nom`.

Noun		Adjective		Verb		Other	
Manual	Auto	Manual	Auto	Manual	Auto	Manual	Auto
gender†	gender	gender†	gender	number‡*	-	-	part:negative
number	number	number‡	number	person‡*			part:subpos
case	case	case‡	case	tense*			punc:type
				mode*			num:type
		declension	synpos				
			degree				

† Transferred with lemma. ‡ Propagated from noun. * Dropped in later work.

Table 4: Salient attributes for English–German.

4. Find the attribute $\hat{m} \in M_i$, such that:

$$\hat{m} = \arg\min_{m\prime \in M_i} \left(\sum_{(\mathbf{s},\mathbf{t}) \in \mathbf{H}_i} \log P(\mathbf{t} \mid \mathbf{s}''^{\,(i)}_m) - \log P(\mathbf{t} \mid \mathbf{s}''^{\,(i)\setminus m'}_m) \right)$$

   where $\mathbf{s}''^{\,(i)\setminus m'}_m$ denotes a sentence with the attributes in M_i minus attribute m'.
5. Merge all values of $\hat{m}$ in $\mathbf{T}_i$ and $\mathbf{H}_i$, producing $\mathbf{T}_{i+1}$ and $\mathbf{H}_{i+1}$.
6. Estimate $P(\mathbf{t} \mid \mathbf{s}''^{\,(i+1)}_m)$: Merge the t-tables containing $\hat{m}$ and perfom IBM Model 1 iteration on $\mathbf{T}_{i+1}$.
7. Repeat from (3) with $i = i + 1$. Stop if no possible merge improves $\mathcal{L}(\mathbf{H}_i)$.

5.2 Intrinsic evaluation

The complexity of the clustering procedure is $O(|M| \times k \times l^2)$ for k sentences of length l. In practice, the procedure runs for several hours on a standard machine. Table 4 shows the attributes determined by the learning procedure. The column *Auto* shows the procedure's selection and the column *Manual* shows the manually determined set of morphological attributes for the same language pair, as used by Fraser et al. (2012).

Quality of the selection From inspection of these attributes, we find that our method learns a reasonable set of salient attributes. The manual and automatic selections differ mainly in the verb attributes, which our learning procedure removed from the final set. Morphological attributes in the manual selection which are marked with ‡, are attributes that in the work of Fraser et al. (2012) were transfered as part of the translated stem by their MT system. The symbol ‡ marks morphological attributes that they propagated from the noun (for example, an adjective's case is copied from the noun it modifies). Finally, the verb attributes, which are marked with * are used by Fraser et al. (2012) but found to be problematic by Cap et al. (2014b) and dropped in later work (Cap et al., 2014a). Likewise, inspection of our model showed that verb attributes perform badly as they may be difficult to predict. Hence, our procedure successfully learnt not to model these attributes while retaining the beneficial noun and adjective attributes.

Granularity of the morphological attributes When simulating the removal of a morphological attribute with this learning algorithm, all of its values are merged. In some language pairs, however, it would be useful to merge the individual values of the attributes instead. For example, from the spelling of German nouns it is usually not recognizable whether the noun is `case=nominative` or `case=accusative`. Hence, the algorithm should ideally be able to also merge individual values. Since this is a straight-forward extension of our current algorithm, we plan to evaluate this aspect in future work.

6 Morphology-informed translation

To leverage the morphology predictions in a machine translation decoder, we integrate this additional information into the translation model. During training and tuning, the translation model is decorated with morphological attributes either projected from the target side or predicted by our dependency chain model.

6.1 Integration of target morphology predictions

In practice, the predicted morphological attributes on the source side can be integrated into the machine translation system as arbitrary features based on source morphology and target strings. In our experiments, we opted for a feature representation in which this information is encoded as source morphology-to-target affix features. We chose this simple representation because it is generic enough to produce improvements on the one hand and it is not prone to overfitting on the other hand. For each phrase candidate on the source side, sparse features fire for a given sequence of source-side morphology tags and target-side string affixes. As an example, consider the sentence *Peter entkam der Polizei* (Peter escaped from the police) from Figure 2. In this case, the morphological attributes gender (*female*), number (*singular*) and grammatical case (*dative*) would have been projected from the target to the source side for the phrase *the police*/*der Polizei*. When translating the source segment *the police*, the feature `gender=fem+number=sing+case=dat X` $\rightarrow$ `-er X` would fire based on the predicted morphology. This hint would help the machine translation system choose the correct German determiner *der*.[4]

6.2 Inference strategies

At test time, the morphological decoration of the source sentence needs to be selected. This decision should ideally take into account both the predictions of our source-side dependency chain model and the content of the phrase table, which may be decorated with projected morphology.

We compare several inference strategies. The major distinction between these strategies is whether the machine translation system is trained and tuned on projected morphology or predicted morphology. Training on predicted morphology has the benefit that it lets the MT system learn how much it can trust the predictions made by the dependency chain model. However, this method is also more laborious in system development, since it requires retraining and tuning the whole translation system for every change in the prediction model.

Training and decoding with Viterbi predictions In the first decoding setup, which is similar to the most common setup used in preordering, we decorate both training and test set with the Viterbi decorations extracted from the dependency chain model. Specifically, for each possible dependency chain in the source dependency tree, we preform standard CRF Viterbi tagging starting from the root of the tree. The full training and tuning set is decorated with these single-best predicted decorations. System training and tuning is then performed on these sets. During test time, only the single-best Viterbi prediction is considered by the MT system.

Training on projected morphology and decoding with Viterbi predictions The projected training setup differs from the previous setup in that the morphological decorations on the training and tuning set are not predicted but projected from the target side via alignments. During test time, the decorations are predicted using single-best Viterbi predictions as in the previous setup. While this strategy is advantageous since it simplifies the system training, the main downside of this strategy is that it cannot take into account possible shortcomings of the prediction model.

[4] This feature example is taken form the weights of the system trained with the automatic morphological attribute set and predicted training and test decoration.

		Translation		Word order	Lexical choice
Morph. attributes	Training decor.	MTR	BLEU	Kendall's τ	BLEU-1
No morphology	-	35.74	15.12	45.26	49.86
Manual selection	Predicted	35.85	15.19	45.43	50.01
	Projected	34.63[A]	14.00[A]	44.07	48.75
Autom. selection	Predicted	35.99[AC]	15.23[B]	45.88	50.27
	Projected	35.98[AC]	15.22[C]	45.89	50.27

[A]Statistically significant against baseline at $p < 0.05$ [B]Statistically significant against baseline at $p < 0.06$
[C]Statistically significant against Manual selection at $p < 0.05$

Table 5: Translation with predicted test decorations.

At training time, only projected decorations are observed, which might not be realistic when taking into account the prediction model.

6.3 Evaluation

Having introduced and evaluated the attribute selection process and the prediction of target-side morphological attributes based on source-side dependency chains, we now turn to the evaluation of the predicted morphological information within a full machine translation pipeline.

Experimental details We use a standard phrase-based machine translation system (Cer et al., 2010) with a 5-gram language model and distortion-based reordering (dl=5). Features based on the source morphology predictions are learnt on either the projected morphology or the predictions of the source dependency chain model. Experiments are conducted on English–German. Source-side dependency trees are predicted based on the HamleDT treebank (Zeman et al., 2012) using TurboParser (Martins et al., 2010). The dependency parser is trained to produce pseudo-projective dependency trees (Nivre and Nilsson, 2005).[5] The system is trained on the full parallel sections of Europarl (Koehn, 2005) and tuned and tested on the WMT 2009 and WMT 2010 newstest sets respectively.

Monolingual morphological tagging is performed using the Marmot CRF-based tagger (Mueller et al., 2013). The tagger is trained on the English and German parts of the HamleDT treebank. The morphological attributes of both languages follow the Interset standard (Zeman, 2008), which contains 45 unique attribute vectors (tags) for English and 958 for German.

Discussion Table 5 shows the outcomes of using the inference strategies presented in Section 6.2. We evaluate translation quality with METEOR and BLEU (Denkowski and Lavie, 2011; Papineni et al., 2002), word order with Kendall's Tau (Kendall, 1938) and lexical choice with unigram BLEU. Statistical significance tests are performed for the translation scores (METEOR and BLEU) using the bootstrap resampling method (Koehn, 2004).

The results show that both attribute selections show improvements over the baseline when training and testing on predicted morphology. On the other hand, when training on projected morphology and performing Viterbi predictions, a visible gap between the manual set and the automatic set can be observed. This gap indicates that with the automatic set, the predictions by the dependency chain model are closer to the projected predictions so that the machine trans-

[5]Projectivization was performed using MaltParser version 1.8; `http://www.maltparser.org/`.

lation system learns realistic weights for the prediction part. Additionally, the system based on the automatic selection produces a significantly better METEOR score than the system using the manual selection. As in the experiments with projected morphology, the results of this evaluation indicate that the improvements stem from both word order choices as well as better lexical selection. In terms of time performance, we found that the additional information does not significantly affect the speed of the translation system. The Viterbi algorithm for predicting the target morphology is efficient and as the information is passed to the MT system as sparse features, no additional complexity is added. While we have focused on the language pair English–German, the methods presented in this paper are applicable to many other language pairs. We therefore aim to perform additional experiments for morphologically-rich target languages such as Turkish, Arabic and Czech.

7 Conclusion

In this paper, we have explored the novel approach of target morphology projection. After testing the idea empirically, we have put forward three proposals to realize this idea: First, we introduced the dependency chain model for predicting arbitrary target morphology attributes based on source dependency trees. Second, we introduced a learning procedure to determine a language pair's set of salient morphological attributes. And finally, we have introduced and compared various strategies for integrating this new information into a machine translation system. The experiments we have performed have provided us with important insights. They have demonstrated that projecting a small subset of morphological attributes to the source side can provide major translation improvements, while reducing the complexity of prediction. Furthermore, the approach for learning the useful subset performs well based both on the intrinsic evaluation and the empirical results during prediction and translation. Given that previous work has found it rather difficult to achieve improvements in German morphology, we consider the improvements in METEOR score and the modest improvements in BLEU score encouraging.

While the prediction performance of the dependency chain model leaves room for improvement, we submit that our experiments sufficiently demonstrate the potential of this approach. We plan to further improve the prediction performance of the dependency chain model with extensions such as the use of (bilingual) word embeddings that could help resolve ambiguous cases. In addition, to let the machine translation system better exploit this new knowledge, deeper integration (e.g. into the language model) is necessary. Both ideas constitute the main topics for extending the current work in the future.

Acknowledgements

We thank the three anonymous reviewers for their constructive comments and suggestions. The first author is supported by the EXPERT (EXPloiting Empirical appRoaches to Translation) Initial Training Network (ITN) of the European Union's Seventh Framework Programme.

References

Avramidis, E. and Koehn, P. (2008). Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 763–770, Columbus, Ohio. Association for Computational Linguistics.

Bojar, O. and Kos, K. (2010). 2010 failures in English-Czech phrase-based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 60–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Cap, F., Fraser, A., Weller, M., and Cahill, A. (2014a). How to produce unseen teddy bears: Improved morphological processing of compounds in SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 579–587, Gothenburg, Sweden. Association for Computational Linguistics.

Cap, F., Weller, M., Ramm, A., and Fraser, A. (2014b). CimS – the CIS and IMS joint submission to WMT 2014 translating from English into German. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 71–78, Baltimore, Maryland, USA. Association for Computational Linguistics.

Carpuat, M. and Wu, D. (2007). Context-dependent phrasal translation lexicons for statistical machine translation. *Proceedings of Machine Translation Summit XI*, pages 73–80.

Cer, D., Galley, M., Jurafsky, D., and Manning, C. D. (2010). Phrasal: A statistical machine translation toolkit for exploring new model. *Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 9–12.

Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687, Seattle, Washington, USA. Association for Computational Linguistics.

Collins, M., Koehn, P., and Kucerova, I. (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, Michigan. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Fraser, A., Weller, M., Cahill, A., and Cap, F. (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. Association for Computational Linguistics.

Jeong, M., Toutanova, K., Suzuki, H., and Quirk, C. (2010). A discriminative lexicon model for complex morphology. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, pages 81–93.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, volume 5, pages 79–86.

Lerner, U. and Petrov, S. (2013). Source-side classifier preordering for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 513–523, Seattle, Washington, USA. Association for Computational Linguistics.

Machacek, M. and Bojar, O. (2014). Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Martins, A., Smith, N., Xing, E., Aguiar, P., and Figueiredo, M. (2010). Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 34–44, Cambridge, MA. Association for Computational Linguistics.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Nivre, J. and Nilsson, J. (2005). Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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. Association for Computational Linguistics.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, Ann Arbor, Michigan. Association for Computational Linguistics.

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio. Association for Computational Linguistics.

Tran, K., Bisazza, A., and Monz, C. (2014). Word translation prediction for morphologically rich languages with bilingual neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Williams, P. and Koehn, P. (2011). Agreement constraints for statistical machine translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 217–226, Edinburgh, Scotland. Association for Computational Linguistics.

Yeniterzi, R. and Oflazer, K. (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. Association for Computational Linguistics.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). Hamledt: To parse or not to parse? In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).