A Distributed Inflection Model for Translating into Morphologically Rich Languages

Ke Tran m.k.tran@uva.nl
Arianna Bisazza a.bisazza@uva.nl
Christof Monz c.monz@uva.nl
Informatics Institute, University of Amsterdam,
Science Park 904, 1098 XH Amsterdam, The Netherlands

Abstract

Lexical sparsity is a major challenge for machine translation into morphologically rich languages. We address this problem by modeling sequences of fine-grained morphological tags in a bilingual context. To overcome the issue of ambiguous word analyses, we introduce *soft tags*, which are under-specified representations retaining all possible morphological attributes of a word. In order to learn distributed representations for the soft tags and their interactions we adopt a neural network approach. This approach allows for the combination of source and target side information to model a wide range of inflection phenomena. Our re-inflection experiments show a substantial increase in accuracy compared to a model trained on morphologically disambiguated data. Integrated into an SMT decoder and evaluated for English-Italian and English-Russian translation, our model yields improvements of up to 1.0 BLEU over a competitive baseline.

1 Introduction

In morphologically rich languages (MRLs), words can have many different surface forms depending on the grammatical context. When translating into MRLs, standard statistical machine translation (SMT) models such as phrase translation models and n-gram language models (LMs) often fail to select the right surface form due to the sparsity of observed word sequences (Minkov et al., 2007; Green and DeNero, 2012). While neural LMs (Bengio et al., 2003; Schwenk, 2007) address lexical sparsity to a certain degree by projecting word sequences to distributed vector representations, they still suffer from the problem of rare words which is particularly exacerbated in MRLs (Botha and Blunsom, 2014; Jean et al., 2015; Luong et al., 2015).

A potential solution to overcome data sparsity in MRLs, is to use word representations that separate the grammatical aspects of a word, i. e. inflection, from the lexical ones. Such word representations already exist for many languages in the form of morphological analyzers or lexicons. However, using these resources for statistical language modeling is far from trivial due to the issue of ambiguous word analyses. Table 1 illustrates this problem in Italian, for which a fine-grained morphological lexicon but no sizable disambiguated corpus exists. These morphological analyses[1] clearly contain information that is useful to encourage grammatical agreement and, in this case, detect the highlighted error. Unfortunately, though, the needed

[1] In this work we use the terms *analysis* and *tag* interchangeably to denote fine-grained word annotations provided by a morphological analyzer or lexicon.

SMT	**idee**	**ribelli**	**che**	**circola**
Gloss	*ideas*	*rebellious*	*that*	*circulate*
	noun-f:p	noun-f:p	con	ver:impr+pres+2+s
		noun-m:p	**wh-che**	**ver:ind+pres+3+s**
Analyses		**adj:pos+f+p**	det-wh:f+p	
		adj:pos+m+p	det-wh:f+s	
		...	...	

Table 1: Example of morphological error in Italian SMT output: the verb form should be plural (*circolano*) and not singular (*circola*) to agree in number with the subject. Most of the words have multiple analyses according to our morphological lexicon of reference (Zanchetta and Baroni, 2005). The correct one in context is highlighted.

information is difficult to access because each word can have multiple analyses. Performing contextual disambiguation during translation is an ill-posed problem because the SMT decoder produces large numbers of ungrammatical word sequences but gold tagged training data is naturally composed of grammatical sentences.[2] Moreover, searching for the optimal tag sequence introduces spurious ambiguity into the SMT decoder. Finally, training a disambiguator requires manually disambiguated data, which is not available in many languages and costly to produce.

In this paper, we address this problem with a novel inflection model that is based on two main ideas: First, morphological ambiguity does not need to be resolved for SMT. Instead, we map words to a space where *all possible* morphological attributes of a word are retained. Rather than enforcing hard tagging decisions, we let the model operate on *soft* word representations. The resulting tag set is larger than the original one, but still effective at reducing the lexical sparsity of purely word-based LM. Second, learning distributed representations for soft morphological tags can help share statistical strength among overlapping tags, i. e. tags that have some attributes in common. To achieve this, we train a neural network that predicts sequences of soft tags conditioned on rich contextual features.

We show that: (i) our soft representation model achieves higher accuracies in re-inflecting translations than a model performing contextual disambiguation, and (ii) our model significantly improves translation quality on two different target MRLs, including a language for which no sizable disambiguated corpora exist.

The paper is organized as follows: after reviewing the previous work (Section 2), we present our distributed inflection model based on soft morphological representations (Section 3). In Section 4 we introduce the general experimental setup, followed by a detailed description of the re-inflection experiments (Section 5) and the end-to-end SMT experiments (Section 6). We conclude with a discussion of SMT output examples and an outlook of future work

2 Previous Work

Previous work on inflection modeling for translation into MRLs has mostly relied on the availability of morphologically disambiguated data to choose the most probable analysis of each word in either a context-independent (Minkov et al., 2007) or context-dependent (Green and DeNero, 2012; Koehn and Hoang, 2007; Subotin, 2011) way. While the former irrevocably discards potentially useful attributes of the words, the latter tasks the inflection model with disambiguating the word sequence under construction, which is difficult given the ill-formedness of SMT output and a cause of spurious ambiguity.

[2] This issue has also been shown to affect syntactic parsing of SMT output (Post and Gildea, 2008; Carter and Monz, 2011).

Considerably less work has focused on MRLs where disambiguated data does not exist, with few exceptions where ambiguity is solved by randomly selecting one analysis per word type (Minkov et al., 2007; Toutanova et al., 2008; Jeong et al., 2010).

As for how inflection models are integrated into the STM system, different strategies have been proposed. Minkov et al. (2007); Toutanova et al. (2008); Fraser et al. (2012) treat inflection as a post-processing task: the SMT model is trained to produce lemmatized target sentences (possibly enhanced with some form of morphological annotation) and afterwards the best surface form for each lemma is chosen by separate inflection models. Some work has focused on the generation of new inflected phrases given the input sentence (Chahuneau et al., 2013) or given the bilingual context during decoding (Koehn and Hoang, 2007; Subotin, 2011). Other inflection models have been integrated to SMT as additional feature functions: e.g. as an additional lexical translation score (Jeong et al., 2010; Tran et al., 2014) or as an additional target language model score (Green and DeNero, 2012). We follow this last strategy, rather than generating new inflections, motivated by previous observations that, when translating into MRLs, a large number of reference inflections are already available in the SMT models but are not selected for Viterbi translation (Green and DeNero, 2012; Tran et al., 2014).

More in general, our work is related to class-based language modeling (Brown et al., 1992) with the major difference that we also condition on source-side context and that we use explicit morphological representations instead of data-driven word clusters (Uszkoreit and Brants, 2008), word suffixes (Müller et al., 2012; Bisazza and Monz, 2014) or coarse-grained part-of-speech tags (Koehn et al., 2008).

Modeling morphology using neural networks has recently shown promising results: in the context of monolingual neural language modeling, Luong et al. (2013); Botha and Blunsom (2014) obtain the vectorial representation of a word by composing the representations of its morphemes. Tran et al. (2014) model translation stem and suffix selection in SMT with a bilingual neural network. Soricut and Och (2015) discover morphological transformation rules from word embeddings learned by a shallow network. We are not aware of work that leveraged fine-grained morphological tags for neural language or translation modeling.

3 A Distributed Inflection Model

In MRLs, the surface form of a word is heavily determined by its grammatical features, such as number, case, tense etc. Choosing the right target word form during translation is a complex problem since some of these features depend on the source context while others depend on the target context (agreement phenomena). We model target language inflection by a Markov process generating a sequence of abstract word representations based on source and target context. This complements previous work focusing on either the former (Avramidis and Koehn, 2008; Chahuneau et al., 2013; Tran et al., 2014) or the latter (Green and DeNero, 2012; Fraser et al., 2012; Botha and Blunsom, 2014; Bisazza and Monz, 2014).

3.1 Soft Morphological Representations

As previously stated, it is common for words in MRLs to admit multiple morphological analyses out of context. Rather than trying to disambiguate the analyses in context using for instance conditional random fields (Green and DeNero, 2012; Fraser et al., 2012), we modify the tagging scheme so that each word corresponds to only one tag. To also avoid the loss of useful information incurred when arbitrarily selecting one analysis per word type (Minkov et al., 2007; Jeong et al., 2010), we introduce soft morphological representations, or simply *soft tags*.

Assume that a morphological analysis μ is a set of morphological attributes $\mathcal{S}(\mu)$ such as masculine or plural. Given a word w, a morphological analyzer or lexicon LEX returns a list

of possible analyses of that word $\mathcal{A}_w = \{\mu : (w, \mu) \in \text{LEX}\}$. Then, we can map word w to a unique soft tag r_w by simply taking the union of all its possible morphological attributes, that is:

$$r_w \stackrel{\text{def}}{=} \bigcup_{\mu_k \in \mathcal{A}_w} \mathcal{S}(\mu_k)$$

For instance, the Italian word *"ribelle"* has four analyses: adj:pos+f+s, adj:pos+m+s, noun-f:s, and noun-m:s. Its corresponding soft tag is adj:pos|adj:f|adj:s|adj:m|noun-f:s|noun-m:s. Hence, soft tags maintain *all* morphological attributes of a word to denote its grammatical dimension while ignoring the lexical content. This new representation scheme compromises between sparsity and ambiguity, and allows for an efficient integration of our model directly into the decoder as no additional cost is incurred for the local tagging search.

Soft tags can also be seen as the marginalization of μ when predicting a surface word w_i given a lemma l_i and its context $\mathbf{C}_i$ (i. e. variables that influence w_i, such as w_{i-1}):

$$\begin{aligned} p(w_i|l_i, \mathbf{C}_i) &= \sum_{\mu_k \in \mathcal{A}_{w_i}} p(w_i|\mu_k, l_i, \mathbf{C}_i) p(\mu_k|l_i, \mathbf{C}_i) \\ &= \sum_{\mu_k \in \mathcal{A}_{w_i}} p(\mu_k|l_i, \mathbf{C}_i) \end{aligned} \quad (1)$$

assuming that any lemma-analysis pair (l, μ) corresponds to at most one inflected form w. Using soft tags, Equation 1 can be approximated by $p(r_w|l_i, \mathbf{C}_i)$.

3.2 Inflection Neural Network

Our inflection model[3], Inf-NN, is trained on word-aligned bilingual data to predict sequences of target soft tags given a fixed-size target history *and* the input source sentence (see Figure 1). We adopt a neural LM approach as learning distributed representations for the soft tags can help to share statistical information among overlapping tags (i. e. tags that share some morphological attributes). Moreover, compared to Maximum Entropy models that use lexical features, neural networks can better exploit sparse input features such as lexicalized source context and target lemma features, as well as their interactions, in high dimensional spaces.

We learn distributed representations for both source words and target soft tags. The source word representations are initialized from pre-trained embeddings, which has been shown to encode certain morphological regularities (Soricut and Och, 2015), whereas target tag representations are initialized randomly.

Inf-NN is a feed-forward neural network whose output is a conditional probability distribution over a set of morphological tags given target history and source context. Formally, let $h_i = (r_{i-1}, \ldots, r_{i-n+1})$ be the $n-1$ tag history of the target word w_i, and $c_j = (s_{j-k}, \ldots, s_{j+k})$ the source context centering at the word s_j aligned to w_i by an automatic aligner. We use simple heuristics similar to the approach by Devlin et al. (2014) to handle null and multiple alignments so that each target word w_i can be mapped to exactly one source word s_j. Let $\mathbf{s}_j \in \mathbb{R}^D$ and $\mathbf{r}_i \in \mathbb{R}^D$ denote the distributed representations of source s_j and target tag r_i respectively. Then, the conditional probability $p_{\text{Inf-NN}}(r_i|h_i, c_j)$ is computed at the output layer $\mathbf{y}$ of the network as follows:

$$\begin{aligned} \mathbf{z}_i &= \phi(\mathbf{W}^{\text{c}}\mathbf{c}_j + \mathbf{W}^{\text{h}}\mathbf{h}_i + \mathbf{b}_z) \\ \mathbf{y} &= \text{softmax}(\mathbf{W}^{\text{m}}\mathbf{z}_i + \mathbf{b}_y) \end{aligned}$$

[3]The implementation is available at `https://bitbucket.org/ketran/soft-tags`

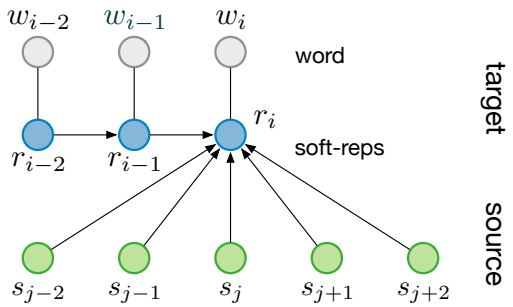
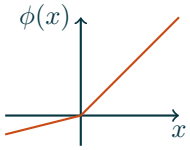


Figure 1: Graphical representation of the Inf-NN model: the current target word's soft tag, r_i, is predicted based on a fixed-size target tag history and a source side context centered around s_j, the translation of w_i. Each target word w_i can be deterministically mapped to a soft tag r_i.

where $\mathbf{W}^{\mathrm{c}}$, $\mathbf{W}^{\mathrm{h}}$, and $\mathbf{W}^{\mathrm{m}}$ are weight matrices, $\mathbf{c}_j$ and $\mathbf{h}_i$ are shorthands for $[\mathbf{s}_{j-k}; \ldots; \mathbf{s}_{j+k}]$ and $[\mathbf{r}_{i-1}; \ldots; \mathbf{r}_{i-n+1}]$ respectively, $[\mathbf{v}; \mathbf{v}']$ denotes vector concatenation, and ϕ is a non-linear transfer prelu. As ϕ, we use in all experiments the channel-shared parametric rectified linear unit (PReLU) introduced by He et al. (2015). PReLU $\phi(x)$ is defined as:

$$\phi(x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{otherwise} \end{cases}$$



where a is a parameter learned during training. To speed up decoding, we train the Inf-NN model with a self-normalized objective (Devlin et al., 2014; Andreas and Klein, 2015). More specifically, we adopt the objective function proposed by Andreas and Klein (2015):

$$\begin{aligned} \ell(\boldsymbol{\theta}) = &- \mathbb{E}\big[\ln p(r_i|h_i, c_j)\big] + \eta \left\|\boldsymbol{\theta}\right\|_2^2 \\ &+ \frac{\gamma}{p}\mathbb{E}\big[\ln^2 Z(h_i, c_j)|(h_i, c_j) \in \mathcal{H}\big] \end{aligned}$$

where $\mathcal{H}$ is a set of random samples on which self-normalization is performed, $\boldsymbol{\theta} = \big\{\{\mathbf{s}_j\}, \{\mathbf{r}_i\}, \mathbf{W}^c, \mathbf{W}^h, \mathbf{W}^m, \mathbf{b}_z, a\big\}$ are the parameters of the networks, and $Z(h_i, c_j)$ is the partition function of the input (h_i, c_j). In practice, we obtain $\mathcal{H}$ by sampling from a Bernoulli distribution Bern(p). This is equivalent to applying dropout (Srivastava et al., 2014) on the loss gradient $\mathbf{1} \in \mathbb{R}^m$ of self-normalization term, where m is the size of a mini-batch. We regularize the networks with ℓ_2 norm.

4 Experimental Setup

We evaluate our approach on two related tasks: re-inflecting reference translations and end-to-end translation from English into MRLs. With the first task, we test the effectiveness of soft morphological representations against (i) a model that randomly assigns one tag per word type (among its possible tags) and (ii) a model that admits multiple tags per word and requires a pre-disambiguated corpus to be trained. With the second task, we measure translation quality when our inflection model is integrated into a state-of-the-art phrase-based SMT decoder, showing its applicability to languages where no disambiguated data exists.

4.1 Data

As target languages, we choose two MRLs belonging to different language families and displaying different inflectional patterns: Russian has very rich nominal, adjectival and verbal inflection, while Italian has moderate nominal and adjectival inflection, but extremely rich verbal inflection. Experiments are performed on the following tasks:

- English-Russian WMT (Bojar et al., 2013): translation of news commentaries with large-scale training data.
- English-Italian IWSLT (Cettolo et al., 2014): translation of speeches with either small-scale training data (TED talks only) or large-scale training data (TED talks and European proceedings).

SMT training data statistics are reported in Table 2. The Russian Inf-NN model is trained on a 1M-sentence subset of the bilingual data, while the Italian one is trained on all the data available in each setting. For each data set, we create automatic word alignments using GIZA++ (Och and Ney, 2003).

		En-Ru	**En-It**	
		large	**small**	**large**
Bilingual	#sentences	2.4M	180K	2.0M
	src/trg #tokens	49.2/47.2M	3.6/3.4M	57.4/57.0M
	src/trg dict.size	774K/1100K	55K/80K	139K/195K
Monoling.	#sentences	21.0M	2.1M	
	trg #tokens	390M	58.4M	
	src/trg dict.size	2.7M	199K	

Table 2: Training corpora statistics.

The ambiguous morphological analyses are obtained from the Russian OpenCorpora lexicon[4] (Bocharov et al., 2013) and from the Italian Morph-it![5] lexicon (Zanchetta and Baroni, 2005). Table 3 shows the number of tags and soft tags occurring in our training data, as well as the expected counts of analyses per word $\mathbb{E}_w[t]$, words per lemma $\mathbb{E}_l[w]$ and analyses per lemma $\mathbb{E}_l[t]$.

Language	#tags	#soft-tags	$\mathbb{E}_w[t]$	$\mathbb{E}_l[w]$	$\mathbb{E}_l[t]$
Russian	892	4431	3.8	7.2	27.4
Italian	450	901	1.9	12.7	24.3

Table 3: Morphological characteristics of the Inf-NN training data: number of tags and soft tags, expected counts of analyses per word $\mathbb{E}_w[t]$, words per lemma $\mathbb{E}_l[w]$ and analyses per lemma $\mathbb{E}_l[t]$.

We find that the Russian tag set and, consequently, the soft tag set are considerably larger than the Italian ones. The average morphological ambiguity is also larger in Russian (3.8 versus

[4] opencorpora.org

[5] sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php

1.9 tags per word). However, somewhat surprisingly, morphological richness is higher in Italian (12.7 versus 7.2 words per lemma). At a closer inspection, we find that most of this richness is due to verbal inflection which goes up to 50 forms for frequently observed verbs.

4.2 Neural network training

The Inf-NN models are trained on a history of 4 target tags and source context of 7 words with the following configuration: Embedding size is set to 200 and the number of hidden units to 768. Target word and soft-tag embeddings are initialized randomly from a Gaussian distribution with mean zero and standard deviation 0.01. Source word embeddings are initialized from pre-trained Glove vectors (Pennington et al., 2014) and rescaled by a factor of 0.1. Weight matrices of linear layers are initialized from a zero-mean Gaussian distribution with standard deviation $\sqrt{2/n_i}$ where n_i is the number of input units (He et al., 2015). We set self-normalization strength $\gamma = 0.02$, Bernoulli parameter $p = 0.1$, and regularization parameter $\eta = 10^{-4}$. All models are trained with a mini-batch size of 128 for 30 epochs. Our stochastic objective functions are optimized using the first-order gradient-based optimizer Adam (Kingma and Ba, 2015). We use the default settings suggested by the authors: $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\lambda = 1 - 10^{-8}$.

5 Re-inflection Experiments

The purpose of this experiment is to simulate the behavior of the inflection model during SMT decoding: Given a reference translation and its corresponding source sentence, we re-inflect the former using a simple beam search and count how many times the model recovers the correct surface word form on a 10K-sentence held-out data set.

Since we do not assume the availability of a disambiguator, we also have to deal with lemma ambiguity. While this issue does not affect the definition and training of our Inf-NN, we do need lemmas to determine the set of candidate surface forms I_w for each word w that is being re-inflected. As a solution, we define I_w as the union of the surface forms of each possible lemma of w or, more formally, as:

$$I_w = \{w_i \,|\, \text{lem}(w_i) \cap \text{lem}(w) \neq \emptyset\}$$

where $\text{lem}(w)$ denotes the set of lemmas returned by the lexicon for word w. For example, the Italian form *baci* has two possible lemmas: *bacio* (noun: *kiss*) and *baciare* (verb: *to kiss*). Its candidate set I_w will then include all the forms of the noun *bacio* and all the forms of the verb *baciare*: that is, *bacio, baci, baciamo, baciate, baciano,* etc.

We compare the proposed soft-tag Inf-NN against an Inf-NN trained on randomly assigned tag per type and to another one trained on tag sequences disambiguated by TreeTagger (Schmid, 1994; Sharoff et al., 2008). The latter model must search through a much larger space of morphological tag sequences. Therefore, to allow for a fair comparison, we set a higher beam size when re-inflecting with this model. As another difference from the other models, the TreeTagger-based inflection model relies on the lemmatization performed by TreeTagger to define the candidate set I_w.

To validate the effectiveness of the neural network approach, we also compare Inf-NN to a simpler MaxEnt model trained on a similar configuration. Finally, we evaluate the importance of source-side context features by experimenting with a series of Inf-NN models that are only conditioned on the target tag history.

Since no morphological disambiguator is available for Italian, we perform this experiment only for Russian. As shown in Table 4, soft tags perform best in all settings and become even more effective when moving from MaxEnt to neural network, demonstrating the impor-

	MaxEnt	Inf-NN		
	with src	w/o src	with src	$beam$
Tree-Tagger: all analyses	56.33	61.19	69.68	200
Random: 1 analysis per word	66.08	72.32	79.92	5
Soft-Reps: 1 soft tag per word	**66.95**	**75.43**	**81.93**	5

Table 4: Token-level re-inflection accuracy (%) on a 10K-sentence English-Russian held-out set. The last column indicates the beam size used when searching for the optimal re-inflected sequence.

tance of learning distributed representations for the soft tags. The notably lower accuracy of the TreeTagger-based model confirms our intuition that morphological disambiguation is not needed to model inflection in SMT, but can actually make the task more difficult. This result can be explained by the fact that, when fixing one tag per word type either by random assignment or with soft tags, the number of tags per lemma becomes substantially smaller (cf. Table 3) and classification easier. On the other hand, the Tree-Tagger based model operating on all word analyses has to deal with spurious ambiguity: that is, a correct sequence of inflected words can correspond to multiple tag sequences that are competing with one another. Solving this problem by marginalizing over the ambiguous analyses (cf. Equation 1) can lead to intractable decoding (Sima'an, 1996; Li et al., 2009).

The model using soft-tags, which capture all possible morphological attributes of words, performs the best. Even without using source context features, our Inf-NN outperforms the MaxEnt model by 8.5% absolute because of the high dimensional space used to capture complex morphological regularities. By adding source context, we further increase accuracy by 6.5%, leading to an overall gain of 15% over the MaxEnt baseline.

Next, we investigate the impact of our most accurate re-inflection model (Soft-Reps Inf-NN) in an end-to-end SMT setting without relying on any disambiguated data.

6 End-to-end SMT Experiments

We integrated our Inf-NN model into a phrase-based SMT decoder similar to Moses (Koehn et al., 2007) as an additional log-probability feature function ($\log p_{\text{Inf-NN}}$).

When a new target phrase $\tilde{w}$ is produced by the decoder, the Inf-NN model returns a probability for each word w_i that composes it, given the previously translated words' soft tags and the source context centered around the source word s_j aligned to w_i. To detect s_j we store phrase-internal word alignments in the phrase table and use simple heuristics to map each target index i to exactly one source index j, as done for the Inf-NN training (Section 3.2). Since every target word corresponds to one soft tag, obtaining the representation of w_i is trivial (by lookup in a word-tag map) and so is maintaining the target tag history. This crucially differs from previous approaches that distinguish between hypotheses with equal surface forms but different morphological analyses (Koehn et al., 2007), thereby introducing spurious ambiguity into what is already a huge search space.[6] As a result, the integration of our Inf-NN does not affect decoding speed.

[6]Green and DeNero (2012) also tag each target phrase in context as it is produced. However, they avoid the spurious ambiguity problem by only preserving the most probable tag sequence for each phrase (incremental greedy decoding).

6.1 Baseline

Our SMT baseline is a competitive phrase-based SMT system including hierarchical lexicalized reordering models (Galley and Manning, 2008) and a 5-gram target LM trained with modified Kneser-Ney smoothing (Chen and Goodman, 1999). Since the large English-Italian data comes from very different sources (TED talks and European proceedings), we construct phrase table and reordering models for this experiment using the fillup technique (Bisazza et al., 2011). Note that our baseline does not include previously proposed inflection models because the main goal of our experiment is to demonstrate the effectiveness of the proposed approach for languages where no sizable disambiguated data exists, which is indeed the case for Italian.

Feature weights are tuned with pairwise ranking optimization (Hopkins and May, 2011) on the union of IWSLT's dev10 and test10 in Italian, and on the first 2000 lines of wmt12 benchmark in Russian (Callison-Burch et al., 2012). During tuning, 14 PRO parameter estimation runs are performed in parallel on different samples of the n-best list after each decoder iteration. The weights of the individual PRO runs are then averaged and passed on to the next decoding iteration. Performing weight estimation independently for a number of samples corrects for some of the instability that can be caused by individual samples.

6.2 Results

Translation quality is measured by case-insensitive BLEU (Papineni et al., 2002) on IWSLT's test12 and test14 in Italian, and on wmt13 and wmt14 for Russian, all provided with one reference translation. To see whether the differences between the approaches we compared in our experiments are statistically significant, we apply approximate randomization (Noreen, 1989).[7]

	Data	Test	Baseline	Inf-NN
en→ru	large	wmt13	19.0	**19.3**▲(+0.3)
		wmt14	26.1	**26.7**▲(+0.6)
en→it	small	iwslt12	24.6	**25.6**▲(+1.0)
		iwslt14	20.4	**20.9**▲(+0.5)
	large	iwslt12	25.0	**25.8**▲(+0.8)
		iwslt14	20.9	**21.4**▲(+0.5)

Table 5: Impact on translation quality of the Inf-NN model. ▲ marks significance level $p < .01$.

Results are presented in Table 5. Our Inf-NN model consistently leads to significant improvements over a competitive baseline, for both language pairs and all test sets, without affecting decoding speed. By comparing the two data conditions in English-Italian, we see that most of the BLEU gain is preserved even after adding a large amount of parallel training data. This suggests that morphological phenomena are not sufficiently captured by phrases and stresses the importance of specifically modeling word inflection. It is possible that adding even more training data would reduce the impact of our inflection model, but currently we do not have access to other data sets that would be relevant to our translation tasks.

To put these results into perspective, our improvements are comparable to those achieved by previous work that generated new phrase inflections using a morphological disambiguator (Chahuneau et al., 2013) on the same large-scale English-Russian task.

[7] Riezler and Maxwell (2005) have shown that approximate randomization is less sensitive to Type-I errors, i. e. less likely to falsely reject the null hypothesis, than bootstrap resampling (Koehn, 2004) in the context of SMT.

(1)	SRC	and if you're wondering about **those other spikes**, those are also fridays
	REF	e se vi state chiedendo cosa sono **questi altri picchi**, sono anche loro dei venerdì
	BASE	e se vi state chiedendo di **queste altre picchi**, sono anche il venerdì
	INFNN	e se vi state chiedendo di **questi altri picchi**, sono anche il venerdì
	Effect:	*Correct number agreement between adjectives and noun*
(2)	SRC	... a three-hour **version** of this that'**s been viewed** four million times
	REF	... una **versione** di tre ore che **è stata vista** 4 milioni di volte
	BASE	... una **versione** di tre ore di ciò che **è stato visto** 4 milioni di volte
	INFNN	... una **versione** di tre ore di questo che **è stata osservata** quattro milioni di volte
	Effect:	*Correct gender agreement between subject and present perfect*
(3)	SRC	**he** died **broken** by history
	REF	morì **distrutto** dalla storia
	BASE	morì **infranta** dalla storia
	INFNN	morì **devastato** dalla storia
	Effect:	*Correct gender agreement between subject and adjective*
(4)	SRC	in one , i was **the classic asian student** ...
	REF	in uno ero **la classica studentessa asiatica** ...
	BASE	in uno stato **il classico asiatica studente** ...
	INFNN	in uno stato **il classico asiatico studente** ...
	Effect:	*Encouraged gender agreement between adjectives and noun, but gender is wrong*
(5)	SRC	in the other , i was **enmeshed** in lives that were precarious
	REF	nell'altro ero **invischiata** tra esistenze precarie
	BASE	tra l'altro, sono stato profondamente **impegnati** in vita che erano più precaria
	INFNN	nell'altro, ero profondamente **impegnati** in vita che erano più precaria
	Effect:	*Failed to encourage gender agreement because surface form is not in the SMT models*

Table 6: Examples of SMT output drawn from IWSLT English-Italian test12 showing the effect of our inflection model on lexical selection.

6.3 Examples

As previously mentioned, most previous approaches to inflection modeling for SMT may not be applied to Italian due to the lack of morphological disambiguated data. It is then particularly interesting to analyze *how* our model affects baseline translations. Table 6 presents a number of English-Italian SMT output examples where the use of our soft-tag Inf-NN either resulted in a better inflection choice (1-3) or not (4-5). Out of the 'good' examples, only (1) resulted in a complete match with the reference translation, while in (2) and (3) the system preferred an equally appropriate lexical choice, showing that automatically evaluating inflection models in an SMT setting is far from trivial.

The usefulness of source-side features is demonstrated by example (3): here, the translation of *broken* should agree in gender with the subject *he* but the baseline system chose instead a feminine form (*infranta*). Since the subject pronoun can be dropped in Italian, this error cannot be detected by the target language model and may only be fixed by translating the sequence *'he died broken'* as a single phrase, which was never observed in the training data. By contrast, Inf-NN successfully exploited the source-side context and preferred a masculine form (*devastato*).

Next are two unsuccessful examples: in (4) Inf-NN encouraged the system to translate the whole phrase *'the classic asian student'* as masculine whereas the baseline translation used

an incoherent mix of masculine and feminine. Unfortunately, though, the *student* in question, i.e., the speaker, happened to be a woman, but this could not be inferred in any way from this sentence. In (5) Inf-NN failed to fix the agreement between adjective and subject pronoun. By inspecting the parallel data we found that the word *enmeshed* always occurred with plural forms of Italian adjectives. This example shows that improving the scoring of the existing translation options is not always sufficient. While we do not address generation of new inflected forms in this work, this is an interesting direction for future work.

7 Conclusions

We have proposed a novel morphological representation scheme combined with a neural network for modeling translation into morphologically rich languages (MRLs). Our approach successfully circumvents the problem of ambiguous word analyses and makes it possible to improve translation into MRLs where morphological lexica but no manually disambiguated corpora exist.

Evaluated in a re-inflection task, the proposed soft tags achieve significantly higher accuracy than (i) a model using standard tags and trained on morphologically disambiguated data and (ii) a Maximum Entropy model that does not learn distributed representations for source words and target tags. When integrated into a state-of-the-art SMT decoder, our inflection model significantly improves translation quality in two different language pairs, without having to disambiguate during decoding. In particular, our positive English-Italian results under both small- and large-scale data conditions demonstrate the applicability of our approach to languages where no disambiguator exists.

As future work, we will consider learning distributed morphology representation directly from the corpus jointly with the inflection model as well as generating unseen word inflections during translation.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 612.001.218.

References

Andreas, J. and Klein, D. (2015). When and why are log-linear models self-normalizing? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 244–249, Denver, Colorado. Association for Computational Linguistics.

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.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bisazza, A. and Monz, C. (2014). Class-based language modeling for translating into morphologically rich languages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1918–1927, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 136–143, San Francisco, CA.

Bocharov, V., Alexeeva, S., Granovsky, D., Protopopova, E., Stepanova, M., and Surikov, A. (2013). Crowdsourcing morphological annotation. In *Proceedings of the International Conference "Dialogue"*, Bekasovo, Russia.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Botha, J. A. and Blunsom, P. (2014). Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China.

Brown, P. F., deSouza, P. V., Mercer, R. L., Della Pietra, V. J., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Carter, S. and Monz, C. (2011). Syntactic discriminative language model rerankers for statistical machine translation. *Machine Translation Journal*, 25(4):317–339.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th IWSLT evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 2–17, Lake Tahoe, California.

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.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 4(13):359–393.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1370–1380, Baltimore, Maryland. 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.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Morristown, NJ, USA. Association for Computational Linguistics.

Green, S. and DeNero, J. (2012). A class-based agreement model for generating accurately inflected translations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 146–155, Jeju Island, Korea. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *ArXiv e-prints*.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015). On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

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

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of ICLR*, San Diego, CA, USA.

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.

Koehn, P., Arun, A., and Hoang, H. (2008). Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio. Association for Computational Linguistics.

Koehn, P. and Hoang, H. (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. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Li, Z., Eisner, J., and Khudanpur, S. (2009). Variational decoding for statistical machine translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 593–601, Singapore.

Luong, T., Socher, R., and Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., and Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Minkov, E., Toutanova, K., and Suzuki, H. (2007). Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135.

Müller, T., Schütze, H., and Schmid, H. (2012). A comparative investigation of morphological language modeling for the languages of the European Union. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 386–395, Montréal, Canada. Association for Computational Linguistics.

Noreen, E. W. (1989). *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley-Interscience.

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

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 of the Association of Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Post, M. and Gildea, D. (2008). Parsers as language models for statistical machine translation. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 172–181.

Riezler, S. and Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Schwenk, H. (2007). Continuous space language models. *Computer Speech and Language*, 21:492–518.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. (2008). Designing and evaluating a russian tagset. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Sima'an, K. (1996). Computational complexity of probabilistic disambiguation by means of tree-grammars. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 1175–1180. Association for Computational Linguistics.

Soricut, R. and Och, F. (2015). Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado. Association for Computational Linguistics.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Subotin, M. (2011). An exponential translation model for target language morphology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238, Portland, Oregon, USA. 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. M., 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 (EMNLP)*, pages 1676–1688, Doha, Qatar. Association for Computational Linguistics.

Uszkoreit, J. and Brants, T. (2008). Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of ACL-08: HLT*, pages 755–762, Columbus, Ohio. Association for Computational Linguistics.

Zanchetta, E. and Baroni, M. (2005). Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, 1(1).