PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 49–60

Learning Morphological Normalization for Translation from and into Morphologically Rich Languages

Franck Burlot, François Yvon

LIMSI, CNRS, Université Paris-Saclay, France

Abstract

When translating between a morphologically rich language (MRL) and English, word forms in the MRL often encode grammatical information that is irrelevant with respect to English, leading to data sparsity issues. This problem can be mitigated by removing from the MRL irrelevant information through normalization. Such preprocessing is usually performed in a deterministic fashion, using hand-crafted rules and yielding suboptimal representations. We introduce here a simple way to automatically compute an appropriate normalization of the MRL and show that it can improve machine translation in both directions.

1. Introduction

Translating from a morphologically rich language (MRL) like Czech or Russian into a more analytical language like English leads to several issues, due to important divergences in their respective linguistic systems. The MRLs considered in this study have synthetic tendencies, which means that they often encode grammatical information in the endings of words, notably case marks which signal the grammatical function of a word in the sentence. There is no such phenomena in English, where the function of a word is instead encoded in a specific word order or expressed in prepositions. This results in an obvious lack of symmetry between those two types of languages. For instance, while on the MRL side adjectives may vary in gender, number and case, their English translation is invariable. Such differences can impact machine translation (MT) quality in several ways:

 Corresponding author: franck.burlot@limsi.fr
Cite as: Franck Burlot, François Yvon. Learning Morphological Normalization for Translation from and into Morphologically Rich Languages. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 49–60.
doi: 10.1515/pralin-2017-0008.

- The increase of word forms in the MRL means that each form has a smaller occurrence count than its English counterpart(s), yielding poor probability estimates for infrequent words;
- An even more extreme case is the translation of word forms unseen in training. Even if other forms of the same lemma are known, the MT system cannot generalize and will produce an erroneous output.

A well-known way to mitigate this problem is to "simplify" the MRL by removing information that is deemed redundant with respect to English. This solution has been repeatedly used to translate into the MRL (eg. in (Ney and Popovic, 2004; Durgar El-Kahlout and Yvon, 2010) for German, (Goldwater and McClosky, 2005) for Czech), and is adopted in recent systems competing at WMT (e.g. (Allauzen et al., 2016; Lo et al., 2016) for Russian), as well as in the reverse direction (Minkov et al., 2007; Toutanova et al., 2008; Fraser et al., 2012) with the additional complexity that the simplified MT output needs to be augmented with the missing information ("reinflected" in the MT jargon). One downside of these procedures is that they are entirely dependent on the language pairs under study, and rely on hand-crafted rules that need to be adapted for each new language. It is also likely that rule-based normalization is suboptimal with respect to the task, as it does not take the peculiarities of the training data into account.

We introduce (Section 3) a new way to automatically perform such normalization, by clustering together MRL forms.[1] Clustering is performed on a per lemma basis and groups together morphological variants that tend to translate into the same target word(s). We show in Section 4 that this normalization helps when translating into English. A second contribution is a new neural reinflection system, which is crucially able to also take advantage of source-side information, yielding significant improvements when translating into a MRL (Section 5).

2. Related Work

The normalization of the vocabulary on the MRL side mostly consists in removing word information that is deemed redundant with respect to English. Most of the time, normalization relies on expert knowledge specifying which MRL words can be merged without generating confusion in English, (see eg. (Ney and Popovic, 2004; Goldwater and McClosky, 2005; Durgar El-Kahlout and Yvon, 2010)). An alternative, which does not require user expertise is introduced by Talbot and Osborne (2006), who proposed to use model selection techniques to identify useful clusters in the MRL vocabulary. Even though we start from the same intuition (to cluster forms having similar translation distributions), our model is much simpler and more explicitly oriented toward morphological variation, which makes it also easier to invert.

[1] Our implementation is available at `https://github.com/franckbrl/bilingual_morph_normalizer`.

The same kind of solution is also useful when translating in the reverse direction; it additionally requires a two-step MT architecture addressing morphology as a post-processing step. Minkov et al. (2007) and Toutanova et al. (2008) translate from English into Russian and Arabic stems, which are used to generate full paradigms, then disambiguated using a classifier. Similarly, Chahuneau et al. (2013) augment the translation model with synthetic phrases obtained by re-inflecting target stems. Bojar (2007) cascade two Statistical MT systems: the first one translates from English into Czech lemmas decorated with source-side information and the second one performs a monotone translation into fully inflected Czech.

Fraser et al. (2012) represent German words as lemmas followed by a sequence of tags and introduce a linguistically motivated selection of these in order to translate from English. The second step consists in predicting the tags that have been previously removed, using a dedicated model for each morphological attribute. Finally, word forms are produced by looking-up in a morphological dictionary. El Kholy and Habash (2012a; 2012b) propose a similar approach for Arabic.

3. Source-side Clustering

3.1. Information Gain

Our goal is to cluster together MRL forms that translate into the same target word(s). We assume that each MRL form f is a combination of a lemma, a part of speech (PoS) and a sequence of morphological tags,[2] and that a word aligned parallel corpus is available, from which lexical translation probabilities $p(e|f)$ and unigram probabilities $p(f)$ can be readily computed. We first consider the simple case where the corpus contains one single lemma for each PoS. We denote respectively $\mathbf{f}$ the set of word forms (or, equivalently, of positions in the paradigm) for this lemma, and $\mathbf{E}$ the complete English vocabulary. The conditional entropy (CE) of the translation model is:

$$H(\mathbf{E}|\mathbf{f}) = \sum_{f \in \mathbf{f}} p(f) H(\mathbf{E}|f) = \sum_{f \in \mathbf{f}} \frac{p(f)}{\log_2 |\mathbf{E}_{a_f}|} \sum_{e \in \mathbf{E}_{a_f}} -p(e|f) \log_2 p(e|f), \quad (1)$$

where $\mathbf{E}_{a_f}$ is the set of words aligned with f. The normalizer ($\log_2 |\mathbf{E}_{a_f}|$) ensures that all the entropy values are comparable, no matter the number of aligned target words.

From an initial state where each form is a singleton cluster, and proceeding bottom-up, we repeatedly try to merge cluster pairs (f_1 and f_2) so as to reduce the CE. We therefore compute the information gain (IG) of the merge operation:

$$IG(f_1, f_2) = p(f_1)H(\mathbf{E}|f_1) + p(f_2)H(\mathbf{E}|f_2) - p(f')H(\mathbf{E}|f') \quad (2)$$

[2] For instance, the Czech *autem* (by car) is represented as: *auto + Noun + neutral + singular + instrumental*.

where f' is the resulting aggregate. IG ($\in [-1, +1]$) measures the difference between the combined CEs of clusters f_1 and f_2 before and after merging in f'. If the corresponding forms have similar translation distributions, the information gain is positive; conversely when their translations are different, it is negative and the merge leads to a loss of information. Note that the total entropy $H(\mathbf{E}|\mathbf{f})$ of the translation model can be recomputed *incrementally* after merging (f_1, f_2) by:

$$H(\mathbf{E}|\mathbf{f}) \leftarrow H(\mathbf{E}|\mathbf{f}) - IG(f_1, f_2) \tag{3}$$

IG can also be interpreted as a measure of similarity between two word forms and can be readily used in any clustering model, such as *k-means*. Doing so would however require to fix the total number of clusters, which we would rather like to determine based on the available data. We have therefore opted for an agglomerative clustering procedure, which we now fully describe.

3.2. Clustering Paradigm Cells

In practice, our algorithm is applied at the level of PoS, rather than individual lemmas: we therefore assume that for a given PoS p, all lemmas have the same number n_p of possible morphological variants (cells in their paradigm). This means that IG computations will be aggregated over all lemmas of a given PoS, based on statistics maintained on a per lemma basis. For each lemma of PoS p, the starting point is a matrix $L_l \in [-1:1]^{n_p \times n_p}$, with $L_l(i,j)$ the IG resulting from merging forms l_i and l_j of lemma $\mathbf{l}$. The average of these matrices over all lemmas defines *the PoS level matrix* $M_p \in [-1:1]^{n_p \times n_p}$ containing the average information gain resulting from merging two cells.

Algorithm 1: A bottom-up clustering algorithm

```
C(p) ← {1, ..., n_p}
i, j ← arg max_{i',j' ∈ C(p)^2} M_p(i', j')
repeat
    Merge i and j in C(p)
    for l ∈ V_lem do
        Remove L_l(i, j), create L_l(ij)
        Compute p(ij), p(E|ij) and H(E|ij)
        Compute L_l(ij, k) for k ∈ C(p)
    M_p ← Σ_{l ∈ V_lem} L_l
    i, j ← arg max_{i',j' ∈ C(p)^2} M_p(i', j')
until M_p(i, j) < m or |C(p)| = 1
```

The clustering procedure is described in Algorithm 1. It starts with n_p classes for each PoS and iteratively performs merge operations, as long as the cumulated information gain for the merge exceeds a minimum threshold m. After each merge,

the statistics for the new cluster (unigram probability, translation probability and entropy) are recomputed *for all lemmas* and used to update the PoS-level IG matrix M_p. When the procedure halts, a clustering $C(p)$ is obtained for PoS p, which can then be applied to normalize the source data in various ways (see Section 4.3).

In practice, we obtained slightly better results and a much better runtime than the exact computation of algorithm 1 with an alternative update regime for the IG Matrix M_p, which dispenses with the costly update of all the matrices L_l (lines 5–8). Once initialized, M_p is treated like a similarity matrix and updated using a procedure reminiscent of the linkage clustering algorithm. The aggregated matrix cell for clusters c_1 and c_2 is thus computed as the average IG of all possible 2-way merging operations:

$$M_p(c_1, c_2) = \frac{\sum_{f_1 \in c_1} \sum_{f_2 \in c_2} M(f_1, f_2)}{|c_1| \times |c_2|}. \quad (4)$$

4. Translating from and into a normalized MRL

We assess the normalization model on MT tasks for three language pairs in both directions: Czech-English, Russian-English and Czech-French; note that the latter involves two MRLs.

4.1. Experimental Setup

Tokenization of English and French uses in-house tools. We used the script from the Moses toolkit (Koehn et al., 2007) for Czech and TreeTagger (Schmid, 1994) for Russian. The MT models are trained using Moses with various datasets from WMT 2016[3] (Table 1). 4-gram language models were trained with KenLM (Heafield, 2011) over the monolingual datasets. These systems are optimized with KB-MIRA (Cherry and Foster, 2012) using WMT Newstest-2015 and tested on Newstest-2016. The Czech-French systems were tuned on Newstest-2014 and tested on Newstest-2013.

	cs2en		**en2cs**		**cs2fr**		**fr2cs**		**ru2en**		**en2ru**	
Setup	parall	mono	parall	mono	parall	mono	parall	mono	parall	mono	parall	mono
`Small`	190k	150M	190k	8.4M	622k	12.3M	622k	8.4M	190k	150M	190k	9.6M
`Larger`	1M	150M	1M	34.4M								
`Largest`	7M	250M	7M	54M								

Table 1. Datasets used to train the MT systems

The source-side normalization is performed independently for each dataset, using the training set of the MT system, except for the `Larger` and `Largest` Czech systems for which the parallel data of the `Larger` system was used. The lemmas and tags are

[3] `http://www.statmt.org/wmt16`

obtained with Morphodita (Straková et al., 2014) for Czech and TreeTagger (Schmid, 1994; Sharoff et al., 2008) for Russian. Filtering the MRL lemmas when performing clustering yields better results and we have excluded lemmas appearing less than 100 times, as well as word forms occurring less than 10 times in the training set in order to mitigate the noise in the initial alignments. When clustering paradigm cells (see Section 3.2), we set the minimum IG value $m = 0$.

4.2. A qualitative assessment of normalized Czech

The clustering learned over the `Small` Czech-English data led to a drastic reduction of the source vocabulary. Starting with 158,914 distinct character strings, corresponding to 237,378 fully disambiguated word forms (represented as lemmas and morphological information), we ended up with a set of 90,170 normalized entries.

The resulting clusters confirm some linguistic intuitions. First, nouns turned out to be distinguished only by their number, a property that is also marked for English nouns. We also observed a small number of singleton noun classes, mainly at the instrumental case which often corresponds to the English prepositions *by* and *with* (including the dual number for *rukama* → *with [my] hands*), as well as the vocative case. All possessive pronouns were distinguished only by their person, as is also the case in English; adjectives were clustered separately according to their degree of comparison, verbs are clustered by time, the third person singular of the present tense being separated, since it is marked in English (*I cluster, he cluster**s***). We only noticed a small residual noise with negative verbs, sometimes clustered with affirmative ones. This might be due to alignment errors where an English negation particle is not linked to a Czech negative verb, a typical issue for this language pair (Rosa, 2013). Our model thus seems to be able to capture subtle linguistic phenomena that would require a large amount of rules if such normalization had to be performed manually.

4.3. MT experiments

The results for all Czech systems are in Table 2 and are reported based on different applications of the normalization model. Indeed, normalization can be used to train both the alignment (ali cx) or the full system (cx2en), yielding a total improvement of 1.36 BLEU in the `Small` conditions. Using it only for alignments or only for the MT system gives worse results, still outperforming the baseline (cs2en). This shows that both tasks take advantage of the source normalization. Another way to apply the clustering model is to exclude from normalization the 100 most frequent lemmas (100 freq), which gives the best result for this setup. For the other direction (en2cs), the Czech normalization was used to train the alignments and gives only a slight improvement over the baseline. Results for the translation into normalized Czech (en2cx) after a reinflection step are reported in Section 5.2.

The same tendency holds for the `Larger` Czech-English system, even though the contrasts in BLEU scores are slightly less visible, due to the larger amount of training

	Small System		Larger System		Largest System	
System	BLEU	OOV	BLEU	OOV	BLEU	OOV
cs2en (ali cs)	21.26	2189	23.85	1878	24.99	1246
cx2en (ali cx)	22.62 (+1.36)	1888	24.57 (+0.72)	1610	24.65 (-0.43)	988
cs2en (ali cx)	22.19 (+0.93)	2152	24.14 (+0.29)	1832	**25.35 (+0.36)**	1212
cx2en (ali cs)	22.34 (+1.08)	1914	24.36 (+0.51)	1627		
cx2en (100 freq)	**22.82 (+1.56)**	1893	**24.85 (+1.00)**	1614		
cx2en ($\mathfrak{m} = -10^{-4}$)			24.44 (+0.59)	1604		
cx2en ($\mathfrak{m} = 10^{-4}$)			24.05 (+0.20)	1761		
cx2en (manual)			24.46 (+0.61)	1623		
en2cs (ali cs)	15.21		17.42		19.14	
en2cs (ali cx)	15.54 (+0.33)		17.55 (+0.13)		19.23 (+0.09)	

Table 2. Czech-English Systems

data, which reduces sparsity. For this setup, we also have tried different values of the minimum IG $\mathfrak{m}$ (see Section 3.2). Our results suggest that the optimal value for $\mathfrak{m}$ is close to 0. Indeed, higher values produce more clusters, which leads to more OOVs (1761 OOVs for 10^{-4}, vs. 1604 for $\mathfrak{m} = -10^{-4}$), thus hurting the overall performance.

In the `Largest` Czech-English setup, using normalization to train both the alignments and the translation system hurts the performance (-0.43 BLEU). On the other hand, using it only to train the alignments does give a small improvement. In the reverse direction (en2cs), training the alignments over normalized Czech does not give any significant improvement.

Results for a manual normalization (manual) are also reported. The normalization rules are close to the ones used in (Burlot et al., 2016) where nouns are distinguished by number and negation, adjectives by negation and degree of comparison, etc. We also applied rules for verb clusters that are distinguished by tense and negation, except the singular third person present tense that is kept. This manual normalization improves the baseline (+0.61), but not as much as our best system (+1.00).

System	BLEU	OOV
cs2fr (ali cs)	19.57	1845
cx2fr (ali cx)	**20.19 (+0.62)**	1592
fr2cs (ali cs)	13.36	
fr2cs (ali cx)	13.18 (-0.18)	

Table 3. Czech-French systems

System	BLEU	OOV
ru-en (ali ru)	19.76	2260
rx-en (ali rx)	**21.02 (+1.26)**	2033
rx-en (ali ru)	20.92 (+1.16)	2033
ru-en (ali rx)	20.53 (+0.77)	2048
rx-en (100 freq)	20.89 (+1.13)	2026
en-ru (ali ru)	16.59	
en-ru (ali rx)	16.95 (+0.36)	

Table 4. Russian-English systems

The results for Russian-English follow the same tendency as Czech-English, except that keeping the word forms for the 100 most frequent lemmas did not improve over

the full normalization of the training set. Finally, we note in Table 3 that the Czech normalization towards French also helps to improve the translation, even though the target language is morphologically richer than English. The improvements are smaller, though, than when translating into English. We assume that this is due to a degree of normalization that is lower when the source shares certain properties with the target, such as adjective inflection, which leads our model to create more classes. Indeed, the model distinguishes nouns by their number, just like with English, but moreover creates separate clusters for each adjective gender. This reduced degree of normalization did not help the training of alignments when translating into Czech (fr2cs).

5. Morphological Reinflection

When translating into a MRL, using normalization to train just the alignments did not prove very helpful (Section 4.3). We now consider using it for the complete translation system. Translating from English into fully inflected Czech however requires a non-trivial post-processing step for reinflection. In this section, we introduce our solution to this problem and provide results for several English-Czech systems.

5.1. A Morphological Reinflection Model

We view the reinflection of the normalized MT output as predicting the fine-grained PoS tag for each output token. Knowing the normalized word and its PoS tag is sufficient to recover the fully inflected word form by dictionary lookup.

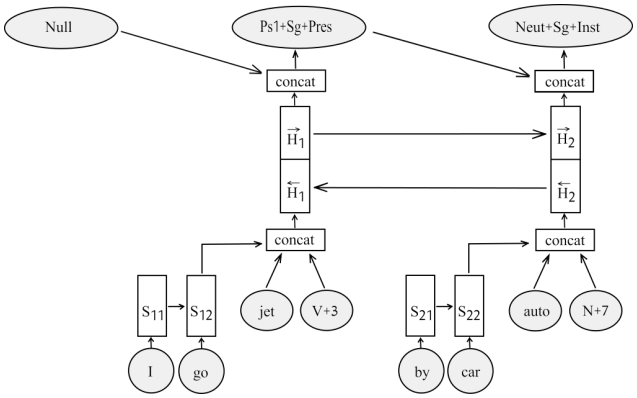

Figure 1. RNN architecture for target-side morphology prediction.

For this sequence labeling task, we used a bidirectional recurrent neural network (RNN) that considers both normalized Czech words as well as source-side English

tokens to make its predictions (see Figure 1). It computes a probability distribution over the output PoS tags $\mathbf{y}$ at time t, given both the Czech ($\mathbf{f}$) and the English ($\mathbf{e}$) sentences, as well as the previous prediction: $p(\mathbf{y}_t|\mathbf{f}, \mathbf{e}, \mathbf{y}_{t-1})$.

For each word f_t in the Czech sentence, we need to encode the English words that generated f_t during the translation process. As there can be an arbitrary number of them (denoted I_t below), we used a RNN layer,[4] where each state S_i inputs a source token representation $a_{t,i}$ and the previous hidden state S_{i-1}. The last state (at time I_t) of that layer is used to represent the sequence of aligned tokens: $S_{t,I} = \mathcal{A}(S_{t,I_t-1}, a_{t,I_t})$.

Each normalized Czech word representation is decomposed into a lemma embedding l_t and a cluster embedding c_t, which are represented in distinct continous spaces. These vectors are concatenated with the source representation S_{t,I_t}, defining the input to the bidirectional RNN[5] performing PoS tagging. A forward layer hidden state H at time t is therefore computed as: $\overrightarrow{H}_t = \mathcal{R}(H_{t-1}, [S_{t,I_t}; l_t; c_t])$. Finally, both forward and backward layers are concatenated with the representation of the preceding PoS tag y_{t-1} [6] and the result is passed through a last feed-forward layer to which a softmax is applied. All the model parameters, including embeddings, are trained jointly in order to maximize the log-likelihood of the training data.

5.2. Experimental Results

The reinflection systems introduced in this section were trained with the parallel English-Czech data used for the `Small` setup (News-Commentary). The fine-grained PoS tags are the same as the ones used to train the normalization in Section 4 (Straková et al., 2014).[7] The word alignments used for the training and validation sets were obtained with fast_align (Dyer et al., 2013). At test time, we use the alignments produced by the MT decoder. Since the Czech side of the parallel data must be normalized prior to training, the results below were obtained with two versions of the RNN model: with the `Small` data normalization and with the `Larger` data one (see Section 4).

Each normalized Czech word is associated with a sequence of source English words that we collect as follows: using word alignments, we take the English words that are linked to the current position, as well as surrounding unaligned words. These unaligned words often contain essential information: as shown in (Burlot and Yvon, 2015), many of them have a grammatical content that is helpful to predict the correct inflection on target side. For instance, the English preposition *of* is an important pre-

[4] Encoding the sequence of aligned tokens with a "bag of words" model, where we just sum the embeddings, performed worse in our experiments.

[5] The RNN layers for English and normalized Czech contain gated recurrent units (Cho et al., 2014).

[6] Representing the full left-side target context with an additional RNN did not bring any improvement.

[7] Our attempts to use the manually annotated data from the Universal Dependency Treebank project (`http://universaldependencies.org`) to train a monolingual variant of our model turned out to give worse results, supposedly because this data is not entirely in-domain.

dictor of the Czech genitive case. This type of grammatical information is the only one that matters for this task, since the lexical content of the Czech words is already computed by the MT system and can not be changed. In fact, replacing the English content words by their PoS and keeping only words in a list of stopwords proved to work better than keeping all the words. Decoding used a beam search of size 5, and the final lookup uses the Morphodita morphological generator.

We consider here three English-Czech MT systems with reinflection. The training data is the same as the `Small`, `Larger` and `Largest` systems described in Section 4, except that the Czech target side is now normalized. The reinflection model can also be used in different ways. One can use it to process the one-best hypothesis of the MT system, or the n-best hypotheses ($n = 300$ in our experiments). A third approach reinflects n-best lists and outputs k-best hypotheses from the reinflection model ($k = 5$ in our experiments). These are finally scored by a language model trained on the same data as the one used in the MT system – albeit with fully inflected words. This score is added to the ones given by the MT system. With nk-best reinflection, we also add the scores given by the reinflection model (log-probability of the predicted sequence). All these scores are finally interpolated using Mira optimization over Newstest-2015 set and produce a single best output sentence.

	Small System			**Larger System**			**Largest System**		
	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓	BLEU ↑	BEER ↑	CTER ↓
en2cs (ali cs)	15.21	0.512	0.624	17.42	0.531	0.602	19.14	0.543	0.582
en2cs (ali cx)	15.54	0.516	0.617	17.55	0.532	0.597	19.23	0.544	0.578
en2cx (1-best)	16.07	0.520	0.606	18.00	0.535	0.589	19.19	0.545	0.573
en2cx (n-best)	16.37	0.521	**0.601**	17.41	0.529	0.591	19.48	0.547	**0.570**
en2cx (nk-best)	**16.93**	**0.525**	0.602	**18.81**	**0.540**	**0.588**	**19.95**	**0.548**	0.572

Table 5. BLEU scores for English-Czech

Results are in Table 5, where we provide, in addition to BLEU, scores computed by BEER (Stanojević and Sima'an, 2014) and CharacTER (Wang et al., 2016). These two metrics proved to be more adapted to MRLs by Bojar et al. (2016). We observe a slight improvement when reinflecting the 1-best hypothesis in the `Small` data conditions. With the `Largest` dataset, the reinflection has nearly no impact on the translation quality according to BLEU and BEER. Like for the reverse direction, the improvements of normalization get lower as the size of the dataset grows. We were nevertheless able to obtain a reasonable improvement of 0.81 BLEU points over the baseline in the `Largest` data conditions, which shows that even when a huge quantity of data is available, a specific handling of morphology on target side can still be useful.

6. Conclusion

We have introduced a simple language agnostic way to automatically infer the normalization of a morphologically rich language with respect to the target language that consists in clustering together words that share the same translation, and have shown that it improves machine translation in both directions. Future work will consist in testing our model on neural machine translation systems.

Acknowledgments

This work has been partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

Bibliography

Allauzen, Alexandre, Lauriane Aufrant, Franck Burlot, Ophélie Lacroix, Elena Knyazeva, Thomas Lavergne, Guillaume Wisniewski, and François Yvon. LIMSI@WMT16: Machine Translation of News. In *Proc. WMT*, pages 239–245, Berlin, Germany, 2016.

Bojar, Ondřej. English-to-Czech Factored Machine Translation. In *Proc. of the 2nd WMT*, pages 232–239, Prague, Czech Republic, 2007.

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. Results of the WMT16 Metrics Shared Task. In *Proc. WMT*, pages 199–231, Berlin, Germany, 2016.

Burlot, Franck and François Yvon. Morphology-Aware Alignments for Translation to and from a Synthetic Language. In *Proc. IWSLT*, pages 188–195, Da Nang, Vietnam, 2015.

Burlot, Franck, Elena Knyazeva, Thomas Lavergne, and François Yvon. Two-Step MT: Predicting Target Morphology. In *Proc. IWSLT*, Seattle, USA, 2016.

Chahuneau, Victor, Eva Schlinger, Noah A. Smith, and Chris Dyer. Translating into Morphologically Rich Languages with Synthetic Phrases. In *EMNLP*, pages 1677–1687, 2013.

Cherry, Colin and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the NAACL-HLT*, pages 427–436, Montreal, Canada, 2012.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proc. SSST@EMNLP*, pages 103–111, Doha, Qatar, 2014.

Durgar El-Kahlout, Ilknur and François Yvon. The pay-offs of preprocessing for German-English Statistical Machine Translation. In *Proc. IWSLT*, pages 251–258, Paris, France, 2010.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proc. NAACL*, pages 644–648, Atlanta, Georgia, 2013.

El Kholy, Ahmed and Nizar Habash. Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation. In *Proc. EAMT*, pages 27–34, Trento, Italy, 2012a.

El Kholy, Ahmed and Nizar Habash. Rich Morphology Generation Using Statistical Machine Translation. In *Proc. INLG*, pages 90–94, 2012b.

Fraser, Alexander, Marion Weller, Aoife Cahill, and Fabienne Cap. Modeling Inflection and Word-Formation in SMT. In *Proc. EACL*, pages 664–674, Avignon, France, 2012.

Goldwater, Sharon and David McClosky. Improving Statistical MT through Morphological Analysis. In *Proc. HLT–EMNLP*, pages 676–683, Vancouver, Canada, 2005.

Heafield, Kenneth. KenLM: Faster and Smaller Language Model Queries. In *Proc. WMT*, pages 187–197, Edinburgh, Scotland, 2011.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical MT. In *Proc. ACL:Systems Demos*, pages 177–180, Prague, Czech Republic, 2007.

Lo, Chi-kiu, Colin Cherry, George Foster, Darlene Stewart, Rabib Islam, Anna Kazantseva, and Roland Kuhn. NRC Russian-English Machine Translation System for WMT 2016. In *Proc. WMT*, pages 326–332, Berlin, Germany, 2016.

Minkov, Einat, Kristina Toutanova, and Hisami Suzuki. Generating Complex Morphology for Machine Translation. In *Proc. ACL*, pages 128–135, Prague, Czech Republic, 2007.

Ney, Hermann and Maja Popovic. Improving Word Alignment Quality using Morpho-syntactic Information. In *Proc. COLING*, pages 310–314, Geneva, Switzerland, 2004.

Rosa, Rudolf. Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Institute of Formal and Applied Linguistics, Charles University, 2013.

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

Sharoff, Serge, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. Designing and Evaluating a Russian Tagset. In *Proc. LREC*, pages 279–285, Marrakech, Marocco, 2008.

Stanojević, Miloš and Khalil Sima'an. Fitting Sentence Level Translation Evaluation with Many Dense Features. In *Proc. EMNLP*, pages 202–206, Doha, Qatar, 2014.

Straková, Jana, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proc. ACL: System Demos*, pages 13–18, Baltimore, MA, 2014.

Talbot, David and Miles Osborne. Modelling Lexical Redundancy for Machine Translation. In *Proc. ACL*, pages 969–976, Sydney, Australia, 2006.

Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp. Applying Morphology Generation Models to Machine Translation. In *Proc. ACL-08: HLT*, pages 514–522, Columbus, OH, 2008.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTer: Translation Edit Rate on Character Level. In *Proc. WMT*, pages 505–510, Berlin, Germany, 2016.

Address for correspondence:
Franck Burlot
franck.burlot@limsi.fr
LIMSI-CNRS
Campus Universitaire Orsay, 91 403 Osay, France