# English-to-Czech Factored Machine Translation

**Ondřej Bojar**
Institute of Formal and Applied Linguistics
ÚFAL MFF UK, Malostranské náměstí 25
CZ-11800 Praha, Czech Republic
bojar@ufal.mff.cuni.cz

## Abstract

This paper describes experiments with English-to-Czech phrase-based machine translation. Additional annotation of input and output tokens (multiple factors) is used to explicitly model morphology. We vary the translation scenario (the setup of multiple factors) and the amount of information in the morphological tags. Experimental results demonstrate significant improvement of translation quality in terms of BLEU.

## 1 Introduction

Statistical phrase-based machine translation (SMT) systems currently achieve top performing results.[1] Known limitations of phrase-based SMT include worse quality when translating to morphologically rich languages as opposed to translating from them (Koehn, 2005). One of the teams at the 2006 summer engineering workshop at Johns Hopkins University[2] attempted to tackle these problems by introducing separate FACTORS in SMT input and/or output to allow explicit modelling of the underlying language structure. The support for factored translation models was incorporated into the Moses open-source SMT system[3].

In this paper, we report on experiments with English-to-Czech multi-factor translation. After a brief overview of factored SMT and our data (Sections 2 and 3), we summarize some possible translating scenarios in Section 4. Section 5 studies the level of detail useful for morphological representation and Section 6 compares the results to a setting with more data available, albeit out of domain. The second part (Section 7) is devoted to a brief analysis of MT output errors.

### 1.1 Motivation for Improving Morphology

Czech is a Slavic language with very rich morphology and relatively free word order. The Czech morphological system (Hajič, 2004) defines 4,000 tags in theory and 2,000 were actually seen in a big tagged corpus. (For comparison, the English Penn Treebank tagset contains just about 50 tags.) In our parallel corpus (see Section 3 below), the English vocabulary size is 35k distinct token types but more than twice as big in Czech, 83k distinct token types.

To further emphasize the importance of morphology in MT to Czech, we compare the standard BLEU (Papineni et al., 2002) of a baseline phrase-based translation with BLEU which disregards word forms (lemmatized MT output is compared to lemmatized reference translation). The theoretical margin for improving MT quality is about 9 BLEU points: the same MT output scores 12 points in standard BLEU and 21 points in lemmatized BLEU.

## 2 Overview of Factored SMT

In statistical MT, the goal is to translate a source (foreign) language sentence $f_1^J = f_1 \dots f_j \dots f_J$ into a target language (Czech) sentence $c_1^I = c_1 \dots c_j \dots c_I$. In phrase-based SMT, the assumption is made that the target sentence can be constructed by segmenting source sentence into phrases, translating each phrase and finally composing the

[1] http://www.nist.gov/speech/tests/mt/
[2] http://www.clsp.jhu.edu/ws2006/
[3] http://www.statmt.org/moses/

*Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239,
Prague, June 2007. ©2007 Association for Computational Linguistics

target sentence from phrase translations, $s_1^K$ denotes the segmentation of the input sentence into $K$ phrases. Among all possible target language sentences, we choose the sentence with the highest probability,

$$\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I,c_1^I,K,s_1^K} \{Pr(c_1^I | f_1^J, s_1^K)\} \tag{1}$$

In a log-linear model, the conditional probability of $c_1^I$ being the translation of $f_1^J$ under the segmentation $s_1^K$ is modelled as a combination of independent feature functions $h_1(\cdot,\cdot,\cdot) \dots h_M(\cdot,\cdot,\cdot)$ describing the relation of the source and target sentences:

$$Pr(c_1^I | f_1^J, s_1^K) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(c_1^I, f_1^J, s_1^K))}{\sum_{c'^{I'}_1} \exp(\sum_{m=1}^{M} \lambda_m h_m(c'^{I'}_1, f_1^J, s_1^K))} \tag{2}$$

The denominator in 2 is used as a normalization factor that depends on the source sentence $f_1^J$ and segmentation $s_1^K$ only and is omitted during maximization. The model scaling factors $\lambda_1^M$ are trained either to the maximum entropy principle or optimized with respect to the final translation quality measure.

Most of our features are phrase-based and we require all such features to operate synchronously on the segmentation $s_1^K$ and independently of neighbouring segments. In other words, we restrict the form of phrase-based features to:

$$h_m(c_1^I, f_1^J, s_1^K) = \sum_{k=1}^{K} \tilde{h}_m(\tilde{c}_k, \tilde{f}_k) \tag{3}$$

where $\tilde{f}_k$ represents the source phrase and $\tilde{c}$ represents the target phrase $k$ given the segmentation $s_1^K$.

### 2.1 Decoding Steps

In factored SMT, source and target words $f$ and $c$ are represented as tuples of $F$ and $C$ FACTORS, resp., each describing a different aspect of the word, e.g. its word form, lemma, morphological tag, role in a verbal frame. The process of translation consists of DECODING steps of two types: MAPPING steps and GENERATION steps. If more steps contribute to the same output factor, they have to agree on the outcome, i.e. partial hypotheses where two decoding steps produce conflicting values in an output factor are discarded.

A MAPPING step from a subset of source factors $S \subseteq \{1 \dots F\}$ to a subset of target factors $T \subseteq \{1 \dots C\}$ is the standard phrase-based model (see e.g. (Koehn, 2004a)) and introduces a feature in the following form:

$$\tilde{h}_m^{\mathrm{map}:S \rightarrow T}(\tilde{c}_k, \tilde{f}_k) = \log p(\tilde{f}_k^S | \tilde{c}_k^T) \tag{4}$$

The conditional probability of $\tilde{f}_k^S$, i.e. the phrase $\tilde{f}_k$ restricted to factors $S$, given $\tilde{c}_k^T$, i.e. the phrase $\tilde{c}_k$ restricted to factors $T$ is estimated from relative frequencies: $p(\tilde{f}_k^S | \tilde{c}_k^T) = N(\tilde{f}^S, \tilde{c}^T)/N(\tilde{c}^T)$ where $N(\tilde{f}^S, \tilde{c}^T)$ denotes the number of co-occurrences of a phrase pair $(\tilde{f}^S, \tilde{c}^T)$ that are consistent with the word alignment. The marginal count $N(\tilde{c}^T)$ is the number of occurrences of the target phrase $\tilde{c}^T$ in the training corpus.

For each mapping step, the model is included in the log-linear combination in source-to-target and target-to-source directions: $p(\tilde{f}^T | \tilde{c}^S)$ and $p(\tilde{c}^S | \tilde{f}^T)$. In addition, statistical single word based lexica are used in both directions. They are included to smooth the relative frequencies used as estimates of the phrase probabilities.

A GENERATION step maps a subset of target factors $T_1$ to a disjoint subset of target factors $T_2$, $T_{1,2} \subset \{1 \dots C\}$. In the current implementation of Moses, generation steps are restricted to word-to-word correspondences:

$$\tilde{h}_m^{\mathrm{gen}:T_1 \rightarrow T_2}(\tilde{c}_k, \tilde{f}_k) = \log \prod_{i=1}^{\mathrm{length}(\tilde{c}_k)} p(\tilde{c}_{k,i}^{T_1} | \tilde{c}_{k,i}^{T_2}) \tag{5}$$

where $\tilde{c}_{k,i}^T$ is the $i$-th words in the $k$-th target phrase restricted to factors $T$. We estimate the conditional probability $p(\tilde{c}_{k,i}^{T_2} | \tilde{c}_{k,i}^{T_1})$ by counting over words in the target-side corpus. Again, the conditional probability is included in the log-linear combination in both directions.

In addition to features for decoding steps, we include arbitrary number of target language models over subsets of target factors, $T \subseteq \{1 \dots C\}$. Typically, we use the standard $n$-gram language model:

$$h^{T}_{\mathrm{LM}_n}(f_1^J, c_1^I) = \log \prod_{i=1}^{I} p(c_i^T | c_{i-1}^T \dots c_{i-n+1}^T) \tag{6}$$

While generation steps are used to enforce "vertical" coherence between "hidden properties" of output words, language models are used to enforce sequential coherence of the output.

Operationally, Moses performs a stack-based beam search very similar to Pharaoh (Koehn, 2004a). Thanks to the synchronous-phrases assumption, all the decoding steps can be performed during a preparatory phase. For each span in the input sentence, all possible translation options are constructed using the mapping and generation steps in a user-specified order. Low-scoring options are pruned already during this phase. Once all translation options are constructed, Moses picks source phrases (all output factors already filled in) in arbitrary order, subject to a reordering limit, producing output in left-to-right fashion and scoring it using the specified language models exactly as Pharaoh does.

## 3 Data Used

The experiments reported in this paper were carried out with the News Commentary (NC) corpus as made available for the SMT workshop[4] of the ACL 2007 conference.[5]

The Czech part of the corpus was tagged and lemmatized using the tool by Hajič and Hladká (1998), the English part was tagged MXPOST (Ratnaparkhi, 1996) and lemmatized using the Morpha tool (Minnen et al., 2001). After some final cleanup, the corpus consists of 55,676 pairs of sentences (1.1M Czech tokens and 1.2M English tokens). We use the designated additional tuning and evaluation sections consisting of 1023, resp. 964 sentences.

In all experiments, word alignment was obtained using the grow-diag-final heuristic for symmetrizing GIZA++ (Och and Ney, 2003) alignments. To reduce data sparseness, the English text was lowercased and Czech was lemmatized for alignment estimation. Language models are based on the target side of the parallel corpus only, unless stated otherwise.

### 3.1 Evaluation Measure and MERT

We evaluate our experiments using the (lowercase, tokenized) BLEU metric and estimate the empirical confidence using the bootstrapping method described in Koehn (2004b).[6] We report the scores obtained on the test section with model parameters tuned using the tuning section for minimum error rate training (MERT, (Och, 2003)).

## 4 Scenarios of Factored Translation English→Czech

We experimented with the following factored translation scenarios.

The baseline scenario (labelled T for translation) is single-factored: input (English) lowercase word forms are directly translated to target (Czech) lowercase forms. A 3-gram language model (or more models based on various corpora) checks the stream of output word forms. The baseline scenario thus corresponds to a plain phrase-based SMT system:

| English | | Czech | |
|---|---|---|---|
| lowercase | ⟶ | lowercase | +LM |
| lemma | | lemma | |
| morphology | | morphology | |

In order to check the output not only for word-level coherence but also for morphological coherence, we add a single generation step: input word forms are first translated to output word forms and each output word form then generates its morphological tag.

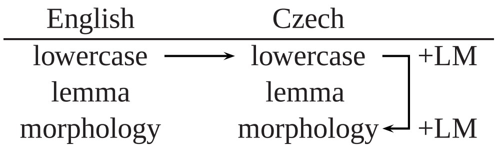
Two types of language models can be used simultaneously: a (3-gram) LM over word forms and a (7-gram) LM over morphological tags.

We used tags with various levels of detail, see section 5. We call this the "T+C" (translate and check) scenario:

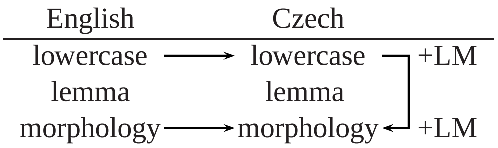
[4] `http://www.statmt.org/wmt07/`

[5] Our preliminary experiments with the Prague Czech-English Dependency Treebank, PCEDT v.1.0 (Čmejrek et al., 2004), 20k sentences, gave similar results, although with a lower level of significance due to a smaller evaluation set.

[6] Given a test set of sentences, we perform 1,000 random selections with repetitions to estimate 1,000 BLEU scores on test sets of the same size. The empirical 90%-confidence upper and lower bounds are obtained after removing top and bottom 5% of scores. For conciseness, we report the average of the distance between to standard BLEU value and the empirical upper and lower bound after the "±" symbol.

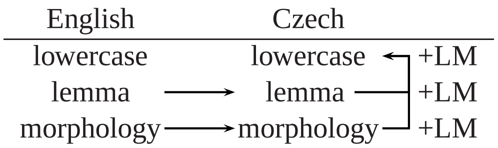


As a refinement of T+C, we also used T+T+C scenario, where the morphological output stream is constructed based on both output word forms and input morphology. This setting should reinforce correct translation of morphological features such as number of source noun phrases. To reduce the risk of early pruning, the generation step operationally precedes the morphology mapping step. Again, two types of language models can be used in this "T+T+C" scenario:



The most complex scenario we used is linguistically appealing: output lemmas (base forms) and morphological tags are generated from input in two independent translation steps and combined in a single generation step to produce output word forms. The input English text was not lemmatized so we used English word forms as the source for producing Czech lemmas.

The "T+T+G" setting allows us to use three types of language models. Trigram models are used for word forms and lemmas and 7-gram language models are used over tags:



### 4.1 Experimental Results: Improved over T

Table 1 summarizes estimated translation quality of the various scenarios. In all cases, a 3-gram LM is used for word forms or lemmas and a 7-gram LM for morphological tags.

The good news is that multi-factored models always outperform the baseline T.

Unfortunately, the more complex multi-factored scenarios do not bring any significant improvement over T+C. Our belief is that this effect is caused by search errors: with multi-factored models, more hypotheses get similar scores and future costs of partial hypotheses might be estimated less reliably. With the limited stack size (not more than 200 hypotheses of the same number of covered input words), the decoder may more often find sub-optimal solutions. Moreover, the more steps are used, the more model weights have to be tuned in the minimum error rate training. Considerably more tuning data might be necessary to tune the weights reliably.

| | BLEU |
|---|---|
| T+T+G | 13.9±0.7 |
| T+T+C | 13.9±0.6 |
| T+C | 13.6±0.6 |
| Baseline: T | 12.9±0.6 |

Table 1: BLEU scores of various translation scenarios.

## 5 Granularity of Czech Part-of-Speech

As stated above, the Czech morphological tag system is very complex: in theory up to 4,000 different tags are possible. In our T+T+C scenario, we experiment with various simplifications of the system to find the best balance between richness and robustness of the statistics available in our corpus. (The more information is retained in the tags, the more severe data sparseness is.)

**Full tags (1200 unique seen in the 56k corpus):** Full Czech positional tags are used. A tag consists of 15 positions, each holding the value of a morphological property (e.g. number, case or gender).[7]

**POS+case (184 unique seen):** We simplify the tag to include only part and subpart of speech (distinguishes also partially e.g. verb tenses). For nouns, pronouns, adjectives and prepositions[8], also the case is included.

**CNG01 (621 unique seen):** CNG01 refines POS. For nouns, pronouns and adjectives we include not only the case but also number and gender.

[7] In principle, each of the 15 positions could be used as a separate factor. The set of necessary generation steps to encode relevant dependencies would have to be carefully determined.

[8] Some Czech prepositions select for a particular case, some are ambiguous. Although the case is never shown on surface of the preposition, the tagset includes this information and Czech taggers are able to infer the case.

**CNG02 (791 unique seen):** Tag for punctuation is refined: the lemma of the punctuation symbol is taken into account; previous models disregarded e.g. the distributional differences between a comma and a question mark. Case, number and gender added to nouns, pronouns, adjectives, prepositions, but also to verbs and numerals (where applicable).

**CNG03 (1017 unique seen):** Optimized tagset:

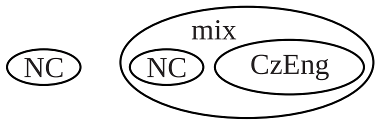
- Tags for nouns, adjectives, pronouns and numerals describe the case, number and gender; the Czech reflexive pronoun *se* or *si* is highlighted by a special flag.
- Tag for verbs describes subpart of speech, number, gender, tense and aspect; the tag includes a special flag if the verb was the auxiliary verb *být (to be)* in any of its forms.
- Tag for prepositions includes the case and also the lemma of the preposition.
- Lemma included for punctuation, particles and interjections.
- Tag for numbers describes the "shape" of the number (all digits are replaced by the digit *5* but number-internal punctuation is kept intact). The tag thus distinguishes between 4- or 5-digit numbers or the precision of floating point numbers.
- Part of speech and subpart of speech for all other words.

### 5.1 Experimental Results: CNG03 Best

Table 2 summarizes the results of T+T+C scenario with varying detail in morphological tag.

| | BLEU |
|---|---|
| Baseline: T (single-factor) | 12.9±0.6 |
| T+T+C, POS+case | 13.2±0.6 |
| T+T+C, CNG01 | 13.4±0.6 |
| T+T+C, CNG02 | 13.5±0.7 |
| T+T+C, full tags | 13.9±0.6 |
| T+T+C, CNG03 | 14.2±0.7 |

Table 2: BLEU scores of various granularities of morphological tags in T+T+C scenario.



$$\text{weighted} = \alpha\text{NC} + \beta\text{mix}$$

| Scenario | Phrases from | LMs | BLEU |
|---|---|---|---|
| T | NC | NC | 12.9±0.6 |
| T | mix | mix | 11.8±0.6 |
| T | mix | weighted | 11.8±0.6 |
| T+C CNG03 | NC | NC | 13.7±0.7 |
| T+C CNG03 | mix | mix | 13.1±0.7 |
| T+C CNG03 | mix | weighted | 13.7±0.7 |
| T+C full tags | NC | NC | 13.6±0.6 |
| T+C full tags | mix | mix | 13.1±0.7 |
| T+C full tags | mix | weighted | 13.8±0.7 |

Figure 1: The effect of additional data in T and T+C scenarios.

Our results confirm improvement over the single-factored baseline. Detailed knowledge of the morphological system also proves its utility: by choosing the most relevant features of tags and lemmas but avoiding sparseness, we can improve on BLEU score by about 0.3 absolute over T+T+C with full tags.

## 6 More Out-of-Domain Data in T and T+C Scenarios

In order to check if the method scales up with more parallel data available, we extend our training data using the CzEng parallel corpus (Bojar and Žabokrtský, 2006). CzEng contains sentence-aligned texts from the European Parliament (about 75%), e-books and stories (15%) and open source documentation. By "Baseline" corpus we denote NC corpus only, by "Large" we denote the combination of training sentences from NC and CzEng (1070k sentences, 13.9M Czech and 15.5 English tokens) where in-domain NC data amounts only to 5.2% sentences.

Figure 1 gives full details of our experiments with the additional data. We varied the scenario (T or T+C), the level of detail in the T+C scenario (full tags vs. CNG03) and the size of the training corpus. We extract phrases from either the in-domain corpus only (NC) or the mixed corpus (mix). We use either one LM per output factor, varying the corpus size (NC or mix), or two LMs per output factors with weights trained independently in the MERT procedure (weighted). Independent weights allow us to take domain difference into account, but we exploit this in the target LM only, not the phrases.

The only significant difference is caused by the scenario: T+C outperforms the baseline T, regardless of corpus size. Other results (insignificantly) indicate the following observations:

- Ignoring the domain difference and using only the mixed domain LM in general performs worse than allowing MERT to optimize LM weights for in-domain and generic data separately.[9]
- CNG03 outperforms full tags only in small data setting, with large data (treating the domain difference properly), full tags perform better.

## 7 Untreated Morphological Errors

The previous sections described improvements gained on small data sets when checking morphological agreement using T+T+C scenario (BLEU raised from 12.9% to 13.9% or up to 14.2% with manually tuned tagset, CNG03). However, the best result achieved is still far below the margin of lemmatized BLEU (21%), as mentioned in Section 1.1.

When we searched for the unexploited morphological errors, visual inspection of MT output suggested that local agreement (within 3-word span) is relatively correct but Verb-Modifier relations are often malformed causing e.g. a bad case for the Modifier. To quantify this observation we performed a micro-study of our best MT output using an intuitive metric. We checked whether Verb-Modifier relations are properly preserved during the translation of 15 sample sentences.

The *source* text of the sample sentences contained 77 Verb-Modifier pairs. Table 3 lists our observations on the two members in each Verb-Modifier pair. We see that only 56% of verbs are translated correctly and 79% of nouns are translated correctly. The system tends to skip verbs quite often (27% of cases).

| **Translation of** | **Verb** | **Modifier** |
|---|---|---|
| . . . preserves meaning | 56% | 79% |
| . . . is disrupted | 14% | 12% |
| . . . is missing | 27% | 1% |
| . . . is unknown (not translated) | 0% | 5% |

Table 3: Analysis of 77 Verb-Modifier pairs in 15 sample sentences.

More importantly, our analysis has shown that even in cases where both the Verb and the Modifier are lexically correct, the relation between them in Czech is either non-grammatical or meaning-disrupted in 56% of these cases. Commented samples of such errors are given in Figure 2 below. The first sample shows that a strong language model can lead to the choice of a grammatical relation that nevertheless does not convey the original meaning. The second sample illustrates a situation where two correct options are available but the system chooses an inappropriate relation, most probably because of backing off to a generic pattern verb-noun$^{accusative}_{plural}$. This pattern is quite common for expressing the object role of many verbs (such as *vydat*, see Correct option 2 in Figure 2), but does not fit well with the verb *vyběhnout*. While the target-side data may be rich enough to learn the generalization vyběhnout–s–*instr*, no such generalization is possible with language models over word forms or morphological tags only. The target side data will be hardly ever rich enough to learn this particular structure in all correct morphological and lexical variants: *vyběhl–s–reklamou, vyběhla–s–reklamami, vyběhl–s–prohlášením, vyběhli–s–oznámením, . . . .* We would need a mixed model that combines verb lemmas, prepositions and case information to properly capture the relations.

Unfortunately, our preliminary experiments that made use of automatic Czech dependency parse trees to construct a factor explicitly highlighting the Verb (lexicalized) its Modifiers (case and the lemma of the preposition, if present) and boundary symbols such as punctuation or conjunctions and using a dummy token for all other words did not bring any improvement over the baseline. A possible reason is that we employed only a standard 7-gram language model to this factor. A more appropriate treatment

[9] In our previous experiments with PCEDT as the domain-specific data, the difference was more apparent because the corpus domains were more distant. In the T scenario reported here, the weighted LMs did not bring any improvement over "mix" and even performed worse than the baseline NC. We attribute this effect to some randomness in the MERT procedure.

is to disregard the dummy tokens in the language model at all and use an n-gram language model that looks at last $n - 1$ non-dummy items.

## 8 Related Research

Class-based LMs (Brown et al., 1992) or factored LMs (Bilmes and Kirchhoff, 2003) are very similar to our T+C scenario. Given the small differences in all T+... scenarios' performance, class-based LM might bring equivalent improvement. Yang and Kirchhoff (2006) have recently documented minor BLEU improvement using factored LMs in single-factored SMT to English. The multi-factored approach to SMT of Moses is however more general.

Many researchers have tried to employ morphology in improving word alignment techniques (e.g. (Popović and Ney, 2004)) or machine translation quality (Nießen and Ney (2001), Koehn and Knight (2003), Zollmann et al. (2006), among others, for various languages; Goldwater and McClosky (2005), Bojar et al. (2006) and Talbot and Osborne (2006) for Czech), however, they focus on translating *from* the highly inflectional language.

Durgar El-Kahlout and Oflazer (2006) report preliminary experiments in English to Turkish single-factored phrase-based translation, gaining significant improvements by splitting root words and their morphemes into a sequence of tokens. In might be interesting to explore multi-factored scenarios for different Turkish morphology representation suggested the paper.

de Gispert et al. (2005) generalize over verb forms and generate phrase translations even for unseen target verb forms. The T+T+G scenario allows a similar extension if the described generation step is replaced by a (probabilistic) morphological generator.

Nguyen and Shimazu (2006) translate from English to Vietnamese but the morphological richness of Vietnamese is comparable to English. In fact the Vietnamese vocabulary size is even smaller than English vocabulary size in one of their corpora. The observed improvement due to explicit modelling of morphology might not scale up beyond small-data setting.

As an alternative option to our verb-modifier experiments, structured language models (Chelba and Jelinek, 1998) might be considered to improve clause coherence, until full-featured syntax-based MT models (Yamada and Knight (2002), Eisner (2003), Chiang (2005) among many others) are tested when translating to morphologically rich languages.

## 9 Conclusion

We experimented with multi-factored phrase-based translation aimed at improving morphological coherence in MT output. We varied the setup of additional factors (translation scenario) and the level of detail in morphological tags. Our results on English-to-Czech translation demonstrate significant improvement in BLEU scores by explicit modelling of morphology and using a separate morphological language model to ensure the coherence. To our knowledge, this is one of the first experiments showing the advantages of using multiple factors in MT.

Verb-modifier errors have been studied and a factor capturing verb-modifier dependencies has been proposed. Unfortunately, this factor has yet to bring any improvement.

## 10 Acknowledgement

The work on this project was partially supported by the grants Collegium Informaticum GAČR 201/05/H014, grants No. ME838 and GA405/06/0589 (PIRE), FP6-IST-5-034291-STP (Euromatrix), and NSF No. 0530118.

## References

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proc. of NAACL 2003*, pages 4–6.

Ondřej Bojar and Zdeněk Žabokrtský. 2006. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62.

Ondřej Bojar, Evgeny Matusov, and Hermann Ney. 2006. Czech-English Phrase-Based Machine Translation. In *Proc. of FinTAL 2006*, pages 214–224, Turku, Finland.

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

Ciprian Chelba and Frederick Jelinek. 1998. Exploiting syntactic structure for language modeling. In *Proc. of ACL 1998*, pages 225–231, San Francisco, California.

| | |
|---|---|
| Input: | Keep on investing. |
| MT output: | Pokračovalo investování. (grammar correct here!) |
| Gloss: | Continued investing. (Meaning: The investing continued.) |
| Correct: | Pokračujte v investování. |

| | | | | |
|---|---|---|---|---|
| Input: | brokerage firms rushed out ads … | | | |
| MT Output: | brokerské | firmy | vyběhl | reklamy |
| Gloss: | brokerage | firms$_{pl.fem}$ | ran$_{sg.masc}$ | ads$^{pl.voc,sg.gen}_{pl.nom,pl.acc}$ |
| Correct option 1: | brokerské | firmy | vyběhly | s reklamami$_{pl.instr}$ |
| Correct option 2: | brokerské | firmy | vydaly | reklamy$_{pl.acc}$ |

Figure 2: Two sample errors in translating Verb-Modifier relation from English to Czech.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of ACL 2005*, pages 263–270.

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proc. of LREC 2004*, Lisbon, Portugal.

Adrià de Gispert, José B. Mariño, and Josep M. Crego. 2005. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Proc. of Eurospeech 2005*, pages 3185–3188, Lisbon, Portugal.

İlknur Durgar El-Kahlout and Kemal Oflazer. 2006. Initial Explorations in English to Turkish Statistical Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation, ACL 2006*, pages 7–14, New York City.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proc. of ACL 2003, Companion Volume*, pages 205–208, Sapporo, Japan.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proc. of HLT/EMNLP 2005*, pages 676–683.

Jan Hajič and Barbora Hladká. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proc. of COLING/ACL 1998*, pages 483–490, Montreal, Canada.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proc. of EACL 2003*, pages 187–193.

Philipp Koehn. 2004a. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proc. of AMTA 2004*, pages 115–124.

Philipp Koehn. 2004b. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of EMNLP 2004*, Barcelona, Spain.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit X*.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

T.P. Nguyen and A. Shimazu. 2006. Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation. In *Proc. of AMTA 2006*, pages 138–147.

Sonja Nießen and Hermann Ney. 2001. Toward hierarchical models for statistical machine translation of inflected languages. In *Proc. of Workshop on Data-driven methods in machine translation, ACL 2001*, pages 1–8.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL 2003*, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL 2002*, pages 311–318.

M. Popović and H. Ney. 2004. Improving Word Alignment Quality using Morpho-Syntactic Information. In *Proc. of COLING 2004*, Geneva, Switzerland.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proc. of EMNLP 1996*, Philadelphia, USA.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proc. of COLING and ACL 2006*, pages 969–976, Sydney, Australia.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proc. of ACL 2002*, pages 303–310.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proc. of EACL 2006*.

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. 2006. Bridging the inflection morphology gap for arabic statistical machine translation. In *Proc. of HLT/NAACL*.