Probes in a Taxonomy of Factored Phrase-Based Models *

Ondřej Bojar, Bushra Jawaid, Amir Kamran
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Praha 1, CZ-118 00, Czech Republic
{bojar,jawaid,kamran}@ufal.mff.cuni.cz

Abstract

We introduce a taxonomy of factored phrase-based translation scenarios and conduct a range of experiments in this taxonomy. We point out several common pitfalls when designing factored setups. The paper also describes our WMT12 submissions CU-BOJAR and CU-POOR-COMB.

1 Introduction

Koehn and Hoang (2007) introduced "factors" to phrase-based MT to explicitly capture arbitrary features in the phrase-based model. In essence, input and output tokens are no longer atomic units but rather vectors of atomic values encoding e.g. the lexical and morphological information separately. Factored translation has been successfully applied to many language pairs and with diverse types of information encoded in the additional factors, i.a. (Bojar, 2007; Avramidis and Koehn, 2008; Stymne, 2008; Badr et al., 2008; Ramanathan et al., 2009; Koehn et al., 2010; Yeniterzi and Oflazer, 2010). On the other hand, it happens quite frequently, that the factored setup causes a loss compared to the phrase-based baseline. The underlying reason is the complexity of the search space which gets boosted when the model explicitly includes detailed information, see e.g. Bojar and Kos (2010) or Toutanova et al. (2008).

* This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic) and the Czech Science Foundation grants P406/11/1499 and P406/10/P259. We are grateful for reviewers' comments but we have to obey the 6 page limit. Thanks also to Aleš Tamchyna for supplementary material on MERT.

Number of Translation Steps	Number of Independent Searches	Structure of Searches	Nickname
One	One	–	Direct
Several	One	–	Single-Step
	Several	Serial	Two-Step
		Complex	Complex

Figure 1: A taxonomy of factored phrase-based models.

In this paper, we first provide a taxonomy of (phrase-based) translation setups and then we examine a range of sample configurations in this taxonomy. We don't state universal rules, because the applicability of each of the setups depends very much on the particular language pair, text domain and amount of data available, but we hope to draw attention to relevant design decisions.

The paper also serves as the description of our WMT12 submissions CU-BOJAR and CU-POOR-COMB between English and Czech.

2 A Taxonomy of Factored P-B Models

Figure 1 suggests a taxonomy of various Moses setups. Following the definitions of Koehn and Hoang (2007), a *search* consists of several *translation* and *generation* steps: translation steps map source factors to target factors and generation steps produce target factors from other target factors.

The taxonomy is vaguely linked to the types of problems that can be expected with a given configuration. Direct translation is likely to suffer from out-of-vocabulary issues (due to insufficient generalization) on either side. Single-step scenarios have

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 253–260,
Montréal, Canada, June 7-8, 2012.

a very high risk of combinatorial explosion of translation options (think cartesian product of all target side factors) and/or of spurious ambiguity (several derivations leading to the same output). Such added ambiguity can lead to n-best lists with way fewer unique items than the given n, which in turn renders MERT unstable, see also Bojar and Tamchyna (2011). Serially connected setups (two as our Two-Step or more) can lose relevant candidates between the searches, unless some ambiguous representation like lattices is passed between the steps.

An independent axis on which Moses setups can be organized consists of the number and function of factors on the source and the target side.

We use a very succint notation for the setups except the "complex" one: *t***X**-**Y** denotes a translation step between the factors **X** in the source language and **Y** in the target language. Generation steps are denoted with *g***Y**-**Z**, where both **Y** and **Z** are target-side factors. Individual mapping steps are combined with a plus, while individual source or target factors are combined with an "*a*".

As a simple example, *tF-F* denotes the direct translation from source form (*F*) to the target form. A linguistically motivated scenario with one search can be written as *tL-L+tT-T+gLaT-F*: translate (1) the lemma (L) to lemma, (2) the morphological tag (T) to tag independently and (3) finally generate the target form from the lemma and the tag.

We use two more operators: "*:*" delimits alternative decoding paths (Birch et al., 2007) used within one search and "=" delimits two independent searches. A plausible setup is e.g. *tF-LaT=tLaT-F:tL-F* motivated as follows: the source word form is translated to the lemma and tag in the target language. Then a second search (whose translation tables can be trained on larger monolingual data) consists of two alternative decoding paths: either the pair of *L* and *T* is translated into the target form, or as a fallback, the tag is disregarded and the target form is guessed only from the lemma (and the context as scored by the language model). The example also illustrated the priorities of the operators.

3 Common Settings

Throughout the experiments, we use the Moses toolkit (Koehn et al., 2007) and GIZA++ (Och and Ney, 2000). The texts were processed using the Treex platform (Popel and Žabokrtský, 2010)[1], which included lemmatization and tagging by Morce (Spoustová et al., 2007). After the tagging, we tokenized further so words like "23-year" or "Aktualne.cz" became three tokens.

Dataset	Sents (cs/en)	Toks (cs/en)	Source
Small	197k parallel	4.2M/4.8M	CzEng 1.0 news
Large	14.8M parallel	205M/236M	CzEng 1.0 all
Mono	18M/50M	317M/1.265G	WMT12 mono

Table 1: Summary of training data.

Decoding Path	Language Models	BLEU
tF-FaLaT	form + lemma + tag	13.05±0.44
tF-FaT	form + tag	13.01±0.44
tF-FaLaT	form + tag	12.99±0.44
tF-F (baseline)	form	12.42±0.44
tF-FaT	form	12.19±0.44
tF-FaLaT	form	12.08±0.45

Table 2: Direct en→cs translation (a single search with one translation step only).

Our training data is summarized in Table 1.[2]

In most experiments reported here, we use the Small dataset only. The language model (LM) for these experiments is a 5-gram one based on the target-side of Small only.

Our WMT12 submissions are based on the Large and Mono data. The language model for the large experiments uses 6-grams of forms and optionally 8-grams of morphological tags. As in previous years, the language models are interpolated (towards the best cross entropy on WMT08 dataset) from domain-specific LMs, e.g. czeng-news, czeng-techdoc, wmtmono-2011, wmtmono-2012.

Except where stated otherwise, we tune on the official WMT10 test set and report BLEU (Papineni et al., 2002) scores on the WMT11 test set.

4 Direct Setups

Table 2 lists our experiments with direct translation, various factors and language models in our notation.

[1] `http://ufal.mff.cuni.cz/treex/`

[2] We did not include the parallel en-cs data made available by the WMT12 organizers. This probably explains our loss compared to UEDIN but allows a direct comparison with CU TECTOMT, a deep syntactic MT based on the same data.

Decoding Paths	LMs	Avg. BLEU	Eff. Nbl. Size
tL-L+tT-T+gLaT-F:tF-FaLaT	F + L + T	13.31±0.06	12.24±1.33
tL-L+tT-T+gLaT-F	F + L + T	13.30±0.05	40.33±3.82
tL-L+tT-T+gLaT-F	F + T	13.17±0.01	39.91±2.58
tL-L+tT-T+gLaT-F:tF-FaLaT, 200-best-list	F + L + T	13.15±0.24	20.47±5.63
tF-FaLaT	F + L + T	13.13±0.06	34.28±3.08
tL-L+tT-T+gLaT-F:tF-FaLaT	L + T	13.09±0.06	16.65±1.07
tF-FaT	F + T	13.08±0.05	39.67±2.21
tL-L+tT-T+gLaT-F:tF-FaT	F + T	13.01±0.43	14.87±5.04
tF-F (baseline)	F	12.38±0.03	43.13±0.48
tL-L+tT-T+gLaT-F:tF-F	F	12.30±0.03	17.83±3.27

Table 3: Results of three MERT runs of several single-step configurations.

Explicit modelling of target-side morphology improves translation quality, compare tF-FaLaT with the baseline tF-F. However, two results document that if some detailed information is distinguished in the output, it introduces target ambiguity and leads to a loss in BLEU, unless the detailed information is actually used in the language model: (1) tF-FaLaT with LM on forms is worse than the baseline tF-F but tF-FaLaT with all the three language models is better, (2) tF-FaLaT with two LMs (forms and tags) is negligibly worse than tF-FaT with the same language models.

5 Single-Step Experiments

Single-step scenarios consist of more than one translation steps within a single search. We do not distinguish whether all the translation steps belong to the same decoding path or to alternative decoding paths.

Table 3 lists several single-step configurations (and three direct translations for a comparison). The single-step configurations always include the linguistically-motivated tL-L+tT-T+gLaT-F with varying language models and optionally with an alternative decoding path to serve as the fallback.

Aware of the low stability of MERT (Clark et al., 2011), we run MERT three times and report the average BLEU score including the standard deviation.

The last column in Table 3 lists the average number of *distinct* candidates per sentence in the n-best lists during MERT, dubbed "effective n-best list size". Unless stated otherwise, we used 100-best lists. We see that due to spurious ambiguity, e.g. various segmentations of the input into phrases, the effective size does not reach even a half of the limit.

We make three observations here:

(1) In this small data setting with a very morphologically rich language, the complex setup tL-L+tT-T+gLaT-F does not even need the alternative decoding path tF-F. Ramanathan et al. (2009) report gains in English-to-Hindi translation and also probably do not use alternative decoding paths.

(2) Reducing the range of language models used leads to worse scores, which is in line with the observation made with direct setups. We are surprised by the relative importance of the lemma-based LM.

(3) Alternative decoding paths significantly reduce effective n-best list size to just 12–18 unique candidates per sentence. However, we don't see an obvious relation to the stability of MERT: the standard deviations of BLEU average are very similar except for two outliers: 13.15±0.24 and 13.01±0.43. One of the outliers, 13.15, is actually a repeated run of the 13.31 with n-best-list size set to 200. Here we see a slight increase in the effective size (20 instead of 12) but also a slight loss in BLEU. We repeated the 13.31 experiment also with $n \in \{300, 400, 500, 600\}$, three MERT runs for each n. All the runs reached BLEU of about 13.30 except for one ($n = 600$) where the score dropped to 11.50. The low result was obtained when MERT ended at 25 iterations, the standard limit. On the other hand, several successful runs also exhausted the limit.

Figure 2 plots the BLEU scores in the 25 iterations of the underperforming run with $n = 600$. The MERT implementation in the Moses toolkit reports at each iteration what we call "predicted BLEU", i.e. the BLEU of translations selected by the current

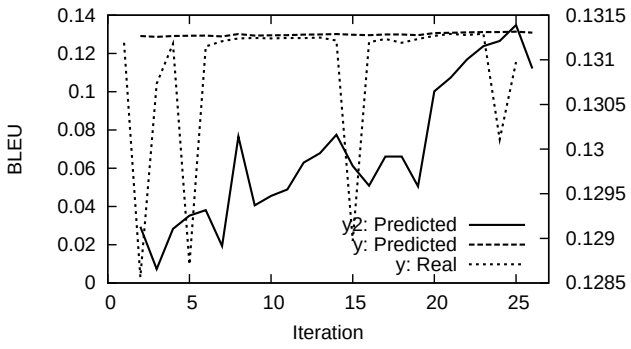

Figure 2: Predicted and real devset BLEU scores.

weight settings *from the (accumulated) n-best list*. We plot this predicted BLEU twice: once on the y2 axis alone and for the second time on the primary y axis together with the real BLEU, i.e. the BLEU of the dev set when Moses is actually run with the weight settings. The real BLEU drops several times, indicating that the prediction was misleading. Similar drops were observed in all runs. With bad luck as here, the iteration limit is reached when the optimization is still recovering from such a drop.

To avoid such a pitfall, one should check the real BLEU and continue or simply rerun the optimization if the iteration limit was reached.

6 Two-Step Experiments

The linguistically motivated setups used in the previous sections are prohibitively expensive for large data, see also Bojar et al. (2009). A number of researchers have thus tried diving the complexity of search into two independent phases: (1) translation and reordering, and (2) conjugation and declination. The most promising results were obtained with the second step predicting individual morphological features using a specialized tool (Toutanova et al., 2008; Fraser et al., 2012). Here, we simply use one more Moses search as Bojar and Kos (2010).

In the first step, source English gets translated to a simplified Czech and in the second step, the simplified Czech gets fully inflected.

6.1 Factors in Two-Step Setups

Two-step setups can use factors in the source, middle or the target language. We experiment with factors only in the middle language (affecting both the first and the second search) and use only the form in both source and target sides.

In the middle language, we experiment with one or two factors. For presentation purposes, we always speak about two factors: "LOF" ("lemma or form", i.e. a representation of the lexical information) and "MOT" ("modified tag", i.e. representing the morphological properties). In the single-factor experiments the LOF and MOT are simply concatenated into a token in the shape LOF+MOT.

Figure 3 illustrates the range of LOFs and MOTs we experimented with. LOF_0 and MOT_0 are identical to the standard Czech lemma and morphological tag as used e.g. in the Prague Dependency Treebank (Hajič et al., 2006).

LOF_1 and MOT_1 together make what Bojar and Kos (2010) call "pluslemma". MOT_1 is less complex than the full tag by disregarding morphological attributes not generally overt in the English source side. For most words, LOF_1 is simply the lemma, but for frequent words, the full form is used. This includes punctuation, pronouns and the verbs "býti" (to be) and "mít" (to have).

MOT_2 uses a more coarse grained part of speech (POS) than MOT_1. Depending on the POS, different attributes are included: gender and number for nouns, pronouns, adjectives and verbs; case for nouns, pronouns, adjectives and prepositions; negation for nouns and adjectives; tense and voice for verbs and finally grade for adjectives. The remaining grammatical categories are encoded using POS, number, grade and negation.

6.2 Decoding Paths in Two-Step Setups

Each of the searches in the two-step setup can be as complex as the various single-step configurations. We test just one decoding path for the one or two factors in the middle language.

All experiments with one middle factor (i.e. "+") follow this config: tF-LOF+MOT = tLOF+MOT-F, i.e. two direct translations where the first one produces the concatenated LOF and MOT tokens and the second one consumes them. The first step uses a 5-gram LOF+MOT language model and the second step uses a 5-gram LM based on forms.

This setup has the capacity to improve translation quality by producing forms of words never seen aligned with a given source form. For example the English word *green* would be needed in the parallel

Word Form	LOF_0	LOF_1	MOT_0	MOT_1	MOT_2	Gloss
lidé	člověk	člověk	NNMP1-----A---1	NPA-	NMP1-A	people
by	být	by	Vc----------	c---	V-----	would
neočekávali	očekávat	očekávat	VpMP---XR-NA---	pPN-	VMP-RA	expect

Figure 3: Examples of LOFs and MOTs used in our experiments.

Middle Factors	1 +	2 |
LOF_0 +/| MOT_0	11.11±0.48	12.42±0.48
LOF_1 +/| MOT_1	12.10±0.48	11.85±0.42
LOF_1 +/| MOT_2	11.87±0.51	12.47±0.51

Table 4: Two-step experiments.

data with all the morphological variants of the Czech word *zelený*. Adding the middle step with appropriately reduced morphological information so that only features overt in the source are represented in the middle tokens (e.g. negation and number but not the case) allows the model to find the necessary form anywhere in the target-side data only:

$$\textit{green} \rightarrow \textit{zelený+NSA-} \rightarrow \begin{cases} \textit{zeleného} \text{ (genitive)} \\ \textit{zelenému} \text{ (dative)} \\ \ldots \end{cases}$$

The experiments with two middle factors (i.e. "|") use this path: tF-LOFaMOT = tLOFaMOT-F:LOF-F. The first step is identical, except that now we use two separate LMs, one for LOFs and one for MOTs. The second step has two alternative decoding paths: (1) as before, producing the form from both the LOF and the MOT, and (2) ignoring the morphological features from the source altogether and using just target-side context to choose an appropriate form of the word. This setup is capable of sacrificing adequacy for a more fluent output.

6.3 Experiments with Two-Step Setups

Table 4 reports the BLEU scores when changing the number of factors ("+" vs. "|") in the middle language and the type of the LOF and MOT.

We see an interesting difference between MOT_1 and $MOT_{0\text{ or }2}$. The more fine-grained $MOT_{0\text{ or }2}$ work better in the two-factor "|" setup that allows to disregard the MOT, while MOT_1 works better in the direct translation "+".

Overall, we see no improvement over the tF-F baseline (BLEU of 12.42) and this is mainly due to to the fact that we used Small data in both steps.

7 A Complex Moses Setup

Obviously, many setups fall under the "complex" category of our taxonomy, including also some system combination approaches. We tried to combine three Moses systems: (1) CU-BOJAR as described below, (2) same setup like CU-BOJAR but optimized towards 1-TER (Snover et al., 2006), and (3) a large-data two-step setup.[3] The system combination is performed using a fourth Moses search that gets a lattice (Dyer et al., 2008) of individual systems' outputs, performs an identity translation and scores the candidates by language models and other features. The lattice is created from the individual system outputs in the ROVER style (Matusov et al., 2008) utilizing the source-to-hypothesis word alignments as produced by the individual systems. We use our simple implementation for constructing the confusion networks and converting them to the lattices. The "combination Moses" was tuned on the WMT11 test set towards BLEU. The resulting system is called CU-POOR-COMB, because we felt it underperformed the individual systems not only in BLEU but also in an informal subjective evaluation.

Surprisingly, CU-POOR-COMB won the WMT12 automatic evaluation in TER. In the retrospect, this is caused by TER overemphasizing word-level precision. CU-POOR-COMB skipped words not confirmed by several systems and its hypotheses are shorter (18.1 toks/sent) than those by CU-BOJAR (20.1 toks/sents) or the reference (21.9 toks/sent). A quick manual inspection of 32 sentences suggests that about one third or quarter of CU-POOR-COMB suffer from some information loss whereas the rest are acceptable or even better paraphrases. Prelim-

[3]The large two-step setup is identical to the one by (Bojar and Kos, 2010), except that we use only the current Large and Mono datasets as described in Section 3.

		Our Scoring				matrix.statmt.org	
	Test Set	newstest-2011		newstest-2012			
	Metric	BLEU	TER*100	BLEU	TER*100	BLEU	TER
→cs	CU-POOR-COMB	–used–for–	–tuning–	14.17±0.53	**64.07±0.53**	14.0	**0.741**
	CU-BOJAR (tFaT-FaT, lex. r.)	**18.10±0.55**	62.84±0.71	**16.07±0.55**	65.52±0.59	**15.9**	0.759
	As ↑ but towards 1-TER	16.10±0.54	**61.64±0.59**	14.13±0.54	64.28±0.55	–	–
	Large Two-Step	17.34±0.57	63.47±0.66	15.37±0.54	65.85±0.57	–	–
	Unused (tFaT-FaT, dist. reord.)	18.07±0.56	62.74±0.70	15.92±0.57	65.50±0.60	–	–
	Unused (tF-FaT, dist. reord.)	17.85±0.58	63.13±0.68	15.73±0.55	65.85±0.58	–	–
	Unused (tF-F, lex. reord.)	17.73±0.58	63.04±0.68	15.61±0.57	65.76±0.58	–	–
	Unused (tFaT-F, dist. reord.)	17.62±0.56	62.97±0.70	15.33±0.58	65.70±0.59	–	–
	Unused (tF-F, dist. reord.)	17.51±0.57	63.32±0.69	15.48±0.56	65.79±0.58	–	–
→en	CU-BOJAR (tF-F:tL-F, dist. reord.)	**24.65±0.60**	**58.54±0.66**	**23.09±0.59**	**61.24±0.68**	**21.5**	**0.726**
	Unused (tF-F, dist. reord.)	24.62±0.59	58.66±0.66	22.90±0.56	61.63±0.67	–	–

Table 5: Summary of large data runs and systems submitted to WMT12 manual evaluation. The upper part lists the two submissions in en→cs translation and two more systems used in CU-POOR-COMB. The lower part of the table shows the scores for CU-BOJAR when translating to English. All systems reported here use the Large and Mono data.

inary results of WMT 12 manual ranking indicate that overall, our system combination performs poor.

8 Overview of Systems Submitted

Table 5 summarizes the scores for our two system submissions. We report the scores in our tokenization on the official test sets of WMT11 and WMT12 and also the scores as measured by `http://matrix.statmt.org`. Note that for the latter, we use the detokenized outputs processed by the recommended normalization script.[4]

8.1 Details of CU-BOJAR for en→cs

We deliberately used only direct setups for the large data and due to time constraints, we ran just a few configurations, see Table 5.

We knew from previous years that including English (source) POS tag improves overall target sentence structure: English words are often ambiguous between noun and verb, so without the POS information, verbs got often translated as nouns, rendering the sentence incomprehensible. Tagging and including the source tag helps, as confirmed by the tFaT-F setup being somewhat better than tF-F.

We also knew that target-side tag LM is helpful (esp. if we can afford up to 8-grams in the LM). This was confirmed by tF-FaT being better than tF-F. Ultimately, we use tags on both sides: tFaT-FaT and get the best scores. This confirms that our parallel data is sufficiently large so that even the added sparsity due to tags does not cause any trouble.

A little gain comes from a lexicalized reordering model (or-bi-fe) based on word forms, see CU-BOJAR reaching 18.10 BLEU on WMT11 test set.

8.2 Details of CU-BOJAR for cs→en

For the translation into English, we tested just two setups: tF-F and tF-F:tL-T. The latter setup falls back to the Czech lemma, if the exact form is not available. The gain is only small, because our parallel data is already quite large.

9 Conclusion

We introduced a simple taxonomy of factored phrase-based setups and conducted several probes for English→Czech translation. We gained small improvements in both small and large data settings.

We also warned about some common pitfalls: (1) all target-side factors should be accompanied with a language model to compensate for the added sparseness, (2) alternative decoding paths significantly reduce the effective n-best list size, and (3) the infamous instability of MERT can be caused by bad luck at exhausted iteration limit.

On a general note, we learnt that a breadth-first search for best configurations should be automated as much as possible so that more human effort can be invested into analysis.

[4] `http://www.statmt.org/wmt11/normalize-punctuation.perl`

References

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

Ibrahim Badr, Rabih Zbib, and James Glass. 2008. Segmentation for english-to-arabic statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 153–156, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

Ondřej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. 2009. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL (Short Papers)*, pages 176–181. The Association for Computer Linguistics.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *Proc. of EACL 2012*. Association for Computational Linguistics.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proc. of EMNLP*.

Philipp Koehn, 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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, 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, June. Association for Computational Linguistics.

Philipp Koehn, Barry Haddow, Philip Williams, and Hieu Hoang. 2010. More linguistic annotation for statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–

304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: addressing the crux of the fluency problem in english-hindi smt. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 800–808, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings AMTA*, pages 223–231, August.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha.

Sara Stymne. 2008. German Compounds in Factored Statistical Machine Translation. In Bengt Nordstrm and Aarne Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 464–475. Springer Berlin / Heidelberg.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June. Association for Computational Linguistics.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 454–464, Uppsala, Sweden, July. Association for Computational Linguistics.