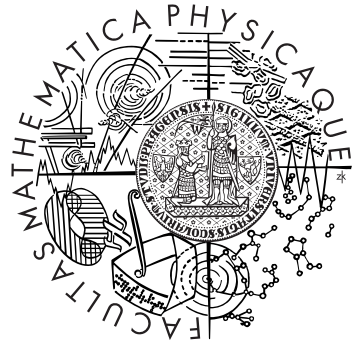
English-to-Czech MT: Large Data and Beyond

Ondřej Bojar

Habilitation Thesis



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, 2017

Contents

Chapter 1

Introduction

This habilitation thesis consists of 12 publications authored or co-authored by Ondřej Bojar. The publications were selected and organized to highlight the author's contribution to the state of the art in machine translation (MT), particularly translation into morphologically rich languages like Czech.

The thesis is structured as follows. Chapter 2 serves as a very brief overview of the task of machine translation, highlighting the core problems that have to be tackled and setting the context for the author's contributions detailed in the rest of this text.

Chapter 3 starts with a quick summary of the author's efforts devoted to the collection and preparation of training data. What may seem a somewhat boring product is nevertheless a valuable resource for many researchers and a critical component necessary to achieve the state of the art in translation quality, as discussed in the following chapters.

Chapter 4 covers the first of the three main contributions of the author: **improving grammaticality, and particularly morphological coherence**, in phrase-based machine translation. While large data are essential for attaining good performance in machine translation, it is not conceivable to collect corpora large enough to cover all possible word forms and provide sufficiently dense statistics about their usage in all possible contexts. Targeting languages with highly productive morphological systems such as Czech thus requires some form of explicit handling of morphology and this chapter summarizes the author's research in this area.

Chapter 5 is focused on the second main contribution, namely **employing deeper linguistic information** to improve translation quality. While statistical methods have had a great success in machine translation, the nature of the handled subject, natural text, belongs to the field of linguistics, and it is therefore interesting to examine to what extent can statistical approaches to MT benefit from linguistic knowledge. The chapter explains the problems faced when trying to organize the statistical models along the linguistic structure of the sentence and describes the author's proposed method that circumvents these problems. The resulting system Chimera outperformed all other MT systems participating in the English-to-Czech news translation task in the years 2013–2015, including Google Translate and other commercial and on-line systems. The setup of Chimera is naturally not limited to translating news text, and adapted

versions of the system served in applied EU projects (QTLeap, HimL) as well as in commercial collaboration of the author's department with IBM.

Finally, evaluation is critical in all applied sciences and evaluating machine translation is particularly intriguing. Chapter 6 is devoted to the third main area of the author's contributions, namely to methods of **manual and automatic MT evaluation**, explaining why MT evaluation is a difficult discipline, revealing the reasons of low performance of an established automatic evaluation measure and proposing modifications to improve the correlation with human judgement.

The last Chapter 7 summarizes the author's service to the community through his contribution to the organization of shared tasks related to machine translation.

The thesis is concluded in Chapter 8. Key papers (co-)authored by Ondřej Bojar and cited throughout the text are reprinted in Appendix A.

Chapter 2

Problems and Solutions in Machine Translation

The goal of machine translation is to translate text from one natural language to another. Machine translation is sometimes dubbed as the "king discipline" of computational linguistics, because translation easily entails almost all aspect of natural language and its meaning: from meaning ambiguity and the relation between the form of an expression and its function in the communication to complex rules of grammatical correctness.

Despite the complexity of language phenomena involved, machine translation has been very successfully tackled by **statistical methods** even in their relatively simple form.

In statistical machine translation, an approach prevalent since 1990s (Brown *et al.*, 1990, 1993; Berger *et al.*, 1994), we search for the most likely target sentence $\hat{e}_1^{\hat{I}}$ (a sequence of target words $\hat{e}_1, \ldots, \hat{e}_{\hat{I}}$) given the source sentence f_1^J:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \tag{2.1}$$

The parameters of the probabilistic distribution $p(e_1^I | f_1^J)$ are estimated automatically from parallel corpora (texts translated previously by humans), subject to various simplifying assumptions.

One of these assumptions, still mainly followed today and reflected also in Eq. 2.1, is that sentences are translated individually, ignoring any contextual information beyond sentence boundaries.

Another critical assumption is that the sentence can be decomposed into a small finite number of translation units which are then translated more or less independently of each other. This assumption has been removed only very recently through the adoption of deep learning methods (neural machine translation, see Section 2.2.3 below). Since the nature of neural MT is also statistical, we will use the qualifier "classical" statistical methods to denote approaches that rely on the decomposition into separate translation units. We will however follow the common usage of abbreviations and use SMT to denote *classical* statistical MT only.

In SMT, additional model components are used to compensate for the independence assumption of translation units and ensure overall coherence of the sentence.

The first step in the classical SMT derivation is to use the Bayes' law and decompose the probability into two components, the **translation model** $p(f_J^I|e_1^I)$ and the **language model** $p(e_1^I)$:

$$p(e_1^I|f_1^J) = \frac{p(f_J^I|e_1^I)p(e_1^I)}{p(f_1^J)} \tag{2.2}$$

Bayes' law reverses the conditional probability in the translation model, but this does not pose any problem: translational equivalence is usually understood as bidirectional and the reversed probability is going to be estimated from the same type of data, parallel texts, anyway.

Furthermore, the denominator is constant in the maximization, so under argmax, we can write:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}}\, p(e_1^I|f_1^J) = \underset{I,e_1^I}{\operatorname{argmax}}\, p(f_J^I|e_1^I)p(e_1^I) \tag{2.3}$$

Eq. 2.3 is called the **noisy channel model** (Brown *et al.*, 1990). Since Och and Ney (2002), the common formal device used in SMT is the more flexible **log-linear model**: The conditional probability of e_1^I being the translation of f_1^J is modelled as a combination of independent feature functions $h_1(\cdot,\cdot),\ldots,h_M(\cdot,\cdot)$ describing the relation of the source and target sentences:

$$p(e_1^I|f_1^J) = \frac{\exp(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^{I'}_1} \exp(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J))} \tag{2.4}$$

Similarly to the noisy channel model (which is in fact a special case of the log-linear model), the denominator in Eq. 2.4 depends on the source sentence f_1^J only and does not affect the selection of the maximum, and neither does the exponential, giving us a simplified formula:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}}\, p(e_1^I|f_1^J) = \underset{I,e_1^I}{\operatorname{argmax}} \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \tag{2.5}$$

The assumption of translation units is formally reflected by defining a joint segmentation s_1^K of the source sentence and the target candidate into K translation units. The majority of features $h_m(\cdot,\cdot)$ are required to decompose along the segmentation, i.e., to take the form:

$$h_m(e_1^I, f_1^J, s_1^K) = \sum_{k=1}^K \tilde{h}_m(\tilde{e}_k, \tilde{f}_k) \tag{2.6}$$

where $\tilde{f}_k$ represents the source side of the translation unit and $\tilde{e}_k$ represents its target side given the segmentation s_1^K.

Feature functions that decompose along this joint segmentation are called **local** and other feature functions are called **non-local**. To distinguish them, we can divide the sum over model components into two parts: M_L local and M_N non-local features:

$$\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) = \sum_{m_L=1}^{M_L} \lambda_{m_L} \sum_{k=1}^{K} \tilde{h}_{m_L}(\tilde{e}_k, \tilde{f}_k) + \sum_{m_N=1}^{M_N} \lambda_{m_N} h_{m_N}(e_1^I, f_1^J) \tag{2.7}$$

Ideally, the segmentation s_1^K should be treated as a hidden parameter and summed over in the maximization in Eq. 2.1. This would be too complicated and too expensive, so it practice, we search for the best derivation, i.e., the pair of segmentation $\hat{s}_1^{\hat{K}}$ and translation $\hat{e}_1^{\hat{I}}$:

$$\begin{aligned}
\hat{e}_1^{\hat{I}}, \hat{s}_1^{\hat{K}} &= \operatorname*{argmax}_{I, e_1^I, K, s_1^K} p(e_1^I | f_1^J) \\
&= \operatorname*{argmax}_{I, e_1^I, K, s_1^K} \sum_{m_L=1}^{M_L} \lambda_{m_L} \sum_{k=1}^{K} \tilde{h}_{m_L}(\tilde{e}_k, \tilde{f}_k) + \sum_{m_N=1}^{M_N} \lambda_{m_N} h_{m_N}(e_1^I, f_1^J) \\
&= \operatorname*{argmax}_{I, e_1^I, K, s_1^K} \sum_{k=1}^{K} \sum_{m_L=1}^{M_L} \lambda_{m_L} \tilde{h}_{m_L}(\tilde{e}_k, \tilde{f}_k) + \sum_{m_N=1}^{M_N} \lambda_{m_N} h_{m_N}(e_1^I, f_1^J)
\end{aligned} \tag{2.8}$$

The component weights λ_m are most commonly optimized with respect to the final translation quality measure. Traditionally, this process is called "tuning" or "model optimization".

2.1 Problems of Machine Translation

Machine translation is a challenging task for several reasons. Adopting the classical statistical MT strategy, we have to choose adequate translation units first and be able to effectively gather them from training data. Then, SMT has to consider a very large search space of possible outputs. And finally, identifying which possible outputs are good and which are bad is difficult.

Defining Translation Units As mentioned above, individual sentences of natural languages are rather complex and up until very recently, they were always decomposed into some smaller units, translating each of these units more or less independently. The various definitions of the units gave rise to word-based (Brown *et al.*, 1990, 1993), phrase-based (PBMT, Koehn *et al.*, 2003) or various

arts of syntax-based (Yamada and Knight, 2001; Zollmann and Venugopal, 2006; Chiang, 2010; Bojar and Hajič, 2008) statistical machine translation.

The choice of a translation unit affects the difficulty in obtaining the "translation dictionary" of these units and the difficulty in decomposing sentences into these units and putting them back together to form the translated sentence.

Shallow units like individual word forms or short sequences of word forms ("phrases" in phrase-based MT, see Section 4.1) are easier to obtain but we very often risk producing a grammatically incorrect output when combining them. Linguistically more adequate units, e.g., some deep-syntactic nodes or treelets, rely on tools for sentence analysis and generation and suffer from their errors.

Larger units (e.g., longer phrases in phrase-based MT) can cover the necessary linguistic dependencies within a single unit, thereby preventing errors at unit combination, but they are obviously much harder to observe in sufficient numbers.

More coarse-grained units such as base forms (lemmas) of words are less prone to data sparsity issues but they imply some information loss which can easily cause a harm to the meaning of the sentence and they are again harder to use correctly.

Managing Huge Search Space As shown already by Knight (1999), picking the right word order and covering source multi-word translation units with entries from translation dictionary are two sub-tasks that render machine translation NP-complete.

When we work with two languages, we can treat target language words as the repertoire of possible "meanings" of source words. It is easy to notice the ambiguity of expressions and its multiplicative effect whenever more occur in a sentence in striking examples like *The plant is next to the bank.* (The *plant* can be a flower or a factory, the *bank* can be a financial institution or a river bank.)

In practice, the number of options to choose from is actually much higher for two main reasons: (1) the input can be often segmented into translation units in many possible ways, and (2) automatically extracted "translation dictionaries" offer many more possible translations (as observed in the translated data) than one would expect. Bojar (2015)[1] reviews how various problems of MT get worse due to morphological richness of languages, including this type of ambiguity: i.e., the translation system has to choose not only the right word but also its morphological form to indicate its relationship to other words in the sentence (e.g., agreement) or to refine its meaning (e.g., plural).

Assessing Translation Quality Given the large space of possible translations, we would need a reliable method for distinguishing good and bad translations. This enterprise is called "machine translation evaluation" (if a reference

[1]See page 63 for the full reference and link to Bojar (2015).

translation is available) or "quality estimation" (if we do not have the reference translation) and it is as old and as complex as MT itself.

Not very surprisingly, small changes in the sentence can drastically change its meaning (e.g., reversing the negation). At the same time, a very different wording can convey the same meaning as the original but we are usually given just one reference translation.

2.2 Complementary Solutions

The history of SMT, see Bojar (2012) or Koehn (2009) for a summary, has seen many complementary methods addressing various aspects of the core problems outlined above. Here we highlight those related our contributions as detailed in the subsequent chapters.

2.2.1 Using Large Data

The success of statistical MT relies on the access to large training data. In fact, some of the problems of MT outlined above lose in their severity as the training data grow. With very large data, we can afford using larger translation units (e.g., longer and longer phrases in phrase-based translation) when covering the input and the phrase-independence assumption will have fewer occasions to do any harm. In the ideal case (which indeed does happen in small and repetitive domains), the whole sentences will be available for reuse.

Precisely for that reason, we have put considerable efforts into collecting large Czech-English parallel data, see Chapter 3.

2.2.2 Adding Linguistic Information

Common approaches to SMT often lack sufficient generalization power and violate many linguistic constraints. For instance, pure phrase-based MT can only produce forms of words as seen in the training data and it has no means to capture the overall sentence structure.

It is therefore interesting to add linguistic knowledge explicitly to the model. In our work, we followed the layered formal description of sentences in natural language defined by the Functional Generative Description (Sgall *et al.*, 1986). We tried to benefit from both relatively shallow morphological layer (information relevant for each token in the linear sequence of words in the sentence) as well as from the syntactic analysis of the sentence.

We were successful in utilizing the token-level information, see Chapter 4 for more details. Our attempts to employ the subsequent layers of linguistic description (shallow and deep syntax) were less successful, mainly because they implicitly *strengthened* the unjustified independence assumption of individual translation units. The deep-syntactic approach to MT was so far best exploited

in the transfer-based system TectoMT (Popel and Žabokrtský, 2010) and despite the system did not perform very well on its own, we managed to incorporate TectoMT to the standard phrase-based system in a way that set the new state of the art in English-to-Czech translation, see Chapter 5.

2.2.3 Removing Independence Assumptions

Our work on the core of machine translation has been carried out in the confines of classical statistical MT that deals with individual translation units. We contributed to attempts at removing this assumption through the supervision of Aleš Tamchyna's PhD studies (2012–2017), where Aleš developed a discriminative model to select translation of phrases considering the whole source-side and a small target-side context (Tamchyna, 2017; Tamchyna *et al.*, 2016a; Huck *et al.*, 2017); more details are provided in Section 4.3.3.

A breakthrough in machine translation quality was achieved recently through deep learning, giving rise to neural machine translation, NMT (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014).

NMT replaces the log-linear model with a model directly predicting target words, one at a time, conditioned on the whole source sentence f_1^J:

$$\begin{aligned} p(e_1^I|f_1^J) &= p(e_1, e_2, \dots e_I|f_1^J) \\ &= p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \dots \\ &= \prod_{i=1}^{I} p(e_i|e_{i-1}, \dots e_1, f_1^J) \end{aligned} \tag{2.9}$$

The similarity of NMT to a standard language model should be highlighted. A language model (see Section 4.1 below) predicts the next word based on the previous words: $p(e_1^I) = \prod_{i=1}^{I} p(e_i|e_1, \dots e_{i-1})$. NMT adds f_1^J to the antecedent.

While the main steering force in PBMT is the translation model and the language model "only" caters for target coherence, the main steering force in NMT is the language model and the source "only" conditions the word choices. The exact consequences of this major shift are yet to be explored but NMT generally performs much better in fluency of translation and somewhat better in adequacy.

2.2.4 Better Evaluation

The outputs of machine translation are evaluated manually and automatically for a number of reasons. From the end users' point of view, we need to be able to select the overall best performing MT system. System developers need to be able to reliably check progress or, with the help of automatic evaluation methods, automatically optimize model parameters.

If our metrics[2] of MT quality do not reflect well the problems in output, we cannot expect any improvements. At the same time, *understanding* what is a good and what is a bad translation is an essential component of machine translation as a field of study.

Our contributions to both the technical and scientific aspects of MT evaluation are summarized in Chapter 6.

[2]The term "metric" traditionally used in the field of MT evaluation does not imply the properties of a metric in the mathematical sense.

Chapter 3

Large Data

The collection and preparation of training data may seem a rather mundane task from the scientific point of view. It is nevertheless undisputably the key prerequisite for statistical methods in NLP in general and MT in particular. We also take the stance that a high-quality training dataset attracts attention to the task and languages concerned. We believe that our long-term work on a large Czech-English parallel corpus CzEng described in this chapter has thus not only allowed our own research in English-to-Czech MT but also considerably contributed to the overall focus on this language pair and its adoption as an interesting research problem. MT into Czech is thus examined to a much deeper extent than what would correspond for example to the number of speakers of Czech or the amount of money spent on NLP research by national funding agencies.

Our main contribution in data collection and preparation is the series of releases of CzEng, summarized in Table 3.1. Every release, aside from including additional training data was devoted to a particular topic.

Three CzEng releases deserve a special remark. The version 0.9 (Bojar and Žabokrtský, 2009) was the first major upgrade when we processed both of the sides of the corpus with the Treex NLP processing platform (Popel and Žabokrtský, 2010; in 2009, the platform was still called TectoMT). CzEng 0.9 with its 8.0 million sentences posed a significant technical challenge to the toolkit. Up until then, Treex has been used in various NLP tasks, but processing time and stability across a wide range of data conditions were never the main focus of its development. CzEng 0.9 served as a very thorough test case and allowed to identify many corner cases and minor bugs in the toolkit. Since there was no time available for any major code rewrites, the goal was achieved through data parallelization and automatic collection of failures. We then processed the bugs from the most frequent to the less common ones.

The second major step in CzEng development was achieved in the version 1.0 (Bojar *et al.*, 2012b).[1] In that release, we not only almost doubled the corpus size again, provided the automatic processing (improved in various aspects) but we also carefully filtered the corpus to avoid low-quality sentence pairs. In CzEng 1.0 for the first time, we exploited the other side of the corpus to enhance the automatic annotation even monolingually. Specifically, the comparison of the

[1] See page 63 for the full reference and link to Bojar *et al.* (2012b).

Ver.	Size	Main Focus	Details in
0.5	0.9M	Sentence alignment, common format	Bojar and Žabokrtský (2006)
0.7	1.0M	Used in WMT06 and WMT07	Bojar *et al.* (2008)
0.9	8.0M	Automatic annotation up to t-layer	Bojar and Žabokrtský (2009)
–	–	Sentence-level filtering	Bojar *et al.* (2010b)
1.0	15.0M	Improving monolingual annotation through parallel data	Bojar *et al.* (2012b)
1.6	62.5M	Processing tools dockered	Bojar *et al.* (2016b)

Table 3.1: Summary of CzEng release versions. Size is reported in millions of sentence pairs.

Czech and English automatic annotation allowed us to (1) improve sentence segmentation by adding dedicated training data and new focus patterns to our trainiable tokenizer (Maršík and Bojar, 2012) and (2) spot and fix several errors in the rules construcing "formemes" (Žabokrtský *et al.*, 2008) due to unexpected formeme mismatches in the aligned sentences.

Finally, the most recent release, CzEng 1.6 (Bojar *et al.*, 2016b) benefited from our supervision of Jakub Kúdela's master thesis and publication (Kúdela *et al.*, 2017): 1.84 billion of web pages of the July 2015 Common Crawl were scanned for parallel Czech-English texts through sentence embeddings and locality-sensitive hashing. The goal was to again extend the CzEng parallel data, but as we described in Kúdela *et al.* (2017), Common Crawl was too "sparse". From each website, Common Crawl usually gets only a handful of pages. We thus could not rely directly on Common Crawl data dump and re-crawled the list of websites with parallel content for CzEng 1.6.

Besides MT, CzEng has been used in research on coreference resolution (Novák *et al.*, 2013), automatic valency frame selection (Dušek *et al.*, 2014), in the development of a valency lexicon (Fučíková *et al.*, 2016), a subjectivity lexicon (Veselovská, 2015), a lexical network (Ševčíková *et al.*, 2016), word-level (Kocmi and Bojar, 2016) and sentence-level (Wieting *et al.*, 2017) embeddings or a spoken corpus of Czech dialects (Michlíková, 2013) and in semi-automatic linking between corpora and lexicons (Bejček, 2015).

Chapter 4

Handling Morphology in Phrase-Based MT

The common topic that threads through this thesis is the difficulty of targetting Czech with its rich morphology. Morphological correctness was undoubtedly the most apparent issue of the PBMT-based systems.

Table 4.1 motivates this research by illustrating the availability of morphological variants of the Czech word *čéška (knee cap)* in plural in training corpora of 50K to 50M sentences. The word is not very frequent, but we are lucky to see it in the nominative case (line 1) already in 50K training sentences. Other morphological variants are seen as we use larger corpora. In 50M sentences, we finally see all morphological variants of the word, although the vocative case (line 5) was actually still not seen and we know the form only thanks to its homonymy with the nominative.

case	surface form	50K	500K	5M	50M
1	čéšky	•	•	•	•
2	čéšek	–	•	•	•
3	čéškám	–	–	•	•
4	čéšky	○	○	•	•
5	čéšky	○	○	○	○
6	čéškách	–	•	•	•
7	čéškami	–	–	–	•

Table 4.1: The seven Czech cases of the word *čéška* (knee cap) in plural as seen in 50K/500K/5M/50M sentences. "•" indicates the word was seen in the particular case, "○" indicates that the surface form was seen but in a different case. Reproduced from Huck *et al.* (2017).

In order to correctly use words in a morphologically rich language, the SMT system has to have the capacity to produce them given the English source in the first place (i.e., to see them in a parallel corpus) and also to select the form that fits the given context. As indicated by the example in Table 4.1, *some* morphological variant of a word may be seen in a relatively small number of sentence pairs, but we can't expect to see all forms.

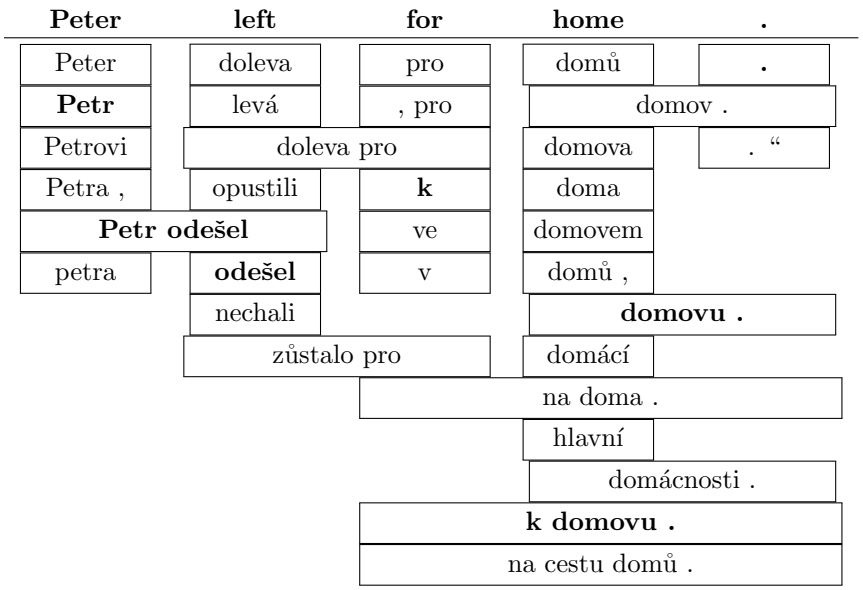


Figure 4.1: Translation options considered by PBMT when translating the sentence “*Peter left for home.*” from English into Czech. Options with a higher translation probability are listed higher, bold indicates options that could be used to construct an acceptable, although not very good translation. Figure simplified from Bojar (2012).

In this chapter, we describe our contributions to producing correct text in morphologically rich languages. We start with a very brief summary of the underlying framework of phrase-based MT (Section 4.1), then focus on improved modelling in situations when the needed target word forms are generally available in the training data (Section 4.2) and conclude by our contributions to producing word forms which were not observed in the parallel or even in the monolingual data (Section 4.3).

4.1 Overview of Phrase-Based MT

Phrase-Based MT (PBMT, Koehn *et al.*, 2003) is one of several classical statistical approaches to MT. Thanks to the availability of open-source implementation of a strong PBMT system Moses (Koehn *et al.*, 2007), phrase-based MT has become the industry standard and remained so until about 2016.

PBMT assumes that the input sentence can be decomposed into *contiguous* sequences of words called “phrases” and each of the phrases can be translated more or less independently. Figure 4.1 illustrates such a decomposition and possible translation units (called **translation options** in PBMT) for the English sentence *Peter left for home.*

The output sentence is constructed left-to-right, selecting phrase translations

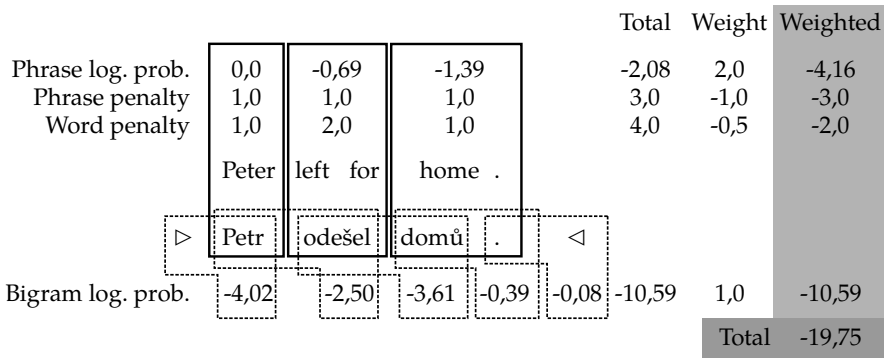


Figure 4.2: Local and non-local features scoring one candidate translation. The solid rectangles indicate individual translation options. Any information available in each of the rectangle can be used for local features. Non-local features cross translation options boundaries and as an example, we illustrate the use of a bigram LM (dotted). The scores are added up for each feature and finally weighted by the weights of the log-linear model. Reproduced from Bojar (2012).

from the source sentence in any order (subject to reordering costs).

Formally, PBMT is implemented as a log-linear model described in Chapter 2, where the key local features are:

- phrase translation probabilities (several variants are used simultaneously, see Koehn, 2009),
- phrase count; its weight is called phrase penalty and moderates if translations are rather literal (word for word) or not,
- word count; its weight is called word penalty and controls the output length.

The only non-local feature is the language model, so any coherence of the selected candidate (e.g., short- or long-range agreements) is to be ensured by the language model. Unfortunately, the language models that were most often used were n-gram LMs. This helped tractability (it was sufficient to keep the previous $n-1$ output tokens in the search state to allow LM evaluation)[1] but it has a serious detrimental effect on overall sentence grammaticality.

Figure 4.2 illustrates one candidate translation, as constructed from the English source using three translation options.

The parameters of these model components are generally estimated using maximum likelihood estimates, usually subject to some form of smoothing or

[1]LMs of unlimited history became possible thanks to deep learning (Bengio *et al.*, 2003) and they indeed brought an improvement to PBMT (Schwenk, 2007) but they never became widely used because the computational costs were too high before the computation was moved to GPUs (Schwenk *et al.*, 2012).

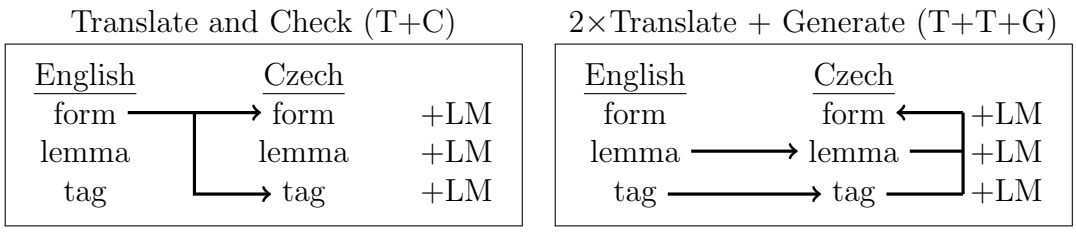


Figure 4.3: Two basic factored translation setups.

interpolation (Chen and Goodman, 1996; Foster *et al.*, 2006) from parallel data (phrase translations) and target-side monolingual data (language models).

As mentioned in Section 2.1, the number of possible translations of a given sentence is exponential to the sentence length, so the space is explored in an approximate search, e.g., **beam search**. Many candidate *partial* translations are considered simultaneously, the more promising ones are further expanded by attaching translation options covering so far untranslated words while the less promising ones are discarded.

In its pure form, PBMT treats word forms as opaque symbols. This is a great advantage for language independence of the method but it comes at the cost of severe data sparsity for morphologically rich languages: the model needs to see all possible forms of all possible translations of a word to have the capacity to produce them. And it should also see each of them in a large number of contexts to be able to select the correct one.

4.2 Factored Setups for Improving Morphological Choices

The implementation of PBMT in the Moses translation system introduced **factors** (Koehn and Hoang, 2007). In short, factors provide additional information for each input and/or output token, and thereby allow to introduce new score components and also to generate output factors based on additional data, not just the parallel corpus.

In Bojar (2007),[2] we thoroughly examined the utility of factored PBMT for targetting Czech.

If we limit ourselves to factors bearing morphological information,[3] two setups immediately come to mind, as illustrated in Figure 4.3 and explored in Bojar (2007):

- T+C (Translate and Check) translates the source word forms into target

[2]See page 63 for the full reference and link to Bojar (2007).

[3]Other options are obviously possible and helpful, see e.g., Avramidis and Koehn (2008), Birch *et al.* (2007), or Niehues and Waibel (2010).

word forms, as baseline PBMT would do, but it also produces target-side morphological tags. This sequence of tags can be then scored with a dedicated language model which operates on a much smaller vocabulary (morphological tags) and therefore can be effectively trained for a much higher n-gram size (e.g., 7 or 10-grams).

- T+T+G (2×Translate and Generate) translates lemmas and morphological tags *independently* and generates the target word form from the lemma and morphological tag; again, multiple language models are used. This setup is linguistically appealing, it correctly strips morphological variance of words from their lexical values. Figure 4.4 explains the benefit from independent learning of translation of lemmas and translation of morphological tags: evidence can be assembled from different sentences, the co-occurrence counts are generally higher and probability estimates more reliable.

In later studies, we wanted to build upon these setups. The T+C setup works very well, as we demonstrated in Bojar (2007) but it is difficult to improve it further, see Section 4.2.1. The T+T+G setup brings serious complications, as described in Section 4.2.2. We proposed several techniques to circumvent the issues, see Section 4.3.

4.2.1 Automatic Exploration of Configurations Infeasible

The content of factors as well as the exact sequence in which they are used on the source side and constructed on the target side is fully configurable. The space of possible configurations is thus very large, especially if we consider also the various meta-parameters such as n-gram size or type of smoothing of each of the language models, and their effectiveness also depends on the amounts of available training data.

In a series of experiments, we largely explored this space of possible configurations:

- In Bojar and Tamchyna (2013),[4] we developed Eman, an experiment manager. Eman, populated with "seed" scripts relevant for machine translation (or any other field of study), allows to manually explore large numbers of configurations, automatically reusing common model parts and rebuilding only what is necessary.

  Eman has been used in the development of almost all our MT experiments and when building shared task systems as well as commercially applied MT systems. While Eman was designed for research and flexibility in experimenting, it also serves as the backbone of a fully automated batch translation online service that we run for IBM to translate into Czech, Hungarian, Arabic and experimentally also into Japanese.

[4] See page 63 for the full reference and link to Bojar and Tamchyna (2013).

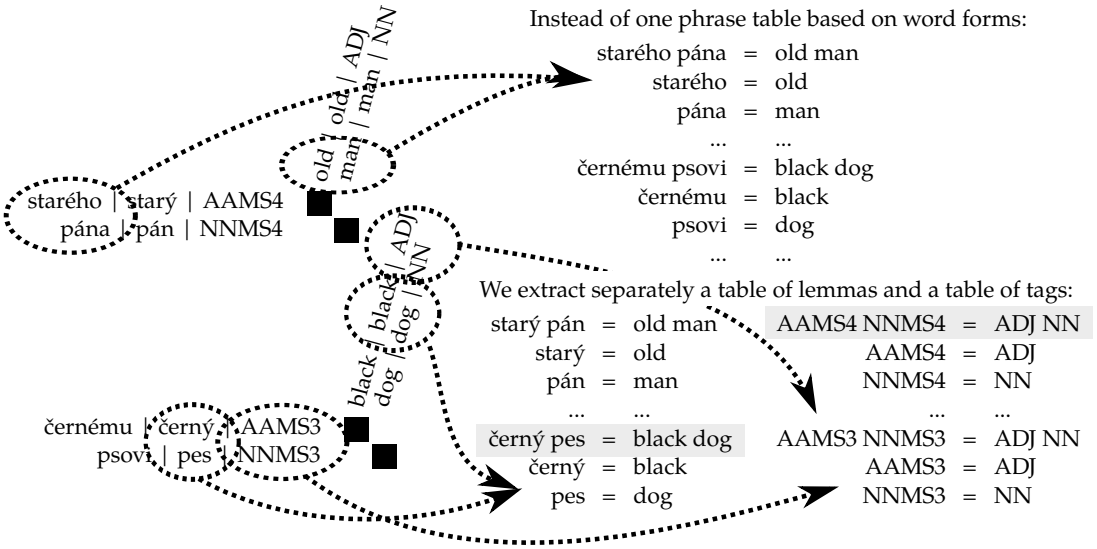


Figure 4.4: Linguistically motivated extraction of factored phrases from a parallel corpus. The corpus, consisting of just two "sentence" pairs: *(viděl jsem) starého pána = (I saw an) old man* and *(dej to tomu) černému psovi = (give it to the) black dog*, does not allow to directly learn the phrase *black dog = černého psa* (the translation of *black dog* into Czech accusative case). In the factored setup T+T+G, this translation is licensed by the combination of the separate lemma (*černý pes*) and tag (*AAMS4 NNMS4*) translations, each of which comes from a different training sentence pair. Reproduced from Bojar (2012).

- In Bojar *et al.* (2012a), we introduced a simple taxonomy for the more common factored setups and further examined which setups work best in various data conditions.

- In Tamchyna and Bojar (2013), Eman served as the underlying engine in an attempt to explore the space of possible PBMT configurations *fully automatically*. While we were able to find a small number of setups that improved the baseline, the main result of that work is negative:

  - The space of possible factored configurations is too large to be explored automatically, i.a. there are exponentially many setups given a number of source-side factors.
  - Evaluating each configuration is computationally demanding (e.g., a few days of computing time with large training data).
  - The automatic evaluation metric (BLEU in that case, see Chapter 6 for more details) is not sufficiently discerning and reliable, many setups receive too similar scores.

– Model optimization is non-deterministic and fragile; several optimization runs of the same setup often differ in their performance more than possible alternative setups.

Across all examined setups, we confirmed that a significant improvement can be expected from essentially only T+C, i.e., a setup that improves target-side morphological coherence by employing an additional language model over morphological tags. This setup does not allow the MT system to produce any word forms that were not seen in both the parallel and monolingual training data, but it improves the probability estimates of word form sequences.

4.2.2 Morphological Explosion on the Fly

The T+T+G setup illustrated in Figure 4.4 unfortunately works only with extremely small datasets (at generally low levels of overall performance). As soon as the parallel corpus becomes reasonably big, T+T+G introduces a loss of essential details and more importantly leads to an explosion of the search space: too many possible word forms have to be generated and scored. Consider our setup where all combinations of lemmas and tags have to be produced and evaluated. For instance for the Czech word *stát* (one of the possible translations of the English word *state*, both the verb and the noun), this amounts to 347 possible Czech word forms (or 182 word forms when dialects and archaic forms are excluded) according to Hajič (2004).

Containing this explosion proved impossible given the design of factored translation models. The models are said to be **synchronous**, i.e., translation options have to be fully generated (all target factors filled) before the main search starts. While we can prune this space by dropping less promising translation options, the scores available at this early stage are only *local*, they cannot consider the context of surrounding words because it will be (gradually) built only later in the main search. At the same time, many morphological features express the relation of words to the context. Dropping some "unlikely" case variations of a noun before the verb is known will inevitably fail because it is the verb that requests a particular case.

In the following section and also later in Chapter 5, we present techniques that avoid these problems.

4.3 Producing Unseen Word Forms

Table 4.1 motivated the need to generate Czech word forms on the fly but in Section 4.2.2, we explained that simply allowing to generate word forms from combinations of lemmas and tags doesn't work.

In this section, we summarize three methods we proposed as possible solutions: two-step translation, reverse self-training and an integrated discriminative model.

Src	after a sharp drop		
Mid	po+6	ASA1.prudký	NSA-.pokles
Gloss	*after+loc*	*adj+sg...sharp*	*noun+sg...drop*
Out	po	prudkém	poklesu

Figure 4.5: An illustration of two-step translation: translating from English to lemmatized Czech (Mid) and only then inflecting.

4.3.1 Two-Step Translation

In Bojar and Kos (2010),[5] we presented the idea of two-step translation to avoid the explosion of variants of words and the difficulties of pruning them before the surrounding context is available. In **two-step translation**, the search is divided into two consecutive phases, see Figure 4.5 for an illustration:

1. **Reordering and lexical choices.** The input sentence is translated into an intermediate "language" that disregards morphological attributes implied solely by the target language. The desired number of tokens, their positions and meaning-bearing morphological features (e.g., plural for nouns or negation) are preserved.[6]

2. **Morphological choices.** The intermediate representation is inflected, preserving the number and order of tokens.

The benefit from phasing the search into two independent steps is that the inflection in Step 2 have full access to the context of surrounding words. Generating all forms is acceptable because they can be effectively pruned without risking serious search errors suffered by T+T+G (Section 4.2.2).

Technically, we realized both steps as factored PBMT setups. Step 1 was trained on parallel data, with standard limits on reordering and target side simplified to lemmas and a hand-picked subset of morphological features.

Step 2 was a monotone word-for-word "translation": the translation model (a phrase table with all phrases limited the length of one token) mapped each simplified Czech word to all possible regular word forms and the standard language model ensured selecting coherent combinations. Since Step 2 was mapping between simplified Czech and regular Czech, we could train it on (large) Czech-only texts.

Compared to the T+C baseline (Section 4.2), our results in Bojar and Kos (2010) were mixed: the two-step translation improved over the baseline in small data setting but not in large data setting.

[5] See page 63 for the full reference and link to Bojar and Kos (2010).

[6] Prior work of Minkov *et al.* (2007), Toutanova *et al.* (2008), or Fraser (2009) disregarded all morphological information and also targeted other languages.

	Source English		Target Czech
Para 126k	a cat chased...	=	**kočka** honila...
			kočka honit... (lem.)
	I saw a cat	=	viděl jsem **kočku**
			vidět být kočka (lem.)
Mono 2M	?		četl jsem o **kočce**
			číst být o kočka (lem.)
			Use reverse translation
	I read about a cat	←	backed-off by lemmas.

⇒ A new phrase learned: "about a cat" = "o **kočce**".

Figure 4.6: The key idea of reverse self-training: The English word *cat* is present in the parallel corpus but its Czech counterparts do not cover all morphological cases of the word, the locative *kočce* is missing. Translating (based on lemmas) a sentence with this particular form from the monolingual data adds this form in its correct context to the translation model.

We continued in exploring two-step setups with our PhD student in Bojar *et al.* (2012a) and Jawaid and Bojar (2014) with no significant gains. The area was also subsequently studied by others, most recently Burlot *et al.* (2016) who explored several other technical realizations of step 2, generally confirming smaller gains as parallel training data grow. At about 1M parallel sentences, there is little or no benefit from the separation.

4.3.2 Reverse Self-Training

In Bojar and Tamchyna (2011a) and further in Bojar and Tamchyna (2011b),[7] we realized that the decision capacity about word forms lies ultimately in the language model. If word form combinations (such as an agreeing pair of an adjective and a noun) are known to the language model, it will promote them. And conversely, any unknowns will force the system to fall back to denser statistics, e.g., to shorter n-grams or (if linguistically-informed models are available) to lemmas or tag sequences. Any cleverness in offering word forms in the translation model is not going to provide any improvement if the language model cannot support the proposed sequence. In other words, it is the intersection of the translation and language model capabilities that is capping the performance of the system.

Assuming that we have trained the LMs of the system the best way we could (used all possible data, used LMs over different linguistic factors), we must ensure that the translation model is not adding further limitations and that it is offering translation candidates that the LM can effectively evaluate. Any *further* candidates, coming for example from a morphological generator, are not going

[7] See page 63 for the full reference and link to Bojar and Tamchyna (2011b).

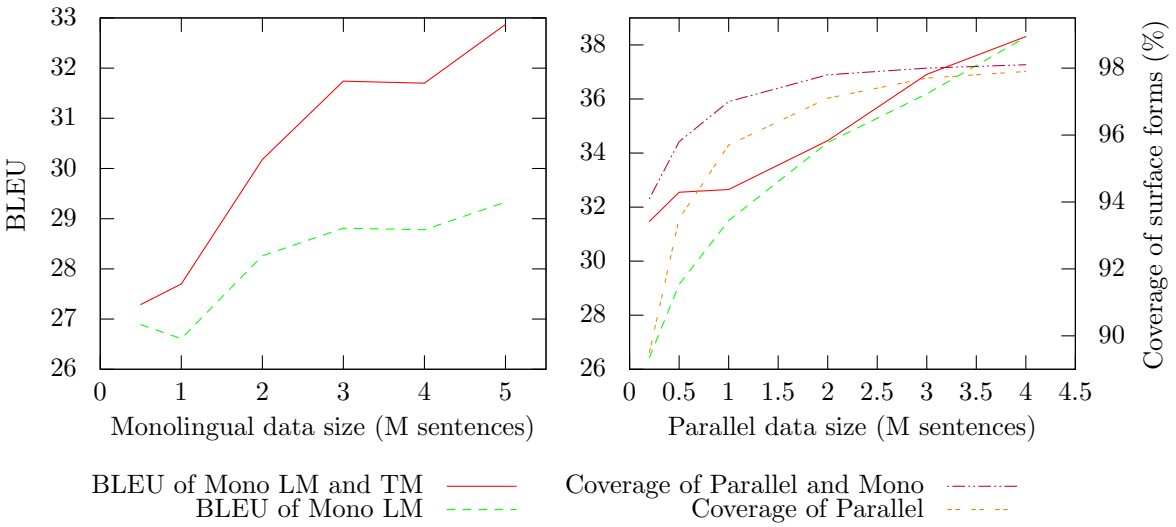


Figure 4.7: Improvements in BLEU score thanks to reverse self-training when adding monolingual data to fixed parallel data (500k sentences, left plot) and when increasing parallel data size with fixed monolingual (5M sentences, right plot). Reproduced from Bojar and Tamchyna (2011b).

to be used anyway because they are not known to the LM (and LM will thus score them lower than other options).

We thus proposed **reverse self-training** as a technique that ensures that the TM is *as capable as* the LM in producing word forms. Given that the LM is trained on generally much larger training data (monolingual texts), we must somehow incorporate these texts into the training of the TM.

The key idea is to use back-translation to translate the target-side monolingual data to the source language and use this synthetic parallel corpus to train the forward system. Back-translation was used previously by Bertoldi and Federico (2009) and became extremely popular recently in neural MT (Sennrich *et al.*, 2016) but one aspect remains unique to our setup.

As illustrated in Figure 4.6 on the preceding page, we back-off the back-translation system to translate from lemmas if the exact word form is not known. If the original source language (English, in our setup) is morphologically less rich, the translation from lemmas will not cause any harm. The forward system will then see a good English sentence or phrase translated to a perfect Czech phrase, containing a word form never seen in the small parallel data. The forward system thus gets the chance to learn a new form of a known word in its correct context.

Figure 4.7 shows the benefits of reverse self-training for English-to-Czech translation. It is well known that increasing LM size is always beneficial (Brants *et al.*, 2007), see the "BLEU of Mono LM" curve in the left plot. Our technique allows to exploit the given monolingual data much better, see the curve "BLEU

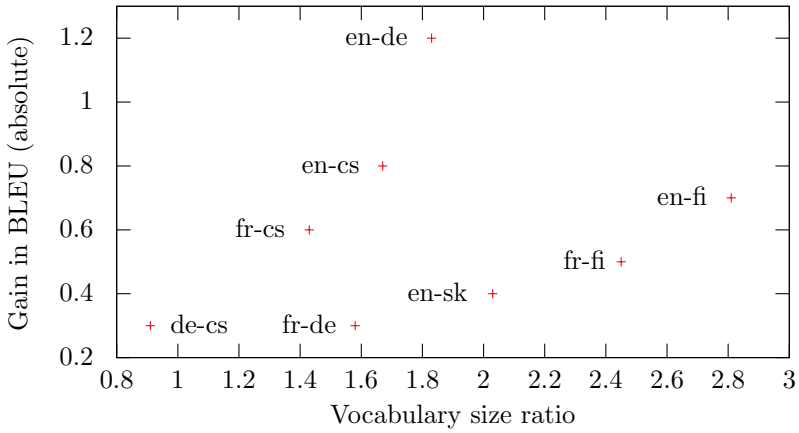


Figure 4.8: Reverse self-training for more language pairs. Reproduced from Bojar and Tamchyna (2011b).

of Mono LM and TM" in the left plot. In the right plot, we can see that the gains diminish as the parallel data grow. The benefit from reverse self-training started at 4 BLEU points but becomes negligible from about 2M of parallel sentences.

Figure 4.8 documents the effectiveness of the method for several language pairs, in relation to their morphological richness. All the underlying experiments used 94–128k parallel sentences and 662–896k monolingual sentences. "Vocabulary size ratio" indicates how many more distinct word forms the target language had in the parallel corpus compared to the source. The extreme is English-Finnish with 2.8× more Finnish forms. The tendency is clear: the richer the target language is compared to the source, the larger the gain. If both languages are rich, such as German-Czech, the benefit is not necessarily big.

4.3.3 Unseen and Discriminatively Trained

As we know from the previous section, having a parallel corpus of 2M sentences for languages like Czech may already be sufficient but arguably, many language pairs suffer from lack of resources much more. Examining methods for particularly low-resource settings is thus interesting.

In the situation when the necessary (target) word forms are not available even in the monolingual data, we have to rely on morphological analyzers and generators, and their dictionaries. Since the dictionaries (naturally) do not provide frequencies or probabilities of the forms in their contexts, we have to rely on a different scoring mechanism.

One option would be to use standard language models in the factored setup (Section 4.2), trained over sequences of morphological tags and (separately) over lemmas. The best form would be selected based on a weighted combination of these scores.

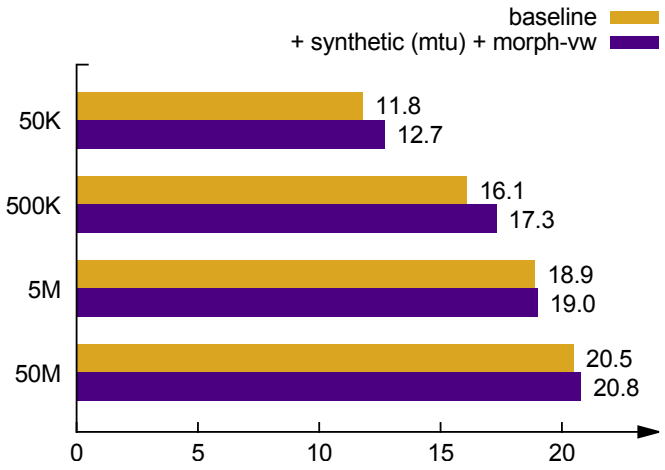


Figure 4.9: The improvement in BLEU thanks to including automatically generated word form variations of translation options ("synthetic (mtu)") and scoring them with the discriminative model ("morph-vw"). Reproduced from Huck *et al.* (2017).

Aleš Tamchyna's thesis examined more fine-grained models, namely **discriminative models** (Tamchyna, 2017). The discriminative model is trained outside of the translation system and allows to include many more features, including fully lexicalized ones (e.g., indicators checking for the presence of individual word forms or lemmas). One of the advantages is that it has the power of learning **valency frames**, that is the requirements of verbs for a particular preposition or case of their arguments.

The integration of such a rich model into the PBMT search is technically challenging because the model is evaluated before pruning for a very large number of translation options. Tamchyna *et al.* (2016a) had to come up with a sequence of optimization tricks to avoid any duplicated calculation. The benefit of this optimization was that the discriminative model could use also a limited context of the *target side*, i.e., the previous word or two of the current partial hypothesis.

In Huck *et al.* (2017), the discriminative model was trained *excluding* the exact word forms and relying only on individual morphological features and the lemma. This allowed to reliably score even word forms generated by the morphological generator; in the case of Czech, Morphodita (Straková *et al.*, 2014) was used. The method is effective especially with corpus sizes of 50k and 500k sentences, small gains are however observed also at 5M and 50M sentence pairs.

Chapter 5

Benefiting from Deep Syntax in MT

The methods and experiments described so far were limited to using relatively shallow linguistic information: lemmatization, tagging, and morphological generation.

In this chapter, we summarize one of our key contributions of this thesis, namely the incorporation of deep-syntactic knowledge to phrase-based MT. We note that we explored this topic already in our PhD thesis (Bojar, 2008), but the approach taken then was not successful.

As we documented for dependency trees used for translation between English and Czech in Bojar and Hajič (2008) and further in Bojar and Týnovský (2009) and as Chiang (2010) described independently for constituency trees for translation from Chinese or Arabic into English, a statistical transfer-based system where the minimum translation units are linguistically-adequate treelets has a considerably harder situation than phrase-based MT or its extension, hierarchical phrase-based translation (Chiang, 2005).

In Section 5.1, we briefly review the problem. Our technique that allows to circumvent it is summarized in Section 5.2, the underlying reasons of its effectiveness are further explained in Section 5.3 and empirical results are provided in Section 5.4.

5.1 Brief Summary of Difficulties with Tree-Based Transfer

In our PhD thesis, we attempted to improve the grammaticality of MT by implementing a transfer-based MT system. Such systems first analyze the input sentence into a formal representation reflecting its syntax and/or semantics, then convert this representation to a corresponding formal representation for the target language and finally generate the plain text in the target language.

The fact that the target string is produced from a formal representation would ideally guarantee that the output will be grammatical and the separation of source linguistic analysis and target generation potentially reduces the need for (large) bilingual training data, benefiting from the generalizations that can be observed monolingually or provided in the form of dictionaries.

In practice, the transfer-based approach fails to surpass shallow methods like phrase-based MT on average, due to especially the following issues (Bojar and Hajič, 2008):

- **Cumulation of errors** when preparing the source and target formal representations of the parallel data. In our case, a tagger was followed by a surface-syntactic parser and then a deep-syntactic parser. If any of them made an error (or if the sentence in the training data was not exactly grammatical, according to the rules embodied in the particular tool or matching the training data behind the tool), the resulting structure contained an error. Shallow methods, on the other hand, suffer only from errors genuinely present in the training data.

- **Mismatching structures** between the source formal representation, the target representation and their alignments prevent extraction of translation counterparts. As outlined above, classical SMT assumes that both source and target can be decomposed into some units, corresponding to one another. If the units follow the syntactic structure of the sentence, as was our case, the decomposition must conform to the structures of both source and target. The underlying grammar formalisms and parsers for the two languages were however built independently and arbitrary decisions as well as natural divergence between languages (Dorr, 1994; Šindlerová *et al.*, 2014) render the sub-structures not matching exactly. Commonly, one accepts only matching sub-structures into the automatically collected "translation dictionary". This means a considerable data loss in comparison to PBMT, where only the word alignment is constraining which pairs of substrings are learned from the data.

- **Increased data sparseness** due to fine-grained details of the deep analysis. As described in Bojar and Týnovský (2009), the core of our approach was a formalism for tree-to-tree transfer (synchronous tree substitution grammars, Eisner, 2003), which assumed operating on trees with atomic nodes. In practice, the nodes of the deep syntactic representation had many attributes, and their values were indeed necessary in order to be able to generate the target sentence correctly. If one combined all the attributes into an atomic unit, the vocabulary size of these units was actually larger than the vocabulary of word forms because the deep representation made finer distinctions. The factorization of translating lemmas and morphological tags separately as discussed for PBMT in Section 4.2 was therefore *necessary*, risking a combinatorial explosion during the translation.

Carefully constructed systems, such as TectoMT (Popel and Žabokrtský, 2010), can to some extent circumvent these shortcomings. For instance, TectoMT still builds upon the assumption that the source and target representations are isomorphic, reducing the transfer to the search for the best labelling of

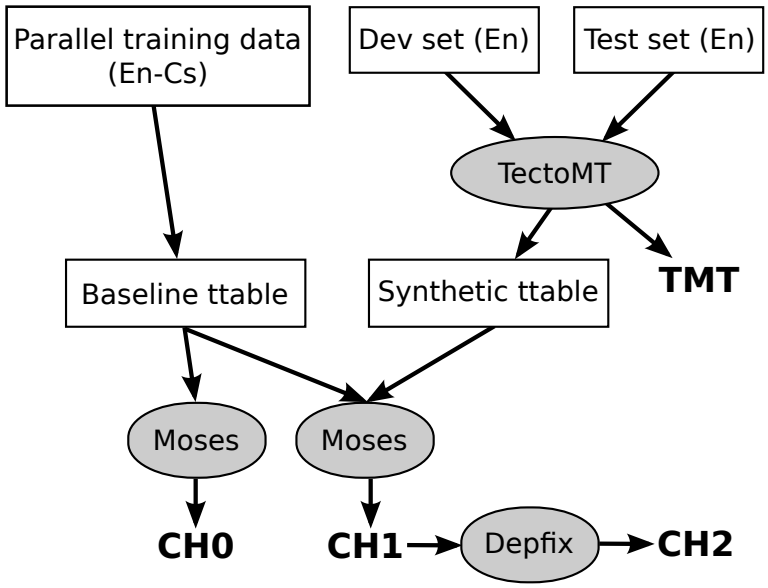


Figure 5.1: Setup of Chimera. Reproduced from Tamchyna and Bojar (2015).

the source-side structure with target-side lemmas and morpho-syntactic labels, so-called "formemes" (Žabokrtský *et al.*, 2008; Dušek *et al.*, 2012). We aimed at a more general data-driven method that would be easier to reuse for other languages, but failed.

While the approach of TectoMT is linguistically appealing and in many cases, it indeed produced grammatically better output than PBMT, it never surpassed PBMT on unconstrained input on average.

5.2 Chimera: Deep-Syntactic and PBMT Systems Combined

In Bojar *et al.* (2013c)[1] and subsequent publications (Tamchyna *et al.*, 2014; Bojar and Tamchyna, 2015; Tamchyna *et al.*, 2016b; Bojar *et al.*, 2017d), we proposed and tested a method that combines the benefits of TectoMT and PBMT. The resulting system was called "Chimera", in reference to the three-headed mythical creature; the third "head" was Depfix (Rosa *et al.*, 2012).

Figure 5.1 schematically illustrates the design of the system combination: the central component is Moses trained on large parallel data and with the best-performing setup (the T+C factored system) as described in Section 4.2. This setup alone is denoted CH0 in the following.

The transfer-based system TectoMT is included in a rather simple but surprisingly effective fashion: TectoMT translates the source side of both the test

[1]See page 63 for the full reference and link to Bojar *et al.* (2013c).

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	**kočky**	
	...	**dvě**	zelené	pruhované	koček	
	viděl	dvou	**zelené**	**pruhované**	kočkám	
	viděla	dvěma	zelení	pruhovaní	kočkách	
	...	dvěmi	zeleného	pruhovaného	kočkami	
	viděl jsem		zelených	pruhovaných		
	viděl jsem		zelenými	pruhovanými		
	viděla jsem	**dvě zelené**		**pruhované kočky**		
		dvě zelené pruhované kočky				

Figure 5.2: Translation options available to CH1: the majority of options come from the corpus and some combination of them hopefully leads to a good translation, underlined. TectoMT provides synthetic options (in bold) that easily match longer sequences of input.

and the development set, leading to a synthetic parallel corpus. The corpus (of a size corresponding to the development and test set, i.e., a few thousand sentence pairs at most) is then processed in the standard "PBMT way": automatic word alignment followed by phrase extraction. We obtain a standard phrase table (the "syntetic ttable" in Figure 5.1) and provide it to Moses, in addition to its standard corpus-based table. Moses has thus the chance to use phrases constructed by TectoMT. Finally, the output is processed by Depfix.

For clarity, we denote the stages of this system TMT (TectoMT alone), CH0 (Moses alone), CH1 (Moses with TectoMT) and CH2 (the full combination). In this chapter, we focus only on the first two components and their interaction.

5.3 Analysis of the Combination

In Tamchyna and Bojar (2015), we carefully analyzed the behavior of the combined system. Technically, the two phrase tables simply provide translation options (as discussed in Section 4.1) to a common pool and the standard search is free to select any of them. Each of the phrase tables comes with its separate phrase penalty, so the model weights can influence whether translation options from one of the tables should be used more often on average.

The nature of the phrases from the CH0 and TMT phrase tables is however rather different. The CH0 table was extracted from a large parallel corpus and, depending of the repetitiveness of the domain and its match with the test data, the source sentence cannot be generally covered with very long phrases, simply because the exact wording is not likely to be seen in the training data.

The TMT table, on the other hand, was created from the source sentences and

all	different?	reachable?	score diff
3003	2665	1741	1601 (<)
			140 (>)
		924	(unreachable)
	338	(identical)	

Table 5.1: Forced decoding—an attempt of CH0 to reach the test set translations produced by CH1. Reproduced from Tamchyna and Bojar (2015).

therefore matches exactly the current source. Much longer phrases can be thus used, as illustrated in Figure 5.2. The CH0 phrase table may have contained all the necessary forms, but they were generally collected from separate sentences. The options by TectoMT may often contain identical words (thus slightly increasing the issue of spurious ambiguity), but it provides them in a longer sequence. The gradual expansion of hypotheses has thus the chance to "jump over" all the combinatorial explosion when searching for a matching combination of word forms.

The language model is applied as usual, giving the combined system the capacity to reject strange parts of the translation that TectoMT may have produced.

Following our discussion on local and non-local features and conflicting structures, our method relieved the language model from being the only source of horizontal coherence of the sentence. Phrases from TectoMT reflect grammatical relations between words locally, within the phrase. The deep-syntactic analysis in TectoMT was useful for producing such phrases but this different structuring along the deep-syntactic tree does not interfere with the simple phrase segmentation of PBMT, thanks to our combination method.

Table 5.1 on the current page documents that TectoMT provided also words not available to CH0. We ran CH0 in the so-called "forced" or "constraint" mode (Schwartz, 2008), checking if it can produce translations created by CH1, i.e., the model with access to TectoMT translations. Out of the 3003 sentences in the WMT14 news test set, CH0 and CH1 produced identical output in 338 cases. In about a third (924) of the remaining sentences, CH0 could not reach the output of CH1, which means that TectoMT either provided a word form never seen in the parallel training data (52M sentences in this experiment), or not seen enough to survive the necessary technical thresholds that disqualify infrequent translations (up to 100 options are considered from each phrase table for each source span).

Figure 5.3 illustrates the complementary benefits of CH0 and TMT, and the ability of CH1 to select the better of each of them. While CH0 makes better lexical choices esp. at the beginning of the sentence when translating the expression *living zone*, it suffers from bad morphological choices at the end of the sentence. The combined system CH1 produces a perfect output for this sentence snippet.

Src	the living zone with the dining room and kitchen section in the household of the young couple .
Ref	obývací zóna s jídelní a kuchyňskou částí v domácnosti mladého páru . *living zone with dining and kitchen section in household* $young_{gen}$ $couple_{gen}$.
CH0	obývací zóna s jídelnou a kuchyní **v sekci** domácnosti **mladý pár** . *living zone with dining_room and kitchen* **in section** $household_{gen}$ $\mathbf{young}_{nom}$ $\mathbf{couple}_{nom}$.
TMT	**živá** zóna **pokoje** s jídelnou a s kuchyňským oddílem v domácnosti mladého páru . **alive** *zone* $\mathbf{room}_{gen}$ *with dining_room and with kitchen section in household* $young_{gen}$ $couple_{gen}$.
CH1	obývací prostor s jídelnou a kuchyní v domácnosti mladého páru . *living space with dining_room and kitchen in household* $young_{gen}$ $couple_{gen}$.

Figure 5.3: Example of translations of Moses (CH0) and TectoMT alone and their phrase-based combination CH1. Errors are in bold, glosses are in italics. Reproduced from Tamchyna and Bojar (2015).

5.4 Empirical Results

We used Chimera in five years of WMT evaluation campaigns, as documented in Table 5.2. During the years 2013–2015, it scored best and it surpassed Google MT significantly in the years 2013–2016.

The table also documents the transition towards neural MT. The first NMT system to join English-to-Czech task was MONTREAL (Jean *et al.*, 2015) and it ended up third or fourth in manual evaluation in 2015. In 2016 and 2017, NMT has proved its superiority.

In Sudarikov *et al.* (2017), we experimented with neural MT but our purely neural approach did not perform well due to various reasons, including the shortage of computing resources (large-memory GPU cards). We nevertheless strongly benefited from NMT outputs by integrating them to our submission in the style of Chimera, adding them in a separate phrase table. Chimera without NMT reached BLEU of 18.3 and NMT allowed an increase to 20.5.

Table 5.2 is sorted by BLEU but it should be noted that this automatic score does not always match human judgements. The most striking difference is seen in WMT17 where our combination including NMT surpassed Google NMT setup in both BLEU and TER but considerably lost in manual scoring. We see this as an indication that humans demand *overall* sentence coherence. This can be achieved by NMT thanks to its avoidance of the assumption of translation units. PBMT, even if provided with well-formed long phrases (from TectoMT or NMT), lacks the capacity to ensure this coherence, and BLEU lacks the capacity to evaluate long-range phenomena.

The disparity between manual and automatic evaluation methods leads naturally to the last large topic in our work, MT evaluation, as described in the next chapter.

	System	BLEU	TER	Manual
WMT13	CH2	20.0	**0.693**	**0.664**
	CH1	**20.1**	0.696	0.637
	CH0	19.5	0.713	–
	Google Translate	18.9	0.720	0.618
	CU-TECTOMT	14.7	0.741	0.455
WMT14	CH2	21.1	0.670	**0.371**
	UEDIN-UNCONSTR.	**21.6**	**0.667**	**0.356**
	CH1	20.9	0.674	**0.333**
	Google Translate	20.2	0.687	0.169
	CU-TECTOMT	15.2	0.716	-0.175
WMT15	CH2	**18.8**	**0.715**	**0.686**
	CH1	18.7	0.717	–
	NMT: MONTREAL	18.3	0.719	0.467
	CH0	17.6	0.730	–
	Google Translate	16.4	0.750	0.515
	CU-TECTOMT	13.4	0.763	0.209
WMT16	NMT: UEDIN-NMT	**26.3**	**0.639**	**0.59**
	CH2	21.7	0.677	0.30
	Google Translate	23.2	0.678	0.19
	CU-TECTOMT	15.2	0.730	-0.03
WMT17	NMT: UEDIN-NMT	**22.8**	**0.667**	**0.308**
	CH2 incl. NMT	20.5	0.696	*0.050*
	NMT: Google Translate	20.1	0.703	0.240
	CH2	18.3	0.719	–

Table 5.2: Automatic scores (BLEU and TER) and results of manual ranking (where available) in WMT13–WMT17. The top other system and Google Translate reported for reference. Bold indicates the best system in each metric, or more systems, if the difference between their manual scores was not sufficiently large for statistical significance.

Chapter 6

Precise MT Evaluation

This chapter summarizes our contributions to the understanding of how to distinguish between good and bad translations.

As mentioned above, MT evaluation serves several purposes and each of them requires a slightly different approach:

- For day-to-day progress check, we need fast and reproducible methods that reflect well overall translation quality *as well as* the problems we want to focus on. Standard automatic evaluation methods may easily neglect our current research target (e.g., translating pronouns or preserving negation), because it is not exhibited on a large portion of the output. Custom targeted methods, on the other hand, can easily overfit, i.e., provide a good score for the aspect their evaluate while ignoring an overall decrease in translation quality.

- For automatic training (model optimization), similarly fast and reproducible methods are necessary. In addition to this, they need to be sufficiently discerning even for very similar candidates (e.g., members of an n-best list). Most importantly, the methods need to be able to rule out poor candidates because otherwise, the optimization could converge to a bad optimum.

- For the selection of the best MT system from a set of fixed possible systems, we have to ask what is the planned use of the MT system: will someone post-edit the translations, or will they be automatically indexed for full-text search, or will someone read them (with or without access or mild understanding of the source)? Each of these uses can lead to a different choice.

  In contrast to the previous situations, most of the compared candidates will be already relatively good machine translations but they can differ considerably on the surface. Methods that work well for selecting the best candidate from an n-best list can fail when the hypotheses become less similar.

In general, both manual and automatic MT output evaluation methods are used. The main benefit of automatic methods is their reproducibility and low

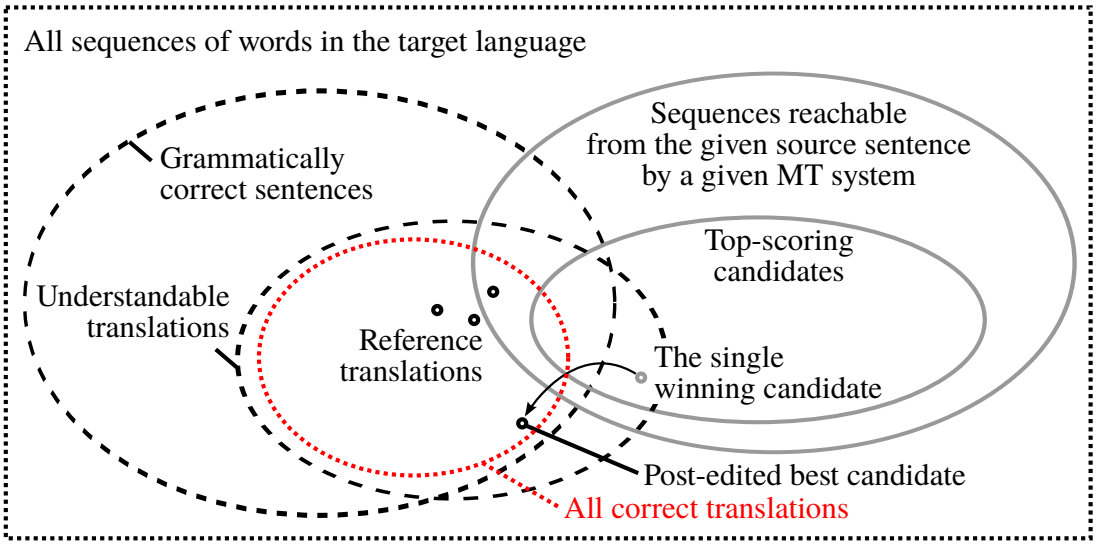


Figure 6.1: Space of possible translations. Reproduced from Bojar *et al.* (2013a).

cost, but they are obviously confined by their inherent assumptions and therefore often overestimate the quality of MT systems based on similar assumptions. Manual evaluation methods are expensive and the main problem is that they are never exactly reproducible because the annotator is affected by the sentences he or she has already evaluated. Reproducibility in manual evaluation can be improved by using large samples with many annotators, however it further increases the cost.

We have contributed to both manual and automatic methods of MT evaluation. In Section 6.1, we explain why MT evaluation is so difficult in general. In Section 6.2, we evaluate the importance of using more references. In Section 6.3, we add a complementary style of manual annotation and notice that PBMT tends to "swallow" words. Finally, Section 6.4 documents that BLEU scores are even less reliable when they are low, and explains why this is the case.

Furthermore, we have contributed to the development of methods of manual MT evaluation that operate along a structured representation of the meaning of the sentence, see Section 6.5.

As a meta-evaluation, automatic MT metrics are evaluated in terms of correlation with human judgements in annual evaluation campaigns, see Section 7.2.

6.1 Why Is MT Evaluation Difficult

It may not be obvious why evaluating MT is so difficult. We contributed to its understanding in Bojar *et al.* (2013a).[1]

[1]See page 63 for the full reference and link to Bojar *et al.* (2013a).

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
Ač ho můžeme prohlásit za politického veterána, reakce radního Karla Březiny byla velmi obdobná.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
A třebaže ho můžeme považovat za politického veterána, reakce Karla Březiny byla velmi podobná.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického veterána.
Radní Karel Březina, navzdory tomu, že ho můžeme označit za politického veterána, reagoval podobně.
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

Figure 6.2: Random sample from 71k possible translations of the English sentence: *And even though he is a political veteran, the Councilor Karel Brezina responded similarly.* Reproduced from Bojar (2012).

Given a fixed input sentence, it is easy to see that there are extremely many possible *erroneous* translations. We can start from any correct translation and modify it by introducing typing errors, altering morphological properties of words (e.g., the number or negation), reordering words or inserting or deleting words. The vast majority of these modifications will damage the translation—and a good MT system should avoid all these errors.

Starting from the other end, considering the set of *all correct translations* is not that straightforward. The situation can be schematically illustrated as in Figure 6.1 on the facing page.

In Bojar *et al.* (2013a), we attempted to quantify the number of *correct* possible translations from English into Czech. Inspired by the work of Dreyer and Marcu (2012), we designed a framework fit for morphologically rich languages and asked several annotators to provide as many good translations of a sentence as possible.

The results, in line with what Dreyer and Marcu (2012) observed for English, are rather interesting. An English sentence of 14 words can easily have 70 *thousands* of correct translations, as illustrated in Figure 6.2.

Each annotator in this exercise was instructed to spend up to two hours per sentence, using our tool to generate and validate sentences semi-automatically. The least prolific annotator provided this sentence with 350 possible translations, the second one created 3192 translations. And the most prolific one reached 67936 translations. Among these, only 8 translations were suggested by all three

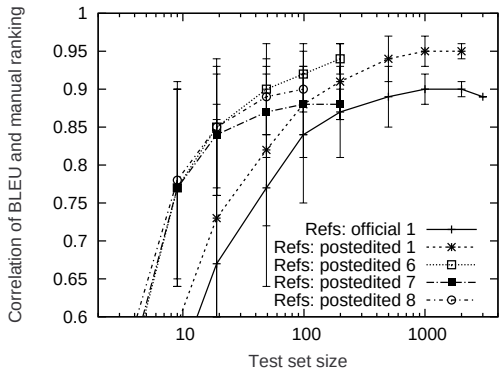


Figure 6.3: Correlation of BLEU and human judgements for varying type and number of reference translations. Reproduced from Bojar *et al.* (2013b).

annotators and only 172 translations were suggested by two of the three annotators. The space of possible translations is thus probably much larger.

The translations are not always 100% literal and they obviously differ in many more or less important aspects, such as register or style, information structure etc. If used in a coherent text and not as isolated sentences, many of these translations may not be acceptable at all, but for the current level of MT quality, all are equally good.

When designing automatic methods of MT evaluation, we thus have to keep in mind that the candidate translation produced by an MT system can be correct but superficially very distant from a given reference translation, or that it can be superficially very similar to the reference translation but suffer from serious errors.

6.2 More and/or Post-Edited References

The most widespread automatic MT evaluation method, BLEU (Papineni *et al.*, 2002), works by validating short fragments (1 to 4-grams) of the candidate translation against a provided reference translation. BLEU has been designed with the assumption that four independent human reference translations will be available, to allow for at least some variance in the MT output. However, BLEU is actually most often used with only one reference.

In Bojar *et al.* (2013b), we extended the manually-collected data of WMT13 with a substantial number of post-edited sentences. Through that experiment, we confirmed that BLEU becomes much more reliable with more references, but also found out that the *nature* of reference translations affects the correlation of BLEU and human judgements. The correlations are generally higher if the reference translations were created by post-editing MT outputs, i.e., if they are (very likely) more similar to the candidate translations.

	Google	Moses-Bojar	PC Translator	TectoMT	Total
Automatic: BLEU	13.59	**14.24**	9.42	7.29	–
Manual: Sentence ranking	0.66	0.61	**0.67**	0.48	–
Manual: Error flags	**2319**	2354	2536	*2895*	10104
Error flags details:					
Words with bad meaning	617	**587**	800	999	3003
Auxiliary word missing	84	111	96	138	429
Content word missing	72	*199*	**42**	108	421
Word form incorrect	783	735	762	713	2993
Superfluous word	381	313	353	394	1441
Non-translated word	51	53	56	97	257
Total serious errors	**1988**	1998	2109	2449	8544
Bad local word order	117	100	157	155	529
Punctuation error	115	117	150	192	574
...	...	...	...	...	...
Tokenization error	7	12	10	6	35

Table 6.1: A comparison of two types of manual evaluation (Sentence ranking and Flagging of errors) and BLEU scores for four English-to-Czech MT systems from WMT09. Noteworthy best results highlighted in bold, noteworthy worst results in italics. Adapted from Bojar (2011).

Figure 6.3 documents the situation. The generally lowest performance is obtained in the standard conditions with 1 "official" reference translation. The error bars reflect the variance due to random subsampling from the full 3k sentences and get narrower as larger and larger portion of the test set is used. With 2k or 3k sentences in the test set, the Spearman's rank correlation coefficient ρ reaches levels of 0.9. Using a single reference created by post-editing randomly selected systems from the set of evaluated systems works clearly better, reaching correlation of 0.95.

We also see from Figure 6.3 that the size of the test set and the number of references can somewhat compensate for each other. Specifically, the common practice of WMT shared translation tasks is to have about 3000 sentences with a single reference translation. A comparable correlation of BLEU and human judgements could be also achieved with just 100–200 sentences and 6–7 reference translations.

6.3 Error Annotations Help to Explain Bad Correlation for BLEU

In Bojar (2011)[2], we experimented with two techniques of detailed error analysis. One was based on semi-automatic interpreting of post edits of candidate

[2]See page 63 for the full reference and link to Bojar (2011).

translations and another relied on manual flagging of errors using some error classification. Here is an example of the error flagging:

Source	Sarkozy meets angry fishermen.
Reference	Sarkozy jde vstříc rozhněvaným rybářům
Moses	Sarkozy se [MISSC: setkává] [MISSA: s] [FORM]rozzlobení rybáři.
TectoMT	Sarkozy [DISAM]splňuje [MISSC: vstříc] naštvané [FORM]rybáře.
Google	Sarkozy [LEX]splňuje [FORM]zlobit [FORM]rybářů.
PC Translator	Sarkozy se setkává [MISSA: s] [FORM]rozhněvané [FORM]rybáře.

In our annotation, we attached flags to individual tokens in MT output (and added tokens for missing words). The example illustrates errors in word form choice ([FORM]), word meaning (source word disambiguation [DISAM] and bad lexical choice [LEX]; the last two are difficult to distinguish and have the highest disagreement rate), as well as missing content ([MISSC]) and auxiliary ([MISSA]) words.

Both post-editing and error flagging led to similar conclusions about the MT systems competing in English-to-Czech translation back then: the traditional commercial system PC Translator was quite bad in lexical choice, TectoMT performed best in picking the right form of the word and phrase-based Google and our Moses were generally good in lexical choice but suffered from errors in morphology.

The flagging of errors also allowed to explain the bad performance of BLEU for this set of systems, see Table 6.1. Our Moses scored best according to BLEU but ended up third in terms of the WMT09 manual sentence ranking. As the detailed error flags reveal, the winning PC Translator made by far the least number of errors in the category of "Content word missing", while our Moses dropped almost five times more content words.

6.4 Low BLEU Scores Unreliable

In the English-to-Czech evaluation campaigns 2009 and 2010, we saw a strikingly low correlation between human judgements about translation quality and BLEU scores, see the left part of Figure 6.4.

While the correlation of BLEU and human judgements for Czech was low, we found in Bojar *et al.* (2010a)[3] a high correlation between the *absolute BLEU scores* and their correlation to human judgements across all language pairs taking part in WMT09, see the right part of Figure 6.4. Put simply, BLEU scores below 20 are not reliable.

Bojar *et al.* (2010a) have also explained the reason for this. The situation is illustrated in Table 6.2 which compares the sets of n-grams in outputs of several MT systems deemed correct according to (1) the presence of the n-gram in the reference translation vs. (2) the absence of manual error flags described above.

[3]See page 63 for the full reference and link to Bojar *et al.* (2010a).

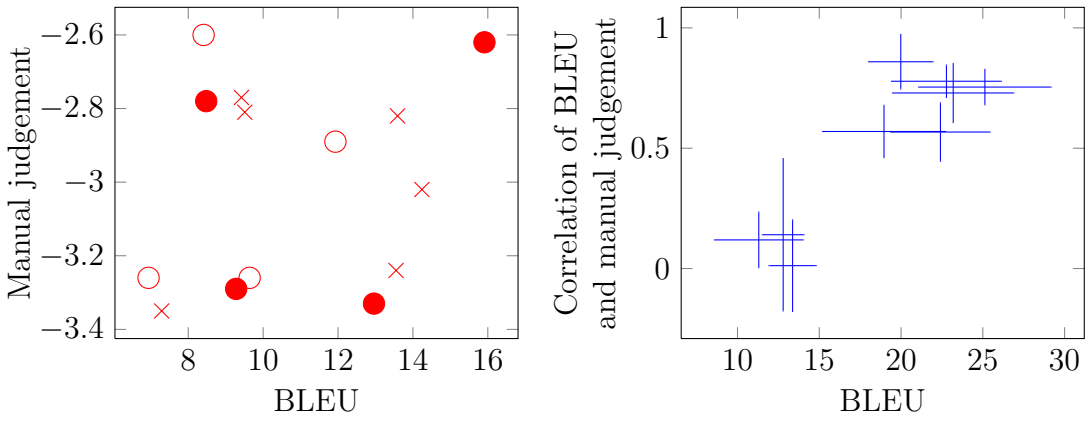


Figure 6.4: Left: Low correlation between BLEU and human judgements. Each point corresponds to one MT system, different point styles indicate a different test conditions. We see no correlation between BLEU and manual judgement. Right: A good correlation between the BLEU scores and their correlation with human judgements, i.e., higher BLEU scores correlate well with humans and lower BLEU scores do not. Each cross corresponds to one language pair, showing the average and standard deviation of BLEU scores and manual judgements across all systems for that language pair. Simplified from Bojar *et al.* (2010a).

Two situations are desirable: when the n-gram does not contain errors and it is confirmed by the reference, and when the n-gram contains errors and the reference does not confirm it. This happens for 59% of unigrams and 56% of bigrams, etc. False positives (n-grams confirmed but containing an error) are luckily rather rare: 6% of unigrams, 2% of bigrams, etc.

The reason for unreliability of BLEU at low scores lies in the fourth case: error-free n-grams that are nevertheless not available in the reference. BLEU does not give any credit to them but the systems can quite differ in the quality of translation in these cases. As seen in Table 6.2, this amounts to more than a third of unigrams, 43% of bigrams etc.

Post-edited references discussed in the previous section are much closer the to candidate translations and don't suffer from this lack of coverage. The unconfirmed n-grams will be only those where the post-editor needed to rephrase the sentence to fix some error or disfluency. Any decrease in BLEU will thus correspond to genuine issues of the candidate translation.

In Bojar *et al.* (2010a), we proposed to increase the coverage of BLEU by matching the candidate with the reference at a coarser level of representation, namely bags of deep-syntactic lemmas (separate for each deep-syntactic part of speech) instead of the common longer n-grams of exact word forms. For English-to-Czech, this increased the correlation in that particular experiment from 0.33 to 0.53.

Confirmed by Ref	Contains Errors	1-grams	2-grams	3-grams	4-grams
Yes	Yes	6,34 %	1,58 %	0,55 %	0,29 %
Yes	No	36,93 %	13,68 %	5,87 %	2,69 %
No	Yes	22,33 %	41,83 %	54,64 %	63,88 %
No	No	**34,40 %**	**42,91 %**	**38,94 %**	**33,14 %**
Total n-grams		35 531	33 891	32 251	30 611

Table 6.2: n-grams as confirmed by the reference and/or by containing or free from errors according to manual error flagging. Lack of coverage of the reference highlighted in bold. Reproduced from Bojar *et al.* (2010a).

In Macháček and Bojar (2011), we further elaborated on that, moving back to the less computationally-demanding shallow but still sufficiently coarse features of words. We also confirmed the applicability of the proposed method in model optimization, performing acceptably in the main manual scoring that rewarded tied results and getting the best score when ties were disfavored (Callison-Burch *et al.*, 2011). See also Section 7.1 for a discussion the manual evaluation method.

6.5 MT Evaluation Focused on Semantics

With the success of neural MT, the focus of MT evaluation has to be changed as well. Multiple studies (Bentivogli *et al.*, 2016a; Bojar *et al.*, 2016a; Castilho *et al.*, 2017b,a) suggest that NMT primarily improves fluency. Adequacy of translations is improved as well, but to a smaller extent. We would therefore expect that, on average, misunderstandings due to MT errors will be less frequent, but at the same time, they will be harder to notice: MT output will be more often seemingly perfect but including a semantic flaw.

For that reason, we have revived our interest in semantic correspondence between the candidate translation and the reference. In Bojar and Wu (2012), we experimented with HMEANT (Lo and Wu, 2011), a manual method of MT evaluation based on aligning the predicate-argument structures of the candidate and the reference. Building upon that, we designed a manual method of MT evaluation that closely follows the semantic structure of the source sentence (and not the reference, thereby avoiding the need to parse the often garbled MT output) in a joint work (Birch *et al.*, 2016).

Chapter 7

Shared Tasks

To reliably measure progress of the field of natural language processing and machine translation in particular, approaches to problems and proposed solutions have to be regularly compared in a rigorous way. Such a comparison is however often difficult to achieve due to many interacting conditions and generally large efforts are needed.

The common practice in NLP resolves this by **shared tasks**: regularly or independently organized events where the organizers specify an exact task description and usually provide training datasets and then collect submissions from participants to evaluate them in a clear and comparable way.

The history of shared tasks related to machine translation has been summarized in Bojar *et al.* (2016c)[1] for WMT (originally Workshop on Statistical Machine Translation which became an ACL-sponsored conference in 2016) and in Bentivogli *et al.* (2016b) for IWSLT, a workshop focused primarily on the translation of spoken language.

Over the years, our contribution to the course of WMT shared task has been twofold: (1) contributing to best practices in MT evaluation, and (2) co-organizing various tasks. We summarize these contributions in the following sections.

7.1 Avoiding Bias in WMT News Translation Task

The main shared task at WMT is translation of news text, see Koehn and Monz (2006) through Bojar *et al.* (2017a). Thanks to our participation in the EU project EuroMatrix[2] and subsequent EU projects within the 6th and 7th Framework Programmes and in H2020, Czech has been included in this task every year since 2007. We also participated in the task with our translation systems of diverse nature.

Up until 2016, the main WMT evaluation measure was derived from annotation screens of up to five systems ranked manually according to the perceived translation quality. The annotators were presented with the source, the reference translation and 5 candidate outputs and they indicated the relative quality of

[1] See page 63 for the full reference and link to Bojar *et al.* (2016c).

[2] `http://www.euromatrix.net/`

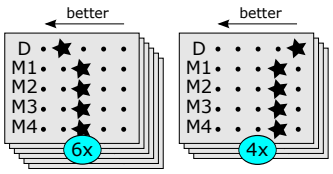


Figure 7.1: Illustration of an artificial collection of manual rankings as used in WMT until 2016. The sample annotation consists of 10 annotation screens in total, in 6 of which the system D wins and in 4 of which it loses. Its four competitors M1...M4 are always on par. Individual annotation screens may be provided by different people.

Interpretation	"≥ Others"	"> Others"	"Ignore Ties"
Formula	$\frac{\text{wins+ties}}{\text{wins+ties+losses}}$	$\frac{\text{wins}}{\text{wins+ties+losses}}$	$\frac{\text{wins}}{\text{wins+losses}}$
Favors	"mainstream"	"distinct"	-
D	$6 \times 4 = 24/40$	**24/40**	24 / 40 = **6/10**
M1	$10 \times 3 + 4 =$ **34/40**	4/40	4/10

Figure 7.2: Various ways of handling ties in WMT ranking. The calculations are based on the sample annotation from Figure 7.1. When ties are rewarded ("≥ Others"), the tying systems M1...M4 "support" each other and each of them thus seems to perform better than D (34/40 over 24/40 wins), unduly favouring similar systems. Penalizing ties ("> Others") promotes distinct systems like D. "Ignore Ties" is a fairer option, for which we advocated in Bojar *et al.* (2011).

these translations; see Figure 7.1 for sample dataset of judgements (the underlying sentences were selected randomly from the test set and were not important when interpreting the evaluation, we thus omit them in the picture). In practice, the exact set of 5 ranked systems differed from screen to screen, sub-sampling five-tuples from all the competing systems.

Observing the performance of our systems in 2010, we noticed that the same collected judgements can be interpreted in subtly different ways, leading to different results. We thus carefully analyzed the discrepancies and reported them in Bojar *et al.* (2011). Here we highlight two of the issues:

Rewarding ties unduly favors similar systems. Figure 7.2 illustrates that depending on the treatment of cases where more systems receive the same rank in an annotation screen, the final ordering of the systems can differ. Specifically, WMT used to rely on a formula that rewards ties ("≥ Others"; "systems ... are ranked based on how frequently they were judged to be better than or equal to any other system", Callison-Burch *et al.*, 2010). This choice can be considered particularly problematic since several system

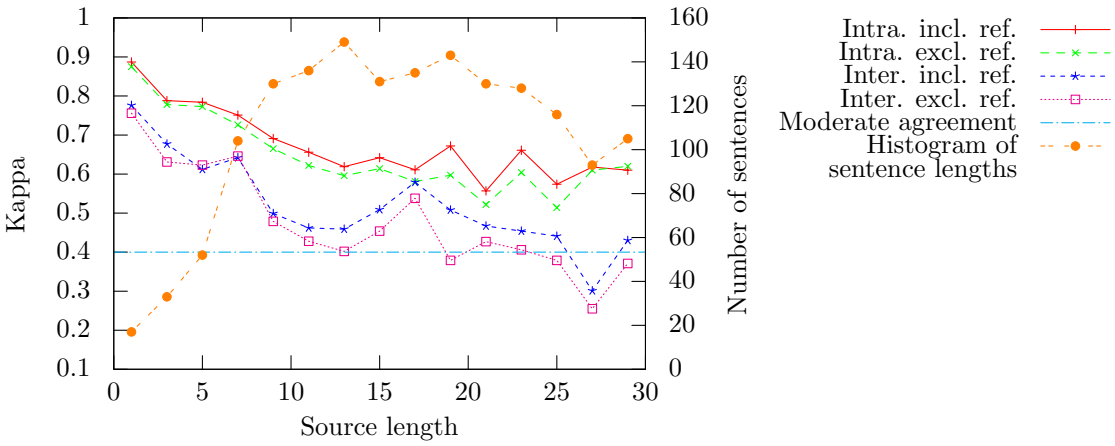


Figure 7.3: Intra- and inter-annotator agreement in terms of the kappa statistic (left axis) of WMT10 evaluation, including of excluding comparisons with reference translations. "Histogram of sentence lengths" (right axis) shows the distribution of sentences in the test set. Adapted from Bojar *et al.* (2011).

submissions were always based on the Moses translation system, where similar translation quality can be expected.

Agreement rates decrease with sentence length. The aggreement rates between different people (inter-) and between annotations of the same person (intra-) have been reported along with the results since Callison-Burch *et al.* (2007), in the form of Cohen's kappa (Bennett *et al.*, 1954). In Bojar *et al.* (2011), we noted that the agreement decreases with sentence length as illustrated in Figure 7.3. Following indicative ranges for the kappa statistic,[3] we see that the inter-annotator agreement when comparing two real systems (as opposed to one system and the reference translation) gets close or below what Landis and Koch (1977) suggest as moderate agreement. Importantly, it turns out that the majority of the evaluated sentences are of this length.

Our discussion sparkled further research and evolution of the method of manual ranking (Lopez, 2012; Koehn, 2012; Hopkins and May, 2013). The current method called "direct assessment" (Graham *et al.*, 2016) simplifies the task by evaluating only one candidate at a time and asking the annotator to provide a score on an effectively continuous *absolute* scale given only the reference translation, not the source. Direct assessment became the official method only in 2017 (Bojar *et al.*, 2017a) so we still anticipate further developments in this area in the coming years.

[3]However, see the discussion in Komagata (2002).

	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17
Participating Teams	-	6	8	14	9	8	12	12	11	9	8
Evaluated Metrics	11	16	38	26	21	12	16	23	46	16	14
Baseline Metrics							5	6	7	7	7
System-level evaluation methods											
Spearman Rank Correlation	●	●	●	●	●	●	●	○			
Pearson Correlation Coefficient							○	●	●	●	●
Segment-level evaluation methods											
Ratio of Concordant Pairs		●	●								
Kendall's τ				❶	❶	❶	❷	❸	❸	❸	◗
Pearson Correlation Coefficient										○	◖
Tuning Task					●				●	●	

● main and ○ secondary score reported for the system-level evaluation.
❶, ❷ and ❸ are slightly different variants regarding ties.

Table 7.1: Summary of metrics and tuning tasks over the years. The vertical bar indicates since when we started co-organizing the task.

7.2 Organizing Shared Tasks

Since 2013, we have been actively involved in the organization of shared tasks of various types:

News Translation Tasks attract the largest number of participants each year. The main goal, translating short news stories, remains unchanged while the underlying set of languages slightly changes every year. The test sets for the task are created anew each year, to provide the participants with genuinely novel text. Huge collective effort is spent on manual evaluation and throughout the years (also due to our analysis presented in Section 7.1 above), the task saw a few modifications to the official method of evaluation.

Our contribution to the organization slightly varied through the years, but every year, we arranged the selection and fixes to the Czech part of the test set (without actually looking at it, to avoid any advantage over other participants in the task), and we organized the evaluation of Czech, relying on a large pool of our Czech colleagues and other annotators.

We were involved in five such campaigns so far (Bojar *et al.*, 2013b, 2014, 2015, 2016a, 2017a).

Metrics Tasks build upon the large pool of manual translation quality judgements collected in the evaluation of News Translation Task and test the performance of automatic metrics against human scoring. Since 2008, two

levels of evaluation are considered: "system-level" (metrics have to predict the quality of a set of sentences) and "segment-level" (metrics have to predict the quality of every sentence).

We were co-organizing five metrics tasks (Macháček and Bojar, 2013; Macháček and Bojar, 2014; Stanojević *et al.*, 2015b; Bojar *et al.*, 2016d, 2017b) and Table 7.1 provides an overview of the full history of the task.

In 2016, we trialled the use of direct assessment as the golden truth in the metrics task and in 2017, it became the official method of news task evaluation, so we switched to it as well. For some language pairs, the direct assessment method did not allow to collect sufficient number of manual judgements and we had to resort to the older style of comparison, as indicated by the symbols ◖ and ◗.

It used to be the case in the past, that successful metrics from one year were never submitted again in the subsequent editions of the task simply because their authors got interested in other topics. To at least partially avoid this loss, we introduced a set of baseline metrics and regularly include them in the task. Accumulating the results over the years (i.e., a varied set of language pairs and evaluated MT systems), we can draw more stable conclusions about the overall performance of these metrics. A first such summary was presented in Bojar *et al.* (2016c).

Tuning Tasks were devoted to the model optimization as mentioned in Section 2: a fixed set of model components for a fixed MT system was provided and task participants had to find the best weight settings. The translations using these settings (run by the task organizers) were then evaluated manually among the News task submissions. The point of the tuning tasks was to assess the applicability of various MT metrics in model optimization and the performance of various model optimization techniques themselves.

After two rounds of the tuning task (Stanojević *et al.*, 2015a; Jawaid *et al.*, 2016), we concluded that the variance among the different submissions in large-data setting (Tuning Task 2016) is small. The results have nevertheless clearly indicated that there was some progress in the optimization algorithms, KBMIRA (Cherry and Foster, 2012) outperforming the prevalent MERT (Och, 2003), but *not* in metrics when used for model optimization: BLEU (Papineni *et al.*, 2002) was still the method that led to the best-performing systems in terms of final manual evaluation.

Neural MT Training Task (Bojar *et al.*, 2017c) is a new type of task we proposed in respose to the shift to neural MT. The performance of neural MT models is affected by several more or less independent aspects: (1) the model structure, (2) the available training data and their pre-processing and (3) the technique used to train the model. In the NMT training task, we fixed (1) and (2), providing task participants with a pre-defined model

> in the Neural Monkey toolkit (Helcl and Libovický, 2017), pre-processed training data and some suggestions what would be interesting to evaluate. As with the tuning tasks, participants did not run the translation themselves, they only provided the trained models. We applied the models to the WMT17 news test set and included these outputs in manual evaluation of WMT17.
>
> The results indicate that statistically-significant differences in translation quality can be obtained by different training techniques, and the more successful submissions shared one particular property: they adapted the training corpus to the news domain by subsampling it or by promoting such sentence pairs. Domain adaptation is thus a critical step in the training of neural MT.

Further long-term observations of the news translation task (esp. its manual evaluation) and the metrics task (a summary of the best performing metrics across the years) are provided in Bojar *et al.* (2016c).[4]

[4]See page 63 for the full reference and link to Bojar *et al.* (2016c).

Chapter 8

Summary

This habilitation thesis summarizes the contributions of Ondřej Bojar in the area of machine translation and machine translation evaluation focused on the translation into morphologically rich languages, mainly from English into Czech.

As documented in the attached publications, the author has:

- created a large automatically-annotated corpus CzEng, allowing a wide audience of researchers to experiment with English-Czech translation and allowing Czech to become a frequent example language in MT research,
- exploited explicit morphological information to improve translation quality into Czech, using several different techniques and different settings: word forms known but less frequent in parallel data, word forms not available in parallel data but covered in monolingual data and word forms not available even in the monolingual data,
- experimented with incorporating deep syntactic processing into machine translation systems, proposing a technique that defined the state of the art for news translation from English to Czech in years 2013–2015,
- contributed to techniques of MT evaluation by analyzing the space of possible translations, difficulties of MT evaluation and issues of the most commonly used MT evaluation method,
- supported the MT community by co-organizing shared task and also significantly contributing to the practices in translation task evaluation.

The original scientific papers detailing these contributions are reproduced in Appendix A, pages 63–65.

Bibliography

Eleftherios Avramidis and Philipp Koehn. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 763–770, Columbus, Ohio, June 2008. Association for Computational Linguistics. Cited on page 20

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. Cited on page 12

Eduard Bejček. *Automatické propojování lexikografických zdrojů a korpusových dat*. PhD thesis, Charles University, Prague, 2015. Cited on page 16

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003. Cited on page 19

E. M. Bennett, R. Alpert, and A. C. Goldstein. Communications through limited questioning. *Public Opinion Quarterly*, 18(3):303–308, 1954. Cited on page 47

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November 2016. Association for Computational Linguistics. Cited on page 44

Luisa Bentivogli, Marcello Federico, Sebastian Stüker, Mauro Cettolo, and Jan Niehues. The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions. In Ondřej Bojar, Aljoscha Burchardt, Christian Dugast, Marcello Federico, Josef Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Georg Rehm, Lucia Specia, Marco Turchi, and Hans Uszkoreit, editors, *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 14–19, Portorož, Slovenia, 2016. [http://www.cracking-the-language-barrier.eu/], LREC. Cited on page 45

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Luboš Ureš. The candide system for machine translation. In *Proceedings of the workshop on Human Language Technology*, HLT '94, pages 157–162, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. Cited on page 7

Nicola Bertoldi and Marcello Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece, March 2009. Association for Computational Linguistics. Cited on page 26

Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 20

Alexandra Birch, Omri Abend, Ondřej Bojar, and Barry Haddow. HUME: Human UCCA-Based Evaluation of Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1274, Austin, Texas, November 2016. Association for Computational Linguistics. peer-reviewed. Cited on page 44

Ondřej Bojar and Jan Hajič. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Cited on page 10, 29, 30

Ondřej Bojar and Kamil Kos. 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 2010. Association for Computational Linguistics. Cited on page 24

Ondřej Bojar and Aleš Tamchyna. Forms Wanted: Training SMT on Monolingual Data. Abstract at Machine Translation and Morphologically-Rich Languages. Research Workshop of the Israel Science Foundation University of Haifa, Israel, January 2011. Cited on page 25

Ondřej Bojar and Aleš Tamchyna. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. Cited on page 25, 26, 27

Ondřej Bojar and Aleš Tamchyna. The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58, 2013. Cited on page 21

Ondřej Bojar and Aleš Tamchyna. CUNI in WMT15: Chimera Strikes Again. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 79–83, Lisboa, Portugal, September 2015. Association for Computational Linguistics. Cited on page 31

Ondřej Bojar and Miroslav Týnovský. Evaluation of Tree Transfer System. Project Euromatrix - Deliverable 3.4, ÚFAL, Charles University, March 2009. Cited on page 29, 30

Ondřej Bojar and Dekai Wu. Towards a Predicate-Argument Evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 30–38, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics. Cited on page 44

Ondřej Bojar and Zdeněk Žabokrtský. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62, 2006. Cited on page 16

Ondřej Bojar and Zdeněk Žabokrtský. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92:63–83, 2009. Cited on page 15, 16

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. ELRA. Cited on page 16

Ondřej Bojar, Kamil Kos, and David Mareček. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Cited on page 42, 43, 44

Ondřej Bojar, Adam Liška, and Zdeněk Žabokrtský. Evaluating Utility of Data Sources in a Large Parallel Czech-English Corpus CzEng 0.9. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 447–452, Valletta, Malta, May 2010. ELRA, European Language Resources Association. Cited on page 16

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. Cited on page 46, 47

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. Probes in a Taxonomy of Factored Phrase-Based Models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 253–260, Montréal, Canada, June 2012. Association for Computational Linguistics. Cited on page 22, 25

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May 2012. ELRA,

European Language Resources Association. Cited on page 15, 16

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. Scratching the Surface of Possible Translations. In *Proc. of TSD 2013*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg, 2013. Západočeská univerzita v Plzni, Springer Verlag. Cited on page 38, 39

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Cited on page 40, 48

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Cited on page 31

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MD, USA, 2014. Association for Computational Linguistics. Cited on page 48

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal, September 2015. Association for Computational Linguistics. Cited on page 48

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation (WMT16). In Ondřej Bojar et al., editors, *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*, volume 2, pages 131–198, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics, Association for Computational Linguistics. Cited on page 44, 48

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London, 2016. Masaryk University, Springer International Publishing. Cited on page 16

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In Ondřej Bojar, Aljoscha Burchardt, Christian Dugast, Marcello Federico, Josef Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Georg Rehm, Lucia Specia, Marco Turchi, and Hans Uszkoreit, editors, *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34, Portorož, Slovenia, 2016. [http://www.cracking-the-language-barrier.eu/], LREC. Cited on page 45, 49, 50

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. Results of the WMT16 Metrics Shared Task. In Ondřej Bojar and et al ., editors, *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*, volume 2, pages 199–231, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics, Association for Computational Linguistics. Cited on page 49

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. Cited on page 45, 47, 48

Ondřej Bojar, Yvette Graham, and Amir Kamran. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. Cited on page 49

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. Results of the WMT17 Neural MT Training Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. Cited on page 49

Ondřej Bojar, Tom Kocmi, David Mareček, Roman Sudarikov, and Dusan Varis. CUNI Submission in WMT17: Chimera Goes Neural. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. Cited on page 31

Ondřej Bojar. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 20, 21

Ondřej Bojar. *Exploiting Linguistic Data in Machine Translation.* PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, October 2008. Cited on page 29

Ondřej Bojar. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, March 2011. Cited on page 41

Ondřej Bojar. *Čeština a strojový překlad (Czech Language and Machine Translation)*, volume 11 of *Studies in Computational and Theoretical Linguistics.* ÚFAL, Praha, Czech Republic, 2012. Cited on page 11, 18, 19, 22, 39

Ondřej Bojar. *Machine translation*, chapter 13, pages 323–347. Oxford Handbooks in Linguistics. Oxford University Press, Oxford, UK, 2015. Cited on page 10

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007. Cited on page 26

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, June 1990. Cited on page 7, 8, 9

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June 1993. Cited on page 7, 9

Franck Burlot, Elena Knyazeva, Thomas Lavergne, and François Yvon. Two-Step MT: Predicting Target Morphology. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT'16, Seattle, USA, 2016. Cited on page 25

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 47

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Revised August 2010. Cited on page 46

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the

2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. Cited on page 44

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, 108:109, Jan 2017. Cited on page 44

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of Machine Translation Summit XVI*, Nagoya, Japan, 2017. Cited on page 44

Stanley F. Chen and Joshua Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. Cited on page 20

Colin Cherry and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. Cited on page 49

David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. Cited on page 29

David Chiang. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452, Uppsala, Sweden, July 2010. Association for Computational Linguistics. Cited on page 10, 29

Bonnie J. Dorr. Machine translation divergences: a formal description and proposed solution. *Comput. Linguist.*, 20(4):597–633, 1994. Cited on page 30

Markus Dreyer and Daniel Marcu. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, June 2012. Association for Computational Linguistics. Cited on page 39

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. Formemes in english-czech deep syntactic MT. In *Proceedings of NAACL 2012 Workshop on Machine Translation*, pages 267–274, Montréal, Canada, 2012. Association for Computational Linguistics. Cited on page 31

Ondřej Dušek, Jan Hajic, and Zdenka Uresova. Verbal valency frame detection and selection in czech and english. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. Cited on page 16

Jason Eisner. Learning Non-Isomorphic Tree Mappings for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, pages 205–208, Sapporo, July 2003. Cited on page 30

George Foster, Roland Kuhn, and Howard Johnson. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 53–61, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. Cited on page 20

Alexander Fraser. Experiments in morphosyntactic processing for translating to and from german. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 115–119, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. Cited on page 24

Eva Fučíková, Jan Hajič, and Zdeňka Urešová. Enriching a Valency Lexicon by Deverbative Nouns. In Eva Hajičová and Igor Boguslavsky, editors, *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces (GramLex)*, pages 71–80, -Osaka, Japan, 2016. ICCL, The COLING 2016 Organizing Committee. Cited on page 16

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1 2016. Cited on page 47

Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague, 2004. Cited on page 23

Jindřich Helcl and Jindřich Libovický. Neural Monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, 107:5–17, 2017. Cited on page 50

Mark Hopkins and Jonathan May. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Cited on page 47

Matthias Huck, Aleš Tamchyna, Ondřej Bojar, and Alexander Fraser. Producing Unseen Morphological Variants in Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 369–375, Stroudsburg, PA, USA, 2017. Universitat Politécnica de València, Association for Computational Linguistics. Cited on page 12, 17, 28

Bushra Jawaid and Ondřej Bojar. Two-step machine translation with lattices. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 682–686, Reykjavík, Iceland, 2014. European Language Resources Association. Cited on page 25

Bushra Jawaid, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. Results of the WMT16 Tuning Shared Task. In Ondřej Bojar and et al ., editors, *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*, volume 2, pages 232–238, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics, Association for Computational Linguistics. Cited on page 49

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal Neural Machine Translation Systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 130–136, Lisboa, Portugal, September 2015. Association for Computational Linguistics. Cited on page 34

Kevin Knight. Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615, 1999. Cited on page 10

Tom Kocmi and Ondřej Bojar. SubGram: Extending Skip-gram Word Representation with Substrings. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 182–189, Cham / Heidelberg / New York / Dordrecht / London, 2016. Masaryk University, Springer International Publishing. Cited on page 16

Philipp Koehn and Hieu Hoang. Factored Translation Models. In *Proc. of EMNLP*, 2007. Cited on page 20

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics. Cited on page 45

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT/NAACL*, 2003. Cited on page 9, 18

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. 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 2007. Association for Computational Linguistics. Cited on page 18

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009. Cited on page 11, 19

Philipp Koehn. Simulating Human Judgment in Machine Translation Evaluation Campaigns. In *Proc. of IWSLT*, pages 179–184, 2012. Cited on page 47

Nobo Komagata. Chance agreement and significance of the kappa statistic. http://nobo.komagata.net/pub/Komagata02-Kappa.pdf (as of Aug 2017), 2002. Cited on page 47

Jakub Kúdela, Irena Holubová, and Ondřej Bojar. Extracting parallel paragraphs from common crawl. *The Prague Bulletin of Mathematical Linguistics*, (107):36–59, 2017. Cited on page 16

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977. Cited on page 47

Chi-kiu Lo and Dekai Wu. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. Cited on page 44

Adam Lopez. Putting Human Assessments of Machine Translation Systems in Order. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 1–9, Montréal, Canada, 2012. Cited on page 47

Matouš Macháček and Ondřej Bojar. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MD, USA, 2014. Association for Computational Linguistics. Cited on page 49

Matouš Macháček and Ondřej Bojar. Approximating a Deep-Syntactic Metric for MT Evaluation and Tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 92–98, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. Cited on page 44

Matouš Macháček and Ondřej Bojar. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Cited on page 49

Jiří Maršík and Ondřej Bojar. TrTok: A Fast and Trainable Tokenizer for Natural Languages. *Prague Bulletin of Mathematical Linguistics*, 98:75–85, September 2012. Cited on page 16

Vendula Michlíková. Výslovnostní rysy češtiny - dialektová analýza. Bachelor Thesis. Charles University, 2013. Cited on page 16

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. Generating Complex Morphology for Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 24

Jan Niehues and Alex Waibel. Domain adaptation in statistical machine translation using factored translation models. In *EAMT*, 2010. Cited on page 20

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Translation of "It" in a Deep Syntax Framework. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 51–59, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Cited on page 16

Franz Josef Och and Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL*, pages 295–302, 2002. Cited on page 8

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc.*

of the Association for Computational Linguistics, Sapporo, Japan, July 6-7 2003. Cited on page 49

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. Cited on page 40, 49

Martin Popel and Zdeněk Žabokrtský. 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, 2010. Iceland Centre for Language Technology (ICLT), Springer. Cited on page 12, 15, 30

Rudolf Rosa, David Mareček, and Ondřej Dušek. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada, June 2012. Association for Computational Linguistics. Cited on page 31

Lane Schwartz. Multi-Source Translation Methods. In *Proc. of AMTA*, 2008. Cited on page 33

Holger Schwenk, Anthony Rousseau, and Mohammed Attik. Large, Pruned or Continuous Space Language Models on a GPU for Statistical Machine Translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, WLM '12, pages 11–19, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. Cited on page 19

Holger Schwenk. Continuous Space Language Models. *Comput. Speech Lang.*, 21(3):492–518, July 2007. Cited on page 19

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. Cited on page 26

Magda Ševčíková, Zdeněk Žabokrtský, Jonáš Vidra, and Milan Straka. Lexikální síť derinet: elektronický zdroj pro výzkum derivace v češtině. *Časopis pro moderní filologii*, 98(1):62–76, 2016. Cited on page 16

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986. Cited on page 11

Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková. Resources in conflict: A bilingual valency lexicon vs. a bilingual treebank vs. a linguistic theory. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2490–2494, Reykjavík, Iceland, 2014. European Language Resources Association. Cited on page 30

Miloš Stanojević, Amir Kamran, and Ondřej Bojar. Results of the WMT15 Tuning Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 274–281, Lisboa, Portugal, September 2015. Association for Computational Linguistics. Cited on page 49

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisboa, Portugal, September 2015. Association for Computational Linguistics. Cited on page 49

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18,

Baltimore, Maryland, June 2014. Association for Computational Linguistics. Cited on page 28

Roman Sudarikov, David Mareček, Tom Kocmi, Dušan Variš, and Ondřej Bojar. CUNI Submission in WMT17: Chimera Goes Neural. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. Cited on page 34

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. pages 3104–3112, 2014. Cited on page 12

Aleš Tamchyna and Ondřej Bojar. No Free Lunch in Factored Phrase-Based Machine Translation. In *Proc. of CICLing 2013*, volume 7817 of *LNCS*, pages 210–223, Samos, Greece, 2013. Springer-Verlag. Cited on page 22

Aleš Tamchyna and Ondřej Bojar. What a Transfer-Based System Brings to the Combination with PBMT. In Bogdan Babych, Kurt Eberle, Patrik Lambert, Reinhard Rapp, Rafael Banchs, and Marta Costa-Jussà, editors, *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 11–20, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics, Association for Computational Linguistics. Cited on page 31, 32, 33, 34

Aleš Tamchyna, Martin Popel, Rudolf Rosa, and Ondřej Bojar. CUNI in WMT14: Chimera still awaits bellerophon. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 195–200, Baltimore, MD, USA, 2014. Association for Computational Linguistics. Cited on page 31

Aleš Tamchyna, Alexander Fraser, Ondřej Bojar, and Marcin Junczys-Dowmunt. Target-Side Context for Discriminative Models in Statistical Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1704–1714. Association for Computational Linguistics, Association for Computational Linguistics, 2016. Cited on page 12, 28

Aleš Tamchyna, Roman Sudarikov, Ondřej Bojar, and Alexander Fraser. CUNI-LMU Submissions in WMT2016: Chimera Constrained and Beaten. In *Proceedings of the First Conference on Machine Translation*, pages 385–390, Berlin, Germany, August 2016. Association for Computational Linguistics. Cited on page 31

Aleš Tamchyna. *Lexical and Morphological Choices in Machine Translation*. PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, June 2017. Cited on page 12, 28

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June 2008. Association for Computational Linguistics. Cited on page 24

Kateřina Veselovská. *On the Linguistic Structure of Emotional Meaning in Czech*. PhD thesis, Charles University, Prague, 2015. Cited on page 16

John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017. Cited on page 16

Kenji Yamada and Kevin Knight. A Syntax-Based Statistical Translation Model. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530, Morristown, NJ, USA, 2001. Association for Computational Linguistics. Cited on page 10

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA, 2008. Cited on page 16, 31

Andreas Zollmann and Ashish Venugopal. Syntax Augmented Machine Translation via Chart Parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June 2006. Association for Computational Linguistics. Cited on page 10

Appendix A

Reprints of Key Papers of the Thesis

The printed version of this habilitation thesis includes the full texts of the key articles and papers published in the years 2007–2016 and supporting the research summary outlined in the main content of the thesis.

For copyright reasons, we do not reproduce the publications here and only give the full reference to them.

1. **Ondřej Bojar**. *Machine translation*, chapter 13, pages 323–347. Oxford Handbooks in Linguistics. Oxford University Press, Oxford, UK, 2015. doi:10.1093/oxfordhb/9780199591428.013.13.
   Citations: 2 (excluding self-citations)

2. **Ondřej Bojar**, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May 2012. ELRA, European Language Resources Association.
   Citations: 35 (excluding self-citations)
   Estimated contribution of the applicant: 40%. Ondřej Bojar organized the team, assembled the corpus and carried out the automatic processing using data and tools provided by co-authors.

3. **Ondřej Bojar**. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 232–239, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. doi:10.3115/1626355.1626390.
   Citations: 21 (excluding self-citations)

4. **Ondřej Bojar** and Aleš Tamchyna. The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58, 2013, doi:10.2478/pralin-2013-0003.
   Citations: 4 (excluding self-citations)
   Estimated contribution of the applicant: 90%. Ondřej Bojar is the main author of Eman and also provided most of the paper text.

5. **Ondřej Bojar** and Kamil Kos. 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 2010. Association for Computational Linguistics.
Citations: 7 (excluding self-citations)

6. **Ondřej Bojar** and Aleš Tamchyna. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.
Citations: 4 (excluding self-citations)

7. **Ondřej Bojar**, Rudolf Rosa, and Aleš Tamchyna. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
Citations: 7 (excluding self-citations)

8. **Ondřej Bojar**. Analyzing Error Types in English-Czech Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 95:63, 2011, doi:10.2478/v10108-011-0005-2.
Citations: 10 (excluding self-citations)

9. **Ondřej Bojar**, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. Scratching the Surface of Possible Translations, pages 465–474. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, doi:10.1007/978-3-642-40585-3_59.
Citations: 5 (excluding self-citations)

10. **Ondřej Bojar**, Kamil Kos, and David Mareček. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden, July 2010. Association for Computational Linguistics.
Citations: 6 (excluding self-citations)

11. **Ondřej Bojar**, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.
Citations: 27 (excluding self-citations)

12. **Ondřej Bojar**, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In Ondřej Bojar, Aljoscha Burchardt, Christian Dugast, Marcello Federico, Josef Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Georg Rehm, Lucia Specia,

Marco Turchi, and Hans Uszkoreit, editors, *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34, Portorož, Slovenia, 2016. [http://www.cracking-the-language-barrier.eu/], LREC.