# Machine Translation and Monolingual Postediting: The AFRL WMT-14 System

**Lane O.B. Schwartz**
Air Force Research Laboratory
lane.schwartz@us.af.mil

**Timothy Anderson**
Air Force Research Laboratory
timothy.anderson.20@us.af.mil

**Jeremy Gwinnup**
SRA International[†]
jeremy.gwinnup.ctr@us.af.mil

**Katherine M. Young**
N-Space Analysis LLC[†]
katherine.young.1.ctr@us.af.mil

## Abstract

This paper describes the AFRL statistical MT system and the improvements that were developed during the WMT14 evaluation campaign. As part of these efforts we experimented with a number of extensions to the standard phrase-based model that improve performance on Russian to English and Hindi to English translation tasks. In addition, we describe our efforts to make use of monolingual English speakers to correct the output of machine translation, and present the results of monolingual postediting of the entire 3003 sentences of the WMT14 Russian-English test set.

## 1 Introduction

As part of the 2014 Workshop on Machine Translation (WMT14) shared translation task, the human language technology team at the Air Force Research Laboratory participated in two language pairs: Russian-English and Hindi-English. Our machine translation system represents enhancements to our system from IWSLT 2013 (Kazi et al., 2013). In this paper, we focus on enhancements to our procedures with regard to data processing and the handling of unknown words.

In addition, we describe our efforts to make use of monolingual English speakers to correct the output of machine translation, and present the results of monolingual postediting of the entire 3003 sentences of the WMT14 Russian-English test set. Using a binary adequacy classification, we evaluate the entire postedited test set for correctness against the reference translations. Using bilingual judges, we further evaluate a substantial subset of the postedited test set using a more fine-grained adequacy metric; using this metric, we show that monolingual posteditors can successfully produce postedited translations that convey all or most of the meaning of the original source sentence in up to 87.8% of sentences.

## 2 System Description

We submitted systems for the Russian-to-English and Hindi-to-English MT shared tasks. In all submitted systems, we use the phrase-based `moses` decoder (Koehn et al., 2007). We used only the constrained data supplied by the evaluation for each language pair for training our systems.

### 2.1 Data Preparation

Before training our systems, a cleaning pass was performed on all data. Unicode characters in the unallocated and private use ranges were all removed, along with C0 and C1 control characters, zero-width and non-breaking spaces and joiners, directionality and paragraph markers.

#### 2.1.1 Hindi Processing

The HindEnCorp corpus (Bojar et al., 2014) is distributed in tokenized form; in order to ensure a uniform tokenization standard across all of our data, we began by detokenized this data using the Moses detokenization scripts. In addition to normalizing various extended Latin punctuation marks to their Basic Latin equivalents, following Bojar et al. (2010) we normalized DEVANAGARI DANDA (U+0964), DOUBLE DANDA (U+0965), and ABBREVIATION SIGN (U+0970) punctuation marks to Latin FULL STOP (U+002E), any DEVANA-

[†]This work is sponsored by the Air Force Research Laboratory under Air Force contract FA-8650-09-D-6939-029.

*Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

gari Digit to the equivalent ASCII Digit, and decomposed all Hindi data into Unicode Normalization Form D (Davis and Whistler, 2013) using `charlint`.[1] In addition, we performed Hindi diacritic and vowel normalization, following Larkey et al. (2003).

Since no Hindi-English development test set was provided in WMT14, we randomly sampled 1500 sentence pairs from the Hindi-English parallel training data to serve this purpose. Upon discovering duplicate sentences in the corpus, 552 sentences that overlapped with the training portion were removed from the sample, leaving a development test set of 948 sentences.

### 2.1.2 Russian Processing

The Russian sentences contained many examples of mixed-character spelling, in which both Latin and Cyrillic characters are used in a single word, relying on the visual similarity of the characters. For example, although the first letter and last letter in the word cейчас appear visually indistinguishable, we find that the former is U+0063 Latin Small Letter C and the latter is U+0441 Cyrillic Small Letter Es. We created a spelling normalization program to convert these words to all Cyrillic or all Latin characters, with a preference for all-Cyrillic conversion if possible. Normalization also removes U+0301 Combining Acute Accent ( ́) and converts U+00F2 Latin Small Letter O with Grave (ò) and U+00F3 Latin Small Letter O with Acute (ó) to the unaccented U+043E Cyrillic Small Letter O (о).

The Russian-English Common Crawl parallel corpus (Smith et al., 2013) is relatively noisy. A number of Russian source sentences are incorrectly encoded using characters in the Latin-1 supplement block; we correct these sentences by shifting these characters ahead by $350_{\text{hex}}$ code points into the correct Cyrillic character range.[2]

We examine the Common Crawl parallel sentences and mark for removal any non-Russian source sentences and non-English target sentences. Target sentences were marked as non-English if more than half of the characters in the sentence were non-Latin, or if more than half of the words were unknown to the `aspell` English spelling correction program, not counting short words, which frequently occur as (possibly false) cognates across languages (English *die* vs. German *die*, English *on* vs. French *on*, for example). Because `aspell` does not recognize some proper names, brand names, and borrowed words as known English words, this method incorrectly flags for removal some English sentences which have a high proportion of these types of words.

Source sentences were marked as non-Russian if less than one-third of the characters were within the Russian Cyrillic range, or if non-Russian characters equal or outnumber Russian characters and the sentence contains no contiguous sequence of at least three Russian characters. Some portions of the Cyrillic character set are not used in typical Russian text; source sentences were therefore marked for removal if they contained Cyrillic extension characters Ukrainian I (і І), Yi(ї Ї), Ghe With Upturn (ґ Ґ) or Ie (є Є) in either upper- or lowercase, with exceptions for U+0406 Ukrainian I (І) in Roman numerals and for U+0491 Ghe With Upturn (ґ) when it occurred as an encoding error artifact.[3]

Sentence pairs where the source was identified as non-Russian or the target was identified as non-English were removed from the parallel corpus. Overall, 12% of the parallel sentences were excluded based on a non-Russian source sentence (94k instances) or a non-English target sentence (11.8k instances).

Our Russian-English parallel training data includes a parallel corpus extracted from Wikipedia headlines (Ammar et al., 2013), provided as part of the WMT14 shared translation task. Two files in this parallel corpus (`wiki.ru-en` and `guessed-names.ru-en`) contained some overlapping data. We removed 6415 duplicate lines within `wiki.ru-en` (about 1.4%), and removed 94 lines of `guessed-names.ru-en` that were already present in wiki.ru-en (about 0.17%).

[1] `http://www.w3.org/International/charlint`

[2] For example: "Ñïðàâêà ïî ãîðîäàì Ðîññèè è ìèðà." becomes "Справка по городам России и мира."

[3] Specifically, we allowed lines containing ґ where it appears as an encoding error in place of an apostrophe within English words. For example: "Песня The Kelly Family Iґm So Happy представлена вам LyricsKeeper."

### 2.2 Machine Translation

Our baseline system is a variant of the MIT-LL/AFRL IWSLT 2013 system (Kazi et al., 2013) with some modifications to the training and decoding processes.

#### 2.2.1 Phrase Table Training

For our Russian-English system, we trained a phrase table using the Moses Experiment Management System (Koehn, 2010b), with `mgiza` (Gao and Vogel, 2008) as the word aligner; this phrase table was trained using the Russian-English Common Crawl, News Commentary, Yandex (Bojar et al., 2013), and Wikipedia headlines parallel corpora.

The phrase table for our Hindi-English system was trained using a similar in-house training pipeline, making use of the HindEnCorp and Wikipedia headlines parallel corpora.

#### 2.2.2 Language Model Training

During the training process we built $n$-gram language models (LMs) for use in decoding and rescoring using the KenLM language modelling toolkit (Heafield et al., 2013). Class-based language models (Brown et al., 1992) were also trained, for later use in $n$-best list rescoring, using the SRILM language modelling toolkit (Stolcke, 2002).We trained a 6-gram language model from the LDC English Gigaword Fifth Edition, for use in both the Hindi-English and Russian-English systems. All language models were binarized in order to reduce model disk usage and loading time.

For the Russian-to-English task, we concatenated the English portion of the parallel training data for the WMT 2014 shared translation task (Common Crawl, News Commentary, Wiki Headlines and Yandex corpora) in addition to the shared task English monolingual training data (Europarl, News Commentary and News Crawl corpora) into a training set for a large 6-gram language model using KenLM. We denote this model as “BigLM”. Individual 6-gram models were also constructed from each respective corpus.

For the Hindi-to-English task, individual 6-gram models were constructed from the respective English portions of the HindEnCorp and Wikipedia headlines parallel corpora, and from the monolingual English sections of the Europarl and News Crawl corpora.

| **Decoding Features** |
|---|
| $\mathsf{P}(\mathbf{f}\,\vert\,\mathbf{e})$ |
| $\mathsf{P}(\mathbf{e}\,\vert\,\mathbf{f})$ |
| $\mathsf{P}_w(\mathbf{f}\,\vert\,\mathbf{e})$ |
| $\mathsf{P}_w(\mathbf{e}\,\vert\,\mathbf{f})$ |
| Phrase Penalty |
| Lexical Backoff |
| Word Penalty |
| Distortion Model |
| Unknown Word Penalty |
| Lexicalized Reordering Model |
| Operation Sequence Model |
| **Rescoring Features** |
| $\mathsf{P}_{class}(\mathbf{E})$ – 7-gram class-based LM |
| $\mathsf{P}_{lex}(\mathbf{F}\,\vert\,\mathbf{E})$ – sentence-level averaged lexical translation score |

Table 1: Models used in log-linear combination

#### 2.2.3 Decoding, $n$-best List Rescoring, and Optimization

We decode using the phrase-based `moses` decoder (Koehn et al., 2007), choosing the best translation for each source sentence according to a linear combination of decoding features:

$$\hat{\mathbf{E}} = \arg\max_{\mathbf{E}} \sum_{\forall r} \lambda_r h_r(\mathbf{E}, \mathbf{F}) \qquad (1)$$

We make use of a standard set of decoding features, listed in Table 1. In contrast to our IWSLT 2013 system, all experiments submitted to this year’s WMT evaluation made use of version 2.1 of `moses`, and incorporated additional decoding features, namely the Operation Sequence Model (Durrani et al., 2011) and Lexicalized Reordering Model (Tillman, 2004; Galley and Manning, 2008).

Following Shen et al. (2006), we use the word-level lexical translation probabilities $\mathsf{P}_w(f_j\,|\,e_i)$ to obtain a sentence-level averaged lexical translation score (Eq. 2), which is added as an additional feature to each $n$-best list entry.

$$\mathsf{P}_{lex}(\mathbf{F}\,|\,\mathbf{E}) = \prod_{j \in 1 \ldots J} \frac{1}{I+1} \sum_{i \in 1 \ldots I} \mathsf{P}_w(f_j\,|\,e_i) \qquad (2)$$

Shen et al. (2006) use the term “IBM model 1 score” to describe the value calculated in Eq. 2. While the lexical probability distribution from IBM Model 1 (Brown et al., 1993) could in fact be used as the $\mathsf{P}_w(f_j \,|\, e_i)$ in Eq. 2, in practice we use a variant of $\mathsf{P}_w(f_j \,|\, e_i)$ defined by Koehn et al. (2003).

We also add a 7-gram class language model score $\mathsf{P}_{class}(\mathbf{E})$ (Brown et al., 1992) as an additional feature of each $n$-best list entry. After adding these features to each translation in an $n$-best list, Eq. 1 is applied, rescoring the entries to extract new 1-best translations.

To optimize system performance we train scaling factors, $\lambda_r$, for both decoding and rescoring features so as to minimize an objective error criterion. In our systems we use DREM (Kazi et al., 2013) or PRO (Hopkins and May, 2011) to perform this optimization. For development data during optimization, we used `newstest2013` for the Russian-to-English task and `newsdev2014` for the Hindi-to-English task supplied by WMT14.

### 2.2.4 Unknown Words

For the Hindi-to-English task, unknown words were marked during the decoding process and were transliterated by the icu4j Devanagari-to-Latin transliterator.[4]

For the Russian-to-English task, we selectively stemmed and inflected input words not found in the phrase table. Each input sentence was examined to identify any source words which did not occur as a phrase of length 1 in the phrase table. For each such unknown word, we used `treetagger` (Schmid, 1994; Schmid, 1995) to identify the part of speech, and then we removed inflectional endings to derive a stem. We applied all possible Russian inflectional endings for the given part of speech; if an inflected form of the unknown word could be found as a stand-alone phrase in the phrase table, that form was used to replace the unknown word in the original Russian file. If multiple candidates were found, we used the one with the highest frequency of occurrence in the training data. This process replaces words that we know we cannot translate with semantically similar words that we can translate, replacing unknown words like фотоном "photon" (instrumental case) with a known morphological variant фотон "photon" (nominative case) that is found in the phrase table. Selective stemming of just the unknown words allows us to retain information that would be lost if we applied stemming to all the data.

| | | | BLEU | BLEU-cased |
|---|---|---|---|---|
| System | 1 | hi-en | 13.1 | 12.1 |
| | 2 | ru-en | 32.0 | 30.8 |
| | 3 | ru-en | 32.2 | 31.0 |
| | 4 | ru-en | 31.5 | 30.3 |
| | 5 | ru-en | 33.0 | 31.1 |

Table 2: Translation results, as measured by BLEU (Papineni et al., 2002).

Any remaining unknown words were transliterated as a post-process, using a simple letter-mapping from Cyrillic characters to Latin characters representing their typical sounds.

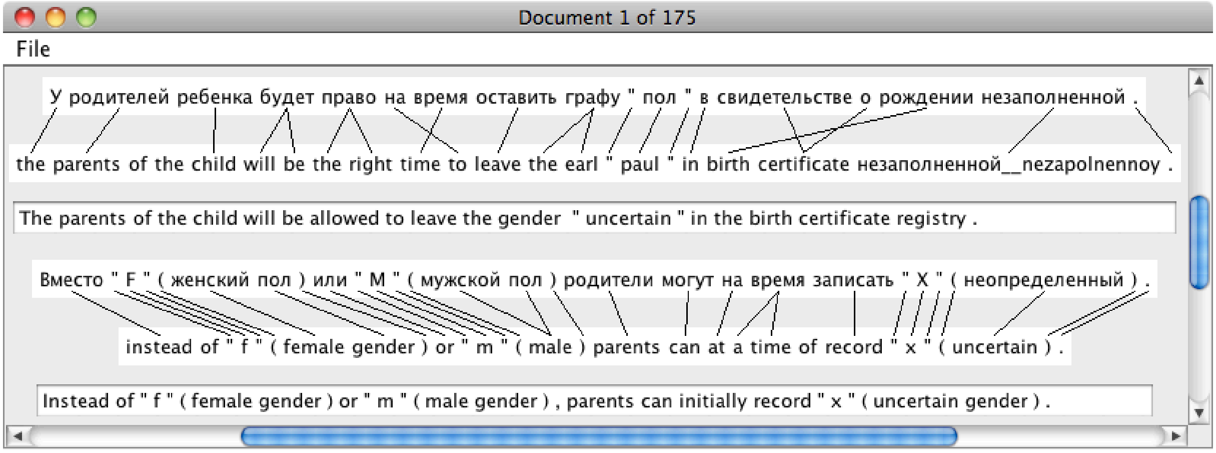
## 2.3 MT Results

Our best Hindi-English system for `newstest2014` is listed in Table 2 as System 1. This system uses a combination of 6-gram language models built from HindEnCorp, News Commentary, Europarl, and News Crawl corpora. Transliteration of unknown words was performed after decoding but before $n$-best list rescoring.

System 2 is Russian-English, and handles unknown words following §2.2.4. We used as independent decoder features separate 6-gram LMs trained respectively on Common Crawl, Europarl, News Crawl, Wiki headlines and Yandex corpora. This system was optimized with DREM. No rescoring was performed. We also tested a variant of System 2 which did perform rescoring. That variant (not listed in Table 2) performed worse than System 2, with scores of 31.2 BLEU and 30.1 BLEU-cased.

System 3, our best Russian-English system for `newstest2014`, used the BigLM and Gigaword language models (see §2.2.2) as independent decoder features and was optimized with DREM. Rescoring was performed after decoding. Instead of following §2.2.4, unknown words were dropped to maximize BLEU score. We note that the optimizer assigned weights of 0.314 and 0.003 to the BigLM and Gigaword models, respectively, suggesting that the optimizer found the BigLM to be much more use-

[4] `http://site.icu-project.org`



У родителей ребенка будет право на время оставить графу " пол " в свидетельстве о рождении незаполненной .

the parents of the child will be the right time to leave the earl " paul " in birth certificate незаполненной__nezapolnennoy .

The parents of the child will be allowed to leave the gender " uncertain " in the birth certificate registry .

Вместо " F " ( женский пол ) или " M " ( мужской пол ) родители могут на время записать " X " ( неопределенный ) .

instead of " f " ( female gender ) or " m " ( male ) parents can at a time of record " x " ( uncertain ) .

Instead of " f " ( female gender ) or " m " ( male gender ) , parents can initially record " x " ( uncertain gender ) .

Figure 1: Posteditor user interface

| | | Documents | Sentences | Words |
|---|---|---|---|---|
| Posteditor | 1 | 44 | 950 | 20086 |
| | 2 | 21 | 280 | 6031 |
| | 3 | 25 | 476 | 10194 |
| | 4 | 25 | 298 | 6164 |
| | 5 | 20 | 301 | 5809 |
| | 6 | 15 | 210 | 4433 |
| | 7 | 10 | 140 | 2650 |
| | 8 | 15 | 348 | 6743 |
| | All | 175 | 3003 | 62110 |

Table 3: Number of documents within the Russian-English test set processed by each monolingual human posteditor. Number of machine translated sentences processed by each posteditor is also listed, along with the total number of words in the corresponding Russian source sentences.

| | | # ✓ | # ✗ | % ✓ |
|---|---|---|---|---|
| Posteditor | 1 | 684 | 266 | 72.0% |
| | 2 | 190 | 90 | 67.9% |
| | 3 | 308 | 168 | 64.7% |
| | 4 | 162 | 136 | 54.4% |
| | 5 | 194 | 107 | 64.5% |
| | 6 | 94 | 116 | 44.8% |
| | 7 | 88 | 52 | 62.9% |
| | 8 | 196 | 152 | 56.3% |
| | All | 1916 | 1087 | 63.8% |

Table 4: For each monolingual posteditor, the number and percentage of sentences judged to be correct (✓) versus incorrect (✗) according to a monolingual human judge.[6]

| | |
|---|---|
| 12 | The postedited translation is superior to the reference translation |
| 10 | The meaning of the Russian source sentence is fully conveyed in the postedited translation |
| 8 | Most of the meaning is conveyed |
| 6 | Misunderstands the sentence in a major way; or has many small mistakes |
| 4 | Very little meaning is conveyed |
| 2 | The translation makes no sense at all |

Table 5: Evaluation guidelines for bilingual human judges, adapted from Albrecht et al. (2009).

| Evaluation Category | | | | | |
|---|---|---|---|---|---|
| 2 | 4 | 6 | 8 | 10 | 12 |
| 0.2% | 2.2% | 9.8% | 24.7% | 60.2% | 2.8% |

Table 6: Percentage of evaluated sentences judged to be in each category by a bilingual judge. Category labels are defined in Table 5.

| | Evaluation Category | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 4 | 6 | 8 | 10 | 12 |
| # ✗ | 2 | 20 | 72 | 89 | 79 | 4 |
| # ✓ | 0 | 1 | 21 | 146 | 493 | 23 |
| % ✓ | 0% | 5% | 23% | 62% | 86% | 85% |

Table 7: Number of sentences in each evaluation category (see Table 5) that were judged as correct (✓) or incorrect (✗) according to a monolingual human judge.

ful than the Gigaword LM. This intuition was confirmed by an experimental variation of System 3 (not listed in Table 2) where we omitted the BigLM; that variant performed substantially worse, with scores of 25.3 BLEU and 24.2 BLEU-cased. We also tested a variant of System 3 which did not perform rescoring; that variant (also not listed in Table 2) performed worse, with scores of 31.7 BLEU and 30.6 BLEU-cased.

The results of monolingual postediting (see §3) of System 4 (a variant of System 2 tuned using PRO) uncased output is System 5. Due to time constraints, the monolingual postediting experiments in §3 were conducted (using the machine translation results from System 4) before the results of Systems 2 and 3 were available. The Moses recaser was applied in all experiments except for System 5.

## 3 Monolingual Postediting

Postediting is the process whereby a human user corrects the output of a machine translation system. The use of basic postediting tools by bilingual human translators has been shown to yield substantial increases in terms of productivity (Plitt and Masselot, 2010) as well as improvements in translation quality (Green et al., 2013) when compared to bilingual human translators working without assistance from machine translation and postediting tools. More sophisticated interactive interfaces (Langlais et al., 2000; Barrachina et al., 2009; Koehn, 2009b; Denkowski and Lavie, 2012) may also provide benefit (Koehn, 2009a).

We hypothesize that for at least some language pairs, monolingual posteditors with no knowledge of the source language can successfully translate a substantial fraction of test sentences. We expect this to be the case especially when the monolingual humans are domain experts with regard to the documents to be translated. If this hypothesis is confirmed, this could allow for multi-stage translation workflows, where less highly skilled monolingual posteditors triage the translation process, postediting many of the sentences, while forwarding on the most difficult sentences to more highly skilled bilingual translators.

Small-scale studies have suggested that monolingual human posteditors, working without knowledge of the source language, can also improve the quality of machine translation output (Callison-Burch, 2005; Koehn, 2010a; Mitchell et al., 2013), especially if well-designed tools provide automated linguistic analysis of source sentences (Albrecht et al., 2009).

In this study, we designed a simple user interface for postediting that presents the user with the source sentence, machine translation, and word alignments for each sentence in a test document (Figure 1). While it may seem counter-intuitive to present monolingual posteditors with the source sentence, we found that the presence of alignment links between source words and target words can in fact aid a monolingual posteditor, especially with regard to correcting word order. For example, in our experiments posteditors encountered some sentences where a word or phrase was enclosed within bracketing punctuation marks (such as quotation marks, commas, or parentheses) in the source sentence, and the machine translation system incorrectly reordered the word or phrase outside the enclosing punctuation; by examining the alignment links the posteditors were able to correct such reordering mistakes.

The Russian-English test set comprises 175 documents in the news domain, totaling 3003 sentences. We assigned each test document to one of 8 monolingual[5] posteditors (Table 3). The postediting tool did not record timing information. However, several posteditors informally reported that they were able to process on average approximately four documents per hour; if accurate, this would indicate a processing speed of around one sentence per minute.

Following Koehn (2010a), we evaluated postedited translation quality according to a binary adequacy metric, as judged by a monolingual English speaker[6] against the En-

[5] All posteditors are native English speakers. Posteditors 2 and 3 know Chinese and Arabic, respectively, but not Russian. Posteditor 8 understands the Cyrillic character set and has a minimal Russian vocabulary from two undergraduate semesters of Russian taken several years ago.

[6] All monolingual adequacy judgements were performed by Posteditor 1. Additional analysis of Posteditor 1's 950 postedited translations were independently judged by bilingual judges against the reference and the source sentence (Table 7).

glish references. In this metric, incorrect spellings of transliterated proper names were not grounds to judge as incorrect an otherwise adequate postedited translation. Binary adequacy results are shown in Table 4; we observe that correctness varied widely between posteditors (44.8–72.0%), and between documents.

Interestingly, several posteditors self-reported that they could tell which documents were originally written in English and were subsequently translated into Russian, and which were originally written in Russian, based on observations that sentences from the latter were substantially more difficult to postedit. Once per-document source language data is released by WMT14 organizers, we intend to examine translation quality on a per-document basis and test whether posteditors did indeed perform worse on documents which originated in Russian.

Using bilingual judges, we further evaluate a substantial subset of the postedited test set using a more fine-grained adequacy metric (Table 5). Because of time constraints, only the first 950 postedited sentences of the test set[6] were evaluated in this manner. Each sentence was evaluated by one of two bilingual human judges. In addition to the 2-10 point scale of Albrecht et al. (2009), judges were instructed to indicate (with a score of 12) any sentences where the postedited machine translation was superior to the reference translation. Using this metric, we show in Table 6 that monolingual posteditors can successfully produce postedited translations that convey all or most of the meaning of the original source sentence in up to 87.8% of sentences; this includes 2.8% which were superior to the reference.

Finally, as part of WMT14, the results of our Systems 1 (hi-en), 3 (ru-en), and 5 (postedited ru-en) were ranked by monolingual human judges against the machine translation output of other WMT14 participants. These judgements are reported in WMT (2014).

Due to time constraints, the machine translations (from System 4) presented to posteditors were not evaluated by human judges, neither using our 12-point evaluation scale nor as part of the WMT human evaluation rankings. However, to enable such evaluation by future researchers, and to enable replication of our experimental evaluation, the System 4 machine translations, the postedited translations, and the monolingual and bilingual evaluation results are released as supplementary data to accompany this paper.

## 4 Conclusion

In this paper, we present data preparation and language-specific processing techniques for our Hindi-English and Russian-English submissions to the 2014 Workshop on Machine Translation (WMT14) shared translation task. Our submissions examine the effectiveness of handling various monolingual target language corpora as individual component language models (System 2) or alternatively, concatenated together into a single big language model (System 3). We also examine the utility of $n$-best list rescoring using class language model and lexicalized translation model rescoring features.

In addition, we present the results of monolingual postediting of the entire 3003 sentences of the WMT14 Russian-English test set. Postediting was performed by monolingual English speakers, who corrected the output of machine translation without access to external resources, such as bilingual dictionaries or online search engines. This system scored highest according to BLEU of all Russian-English submissions to WMT14.

Using a binary adequacy classification, we evaluate the entire postedited test set for correctness against the reference translations. Using bilingual judges, we further evaluate a substantial subset of the postedited test set using a more fine-grained adequacy metric; using this metric, we show that monolingual posteditors can successfully produce postedited translations that convey all or most of the meaning of the original source sentence in up to 87.8% of sentences.

## Acknowledgements

We would like to thank the members of the SCREAM group at Wright-Patterson AFB.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 1 Apr 2014. Originator reference number RH-14-112150. Case number 88ABW-2014-1328.

## References

Joshua S. Albrecht, Rebecca Hwa, and G. Elisabeta Marai. 2009. Correcting automatic translations through collaborations between MT and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pages 60–68, Athens, Greece, March–April.

Waleed Ammar, Victor Chahuneau, Michael Denkowski, Greg Hanneman, Wang Ling, Austin Matthews, Kenton Murray, Nicola Segall, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2013. The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT '13)*, pages 70–77, Sofia, Bulgaria, August.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28, March.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2010. Data issues in English-to-Hindi machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, pages 1771–1777, Valletta, Malta, May.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT '13)*, pages 1–44, Sofia, Bulgaria, August.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Aleš Tamchyna, and Dan Zeman. 2014. Hindi-English and Hindi-only corpus for machine translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC '14)*, Reykjavik, Iceland, May. ELRA, European Language Resources Association.

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

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Chris Callison-Burch. 2005. Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Mark Davis and Ken Whistler. 2013. Unicode normalization forms. Technical Report UAX #15, The Unicode Consortium, September. Rev. 39.

Michael Denkowski and Alon Lavie. 2012. TransCenter: Web-based translation research suite. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*, November.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pages 1045–1054, Portland, Oregon, June.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pages 848–856, Honolulu, Hawai'i, October.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 439–448, Paris, France, April–May.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 690–696, Sofia, Bulgaria, August.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1352–1362, Edinburgh, Scotland, U.K.

Michaeel Kazi, Michael Coury, Elizabeth Salesky, Jessica Ray, Wade Shen, Terry Gleason, Tim Anderson, Grant Erdmann, Lane Schwartz, Brian Ore, Raymond Slyh, Jeremy Gwinnup, Katherine Young, and Michael Hutt. 2013. The MIT-LL/AFRL IWSLT-2013 MT system. In *The 10th International Workshop on Spoken Language Translation (IWSLT '13)*, pages 136–143, Heidelberg, Germany, December.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '13)*, pages 48–54, Edmonton, Canada, May–June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07) Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Philipp Koehn. 2009a. A process study of computer aided translation. *Machine Translation*, 23(4):241–263, November.

Philipp Koehn. 2009b. A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore, August.

Philipp Koehn. 2010a. Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '10)*, pages 537–545, Los Angeles, California, June.

Philipp Koehn. 2010b. An experimental management system. *The Prague Bulletin of Mathematical Linguistics*, 94:87–96, December.

Philippe Langlais, George Foster, and Guy Lapalme. 2000. TransType: A computer-aided translation typing system. In *Proceedings of the ANLP/NAACL 2000 Workshop on Embedded Machine Translation Systems*, pages 46–51, Seattle, Washington, May.

Leah S. Larkey, Margaret E. Connell, and Nasreen Abduljaleel. 2003. Hindi CLIR in thirty days. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):130–142, June.

Linda Mitchell, Johann Roturier, and Sharon O'Brien. 2013. Community-based post-editing of machine translation content: monolingual vs. bilingual. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 35–43, Nice, France, September. EAMT.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 311–318, Philadelphia, Pennsylvania, July.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16, January.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, September.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT Workshop*, Dublin, Ireland, March.

Wade Shen, Brian Delaney, and Tim Anderson. 2006. The MIT-LL/AFRL IWSLT-2006 MT system. In *The 3rd International Workshop on Spoken Language Translation (IWSLT '06)*, Kyoto, Japan.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*, pages 1374–1383, Sofia, Bulgaria, August.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pages 901–904, Denver, Colorado, September.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '04), Companion Volume*, pages 101–104, Boston, Massachusetts, May.

WMT. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT '14)*, Baltimore, Maryland, June.

# CUNI in WMT14: Chimera Still Awaits Bellerophon

**Aleš Tamchyna, Martin Popel, Rudolf Rosa, Ondřej Bojar**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
surname@ufal.mff.cuni.cz

## Abstract

We present our English→Czech and English→Hindi submissions for this year's WMT translation task. For English→Czech, we build upon last year's CHIMERA and evaluate several setups. English→Hindi is a new language pair for this year. We experimented with reverse self-training to acquire more (synthetic) parallel data and with modeling target-side morphology.

## 1 Introduction

In this paper, we describe translation systems submitted by Charles University (CU or CUNI) to the Translation task of the Ninth Workshop on Statistical Machine Translation (WMT) 2014.

In §2, we present our English→Czech systems, CU-TECTOMT, CU-BOJAR, CU-DEPFIX and CU-FUNKY. The systems are very similar to our submissions (Bojar et al., 2013) from last year, the main novelty being our experiments with domain-specific and document-specific language models.

In §3, we describe our experiments with English→Hindi translation, which is a translation pair new both to us and to WMT. We unsuccessfully experimented with reverse self-training and a morphological-tags-based language model, and so our final submission, CU-MOSES, is only a basic instance of Moses.

## 2 English→Czech

Our submissions for English→Czech build upon last year's successful CHIMERA system (Bojar et al., 2013). We combine several different approaches:

- factored phrase-based Moses model (§2.1),
- domain-adapted language model (§2.2),
- document-specific language models (§2.3),
- deep-syntactic MT system TectoMT (§2.4),
- automatic post-editing system Depfix (§2.5).

We combined the approaches in several ways into our four submissions, as made clear by Table 1. CU-TECTOMT is the stand-alone TectoMT translation system, while the other submissions are Moses-based, using TectoMT indirectly to provide an additional phrase-table. CU-BOJAR uses a factored model and a domain-adapted language model; in CU-DEPFIX, Depfix post-processing is added; and CU-FUNKY also employs document-specific language models.

| | CU-TECTOMT | CU-BOJAR | CU-DEPFIX | CU-FUNKY |
|---|---|---|---|---|
| TectoMT (§2.4) | ✓ | ✓ | ✓ | ✓ |
| Factored Moses (§2.1) | | ✓ | ✓ | ✓ |
| Adapted LM (§2.2) | | ✓ | ✓ | ✓ |
| Document-specific LMs (§2.3) | | | | ✓ |
| Depfix (§2.5) | | | ✓ | ✓ |

Table 1: EN→CS systems submitted to WMT.

### 2.1 Our Baseline Factored Moses System

Our baseline translation system (denoted "Baseline" in the following) is similar to last year – we trained a factored Moses model on the concatenation of CzEng (Bojar et al., 2012) and Europarl (Koehn, 2005), see Table 2. We use two factors: tag, which is the part-of-speech tag, and stc, which is "supervised truecasing", i.e. the surface form with letter case set according to the lemma; see (Bojar et al., 2013). Our factored Moses system translates from English stc to Czech stc | tag in one translation step.

Our basic language models are identical to last year's submission. We added an adapted language

*Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 195–200,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics