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

Lane O.B. Schwartz*, **Timothy Anderson***, **Jeremy Gwinnup**†, **Katherine M. Young**‡

*Air Force Research Laboratory, †SRA International, ‡N-Space Analysis LLC

lane.schwartz@us.af.mil, timothy.anderson.20@us.af.mil, jeremy.gwinnup.ctr@us.af.mil, 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. System Description

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.

Data Preparation

Clean data by removing certain Unicode characters:

- Characters in unallocated ranges
- Characters in private use ranges
- C0 and C1 control characters
- Zero-width and non-breaking spaces and joiners
- Directionality and paragraph markers

Hindi Processing

The HindEnCorp corpus is distributed in tokenized form; in order to ensure a uniform tokenization standard across all of our data, after cleaning the data we detokenize using the Moses detokenization scripts.

We normalize punctuation:

Devanagari Danda
Devanagari Double Danda } ⇒ Latin Full Stop
Devanagari Abbreviation Sign

Hindi data was decomposed into Unicode Normalization Form D. Finally, we performed spelling and number normalization.

Devanagari Digits ⇒ ASCII Digits

Transliterate Hindi OOVs

Unknown Hindi words were marked during the decoding process and were transliterated by the icu4j Devanagari-to-Latin transliterator.

Russian Processing

For mixed Cyrillic-Latin words in the input, a spelling map was applied to convert to either all-Cyrillic or all-Latin letters depending on the majority of the letters in that word. Instances of Combining Acute Accent were also removed.

Latin Small Letter O with Grave
Latin Small Letter O with Acute } ⇒ Cyrillic Small Letter O

Stem, then Inflect

We selectively stemmed and inflected Russian 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 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. By replacing unknown words with morphological variants from the phrase table, we replace words that we know we cannot translate with semantically similar words that we can translate. Selective stemming of just the unknown words allows us to retain information that would be lost if we applied stemming to all the data.

Transliterate Russian OOVs

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. MT Results

			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 1: *Translation results, as measured by BLEU.*

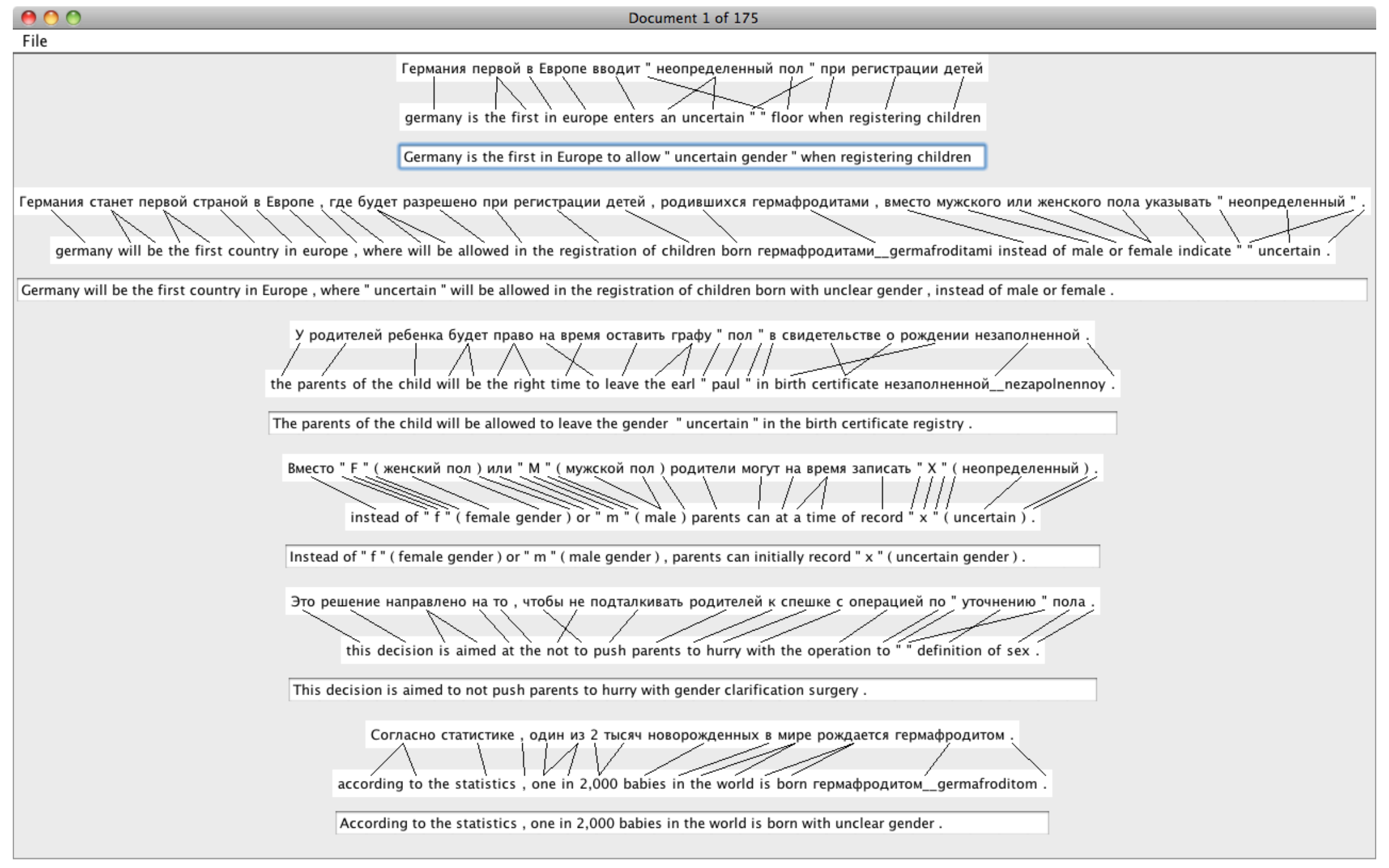
Our best Hindi-English system for newstest2014 is listed in Table 1 as **System 1**. This system uses a combination of 6-gram language models built from HindEnCorp, News Commentary, Europarl, and News Crawl corpora. Transliteration of unknown words was performed after decoding but before n-best list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

System 2 is Russian-English, and includes selective stem-and-inflect processing of unknown words. 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.

System 3 , our best Russian-English system for newstest2014, used the BigLM (a 6-gram model trained with KenLM on the concatenation of all available English monolingual data as well as the English portion of the parallel training data) and Gigaword language models as independent decoder features and was optimized with DREM. Rescoring was performed after decoding. 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 useful than the Gigaword LM.

System 4 and **System 5** reflect before and after monolingual processing, illustrating the positive impact of this approach. These systems are variants of **System 2** tuned using PRO instead of DREM, do not include rescoring and only utilize the BigLM.

3. Monolingual Postediting



Германия первой в Европе вводит " неопределенный пол " при регистрации детей
germany is the first in europe enters an uncertain " " floor when registering children
Germany is the first in Europe to allow " uncertain gender " when registering children

Германия станет первой страной в Европе , где будет разрешено при регистрации детей , родившихся гермафродитами , вместо мужского или женского пола указывать " неопределенный " .
germany will be the first country in europe , where will be allowed in the registration of children born гермафродитами__germafroditami instead of male or female indicate " " uncertain .
Germany will be the first country in Europe , where " uncertain " will be allowed in the registration of children born with unclear gender , instead of male or female .

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 .

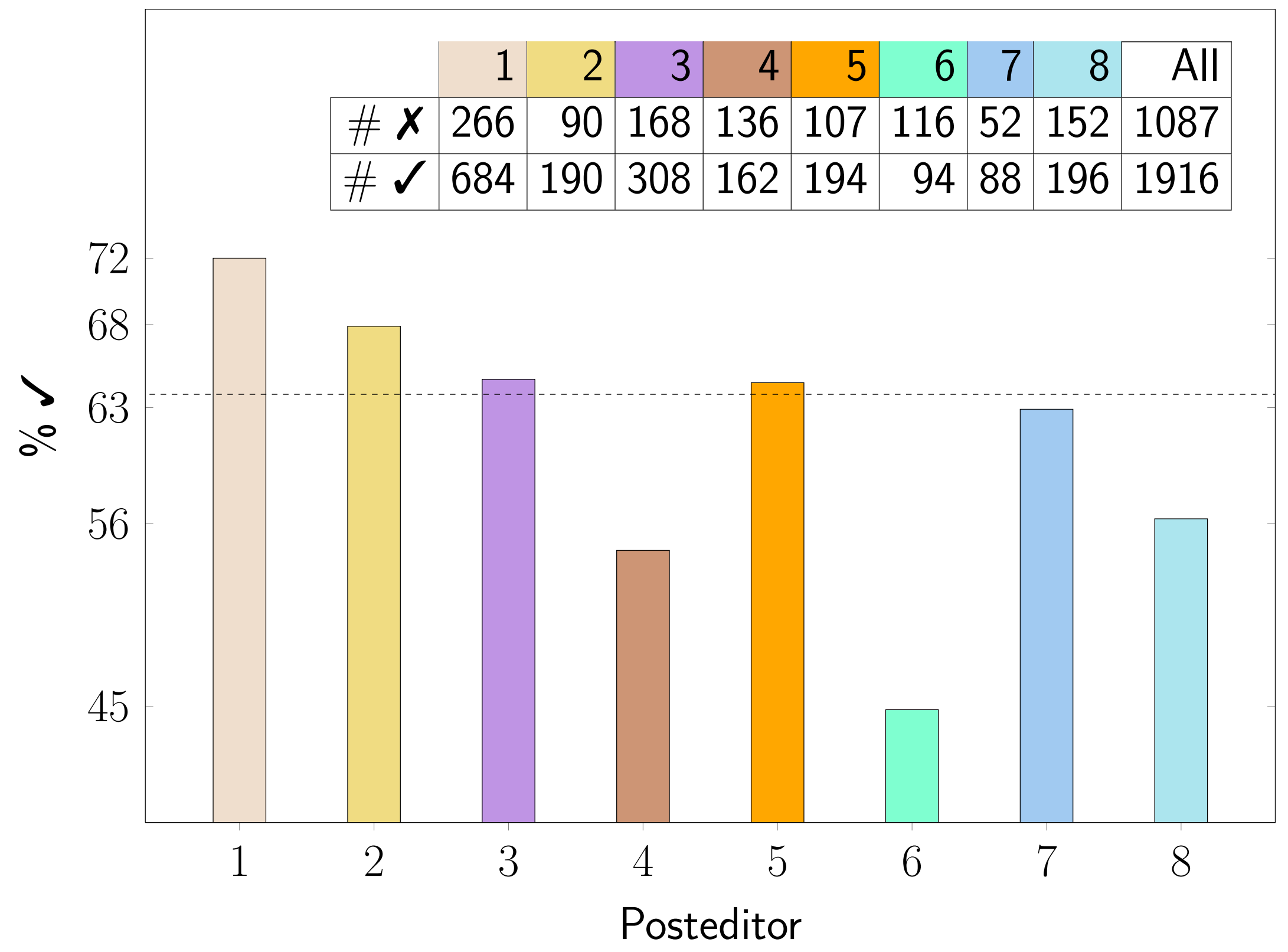
Вместо " F " (женский пол) или " M " (мужской пол) родители могут на время записать " X " (неопределенный) .
instead of " f " (female gender) or " m " (male) parents can at a time of record " x " (uncertain) .
Instead of " f " (female gender) or " m " (male gender) , parents can initially record " x " (uncertain gender) .

Это решение направлено на то , чтобы не подталкивать родителей к спешке с операцией по " уточнению " пола .
this decision is aimed at the not to push parents to hurry with the operation to " " definition of sex .
This decision is aimed to not push parents to hurry with gender clarification surgery .

Согласно статистике , один из 2 тысяч новорожденных в мире рождается гермафродитом .
according to the statistics , one in 2,000 babies in the world is born гермафродитом__germafroditom .
According to the statistics , one in 2,000 babies in the world is born with unclear gender .

Figure 1: *Posteditor user interface*

Monolingual English speakers corrected the output of Russian-English machine translation, postediting 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.



	1	2	3	4	5	6	7	8	All
# ✗	266	90	168	136	107	116	52	152	1087
# ✓	684	190	308	162	194	94	88	196	1916

Figure 2: *For each monolingual posteditor, the number (#) and percentage (%) of postedited sentence translations judged to be correct (✓) versus incorrect (✗) according to a monolingual human judge. Dashed line indicates the overall percentage of all postedited sentences judged to be correct.*

12	The postedited translation is superior to the reference translation
10	The meaning of the Russian source sentence is fully conveyed in the post-edited 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 2: *Evaluation guidelines for bilingual human judges, adapted from Albrecht et al (2009).*

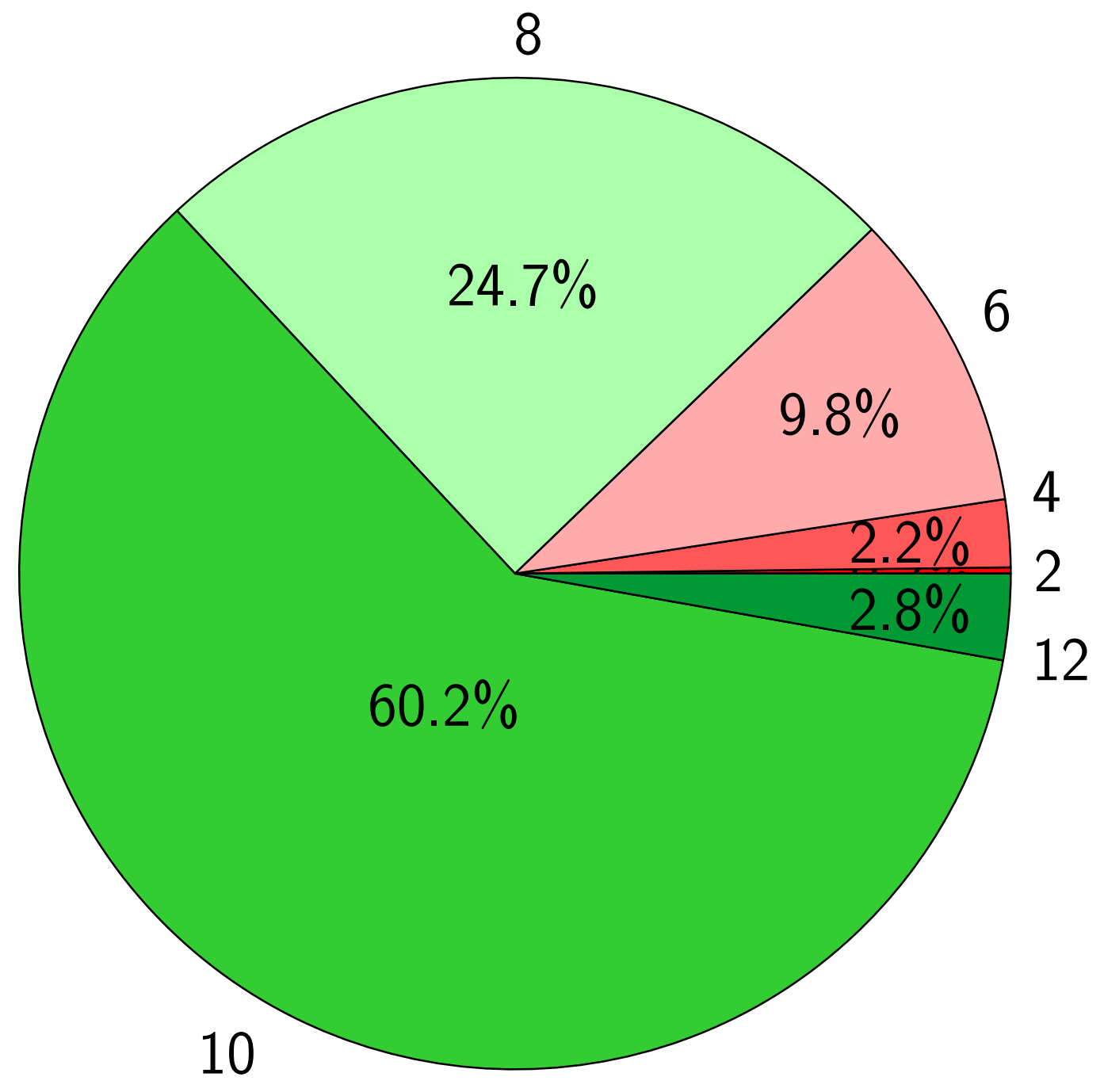
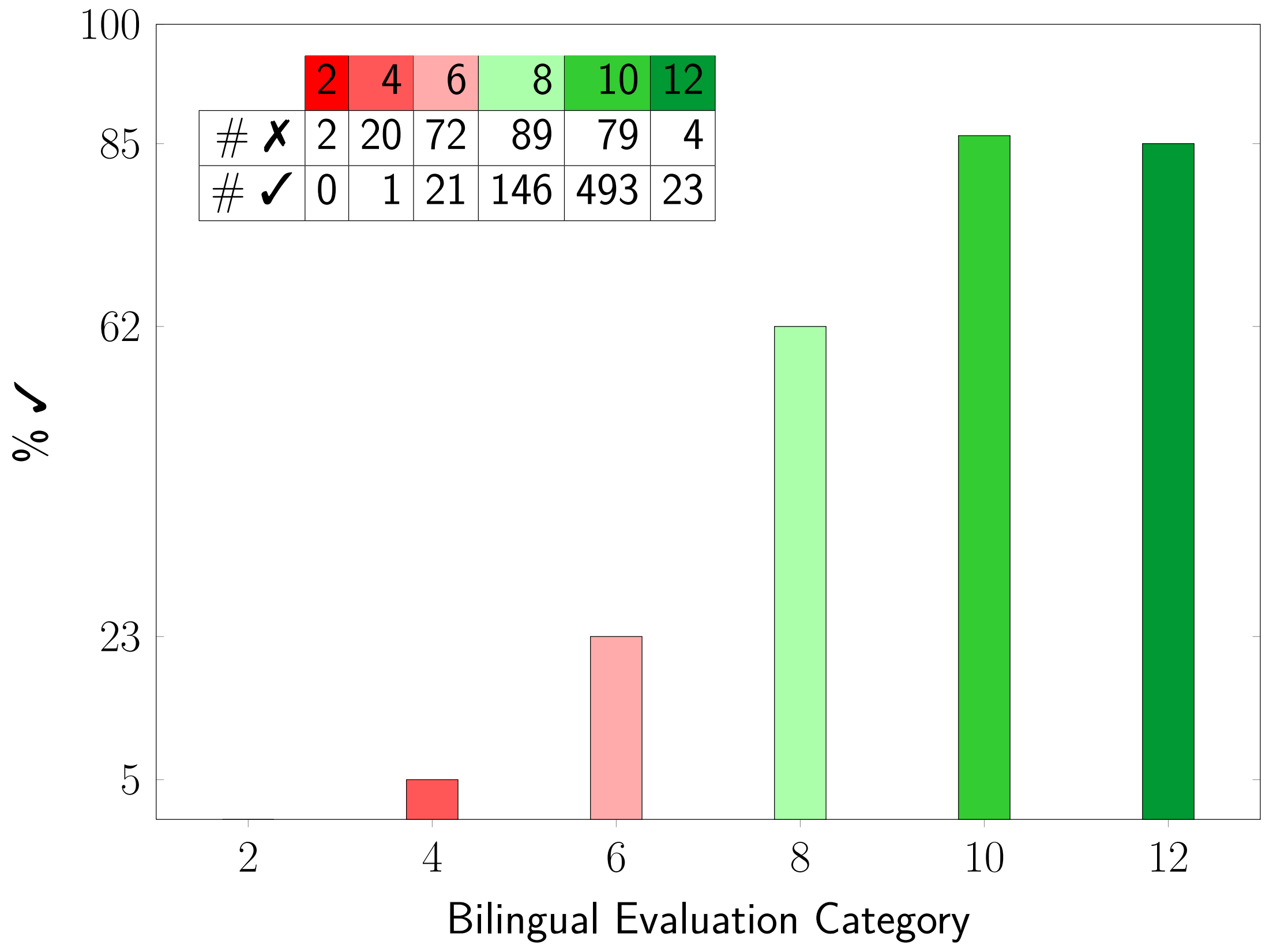


Figure 3: *Percentage of evaluated sentences judged to be in each category by a bilingual judge. Category labels are defined in Table 2.*



	2	4	6	8	10	12
# ✗	2	20	72	89	79	4
# ✓	0	1	21	146	493	23

Figure 4: *For each bilingual evaluation category (see Table 2), the number (#) and percentage (%) of postedited sentence translations that were judged to be correct (✓) versus incorrect (✗) according to a monolingual human judge.*

WMT 2014, Ninth Workshop on Statistical Machine Translation, 26–27 June 2014, Baltimore, Maryland, USA