(Meta-) Evaluation of Machine Translation

Chris Callison-Burch
Johns Hopkins University
ccb@clsp.jhu.edu

Cameron Fordyce
CELCT
fordyce@celct.it

Philipp Koehn
University of Edinburgh
pkoehn@inf.ed.ac.uk

Christof Monz
Queen Mary, University of London
christof@dcs.qmul.ac.uk

Josh Schroeder
University of Edinburgh
j.schroeder@ed.ac.uk

Abstract

This paper evaluates the translation quality of machine translation systems for 8 language pairs: translating French, German, Spanish, and Czech to English and back. We carried out an extensive human evaluation which allowed us not only to rank the different MT systems, but also to perform higher-level analysis of the evaluation process. We measured timing and intra- and inter-annotator agreement for three types of subjective evaluation. We measured the correlation of automatic evaluation metrics with human judgments. This meta-evaluation reveals surprising facts about the most commonly used methodologies.

1 Introduction

This paper presents the results for the shared translation task of the 2007 ACL Workshop on Statistical Machine Translation. The goals of this paper are twofold: First, we evaluate the shared task entries in order to determine which systems produce translations with the highest quality. Second, we analyze the evaluation measures themselves in order to try to determine "best practices" when evaluating machine translation research.

Previous ACL Workshops on Machine Translation were more limited in scope (Koehn and Monz, 2005; Koehn and Monz, 2006). The 2005 workshop evaluated translation quality only in terms of Bleu score. The 2006 workshop additionally included a limited manual evaluation in the style of NIST machine translation evaluation workshop. Here we apply eleven different automatic evaluation metrics, and conduct three different types of manual evaluation.

Beyond examining the quality of translations produced by various systems, we were interested in examining the following questions about evaluation methodologies: How consistent are people when they judge translation quality? To what extent do they agree with other annotators? Can we improve human evaluation? Which automatic evaluation metrics correlate most strongly with human judgments of translation quality?

This paper is organized as follows:

- Section 2 gives an overview of the shared task. It describes the training and test data, reviews the baseline system, and lists the groups that participated in the task.
- Section 3 describes the manual evaluation. We performed three types of evaluation: scoring with five point scales, relative ranking of translations of sentences, and ranking of translations of phrases.
- Section 4 lists the eleven different automatic evaluation metrics which were also used to score the shared task submissions.
- Section 5 presents the results of the shared task, giving scores for each of the systems in each of the different conditions.
- Section 6 provides an evaluation of the different types of evaluation, giving intra- and

Proceedings of the Second Workshop on Statistical Machine Translation, pages 136–158,
Prague, June 2007.

inter-annotator agreement figures for the manual evaluation, and correlation numbers for the automatic metrics.

2 Shared task overview

This year's shared task changed in some aspects from last year's:

- We gave preference to the manual evaluation of system output in the ranking of systems. Manual evaluation was done by the volunteers from participating groups and others. Additionally, there were three modalities of manual evaluation.
- Automatic metrics were also used to rank the systems. In total eleven metrics were applied, and their correlation with the manual scores was measured.
- As in 2006, translation was *from* English, and *into* English. English was again paired with German, French, and Spanish. We additionally included Czech (which was fitting given the location of the WS).

Similar to the IWSLT International Workshop on Spoken Language Translation (Eck and Hori, 2005; Paul, 2006), and the NIST Machine Translation Evaluation Workshop (Lee, 2006) we provide the shared task participants with a common set of training and test data for all language pairs. The major part of data comes from current and upcoming full releases of the Europarl data set (Koehn, 2005).

2.1 Description of the Data

The data used in this year's shared task was similar to the data used in last year's shared task. This year's data included training and development sets for the News Commentary data, which was the surprise out-of-domain test set last year.

The majority of the training data for the Spanish, French, and German tasks was drawn from a new version of the Europarl multilingual corpus. Additional training data was taken from the News Commentary corpus. Czech language resources were drawn from the News Commentary data. Additional resources for Czech came from the CzEng Parallel Corpus (Bojar and Žabokrtský, 2006). Overall, there are over 30 million words of training data per language from the Europarl corpus and 1 million words from the News Commentary corpus. Figure 1 provides some statistics about the corpora used this year.

2.2 Baseline system

To lower the barrier of entrance to the competition, we provided a complete baseline MT system, along with data resources. To summarize, we provided:

- sentence-aligned training corpora
- development and dev-test sets
- language models trained for each language
- an open source decoder for phrase-based SMT called Moses (Koehn et al., 2006), which replaces the Pharaoh decoder (Koehn, 2004)
- a training script to build models for Moses

The performance of this baseline system is similar to the best submissions in last year's shared task.

2.3 Test Data

The test data was again drawn from a segment of the Europarl corpus from the fourth quarter of 2000, which is excluded from the training data. Participants were also provided with three sets of parallel text to be used for system development and tuning.

In addition to the Europarl test set, we also collected editorials from the Project Syndicate website[1], which are published in all the five languages of the shared task. We aligned the texts at a sentence level across all five languages, resulting in 2,007 sentences per language. For statistics on this test set, refer to Figure 1.

The News Commentary test set differs from the Europarl data in various ways. The text type are editorials instead of speech transcripts. The domain is general politics, economics and science. However, it is also mostly political content (even if not focused on the internal workings of the European Union) and opinion.

2.4 Participants

We received submissions from 15 groups from 14 institutions, as listed in Table 1. This is a slight

[1] `http://www.project-syndicate.com/`

Europarl Training corpus

	Spanish ↔ English	French ↔ English	German ↔ English
Sentences	1,259,914	1,288,901	1,264,825
Foreign words	33,159,337	33,176,243	29,582,157
English words	31,813,692	32,615,285	31,929,435
Distinct foreign words	345,944	344,287	510,544
Distinct English words	266,976	268,718	250,295

News Commentary Training corpus

	Spanish ↔ English	French ↔ English	German ↔ English	Czech ↔ English
Sentences	51,613	43,194	59,975	57797
Foreign words	1,263,067	1,028,672	1,297,673	1,083,122
English words	1,076,273	906,593	1,238,274	1,188,006
Distinct foreign words	84,303	68,214	115,589	142,146
Distinct English words	70,755	63,568	76,419	74,042

Language model data

	English	Spanish	French	German
Sentence	1,407,285	1,431,614	1,435,027	1,478,428
Words	34,539,822	36,426,542	35,595,199	32,356,475
Distinct words	280,546	385,796	361,205	558,377

Europarl test set

	English	Spanish	French	German
Sentences	2,000			
Words	53,531	55,380	53,981	49,259
Distinct words	8,558	10,451	10,186	11,106

News Commentary test set

	English	Spanish	French	German	Czech
Sentences	2,007				
Words	43,767	50,771	49,820	45,075	39,002
Distinct words	10,002	10,948	11,244	12,322	15,245

Figure 1: Properties of the training and test sets used in the shared task. The training data is drawn from the Europarl corpus and from the Project Syndicate, a web site which collects political commentary in multiple languages.

ID	Participant
cmu-uka	Carnegie Mellon University, USA (Paulik et al., 2007)
cmu-syntax	Carnegie Mellon University, USA (Zollmann et al., 2007)
cu	Charles University, Czech Republic (Bojar, 2007)
limsi	LIMSI-CNRS, France (Schwenk, 2007)
liu	University of Linköping, Sweden(Holmqvist et al., 2007)
nrc	National Research Council, Canada (Ueffing et al., 2007)
pct	a commercial MT provider from the Czech Republic
saar	Saarland University & DFKI, Germany (Chen et al., 2007)
systran	SYSTRAN, France & U. Edinburgh, UK (Dugast et al., 2007)
systran-nrc	National Research Council, Canada (Simard et al., 2007)
ucb	University of California at Berkeley, USA (Nakov and Hearst, 2007)
uedin	University of Edinburgh, UK (Koehn and Schroeder, 2007)
umd	University of Maryland, USA (Dyer, 2007)
upc	University of Catalonia, Spain (Costa-Jussà and Fonollosa, 2007)
upv	University of Valencia, Spain (Civera and Juan, 2007)

Table 1: Participants in the shared task. Not all groups participated in all translation directions.

increase over last year's shared task where submissions were received from 14 groups from 11 institutions. Of the 11 groups that participated in last year's shared task, 6 groups returned this year.

This year, most of these groups follow a phrase-based statistical approach to machine translation. However, several groups submitted results from systems that followed a hybrid approach.

While building a machine translation system is a serious undertaking we hope to attract more newcomers to the field by keeping the barrier of entry as low as possible. The creation of parallel corpora such as the Europarl, the CzEng, and the News Commentary corpora should help in this direction by providing freely available language resources for building systems. The creation of an open source baseline system should also go a long way towards achieving this goal.

For more on the participating systems, please refer to the respective system description in the proceedings of the workshop.

3 Human evaluation

We evaluated the shared task submissions using both manual evaluation and automatic metrics. While automatic measures are an invaluable tool for the day-to-day development of machine translation systems, they are an imperfect substitute for human assessment of translation quality. Manual evaluation is time consuming and expensive to perform, so comprehensive comparisons of multiple systems are rare. For our manual evaluation we distributed the workload across a number of people, including participants in the shared task, interested volunteers, and a small number of paid annotators. More than 100 people participated in the manual evaluation, with 75 of those people putting in at least an hour's worth of effort. A total of 330 hours of labor was invested, nearly doubling last year's all-volunteer effort which yielded 180 hours of effort.

Beyond simply ranking the shared task submissions, we had a number of scientific goals for the manual evaluation. Firstly, we wanted to collect data which could be used to assess how well automatic metrics correlate with human judgments. Secondly, we wanted to examine different types of manual evaluation and assess which was the best. A number of criteria could be adopted for choosing among different types of manual evaluation: the ease with which people are able to perform the task, their agreement with other annotators, their reliability when asked to repeat judgments, or the number of judgments which can be collected in a fixed time period.

There are a range of possibilities for how human evaluation of machine translation can be done. For instance, it can be evaluated with reading comprehension tests (Jones et al., 2005), or by assigning subjective scores to the translations of individual sentences (LDC, 2005). We examined three different ways of manually evaluating machine translation quality:

- Assigning scores based on five point adequacy and fluency scales
- Ranking translated sentences relative to each other
- Ranking the translations of syntactic constituents drawn from the source sentence

3.1 Fluency and adequacy

The most widely used methodology when manually evaluating MT is to assign values from two five point scales representing *fluency* and *adequacy*. These scales were developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistics Data Consortium (LDC, 2005).

The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation:

5 = All
4 = Most
3 = Much
2 = Little
1 = None

The second five point scale indicates how fluent the translation is. When translating into English the values correspond to:

5 = Flawless English
4 = Good English
3 = Non-native English
2 = Disfluent English
1 = Incomprehensible

Separate scales for fluency and adequacy were developed under the assumption that a translation might be disfluent but contain all the information from the source. However, in principle it seems that people have a hard time separating these two aspects of translation. The high correlation between people's fluency and adequacy scores (given in Tables 17 and 18) indicate that the distinction might be false.

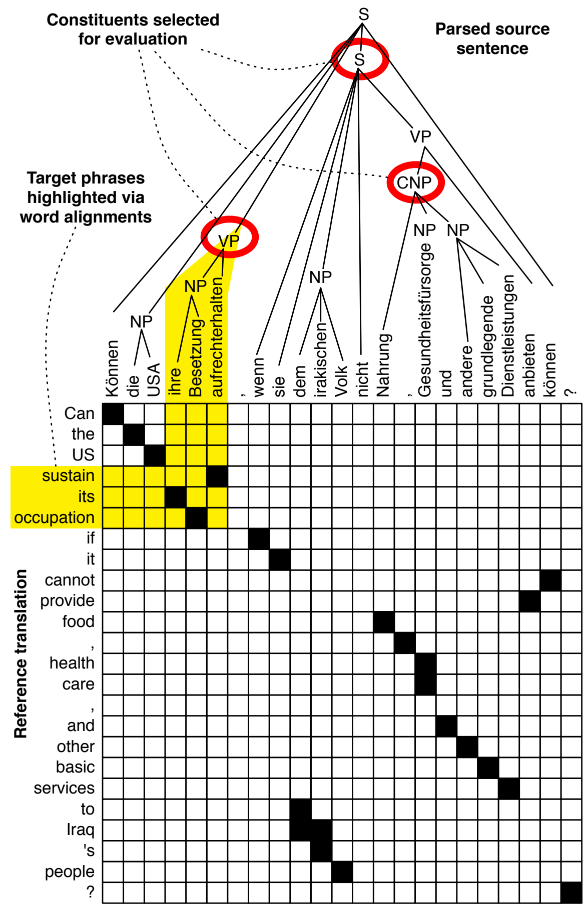

Figure 2: In constituent-based evaluation, the source sentence was parsed, and automatically aligned with the reference translation and systems' translations

Another problem with the scores is that there are no clear guidelines on how to assign values to translations. No instructions are given to evaluators in terms of how to quantify meaning, or how many grammatical errors (or what sort) separates the different levels of fluency. Because of this many judges either develop their own rules of thumb, or use the scales as relative rather than absolute. These are borne out in our analysis of inter-annotator agreement in Section 6.

3.2 Ranking translations of sentences

Because fluency and adequacy were seemingly difficult things for judges to agree on, and because many people from last year's workshop seemed to be using them as a way of ranking translations, we decided to try a separate evaluation where people were simply asked to rank translations. The instructions for this task were:

> Rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed).

These instructions were just as minimal as for fluency and adequacy, but the task was considerably simplified. Rather than having to assign each translation a value along an arbitrary scale, people simply had to compare different translations of a single sentence and rank them.

3.3 Ranking translations of syntactic constituents

In addition to having judges rank the translations of whole sentences, we also conducted a pilot study of a new type of evaluation methodology, which we call *constituent-based evaluation*. In our constituent-based evaluation we parsed the source language sentence, selected constituents from the tree, and had people judge the translations of those syntactic phrases. In order to draw judges' attention to these regions, we highlighted the selected source phrases and the corresponding phrases in the translations. The corresponding phrases in the translations were located via automatic word alignments.

Figure 2 illustrates the constituent based evaluation when applied to a German source sentence. The German source sentence is parsed, and various phrases are selected for evaluation. Word alignments are created between the source sentence and the reference translation (shown), and the source sentence and each of the system translations (not shown). We parsed the test sentences for each of the languages aside from Czech. We used Cowan and Collins (2005)'s parser for Spanish, Arun and Keller (2005)'s for French, Dubey (2005)'s for German, and Bikel (2002)'s for English.

The word alignments were created with Giza++ (Och and Ney, 2003) applied to a parallel corpus containing 200,000 sentence pairs of the training data, plus sets of 4,007 sentence pairs created by pairing the test sentences with the reference translations, and the test sentences paired with each of the system translations. The phrases in the translations were located using techniques from phrase-based statistical machine translation which extract phrase pairs from word alignments (Koehn et al., 2003; Och and Ney, 2004). Because the word-alignments were created automatically, and because the phrase extraction is heuristic, the phrases that were selected may not exactly correspond to the translations of the selected source phrase. We noted this in the instructions to judges:

> Rank each constituent translation from Best to Worst relative to the other choices (ties are allowed). Grade **only the highlighted part** of each translation.
>
> *Please note that segments are selected automatically, and they should be taken as an approximate guide. They might include extra words that are not in the actual alignment, or miss words on either end.*

The criteria that we used to select which constituents were to be evaluated were:

- The constituent could not be the whole source sentence
- The constituent had to be longer three words, and be no longer than 15 words
- The constituent had to have a corresponding phrase with a consistent word alignment in each of the translations

The final criterion helped reduce the number of alignment errors.

3.4 Collecting judgments

We collected judgments using a web-based tool. Shared task participants were each asked to judge 200 sets of sentences. The sets consisted of 5 system outputs, as shown in Figure 3. The judges were presented with batches of each type of evaluation. We presented them with five screens of adequacy/fluency scores, five screens of sentence rankings, and ten screens of constituent rankings. The order of the types of evaluation were randomized.

In order to measure intra-annotator agreement 10% of the items were repeated and evaluated twice by each judge. In order to measure inter-annotator agreement 40% of the items were randomly drawn from a common pool that was shared across all

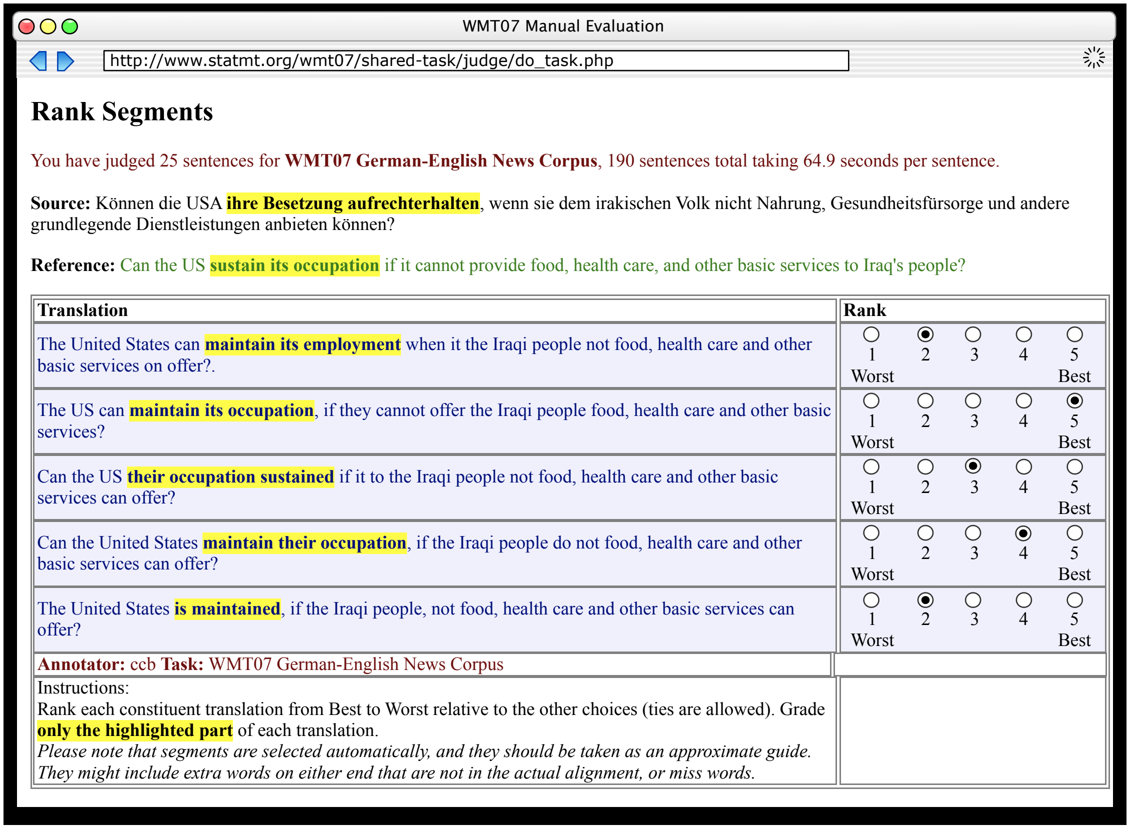

Figure 3: For each of the types of evaluation, judges were shown screens containing up to five different system translations, along with the source sentence and reference translation.

annotators so that we would have items that were judged by multiple annotators.

Judges were allowed to select whichever data set they wanted, and to evaluate translations into whatever languages they were proficient in. Shared task participants were excluded from judging their own systems.

Table 2 gives a summary of the number of judgments that we collected for translations of individual sentences. Since we had 14 translation tasks and four different types of scores, there were 55 different conditions.[2] In total we collected over 81,000 judgments. Despite the large number of conditions we managed to collect more than 1,000 judgments for most of them. This provides a rich source of data for analyzing the quality of translations produced by different systems, the different types of human evaluation, and the correlation of automatic metrics with human judgments.[3]

[2]We did not perform a constituent-based evaluation for Czech to English because we did not have a syntactic parser for Czech. We considered adapting our method to use Bojar (2004)'s dependency parser for Czech, but did not have the time.

[3]The judgment data along with all system translations are available at `http://www.statmt.org/wmt07/`

4 Automatic evaluation

The past two ACL workshops on machine translation used Bleu as the sole automatic measure of translation quality. Bleu was used exclusively since it is the most widely used metric in the field and has been shown to correlate with human judgments of translation quality in many instances (Doddington, 2002; Coughlin, 2003; Przybocki, 2004). However, recent work suggests that Bleu's correlation with human judgments may not be as strong as previously thought (Callison-Burch et al., 2006). The results of last year's workshop further suggested that Bleu systematically underestimated the quality of rule-based machine translation systems (Koehn and Monz, 2006).

We used the manual evaluation data as a means of testing the correlation of a range of automatic metrics in addition to Bleu. In total we used eleven different automatic evaluation measures to rank the shared task submissions. They are:

- Meteor (Banerjee and Lavie, 2005)—Meteor measures precision and recall of unigrams when comparing a hypothesis translation

Language Pair	Test Set	Adequacy	Fluency	Rank	Constituent
English-German	Europarl	1,416	1,418	1,419	2,626
	News Commentary	1,412	1,413	1,412	2,755
German-English	Europarl	1,525	1,521	1,514	2,999
	News Commentary	1,626	1,620	1,601	3,084
English-Spanish	Europarl	1,000	1,003	1,064	1,001
	News Commentary	1,272	1,272	1,238	1,595
Spanish-English	Europarl	1,174	1,175	1,224	1,898
	News Commentary	947	949	922	1,339
English-French	Europarl	773	772	769	1,456
	News Commentary	729	735	728	1,313
French-English	Europarl	834	833	830	1,641
	News Commentary	1,041	1,045	1,035	2,036
English-Czech	News Commentary	2,303	2,304	2,331	3,968
Czech-English	News Commentary	1,711	1,711	1,733	0
	Totals	**17,763**	**17,771**	**17,820**	**27,711**

Table 2: The number of items that were judged for each task during the manual evaluation

against a reference. It flexibly matches words using stemming and WordNet synonyms. Its flexible matching was extended to French, Spanish, German and Czech for this workshop (Lavie and Agarwal, 2007).

- Bleu (Papineni et al., 2002)—Bleu is currently the *de facto* standard in machine translation evaluation. It calculates n-gram precision and a brevity penalty, and can make use of multiple reference translations as a way of capturing some of the allowable variation in translation. We use a single reference translation in our experiments.

- GTM (Melamed et al., 2003)—GTM generalizes precision, recall, and F-measure to measure overlap between strings, rather than overlap between bags of items. An "exponent" parameter which controls the relative importance of word order. A value of 1.0 reduces GTM to ordinary unigram overlap, with higher values emphasizing order.[4]

- Translation Error Rate (Snover et al., 2006)—TER calculates the number of edits required to change a hypothesis translation into a reference translation. The possible edits in TER include insertion, deletion, and substitution of single words, and an edit which moves sequences of contiguous words.

- ParaEval precision and ParaEval recall (Zhou et al., 2006)–ParaEval matches hypothesis and reference translations using paraphrases that are extracted from parallel corpora in an unsupervised fashion (Bannard and Callison-Burch, 2005). It calculates precision and recall using a unigram counting strategy.

- Dependency overlap (Amigó et al., 2006)—This metric uses dependency trees for the hypothesis and reference translations, by computing the average overlap between words in the two trees which are dominated by grammatical relationships of the same type.

- Semantic role overlap (Giménez and Màrquez, 2007)—This metric calculates the lexical overlap between semantic roles (i.e., semantic arguments or adjuncts) of the same type in the the hypothesis and reference translations. It uniformly averages lexical overlap over all semantic role types.

[4]The GTM scores presented here are an F-measure with a weight of 0.1, which counts recall at 10x the level of precision. The exponent is set at 1.2, which puts a mild preference towards items with words in the correct order. These parameters could be optimized empirically for better results.

- Word Error Rate over verbs (Popovic and Ney, 2007)—WER' creates a new reference and a new hypothesis for each POS class by extracting all words belonging to this class, and then to calculate the standard WER. We show results for this metric over verbs.

- Maximum correlation training on adequacy and on fluency (Liu and Gildea, 2007)—a linear combination of different evaluation metrics (Bleu, Meteor, Rouge, WER, and stochastic iterative alignment) with weights set to maximize Pearson's correlation with adequacy and fluency judgments. Weights were trained on WMT-06 data.

The scores produced by these are given in the tables at the end of the paper, and described in Section 5. We measured the correlation of the automatic evaluation metrics with the different types of human judgments on 12 data conditions, and report these in Section 6.

5 Shared task results

The results of the human evaluation are given in Tables 9, 10, 11 and 12. Each of those tables present four scores:

- FLUENCY and ADEQUACY are normalized versions of the five point scores described in Section 3.1. The tables report an average of the normalized scores.[5]

- RANK is the average number of times that a system was judged to be better than any other system in the sentence ranking evaluation described in Section 3.2.

- CONSTITUENT is the average number of times that a system was judged to be better than any other system in the constituent-based evaluation described in Section 3.3.

There was reasonably strong agreement between these four measures at which of the entries was the best in each data condition. There was complete agreement between them in 5 of the 14 conditions, and agreement between at least three of them in 10 of the 14 cases.

[5]Since different annotators can vary widely in how they assign fluency and adequacy scores, we normalized these scores on a per-judge basis using the method suggested by Blatz et al. (2003) in Chapter 5, page 97.

SYSTRAN (systran)	32%
University of Edinburgh (uedin)	20%
University of Catalonia (upc)	15%
LIMSI-CNRS (limsi)	13%
University of Maryland (umd)	5%
National Research Council of Canada's joint entry with SYSTRAN (systran-nrc)	5%
Commercial Czech-English system (pct)	5%
University of Valencia (upv)	2%
Charles University (cu)	2%

Table 3: The proportion of time that participants' entries were top-ranked in the human evaluation

University of Edinburgh (uedin)	41%
University of Catalonia (upc)	12%
LIMSI-CNRS (limsi)	12%
University of Maryland (umd)	9%
Charles University (cu)	4%
Carnegie Mellon University (cmu-syntax)	4%
Carnegie Mellon University (cmu-uka)	4%
University of California at Berkeley (ucb)	3%
National Research Council's joint entry with SYSTRAN (systran-nrc)	2%
SYSTRAN (systran)	2%
Saarland University (saar)	0.8%

Table 4: The proportion of time that participants' entries were top-ranked by the automatic evaluation metrics

Table 3 gives a summary of how often different participants' entries were ranked #1 by any of the four human evaluation measures. SYSTRAN's entries were ranked the best most often, followed by University of Edinburgh, University of Catalonia and LIMSI-CNRS.

The following systems were the best performing for the different language pairs: SYSTRAN was ranked the highest in German-English, University of Catalonia was ranked the highest in Spanish-English, LIMSI-CNRS was ranked highest in French-English, and the University of Maryland and a commercial system were the highest for

Evaluation type	$P(A)$	$P(E)$	K
Fluency (absolute)	.400	.2	.250
Adequacy (absolute)	.380	.2	.226
Fluency (relative)	.520	.333	.281
Adequacy (relative)	.538	.333	.307
Sentence ranking	.582	.333	.373
Constituent ranking	.693	.333	.540
Constituent ranking (w/identical constituents)	.712	.333	.566

Table 5: Kappa coefficient values representing the inter-annotator agreement for the different types of manual evaluation

Evaluation type	$P(A)$	$P(E)$	K
Fluency (absolute)	.630	.2	.537
Adequacy (absolute)	.574	.2	.468
Fluency (relative)	.690	.333	.535
Adequacy (relative)	.696	.333	.544
Sentence ranking	.749	.333	.623
Constituent ranking	.825	.333	.738
Constituent ranking (w/identical constituents)	.842	.333	.762

Table 6: Kappa coefficient values for intra-annotator agreement for the different types of manual evaluation

Czech-English.

While we consider the human evaluation to be primary, it is also interesting to see how the entries were ranked by the various automatic evaluation metrics. The complete set of results for the automatic evaluation are presented in Tables 13, 14, 15, and 16. An aggregate summary is provided in Table 4. The automatic evaluation metrics strongly favor the University of Edinburgh, which garners 41% of the top-ranked entries (which is partially due to the fact it was entered in every language pair). Significantly, the automatic metrics disprefer SYSTRAN, which was strongly favored in the human evaluation.

6 Meta-evaluation

In addition to evaluating the translation quality of the shared task entries, we also performed a "meta-evaluation" of our evaluation methodologies.

6.1 Inter- and Intra-annotator agreement

We measured pairwise agreement among annotators using the kappa coefficient (K) which is widely used in computational linguistics for measuring agreement in category judgments (Carletta, 1996). It is defined as

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would agree by chance. We define chance agreement for fluency and adequacy as $\frac{1}{5}$, since they are based on five point scales, and for ranking as $\frac{1}{3}$ since there are three possible out comes when ranking the output of a pair of systems: $A > B$, $A = B$, $A < B$.

For inter-annotator agreement we calculated $P(A)$ for fluency and adequacy by examining all items that were annotated by two or more annotators, and calculating the proportion of time they assigned identical scores to the same items. For the ranking tasks we calculated $P(A)$ by examining all pairs of systems which had been judged by two or more judges, and calculated the proportion of time that they agreed that $A > B$, $A = B$, or $A < B$. For intra-annotator agreement we did similarly, but gathered items that were annotated on multiple occasions by a single annotator.

Table 5 gives K values for inter-annotator agreement, and Table 6 gives K values for intra-annoator agreement. These give an indication of how often different judges agree, and how often single judges are consistent for repeated judgments, respectively. The interpretation of Kappa varies, but according to Landis and Koch (1977) $0 - -.2$ is slight, $.21 - -.4$ is fair, $.41 - -.6$ is moderate, $.61 - -.8$ is substantial and the rest almost perfect.

The K values for fluency and adequacy should give us pause about using these metrics in the future. When we analyzed them as they are intended to be—scores classifying the translations of sentences into different types—the inter-annotator agreement was barely considered *fair*, and the intra-annotator agreement was only *moderate*. Even when we reassessed fluency and adequacy as relative ranks the agreements increased only minimally.

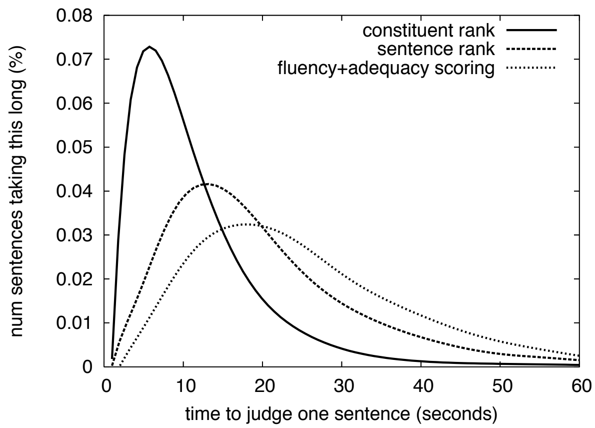

Figure 4: Distributions of the amount of time it took to judge single sentences for the three types of manual evaluation

The agreement on the other two types of manual evaluation that we introduced were considerably better. The both the sentence and constituent ranking had *moderate* inter-annotator agreement and *substantial* intra-annotator agreement. Because the constituent ranking examined the translations of short phrases, often times all systems produced the same translations. Since these trivially increased agreement (since they would always be equally ranked) we also evaluated the inter- and intra-annotator agreement when those items were excluded. The agreement remained very high for constituent-based evaluation.

6.2 Timing

We used the web interface to collect timing information. The server recorded the time when a set of sentences was given to a judge and the time when the judge returned the sentences. We divided the time that it took to do a set by the number of sentences in the set. The average amount of time that it took to assign fluency and adequacy to a single sentence was 26 seconds.[6] The average amount of time it took to rank a sentence in a set was 20 seconds. The average amount of time it took to rank a highlighted constituent was 11 seconds. Figure 4 shows the distribution of times for these tasks.

These timing figures are promising because they indicate that the tasks which the annotators were the most reliable on (constituent ranking and sentence ranking) were also much quicker to complete than the ones that they were unreliable on (assigning fluency and adequacy scores). This suggests that fluency and adequacy should be replaced with ranking tasks in future evaluation exercises.

6.3 Correlation between automatic metrics and human judgments

To measure the correlation of the automatic metrics with the human judgments of translation quality we used Spearman's rank correlation coefficient ρ. We opted for Spearman rather than Pearson because it makes fewer assumptions about the data. Importantly, it can be applied to ordinal data (such as the fluency and adequacy scales). Spearman's rank correlation coefficient is equivalent to Pearson correlation on ranks.

After the raw scores that were assigned to systems by an automatic metric and by one of our manual evaluation techniques have been converted to ranks, we can calculate ρ using the simplified equation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the rank for system$_i$ and n is the number of systems. The possible values of ρ range between 1 (where all systems are ranked in the same order) and -1 (where the systems are ranked in the reverse order). Thus an automatic evaluation metric with a higher value for ρ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower ρ.

Table 17 reports ρ for the metrics which were used to evaluate translations into English.[7]. Table 7 summarizes the results by averaging the correlation numbers by equally weighting each of the data conditions. The table ranks the automatic evaluation metrics based on how well they correlated with human judgments. While these are based on a relatively few number of items, and while we have not performed any tests to determine whether the differences in ρ are statistically significant, the results

[6]Sets which took longer than 5 minutes were excluded from these calculations, because there was a strong chance that annotators were interrupted while completing the task.

[7]The Czech-English conditions were excluded since there were so few systems

are nevertheless interesting, since three metrics have higher correlation than Bleu:

- Semantic role overlap (Giménez and Màrquez, 2007), which makes its debut in the proceedings of this workshop
- ParaEval measuring recall (Zhou et al., 2006), which has a model of allowable variation in translation that uses automatically generated paraphrases (Callison-Burch, 2007)
- Meteor (Banerjee and Lavie, 2005) which also allows variation by introducing synonyms and by flexibly matches words using stemming.

Tables 18 and 8 report ρ for the six metrics which were used to evaluate translations into the other languages. Here we find that Bleu and TER are the closest to human judgments, but that overall the correlations are much lower than for translations into English.

7 Conclusions

Similar to last year's workshop we carried out an extensive manual and automatic evaluation of machine translation performance for translating from four European languages into English, and vice versa. This year we substantially increased the number of automatic evaluation metrics and were also able to nearly double the efforts of producing the human judgments.

There were substantial differences in the results results of the human and automatic evaluations. We take the human judgments to be authoritative, and used them to evaluate the automatic metrics. We measured correlation using Spearman's coefficient and found that three less frequently used metrics were stronger predictors of human judgments than Bleu. They were: semantic role overlap (newly introduced in this workshop) ParaEval-recall and Meteor.

Although we do not claim that our observations are indisputably conclusive, they again indicate that the choice of automatic metric can have a significant impact on comparing systems. Understanding the exact causes of those differences still remains an important issue for future research.

metric	ADEQUACY	FLUENCY	RANK	CONSTITUENT	OVERALL
Semantic role overlap	.774	.839	.803	.741	.789
ParaEval-Recall	.712	.742	.768	.798	.755
Meteor	.701	.719	.745	.669	.709
Bleu	.690	.722	.672	.602	.671
1-TER	.607	.538	.520	.514	.644
Max adequ-correlation	.651	.657	.659	.534	.626
Max fluency correlation	.644	.653	.656	.512	.616
GTM	.655	.674	.616	.495	.610
Dependency overlap	.639	.644	.601	.512	.599
ParaEval-Precision	.639	.654	.610	.491	.598
1-WER of verbs	.378	.422	.431	.297	.382

Table 7: Average corrections for the different automatic metrics when they are used to evaluate translations into English

metric	ADEQUACY	FLUENCY	RANK	CONSTITUENT	OVERALL
Bleu	.657	.445	.352	.409	.466
1-TER	.589	.419	.361	.380	.437
Max fluency correlation	.534	.419	.368	.400	.430
Max adequ-correlation	.498	.414	.385	.409	.426
Meteor	.490	.356	.279	.304	.357
1-WER of verbs	.371	.304	.359	.359	.348

Table 8: Average corrections for the different automatic metrics when they are used to evaluate translations into the other languages

This year's evaluation also measured the agreement between human assessors by computing the Kappa coefficient. One striking observation is that inter-annotator agreement for fluency and adequacy can be called 'fair' at best. On the other hand, comparing systems by ranking them manually (constituents or entire sentences), resulted in much higher inter-annotator agreement.

Acknowledgments

This work was supported in part by the EuroMatrix project funded by the European Commission (6th Framework Programme), and in part by the GALE program of the US Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

We are grateful to Jesús Giménez, Dan Melamed, Maja Popvic, Ding Liu, Liang Zhou, and Abhaya Agarwal for scoring the entries with their automatic evaluation metrics. Thanks to Brooke Cowan for parsing the Spanish test sentences, to Josh Albrecht for his script for normalizing fluency and adequacy on a per judge basis, and to Dan Melamed, Rebecca Hwa, Alon Lavie, Colin Bannard and Mirella Lapata for their advice about statistical tests.

References

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In *Proceedings of COLING-ACL06*.

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of ACL*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL-2005*.

Dan Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of HLT*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. CLSP Summer Workshop Final Report WS2003, Johns Hopkins University.

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

Ondřej Bojar. 2004. Problems of inducing large coverage constraint-based dependency grammar for Czech. In *Constraint Solving and Language Processing, CSLP 2004*, volume LNAI 3438. Springer.

Ondřej Bojar. 2007. English-to-Czech factored machine translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of EACL*.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Scotland.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source decoder for statistical machine translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Marta R. Costa-Jussà and José A.R. Fonollosa. 2007. Analysis of statistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*.

Brooke Cowan and Michael Collins. 2005. Morphology and reranking for the statistical parsing of Spanish. In *Proceedings of EMNLP 2005*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132, San Diego.

Amit Dubey. 2005. What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *Proceedings of ACL*.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Christopher J. Dyer. 2007. The 'noisier channel': translation from morphologically complex languages. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation*.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of ACL Workshop on Statistical Machine Translation*.

Maria Holmqvist, Sara Stymne, and Lars Ahrenberg. 2007. Getting to know Moses: Initial experiments on German-English factored translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Douglas Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005. Measuring translation quality by testing english speakers with a new defense language proficiency test for arabic. In *Proceedings of the 2005 International Conference on Intelligence Analysis*.

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between European languages. In *Proceedings of ACL 2005 Workshop on Parallel Text Translation*.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Philipp Koehn, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Alexandra Constantin, Brooke Cowan, Chris Dyer, Marcello Federico, Evan Herbst, Hieu Hoang, Christine Moran, Wade Shen, and Richard Zens. 2006. Factored translation models. CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*.

Philipp Koehn. 2005. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Workshop on Statistical Machine Translation*, Prague, June. Association for Computational Linguistics.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

Audrey Lee. 2006. NIST 2006 machine translation evaluation official results. Official release of automatic evaluation scores for all submissions, November.

Ding Liu and Daniel Gildea. 2007. Source-language features and maximum correlation training for machine translation evaluation. In *Proceedings of NAACL*.

Dan Melamed, Ryan Green, and Jospeh P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of HLT/NAACL*.

Preslav Nakov and Marti Hearst. 2007. UCB system description for the WMT 2007 shared task. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

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

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Michael Paul. 2006. Overview of the IWSLT 2006 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation*.

Matthias Paulik, Kay Rottmann, Jan Niehues, Silja Hildebrand, and Stephan Vogel. 2007. The ISL phrase-based MT system for the 2007 ACL Workshop on Statistical Machine Translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Maja Popovic and Hermann Ney. 2007. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of ACL Workshop on Statistical Machine Translation*.

Mark Przybocki. 2004. NIST 2004 machine translation evaluation results. Confidential e-mail to workshop participants, May.

Holger Schwenk. 2007. Building a statistical machine translation system for French using the Europarl corpus. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Statistical Machine Translation in the Americas*.

Nicola Ueffing, Michel Simard, Samuel Larkin, and Howard Johnson. 2007. NRC's PORTAGE system for WMT 2007. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Reevaluating machine translation results with paraphrase support. In *Proceedings of EMNLP*.

Andreas Zollmann, Ashish Venugopal, Matthias Paulik, and Stephan Vogel. 2007. The syntax augmented MT (SAMT) system for the shared task in the 2007 ACL Workshop on Statistical Machine Translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation (WMT-07)*, Prague.

system	ADEQUACY	FLUENCY	RANK	CONSTITUENT
German-English Europarl				
cmu-uka	0.511	0.496	0.395	0.206
liu	0.541	0.55	0.415	0.234
nrc	0.474	0.459	0.354	0.214
saar	0.334	0.404	0.119	0.104
systran	**0.562**	**0.594**	**0.530**	**0.302**
uedin	0.53	0.554	0.43	0.187
upc	0.534	0.533	0.384	0.214
German-English News Corpus				
nrc	0.459	0.429	0.325	0.245
saar	0.278	0.341	0.108	0.125
systran	**0.552**	**0.56**	**0.563**	**0.344**
uedin	0.508	0.536	0.485	0.332
upc	0.536	0.512	0.476	0.330
English-German Europarl				
cmu-uka	0.557	0.508	0.416	0.333
nrc	0.534	0.511	0.328	0.321
saar	0.369	0.383	0.172	0.196
systran	0.543	0.525	**0.511**	0.295
uedin	**0.569**	**0.576**	0.389	**0.350**
upc	0.565	0.522	0.438	0.3
English-German News Corpus				
nrc	0.453	0.4	0.437	0.340
saar	0.186	0.273	0.108	0.121
systran	**0.542**	**0.556**	**0.582**	**0.351**
ucb	0.415	0.403	0.332	0.289
uedin	0.472	0.445	0.455	0.303
upc	0.505	0.475	0.377	0.349

Table 9: Human evaluation for German-English submissions

system	ADEQUACY	FLUENCY	RANK	CONSTITUENT
Spanish-English Europarl				
cmu-syntax	0.552	0.568	0.478	0.152
cmu-uka	0.557	0.564	0.392	0.139
nrc	0.477	0.489	0.382	0.143
saar	0.328	0.336	0.126	0.075
systran	0.525	0.566	0.453	0.156
uedin	**0.593**	**0.610**	0.419	0.14
upc	0.587	0.604	**0.5**	**0.188**
upv	0.562	0.573	0.326	0.154
Spanish-English News Corpus				
cmu-uka	0.522	0.495	0.41	0.213
nrc	0.479	0.464	0.334	0.243
saar	0.446	0.46	0.246	0.198
systran	0.525	0.503	0.453	0.22
uedin	0.546	0.534	0.48	0.268
upc	**0.566**	**0.543**	**0.537**	**0.312**
upv	0.435	0.459	0.295	0.151
English-Spanish Europarl				
cmu-uka	0.563	0.581	0.391	0.23
nrc	0.546	0.548	0.323	0.22
systran	0.495	0.482	0.329	0.224
uedin	**0.586**	**0.638**	**0.468**	0.225
upc	0.584	0.578	0.444	0.239
upv	0.573	0.587	0.406	**0.246**
English-Spanish News Corpus				
cmu-uka	0.51	0.492	0.45	0.277
nrc	0.408	0.392	0.367	0.224
systran	0.501	**0.507**	**0.481**	**0.352**
ucb	0.449	0.414	0.390	0.307
uedin	0.429	0.419	0.389	0.266
upc	**0.51**	0.488	0.404	0.311
upv	0.405	0.418	0.250	0.217

Table 10: Human evaluation for Spanish-English submissions

system	ADEQUACY	FLUENCY	RANK	CONSTITUENT
French-English Europarl				
limsi	**0.634**	0.618	0.458	**0.290**
nrc	0.553	0.551	0.404	0.253
saar	0.384	0.447	0.176	0.157
systran	0.494	0.484	0.286	0.202
systran-nrc	0.604	0.6	0.503	0.267
uedin	0.616	**0.635**	**0.514**	0.283
upc	0.616	0.619	0.448	0.267
French-English News Corpus				
limsi	0.575	**0.596**	**0.494**	**0.312**
nrc	0.472	0.442	0.306	0.241
saar	0.280	0.372	0.183	0.159
systran	0.553	0.534	0.469	0.288
systran-nrc	0.513	0.49	0.464	0.290
uedin	0.556	0.586	0.493	0.306
upc	**0.576**	0.587	0.493	0.291
English-French Europarl				
limsi	**0.635**	**0.627**	0.505	0.259
nrc	0.517	0.518	0.359	0.206
saar	0.398	0.448	0.155	0.139
systran	0.574	0.526	0.353	0.179
systran-nrc	0.575	0.58	**0.512**	0.225
uedin	0.620	0.608	0.485	**0.273**
upc	0.599	0.566	0.45	0.256
English-French News Corpus				
limsi	0.537	0.495	0.44	0.363
nrc	0.481	0.484	0.372	0.324
saar	0.243	0.276	0.086	0.121
systran	0.536	0.546	**0.634**	**0.440**
systran-nrc	**0.557**	**0.572**	0.485	0.287
ucb	0.401	0.391	0.316	0.245
uedin	0.466	0.447	0.485	0.375
upc	0.509	0.469	0.437	0.326

Table 11: Human evaluation for French-English submissions

system	ADEQUACY	FLUENCY	RANK	CONSTITUENT
Czech-English News Corpus				
cu	0.468	0.478	0.362	—
pct	0.418	0.388	0.220	—
uedin	0.458	0.471	0.353	—
umd	**0.550**	**0.592**	**0.627**	—
English-Czech News Corpus				
cu	0.523	0.510	0.405	**0.440**
pct	**0.542**	**0.541**	**0.499**	0.381
uedin	0.449	0.433	0.249	0.258

Table 12: Human evaluation for Czech-English submissions

system	METEOR	BLEU	1-TER	GTM	PARAEVAL-RECALL	PARAEVAL-PRECISION	DEPENDENCY-OVERLAP	SEMANTIC-ROLE-OVERLAP	1-WER-OF-VERBS	MAX-CORR-FLUENCY	MAX-CORR-ADEQUACY
German-English Europarl											
cmu-uka	0.559	0.247	0.326	0.455	0.528	0.531	0.259	0.182	0.848	1.91	1.910
liu	0.559	0.263	0.329	0.460	0.537	0.535	0.276	0.197	0.846	1.91	1.910
nrc	0.551	0.253	0.324	0.454	0.528	0.532	0.263	0.185	0.848	1.88	1.88
saar	0.477	0.198	0.313	0.447	0.44	0.527	0.228	0.157	0.846	1.76	1.710
systran	**0.560**	0.268	0.342	0.463	**0.543**	0.541	0.261	0.21	0.849	1.91	1.91
systran-2	0.501	0.154	0.238	0.376	0.462	0.448	0.237	0.154	—	1.71	1.73
uedin	0.56	**0.277**	0.319	**0.480**	0.536	**0.562**	**0.298**	**0.217**	**0.855**	**1.96**	**1.940**
upc	0.541	0.250	**0.343**	0.470	0.506	0.551	0.27	0.193	0.846	1.89	1.88
German-English News Corpus											
nrc	0.563	0.221	0.333	0.454	0.514	0.514	0.246	0.157	0.868	1.920	1.91
saar	0.454	0.159	0.288	0.413	0.405	0.467	0.193	0.120	0.86	1.700	1.64
systran	0.570	0.200	0.275	0.418	0.531	0.472	0.274	0.18	0.858	1.910	1.93
systran-2	0.556	0.169	0.238	0.397	0.511	0.446	0.258	0.163	—	1.86	1.88
uedin	**0.577**	**0.242**	**0.339**	**0.459**	**0.534**	**0.524**	**0.287**	**0.181**	**0.871**	**1.98**	**1.970**
upc	0.575	0.233	**0.339**	0.455	0.527	0.516	0.265	0.171	0.865	1.96	1.96
English-German Europarl											
cmu-uka	0.268	0.189	0.251	—	—	—	—	—	0.884	1.66	1.63
nrc	0.272	0.185	0.221	—	—	—	—	—	0.882	1.660	1.630
saar	0.239	0.174	0.237	—	—	—	—	—	0.881	1.61	1.56
systran	0.198	0.123	0.178	—	—	—	—	—	0.866	1.46	1.42
uedin	**0.277**	**0.201**	**0.273**	—	—	—	—	—	**0.889**	**1.690**	**1.66**
upc	0.266	0.177	0.195	—	—	—	—	—	0.88	1.640	1.62
English-German News Corpus											
nrc	0.257	0.157	0.25	—	—	—	—	—	**0.891**	1.590	1.560
saar	0.162	0.098	0.212	—	—	—	—	—	0.881	1.400	1.310
systran	0.223	0.143	0.266	—	—	—	—	—	0.887	1.55	1.500
ucb	0.256	0.156	0.249	—	—	—	—	—	0.889	1.59	1.56
ucb-2	0.252	0.152	0.229	—	—	—	—	—	—	1.57	1.55
uedin	**0.266**	0.166	**0.266**	—	—	—	—	—	**0.891**	**1.600**	**1.58**
upc	0.256	**0.167**	**0.266**	—	—	—	—	—	0.89	1.590	1.56

Table 13: Automatic evaluation scores for German-English submissions

system	METEOR	BLEU	1-TER	GTM	PARAEVAL-REC	PARAEVAL-PREC	DEPENDENCY	SEMANTIC-ROLE	1-WER-OF-VERBS	MAX-CORR-FLU	MAX-CORR-ADEQ
Spanish-English Europarl											
cmu-syntax	0.602	0.323	0.414	0.499	0.59	0.588	0.338	**0.254**	0.866	**2.10**	**2.090**
cmu-syntax-2	**0.603**	0.321	0.408	0.494	**0.593**	0.584	0.336	0.249	—	2.09	2.09
cmu-uka	0.597	0.32	**0.42**	**0.501**	0.581	**0.595**	0.336	0.247	0.867	2.09	2.080
nrc	0.596	0.313	0.402	0.484	0.581	0.581	0.321	0.227	0.867	2.04	2.04
saar	0.542	0.245	0.32	0.432	0.531	0.511	0.272	0.198	0.854	1.870	1.870
systran	0.593	0.290	0.364	0.469	0.586	0.550	0.321	0.238	0.858	2.02	2.03
systran-2	0.535	0.202	0.288	0.406	0.524	0.49	0.263	0.187	—	1.81	1.84
uedin	0.6	**0.324**	0.414	0.499	0.584	0.589	**0.339**	0.252	**0.868**	2.09	2.080
upc	0.600	0.322	0.407	0.492	0.593	0.583	0.334	0.253	0.865	2.08	2.08
upv	0.594	0.315	0.400	0.493	0.582	0.581	0.329	0.249	0.865	2.060	2.06
Spanish-English News Corpus											
cmu-uka	0.64	0.299	0.428	0.497	0.617	0.575	0.339	0.246	0.89	2.17	2.17
cmu-uka-2	0.64	0.297	0.427	0.496	0.616	0.574	0.339	0.246	—	2.17	2.17
nrc	0.641	0.299	0.434	0.499	0.615	0.584	0.329	0.238	0.892	2.160	2.160
saar	0.607	0.244	0.338	0.447	0.587	0.512	0.303	0.208	0.879	2.04	2.05
systran	0.628	0.259	0.35	0.453	0.611	0.523	0.325	0.221	0.877	2.08	2.10
systran-2	0.61	0.233	0.321	0.438	0.602	0.506	0.311	0.209	—	2.020	2.050
uedin	**0.661**	0.327	0.457	0.512	**0.634**	0.595	**0.363**	0.264	0.893	**2.25**	**2.24**
upc	0.654	**0.346**	**0.480**	**0.528**	0.629	**0.616**	**0.363**	**0.265**	**0.895**	2.240	2.23
upv	0.638	0.283	0.403	0.485	0.614	0.562	0.334	0.234	0.887	2.15	2.140
English-Spanish Europarl											
cmu-uka	**0.333**	0.311	0.389	—	—	—	—	—	0.889	1.98	**2.00**
nrc	0.322	0.299	0.376	—	—	—	—	—	0.886	1.92	1.940
systran	0.269	0.212	0.301	—	—	—	—	—	0.878	1.730	1.760
uedin	0.33	**0.316**	**0.399**	—	—	—	—	—	**0.891**	**1.980**	1.990
upc	0.327	0.312	0.393	—	—	—	—	—	0.89	1.960	1.98
upv	0.323	0.304	0.379	—	—	—	—	—	0.887	1.95	1.97
English-Spanish News Corpus											
cmu-uka	0.368	0.327	0.469	—	—	—	—	—	0.903	2.070	2.090
cmu-uka-2	0.355	0.306	0.461	—	—	—	—	—	—	2.04	2.060
nrc	0.362	0.311	0.448	—	—	—	—	—	0.904	2.04	2.060
systran	0.335	0.281	0.439	—	—	—	—	—	0.906	1.970	2.010
ucb	0.374	**0.331**	0.464	—	—	—	—	—	—	**2.09**	2.11
ucb-2	**0.375**	0.325	0.456	—	—	—	—	—	—	**2.09**	**2.110**
ucb-3	0.372	0.324	0.457	—	—	—	—	—	—	2.08	2.10
uedin	0.361	0.322	**0.479**	—	—	—	—	—	**0.907**	2.08	2.09
upc	0.361	0.328	0.467	—	—	—	—	—	0.902	2.06	2.08
upv	0.337	0.285	0.432	—	—	—	—	—	0.900	1.98	2.000

Table 14: Automatic evaluation scores for Spanish-English submissions

system	METEOR	BLEU	1-TER	GTM	PARAEVAL-REC	PARAEVAL-PREC	DEPENDENCY	SEMANTIC-ROLE	1-WER-OF-VERBS	MAX-CORR-FLU	MAX-CORR-ADEQ
French-English Europarl											
limsi	**0.604**	**0.332**	0.418	0.504	**0.589**	0.591	**0.344**	**0.259**	**0.865**	**2.100**	**2.10**
limsi-2	0.602	0.33	0.417	0.504	0.587	0.592	0.302	0.257	—	2.05	2.05
nrc	0.594	0.312	0.403	0.488	0.578	0.58	0.324	0.244	0.861	2.05	2.050
saar	0.534	0.249	0.354	0.459	0.512	0.546	0.279	0.202	0.856	1.880	1.88
systran	0.549	0.211	0.308	0.417	0.525	0.501	0.277	0.201	0.849	1.850	1.890
systran-nrc	0.594	0.313	0.404	0.492	0.578	0.580	0.330	0.248	0.862	2.06	2.060
uedin	0.595	0.318	**0.424**	**0.505**	0.574	**0.599**	0.338	0.254	**0.865**	2.08	2.08
upc	0.6	0.319	0.409	0.495	0.588	0.583	0.337	0.255	0.861	2.08	2.080
French-English News Corpus											
limsi	0.595	**0.279**	**0.405**	**0.478**	0.563	**0.555**	0.289	**0.235**	0.875	2.030	2.020
nrc	0.587	0.257	0.389	0.470	0.557	0.546	0.301	0.22	**0.876**	2.020	2.020
saar	0.503	0.206	0.301	0.418	0.475	0.476	0.245	0.169	0.864	1.80	1.78
systran	0.568	0.202	0.28	0.415	0.554	0.472	0.292	0.198	0.866	1.930	1.96
systran-nrc	0.591	0.269	0.398	0.475	0.558	0.547	0.323	0.226	0.875	2.050	2.06
uedin	**0.602**	0.27	0.392	0.471	**0.569**	0.545	**0.326**	0.233	0.875	**2.07**	**2.07**
upc	0.596	0.275	0.400	0.476	0.567	0.552	0.322	0.233	**0.876**	2.06	2.06
English-French Europarl											
limsi	**0.226**	**0.306**	0.366	—	—	—	—	—	0.891	**1.940**	1.96
nrc	0.218	0.294	0.354	—	—	—	—	—	0.888	1.930	1.96
saar	0.190	0.262	0.333	—	—	—	—	—	**0.892**	1.86	1.87
systran	0.179	0.233	0.313	—	—	—	—	—	0.885	1.79	1.83
systran-nrc	0.220	0.301	0.365	—	—	—	—	—	**0.892**	**1.940**	**1.960**
uedin	0.207	0.262	0.301	—	—	—	—	—	0.886	1.930	1.950
upc	0.22	0.299	**0.379**	—	—	—	—	—	**0.892**	**1.940**	**1.960**
English-French News Corpus											
limsi	0.206	0.255	0.354	—	—	—	—	—	0.897	1.84	1.87
nrc	0.208	0.257	0.369	—	—	—	—	—	0.9	1.87	1.900
saar	0.151	0.188	0.308	—	—	—	—	—	0.896	1.65	1.65
systran	0.199	0.243	0.378	—	—	—	—	—	0.901	1.860	1.90
systran-nrc	**0.23**	**0.290**	**0.408**	—	—	—	—	—	**0.903**	**1.940**	**1.98**
ucb	0.201	0.237	0.366	—	—	—	—	—	0.897	1.830	1.860
uedin	0.197	0.234	0.340	—	—	—	—	—	0.899	1.87	1.890
upc	0.212	0.263	0.391	—	—	—	—	—	0.900	1.87	1.90

Table 15: Automatic evaluation scores for French-English submissions

system	METEOR	BLEU	1-TER	GTM	PARAEVAL-REC	PARAEVAL-PREC	DEPENDENCY	SEMANTIC-ROLE	1-WER-OF-VERBS	MAX-CORR-FLU	MAX-CORR-ADEQ
Czech-English News Corpus											
cu	0.545	0.215	0.334	0.441	0.502	0.504	0.245	0.165	0.867	1.87	1.88
cu-2	0.558	0.223	0.344	0.447	0.510	0.514	0.254	0.17	—	1.90	1.910
uedin	0.54	0.217	0.340	0.445	0.497	0.51	0.243	0.160	0.865	1.860	1.870
umd	**0.581**	**0.241**	**0.355**	**0.460**	**0.531**	**0.526**	**0.273**	**0.184**	**0.868**	**1.96**	**1.97**
English-Czech News Corpus											
cu	0.429	**0.134**	**0.231**	—	—	—	—	—	—	**1.580**	**1.53**
cu-2	**0.430**	0.132	0.219	—	—	—	—	—	—	1.58	1.520
uedin	0.42	0.119	0.211	—	—	—	—	—	—	1.550	1.49

Table 16: Automatic evaluation scores for Czech-English submissions

	ADEQUACY	FLUENCY	RANK	CONSTITUENT	METEOR	BLEU	1-TER	GTM	PARAEVAL-REC	PARAEVAL-PREC	DEPENDENCY	SEMANTIC-ROLE	1-WER-OF-VS	MAX-CORR-FLU	MAX-CORR-ADEQ
German-English News Corpus															
adequacy	1	0.900	0.900	0.900	0.600	0.300	-0.025	0.300	**0.700**	0.300	**0.700**	**0.700**	-0.300	0.300	0.600
fluency	—	1	1.000	1.000	0.700	0.400	-0.025	0.400	**0.900**	0.400	**0.900**	**0.900**	-0.100	0.400	0.700
rank	—	—	1	1.000	0.700	0.400	-0.025	0.400	**0.900**	0.400	**0.900**	**0.900**	-0.100	0.400	0.700
constituent	—	—	—	1	0.700	0.400	-0.025	0.400	**0.900**	0.400	**0.900**	**0.900**	-0.100	0.400	0.700
German-English Europarl															
adequacy	1	0.893	0.821	0.750	0.599	0.643	**0.787**	0.68	0.750	0.643	0.464	0.750	0.206	0.608	0.447
fluency	—	1	0.964	0.537	0.778	0.858	0.500	0.821	0.821	0.787	0.571	**0.93**	0.562	0.821	0.661
rank	—	—	1	0.500	**0.902**	0.821	0.393	0.714	0.858	0.643	0.464	0.858	0.652	0.893	0.769
constituent	—	—	—	1	0.456	0.464	0.714	0.18	**0.750**	0.250	0.214	0.43	0.117	0.214	0.126
Spanish-English News Corpus															
adequacy	1	1.000	0.964	0.893	0.643	0.68	0.68	0.68	0.68	0.68	0.634	**0.714**	0.571	0.68	0.68
fluency	—	1	0.964	0.893	0.643	0.68	0.68	0.68	0.68	0.68	0.634	**0.714**	0.571	0.68	0.68
rank	—	—	1	0.858	0.714	0.750	0.750	0.750	0.750	0.750	0.741	**0.787**	0.608	0.750	0.750
constituent	—	—	—	1	0.787	**0.821**	**0.821**	**0.821**	0.714	**0.821**	0.599	0.750	0.750	0.714	0.714
Spanish-English Europarl															
adequacy	1	0.93	0.452	0.333	0.596	**0.810**	0.62	0.690	0.542	0.714	0.762	0.739	0.489	0.638	0.638
fluency	—	1	0.571	0.524	0.596	0.787	0.43	0.500	0.732	0.524	0.690	**0.810**	0.346	0.566	0.566
rank	—	—	1	0.643	0.739	0.596	0.43	0.262	**0.923**	0.406	0.500	0.739	0.168	0.542	0.542
constituent	—	—	—	1	0.262	0.143	-0.143	-0.143	**0.816**	-0.094	0.000	0.477	-0.226	0.042	0.042
French-English News Corpus															
adequacy	1	0.964	0.964	0.858	**0.787**	0.750	0.68	0.68	**0.787**	0.571	0.321	**0.787**	0.456	0.68	0.554
fluency	—	1	1.000	0.93	0.750	0.787	0.714	0.714	0.750	0.608	0.214	**0.858**	0.367	0.608	0.482
rank	—	—	1	0.93	0.750	0.787	0.714	0.714	0.750	0.608	0.214	**0.858**	0.367	0.608	0.482
constituent	—	—	—	1	0.858	0.858	0.787	0.787	0.858	0.643	0.393	**0.964**	0.349	0.750	0.661
French-English Europarl															
adequacy	1	0.884	0.778	0.991	0.982	0.956	0.902	0.902	0.812	0.902	0.956	0.956	0.849	0.964	**0.991**
fluency	—	1	0.858	0.893	0.849	0.821	**0.93**	**0.93**	0.571	**0.93**	0.858	0.821	0.787	0.849	0.858
rank	—	—	1	0.821	0.670	0.68	0.858	0.858	0.43	0.858	0.787	0.68	**0.893**	0.741	0.714
constituent	—	—	—	1	0.956	0.93	0.93	0.93	0.750	0.93	**0.964**	0.93	0.893	0.956	**0.964**

Table 17: Correlation of the automatic evaluation metrics with the human judgments when translating into English

	ADEQUACY	FLUENCY	RANK	CONSTITUENT	METEOR	BLEU	1-TER	1-WER-OF-VS	MAX-CORR-FLU	MAX-CORR-ADEQ
English-German News Corpus										
adequacy	1	0.943	0.83	0.943	0.187	0.43	**0.814**	0.243	0.33	0.187
fluency	—	1	0.714	0.83	0.100	0.371	**0.758**	0.100	0.243	0.100
rank	—	—	1	0.771	0.414	0.258	**0.671**	0.414	0.414	0.414
constituent	—	—	—	1	0.13	0.371	**0.671**	0.243	0.243	0.13
English-German Europarl										
adequacy	1	0.714	0.487	0.714	0.487	**0.600**	0.314	0.371	0.487	0.487
fluency	—	1	0.543	0.43	**0.258**	0.200	-0.085	0.03	**0.258**	**0.258**
rank	—	—	1	0.03	-0.37	-0.256	-0.543	-0.485	-0.37	-0.37
constituent	—	—	—	1	0.887	**0.943**	0.658	0.83	0.887	0.887
English-Spanish News Corpus										
adequacy	1	0.714	0.771	0.83	0.314	**0.658**	0.487	0.03	0.314	0.600
fluency	—	1	0.943	0.887	-0.200	0.03	0.143	0.200	-0.085	**0.258**
rank	—	—	1	0.943	-0.029	0.087	0.258	**0.371**	-0.029	**0.371**
constituent	—	—	—	1	-0.143	0.143	0.200	**0.314**	-0.085	0.258
English-Spanish Europarl										
adequacy	1	0.83	0.943	0.543	0.658	**0.943**	**0.943**	**0.943**	0.83	0.658
fluency	—	1	0.771	0.543	0.714	0.771	0.771	0.771	**0.83**	0.714
rank	—	—	1	0.600	0.600	**0.887**	**0.887**	**0.887**	0.771	0.600
constituent	—	—	—	1	**0.43**	**0.43**	**0.43**	**0.43**	0.371	**0.43**
English-French News Corpus										
adequacy	1	0.952	0.762	0.452	0.690	**0.787**	0.690	0.709	0.596	0.686
fluency	—	1	0.810	0.477	0.62	0.739	0.714	**0.792**	0.62	0.780
rank	—	—	1	0.762	0.239	0.381	0.500	**0.757**	0.596	0.601
constituent	—	—	—	1	-0.048	0.096	0.143	**0.411**	0.333	0.304
English-French Europarl										
adequacy	1	0.964	0.750	0.93	0.608	0.528	0.287	-0.07	**0.652**	0.376
fluency	—	1	0.858	0.893	0.643	0.562	0.214	-0.07	**0.652**	0.376
rank	—	—	1	0.750	0.821	0.76	0.393	0.214	**0.830**	0.697
constituent	—	—	—	1	0.571	0.473	0.18	-0.07	**0.652**	0.447

Table 18: Correlation of the automatic evaluation metrics with the human judgments when translating out of English