Findings of the 2009 Workshop on Statistical Machine Translation

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

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

Christof Monz
University of Amsterdam
christof@science.uva.nl

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

Abstract

This paper presents the results of the WMT09 shared tasks, which included a translation task, a system combination task, and an evaluation task. We conducted a large-scale manual evaluation of 87 machine translation systems and 22 system combination entries. We used the ranking of these systems to measure how strongly automatic metrics correlate with human judgments of translation quality, for more than 20 metrics. We present a new evaluation technique whereby system output is edited and judged for correctness.

1 Introduction

This paper presents the results of the shared tasks of the 2009 EACL Workshop on Statistical Machine Translation, which builds on three previous workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007; Callison-Burch et al., 2008). There were three shared tasks this year: a translation task between English and five other European languages, a task to combine the output of multiple machine translation systems, and a task to predict human judgments of translation quality using automatic evaluation metrics. The performance on each of these shared task was determined after a comprehensive human evaluation.

There were a number of differences between this year's workshop and last year's workshop:

- **Larger training sets** – In addition to annual increases in the Europarl corpus, we released a French-English parallel corpus verging on 1 billion words. We also provided large monolingual training sets for better language modeling of the news translation task.
- **Reduced number of conditions** – Previous workshops had many conditions: 10 language pairs, both in-domain and out-of-domain translation, and three types of manual evaluation. This year we eliminated the in-domain Europarl test set and defined sentence-level ranking as the primary type of manual evaluation.
- **Editing to evaluate translation quality** – Beyond ranking the output of translation systems, we evaluated translation quality by having people edit the output of systems. Later, we asked annotators to judge whether those edited translations were correct when shown the source and reference translation.

The primary objectives of this workshop are to evaluate the state of the art in machine translation, to disseminate common test sets and public training data with published performance numbers, and to refine evaluation methodologies for machine translation. All of the data, translations, and human judgments produced for our workshop are publicly available.[1] We hope they form a valuable resource for research into statistical machine translation, system combination, and automatic evaluation of translation quality.

2 Overview of the shared translation and system combination tasks

The workshop examined translation between English and five other languages: German, Spanish, French, Czech, and Hungarian. We created a test set for each language pair by translating newspaper articles. We additionally provided training data and a baseline system.

[1] http://statmt.org/WMT09/results.html

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 1–28,
Athens, Greece, 30 March – 31 March 2009.

2.1 Test data

The test data for this year's task was created by hiring people to translate news articles that were drawn from a variety of sources during the period from the end of September to mid-October of 2008. A total of 136 articles were selected, in roughly equal amounts from a variety of Czech, English, French, German, Hungarian, Italian and Spanish news sites:[2]

Hungarian: hvg.hu (10), Napi (2), MNO (4), Népszabadság (4)

Czech: iHNed.cz (3), iDNES.cz (4), Lidovky.cz (3), aktuálně.cz (2), Novinky (1)

French: dernieresnouvelles (1), Le Figaro (2), Les Echos (4), Liberation (4), Le Devoir (9)

Spanish: ABC.es (11), El Mundo (12)

English: BBC (11), New York Times (6), Times of London (4),

German: Süddeutsche Zeitung (3), Frankfurter Allgemeine Zeitung (3), Spiegel (8), Welt (3)

Italian: ADN Kronos (5), Affari Italiani (2), ASCA (1), Corriere della Sera (4), Il Sole 24 ORE (1), Il Quotidiano (1), La Republica (8)

Note that Italian translation was not one of this year's official translation tasks.

The translations were created by the members of EuroMatrix consortium who hired a mix of professional and non-professional translators. All translators were fluent or native speakers of both languages. Although we made efforts to proofread all translations, many sentences still contain minor errors and disfluencies. All of the translations were done directly, and not via an intermediate language. For instance, each of the 20 Hungarian articles were translated directly into Czech, English, French, German, Italian and Spanish. The total cost of creating the test sets consisting of roughly 80,000 words across 3027 sentences in seven languages was approximately 31,700 euros (around 39,800 dollars at current exchange rates, or slightly more than $0.08/word).

Previous evaluations additionally used test sets drawn from the Europarl corpus. Our rationale behind discontinuing the use of Europarl as a test set was that it overly biases towards statistical systems that were trained on this particular domain, and that European Parliament proceedings were less of general interest than news stories. We focus on a single task since the use of multiple test sets in the past spread our resources too thin, especially in the manual evaluation.

[2]For more details see the XML test files. The `docid` tag gives the source and the date for each document in the test set, and the `origlang` tag indicates the original source language.

2.2 Training data

As in past years we provided parallel corpora to train translation models, monolingual corpora to train language models, and development sets to tune parameters. Some statistics about the training materials are given in Figure 1.

10^9 word parallel corpus

To create the large French-English parallel corpus, we conducted a targeted web crawl of bilingual web sites. These sites came from a variety of sources including the Canadian government, the European Union, the United Nations, and other international organizations. The crawl yielded on the order of 40 million files, consisting of more than 1TB of data. Pairs of translated documents were identified using a set of simple heuristics to transform French URLs into English URLs (for instance, by replacing *fr* with *en*). Documents that matched were assumed to be translations of each other.

All HTML and PDF documents were converted into plain text, which yielded 2 million French files paired with their English equivalents. Text files were split so that they contained one sentence per line and had markers between paragraphs. They were sentence-aligned in batches of 10,000 document pairs, using a sentence aligner that incorporates IBM Model 1 probabilities in addition to sentence lengths (Moore, 2002). The document-aligned corpus contained 220 million segments with 2.9 billion words on the French side and 215 million segments with 2.5 billion words on the English side. After sentence alignment, there were 177 million sentence pairs with 2.5 billion French words and 2.2 billion English words.

The sentence-aligned corpus was cleaned to remove sentence pairs which consisted only of numbers or paragraph markers, or where the French and English sentences were identical. The later step helped eliminate documents that were not actually translated, which was necessary because we did not perform language identification. After cleaning, the parallel corpus contained 105 million sentence pairs with 2 billion French words and 1.8 billion English words.

Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English	
Sentences	1,411,589		1,428,799		1,418,115	
Words	40,067,498	41,042,070	44,692,992	40,067,498	39,516,645	37,431,872
Distinct words	154,971	108,116	129,166	107,733	320,180	104,269

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	74,512		64,223		82,740		79,930	
Words	2,052,186	1,799,312	1,831,149	1,560,274	2,051,369	1,977,200	1,733,865	1,891,559
Distinct words	56,578	41,592	46,056	38,821	92,313	43,383	105,280	41,801

10^9 Word Parallel Corpus

	French ↔ English	
Sentences	22,520,400	
Words	811,203,407	668,412,817
Distinct words	2,738,882	2,861,836

Hunglish Training Corpus

	Hungarian ↔ English	
Sentences	1,517,584	
Words	26,114,985	31,467,693
Distinct words	717,198	192,901

CzEng Training Corpus

	Czech ↔ English	
Sentences	1,096,940	
Words	15,336,783	17,909,979
Distinct words	339,683	129,176

Europarl Language Model Data

	English	Spanish	French	German
Sentence	1,658,841	1,607,419	1,676,435	1,713,715
Words	44,983,136	45,382,287	50,577,097	41,457,414
Distinct words	117,577	162,604	138,621	348,197

News Language Model Data

	English	Spanish	French	German	Czech	Hungarian
Sentence	21,232,163	1,626,538	6,722,485	10,193,376	5,116,211	4,209,121
Words	504,094,159	48,392,418	167,204,556	185,639,915	81,743,223	86,538,513
Distinct words	1,141,895	358,664	660,123	1,668,387	929,318	1,313,578

News Test Set

	English	Spanish	French	German	Czech	Hungarian	Italian
Sentences	2525						
Words	65,595	68,092	72,554	62,699	55,389	54,464	64,906
Distinct words	8,907	10,631	10,609	12,277	15,387	16,167	11,046

News System Combination Development Set

	English	Spanish	French	German	Czech	Hungarian	Italian
Sentences	502						
Words	11,843	12,499	12,988	11,235	9,997	9,628	11,833
Distinct words	2,940	3,176	3,202	3,471	4,121	4,133	3,318

Figure 1: Statistics for the training and test sets used in the translation task. The number of words is based on the provided tokenizer and the number of distinct words is the based on lowercased tokens.

In addition to cleaning the sentence-aligned parallel corpus we also de-duplicated the corpus, removing all sentence pairs that occured more than once in the parallel corpus. Many of the documents gathered in our web crawl were duplicates or near duplicates, and a lot of the text is repeated, as with web site navigation. We further eliminated sentence pairs that varied from previous sentences by only numbers, which helped eliminate template web pages such as expense reports. We used a Bloom Filter (Talbot and Osborne, 2007) to do de-duplication, so it may have discarded more sentence pairs than strictly necessary. After de-duplication, the parallel corpus contained 28 million sentence pairs with 0.8 billion French words and 0.7 billion English words.

Monolingual news corpora

We have crawled the news sources that were the basis of our test sets (and a few more additional sources) since August 2007. This allowed us to assemble large corpora in the target domain to be mainly used as training data for language modeling. We collected texts from the beginning of our data collection period to one month before the test set period, segmented these into sentences and randomized the order of the sentences to obviate copyright concerns.

2.3 Baseline system

To lower the barrier of entry for newcomers to the field, we provided Moses, an open source toolkit for phrase-based statistical translation (Koehn et al., 2007). The performance of this baseline system is similar to the best submissions in last year's shared task. Twelve participating groups used the Moses toolkit for the development of their system.

2.4 Submitted systems

We received submissions from 22 groups from 20 institutions, as listed in Table 1, a similar turnout to last year's shared task. Of the 20 groups that participated with regular system submissions in last year's shared task, 12 groups returned this year. A major hurdle for many was a DARPA/GALE evaluation that occurred at the same time as this shared task.

We also evaluated 7 commercial rule-based MT systems, and Google's online statistical machine translation system. We note that Google did not submit an entry itself. Its entry was created by the WMT09 organizers using Google's online system.[3] In personal correspondence, Franz Och clarified that the online system is different from Google's research system in that it runs at faster speeds at the expense of somewhat lower translation quality. On the other hand, the training data used by Google is unconstrained, which means that it may have an advantage compared to the research systems evaluated in this workshop, since they were trained using only the provided materials.

2.5 System combination

In total, we received 87 primary system submissions along with 42 secondary submissions. These were made available to participants in the system combination shared task. Based on feedback that we received on last year's system combination task, we provided two additional resources to participants:

- Development set: We reserved 25 articles to use as a dev set for system combination (details of the set are given in Table 1). These were translated by all participating sites, and distributed to system combination participants along with reference translations.

- n-best translations: We requested n-best lists from sites whose systems could produce them. We received 25 100-best lists accompanying the primary system submissions, and 5 accompanying the secondary system submissions.

In addition to soliciting system combination entries for each of the language pairs, we treated system combination as a way of doing *multi-source* translation, following Schroeder et al. (2009). For the multi-source system combination task, we provided all 46 primary system submissions from any language into English, along with an additional 32 secondary systems.

Table 2 lists the six participants in the system combination task.

3 Human evaluation

As with past workshops, we placed greater emphasis on the human evaluation than on the automatic evaluation metric scores. It is our contention

[3] `http://translate.google.com`

ID	Participant
CMU-STATXFER	Carnegie Mellon University's statistical transfer system (Hanneman et al., 2009)
COLUMBIA	Columbia University (Carpuat, 2009)
CU-BOJAR	Charles University Bojar (Bojar et al., 2009)
CU-TECTOMT	Charles University Tectogramatical MT (Bojar et al., 2009)
DCU	Dublin City University (Du et al., 2009)
EUROTRANXP	commercial MT provider from the Czech Republic
GENEVA	University of Geneva (Wehrli et al., 2009)
GOOGLE	Google's production system
JHU	Johns Hopkins University (Li et al., 2009)
JHU-TROMBLE	Johns Hopkins University Tromble (Eisner and Tromble, 2006)
LIMSI	LIMSI (Allauzen et al., 2009)
LIU	Linköping University (Holmqvist et al., 2009)
LIUM-SYSTRAN	University of Le Mans / Systran (Schwenk et al., 2009)
MORPHO	Morphologic (Novák, 2009)
NICT	National Institute of Information and Comm. Tech., Japan (Paul et al., 2009)
NUS	National University of Singapore (Nakov and Ng, 2009)
PCTRANS	commercial MT provider from the Czech Republic
RBMT1-5	commercial systems from Learnout&Houspie, Lingenio, Lucy, PROMT, SDL
RWTH	RWTH Aachen (Popovic et al., 2009)
STUTTGART	University of Stuttgart (Fraser, 2009)
SYSTRAN	Systran (Dugast et al., 2009)
TALP-UPC	Universitat Politecnica de Catalunya, Barcelona (R. Fonollosa et al., 2009)
UEDIN	University of Edinburgh (Koehn and Haddow, 2009)
UKA	University of Karlsruhe (Niehues et al., 2009)
UMD	University of Maryland (Dyer et al., 2009)
USAAR	University of Saarland (Federmann et al., 2009)

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

ID	Participant
BBN-COMBO	BBN system combination (Rosti et al., 2009)
CMU-COMBO	Carnegie Mellon University system combination (Heafield et al., 2009)
CMU-COMBO-HYPOSEL	CMU system comb. with hyp. selection (Hildebrand and Vogel, 2009)
DCU-COMBO	Dublin City University system combination
RWTH-COMBO	RWTH Aachen system combination (Leusch et al., 2009)
USAAR-COMBO	University of Saarland system combination (Chen et al., 2009)

Table 2: Participants in the system combination task.

Language Pair	Sentence Ranking	Edited Translations	Yes/No Judgments
German-English	3,736	1,271	4,361
English-German	3,700	823	3,854
Spanish-English	2,412	844	2,599
English-Spanish	1,878	278	837
French-English	3,920	1,145	4,491
English-French	1,968	332	1,331
Czech-English	1,590	565	1,071
English-Czech	7,121	2,166	9,460
Hungarian-English	1,426	554	1,309
All-English	4,807	0	0
Multisource-English	2,919	647	2184
Totals	**35,786**	**8,655**	**31,524**

Table 3: The number of items that were judged for each task during the manual evaluation.

that automatic measures are an imperfect substitute for human assessment of translation quality. Therefore, we define the manual evaluation to be primary, and use the human judgments to validate automatic metrics.

Manual evaluation is time consuming, and it requires a large effort to conduct it on the scale of our workshop. We distributed the workload across a number of people, including shared-task participants, interested volunteers, and a small number of paid annotators. More than 160 people participated in the manual evaluation, with 100 people putting in more than an hour's worth of effort, and 30 putting in more than four hours. A collective total of 479 hours of labor was invested.

We asked people to evaluate the systems' output in two different ways:

- Ranking translated sentences relative to each other. This was our official determinant of translation quality.

- Editing the output of systems without displaying the source or a reference translation, and then later judging whether edited translations were correct.

The total number of judgments collected for the different modes of annotation is given in Table 3.

In all cases, the output of the various translation outputs were judged on equal footing; the output of system combinations was judged alongside that of the individual system, and the constrained and unconstrained systems were judged together.

3.1 Ranking translations of sentences

Ranking translations relative to each other is a reasonably intuitive task. We therefore kept the instructions simple:

> *Rank translations from Best to Worst relative to the other choices (ties are allowed).*

In our the manual evaluation, annotators were shown at most five translations at a time. For most language pairs there were more than 5 systems submissions. We did not attempt to get a complete ordering over the systems, and instead relied on random selection and a reasonably large sample size to make the comparisons fair.

Relative ranking is our official evaluation metric. Individual systems and system combinations are ranked based on how frequently they were judged to be better than or equal to any other system. The results of this are reported in Section 4. Appendix A provides detailed tables that contain pairwise comparisons between systems.

3.2 Editing machine translation output

We experimented with a new type of evaluation this year where we asked judges to edit the output of MT systems. We did not show judges the reference translation, which makes our edit-based evaluation different than the Human-targeted Translation Error Rate (HTER) measure used in the DARPA GALE program (NIST, 2008). Rather than asking people to make the minimum number of changes to the MT output in order capture the same meaning as the reference, we asked them to

edit the translation to be as fluent as possible without seeing the reference. Our hope was that this would reflect people's understanding of the output.

The instructions that we gave our judges were the following:

> *Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select "No corrections needed." If you cannot understand the sentence well enough to correct it, select "Unable to correct."*

Each translated sentence was shown in isolation without any additional context. A screenshot is shown in Figure 2.

Since we wanted to prevent judges from seeing the reference before editing the translations, we split the test set between the sentences used in the ranking task and the editing task (because they were being conducted concurrently). Moreover, annotators edited only a single system's output for one source sentence to ensure that their understanding of it would not be influenced by another system's output.

3.3 Judging the acceptability of edited output

Halfway through the manual evaluation period, we stopped collecting edited translations, and instead asked annotators to do the following:

> *Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is* **bold**.

In addition to edited translations, unedited items that were either marked as acceptable or as incomprehensible were also shown. Judges gave a simple yes/no indication to each item. A screenshot is shown in Figure 3.

3.4 Inter- and Intra-annotator agreement

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

INTER-ANNOTATOR AGREEMENT

Evaluation type	$P(A)$	$P(E)$	K
Sentence ranking	.549	.333	.323
Yes/no to edited output	.774	.5	.549

INTRA-ANNOTATOR AGREEMENT

Evaluation type	$P(A)$	$P(E)$	K
Sentence ranking	.707	.333	.561
Yes/no to edited output	.866	.5	.732

Table 4: Inter- and intra-annotator agreement for the two types of manual evaluation

We measured pairwise agreement among annotators using the kappa coefficient (K) which 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.

For inter-annotator agreement we calculated $P(A)$ for the yes/no judgments 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 $A < B$. Intra-annotator agreement was computed similarly, but we gathered items that were annotated on multiple occasions by a single annotator.

Table 4 gives K values for inter-annotator and intra-annotator 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, $.2 - .4$ is fair, $.4 - .6$ is moderate, $.6 - .8$ is substantial and the rest almost perfect.

Based on these interpretations the agreement for yes/no judgments is *moderate* for inter-annotator agreement and *substantial* for intra-annotator agreement, but the inter-annotator agreement for sentence level ranking is only *fair*.

We analyzed two possible strategies for improving inter-annotator agreement on the ranking task: First, we tried discarding initial judgments to give

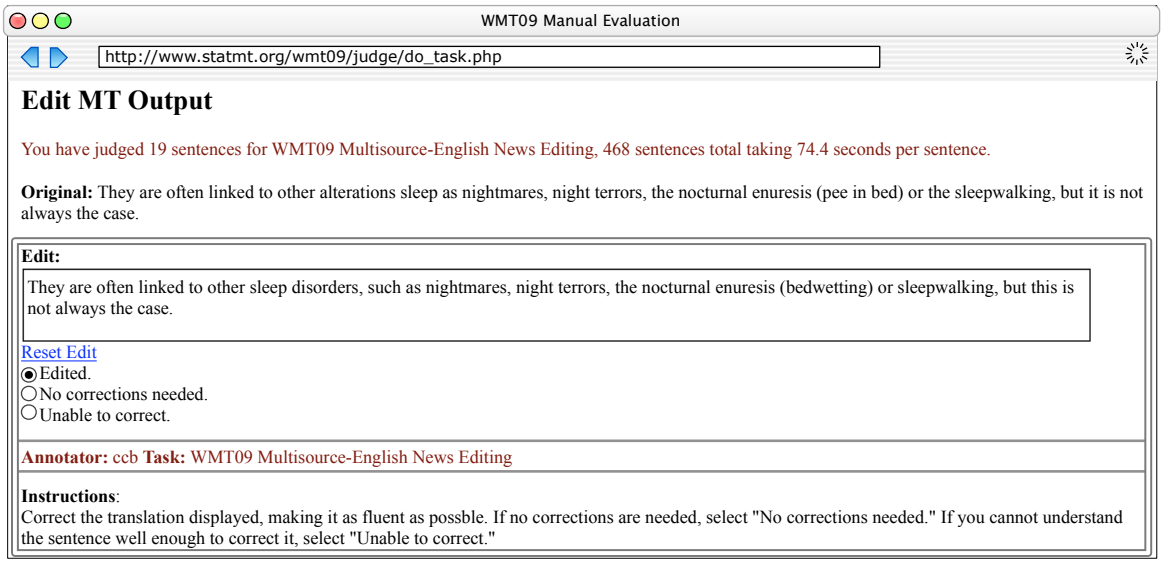

Figure 2: This screenshot shows an annotator editing the output of a machine translation system.

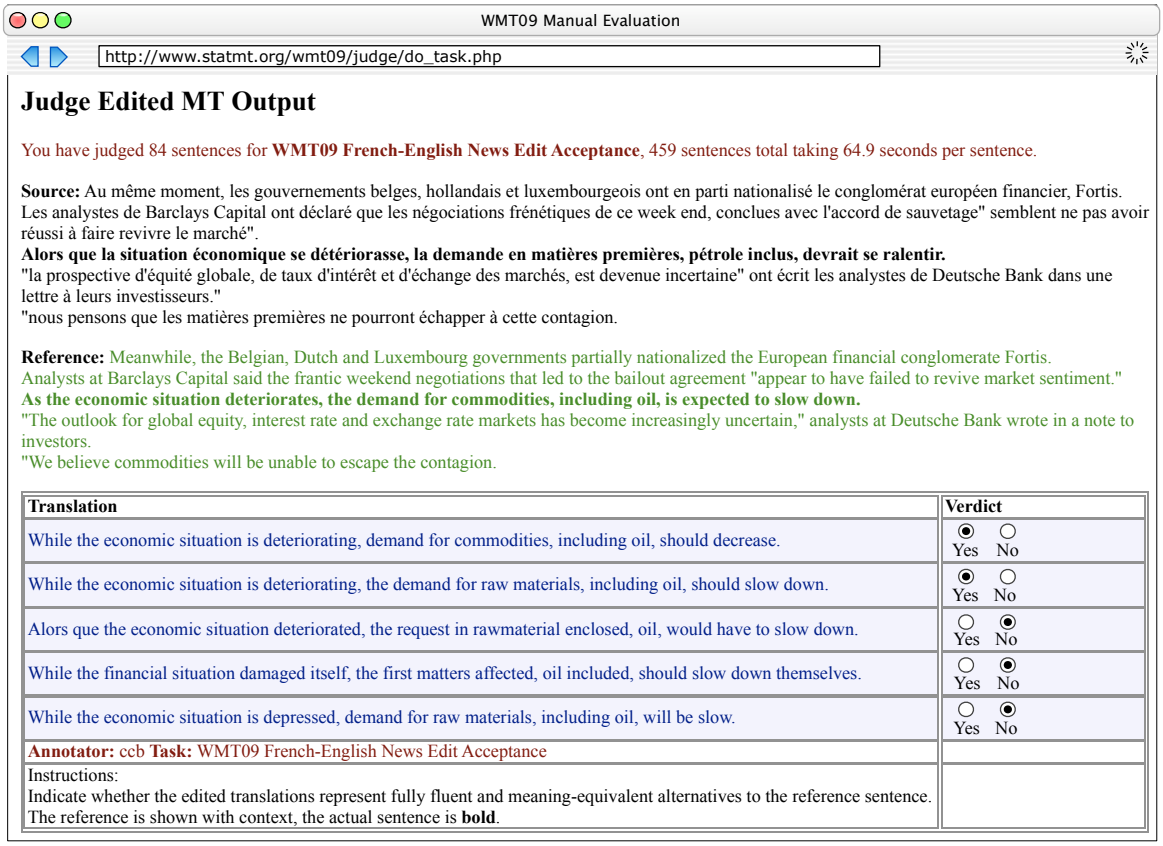

Figure 3: This screenshot shows an annotator judging the acceptability of edited translations.

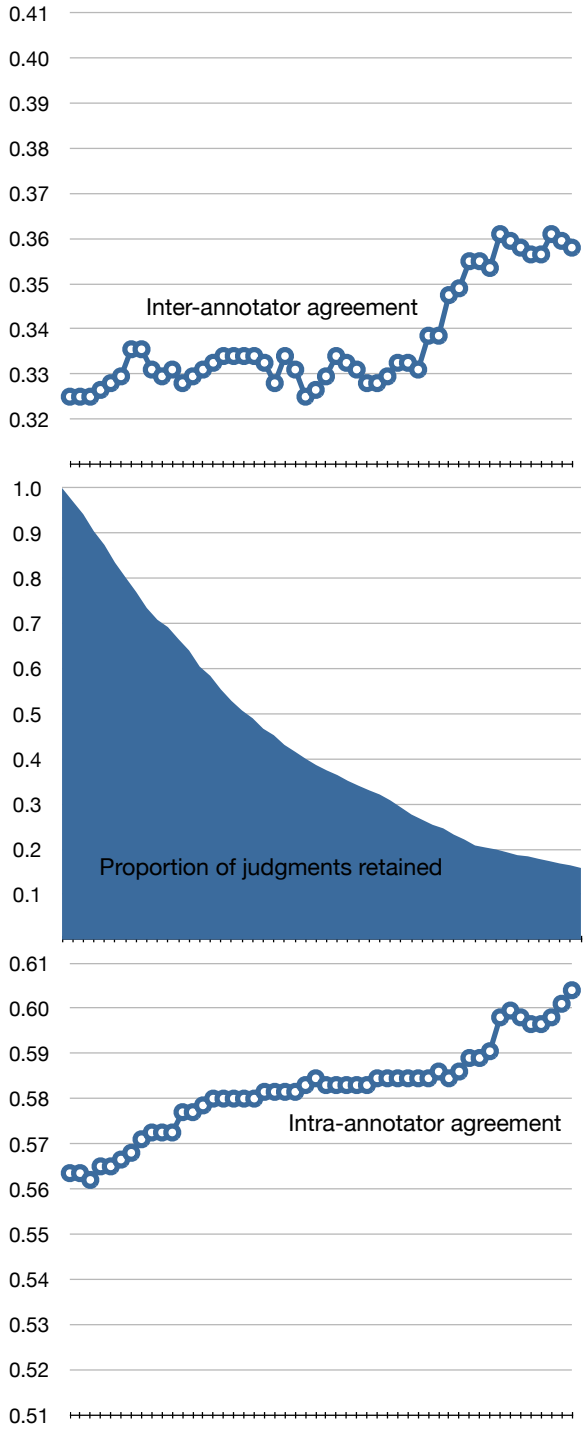

Figure 4: The effect of discarding every annotators' initial judgments, up to the first 50 items

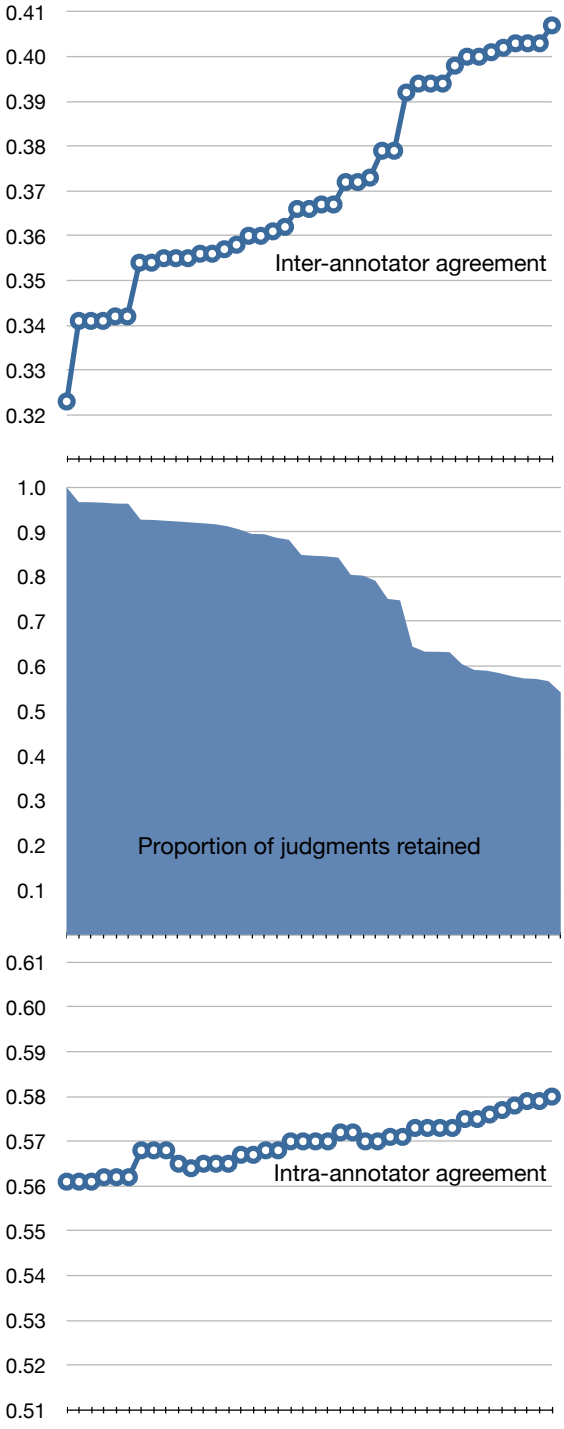

Figure 5: The effect of removing annotators with the lowest agreement, disregarding up to 40 annotators

annotators a chance to learn to how to perform the task. Second, we tried disregarding annotators who have very low agreement with others, by throwing away judgments for the annotators with the lowest judgments.

Figures 4 and 5 show how the K values improve for intra- and inter-annotator agreement under these two strategies, and what percentage of the judgments are retained as more annotators are removed, or as the initial learning period is made longer. It seems that the strategy of removing the worst annotators is the best in terms of improving inter-annotator K, while retaining most of the judgments. If we remove the 33 judges with the worst agreement, we increase the inter-annotator K from *fair* to *moderate*, and still retain 60% of the data.

For the results presented in the rest of the paper, we retain all judgments.

4 Translation task results

We used the results of the manual evaluation to analyze the translation quality of the different systems that were submitted to the workshop. In our analysis, we aimed to address the following questions:

- Which systems produced the best translation quality for each language pair?
- Did the system combinations produce better translations than individual systems?
- Which of the systems that used only the provided training materials produced the best translation quality?

Table 6 shows best individual systems. We define the best systems as those which had no other system that was statistically significantly better than them under the Sign Test at $p \leq 0.1$.[4] Multiple systems are listed for many language pairs because it was not possible to draw a statistically significant difference between the systems. Commercial translation software (including Google, Systran, Morphologic, PCTrans, Eurotran XP, and anonymized RBMT providers) did well in each of the language pairs. Research systems that utilized only the provided data did as well as commercial vendors in half of the language pairs.

The table also lists the best systems among those which used only the provided materials. To determine this decision we excluded *unconstrained systems* which employed significant external resources. Specifically, we ruled out all of the commercial systems, since Google has access to significantly greater data sources for its statistical system, and since the commercial RBMT systems utilize knowledge sources not available to other workshop participants. The remaining systems were research systems that employ statistical models. We were able to draw distinctions between half of these for each of the language pairs. There are some borderline cases, for instance LIMSI only used additional monolingual training resources, and LIUM/Systran used additional translation dictionaries as well as additional monolingual resources.

Table 5 summarizes the performance of the system combination entries by listing the best ranked combinations, and by indicating whether they have a statistically significant difference with the best individual systems. In general, system combinations performed as well as the best individual systems, but not statistically significantly better than them. Moreover, it was hard to draw a distinction between the different system combination strategies themselves. There are a number of possibilities as to why we failed to find significant differences:

- The number of judgments that we collected were not sufficient to find a difference. Although we collected several thousand judgments for each language pair, most pairs of systems were judged together fewer than 100 times.
- It is possible that the best performing individual systems were sufficiently better than the other systems and that it is difficult to improve on them by combining them.
- Individual systems could have been weighted incorrectly during the development stage, which could happen if the automatic evaluation metrics scores on the dev set did not strongly correlate with human judgments.
- The lack of distinction between different combinations could be due to the fact that

[4] In one case this definition meant that the system that was ranked the highest overall was not considered to be one of the best systems. For German-English translation RBMT5 was ranked highest overall, but was statistically significantly worse than RBMT2.

Language Pair	Best system combinations	Entries	Significantly different than best individual systems?
German-English	RWTH-COMBO, BBN-COMBO, CMU-COMBO, USAAR-COMBO	5	BBN-COMBO>GOOGLE, SYSTRAN, USAAR-COMBO<RMBT2, no difference for others
English-German	USAAR-COMBO	1	worse than 3 best systems
Spanish-English	CMU-COMBO, USAAR-COMBO, BBN-COMBO	3	each better than one of the RBMT systems, but there was no difference with GOOGLE, TALP-UPC
English-Spanish	USAAR-COMBO	1	no difference
French-English	CMU-COMBO-HYPOSEL, DCU-COMBO, CMU-COMBO	5	no difference
English-French	USAAR-COMBO, DCU-COMBO	2	USAAR-COMBO>UKA, DCU-COMBO>SYSTRAN, LIMSI, no difference with others
Czech-English	CMU-COMBO	2	no difference
Hungarian-English	CMU-COMBO-HYPOSEL, CMU-COMBO	3	both worse than MORPHO
Multisource-English	RWTH-COMBO	3	n/a

Table 5: A comparison between the best system combinations and the best individual systems. It was generally difficult to draw a statistically significant differences between the two groups, and between the combinations themselves.

there is significant overlap in the strategies that they employ.

Improved system combination warrants further investigation. We would suggest collecting additional judgments, and doing oracle experiments where the contributions of individual systems are weighted according to human judgments of their quality.

Understandability

Our hope is that judging the acceptability of edited output as discussed in Section 3 gives some indication of how often a system's output was understandable. Figure 6 gives the percentage of times that each system's edited output was judged to be acceptable (the percentage also factors in instances when judges were unable to improve the output because it was incomprehensible).

The edited output of the best performing systems under this evaluation model were deemed acceptable around 50% of the time for French-English, English-French, English-Spanish, German-English, and English-German. For Spanish-English the edited output of the best system was acceptable around 40% of the time, for English-Czech it was 30% and for Czech-English and Hungarian-English it was around 20%.

This style of manual evaluation is experimental and should not be taken to be authoritative. Some caveats about this measure:

- Editing translations without context is difficult, so the acceptability rate is probably an underestimate of how understandable a system actually is.
- There are several sources of variance that are difficult to control for: some people are better at editing, and some sentences are more difficult to edit. Therefore, variance in the understandability of systems is difficult to pin down.
- The acceptability measure does not strongly correlate with the more established method of ranking translations relative to each other for all the language pairs.[5]

Please also note that the number of corrected translations per system are very low for some language pairs, as low as 23 corrected sentences per system for the language pair English–French.

[5]The Spearman rank correlation coefficients for how the two types of manual evaluation rank systems are .67 for de-en, .67 for fr-en, .06 for es-en, .50 for cz-en, .36 for hu-en, .65 for en-de, .02 for en-fr, -.6 for en-es, and .94 for en-cz.

French–English
625–836 judgments per system

System	C?	≥others
GOOGLE ●	no	.76
DCU ⋆	yes	.66
LIMSI ●	no	.65
JHU ⋆	yes	.62
UEDIN ⋆	yes	.61
UKA	yes	.61
LIUM-SYSTRAN	no	.60
RBMT5	no	.59
CMU-STATXFER ⋆	yes	.58
RBMT1	no	.56
USAAR	no	.55
RBMT3	no	.54
RWTH ⋆	yes	.52
COLUMBIA	yes	.50
RBMT4	no	.47
GENEVA	no	.34

English–French
422–517 judgments per system

System	C?	≥others
LIUM-SYSTRAN ●	no	.73
GOOGLE ●	no	.68
UKA ●⋆	yes	.66
SYSTRAN ●	no	.65
RBMT3 ●	no	.65
DCU ●⋆	yes	.65
LIMSI ●	no	.64
UEDIN ⋆	yes	.60
RBMT4	no	.59
RWTH	yes	.58
RBMT5	no	.57
RBMT1	no	.54
USAAR	no	.48
GENEVA	no	.38

Hungarian–English
865–988 judgments per system

System	C?	≥others
MORPHO ●	no	.75
UMD ⋆	yes	.66
UEDIN	yes	.45

German–English
651–867 judgments per system

System	C?	≥others
RBMT5	no	.66
USAAR ●	no	.65
GOOGLE ●	no	.65
RBMT2 ●	no	.64
RBMT3	no	.64
RBMT4	no	.62
STUTTGART ●⋆	yes	.61
SYSTRAN ●	no	.60
UEDIN ⋆	yes	.59
UKA ⋆	yes	.58
UMD ⋆	yes	.56
RBMT1	no	.54
LIU ⋆	yes	.50
RWTH	yes	.50
GENEVA	no	.33
JHU-TROMBLE	yes	.13

English–German
977–1226 judgments per system

System	C?	≥others
RBMT2 ●	no	.66
RBMT3 ●	no	.64
RBMT5 ●	no	.64
USAAR	no	.58
RBMT4	no	.58
RBMT1	no	.57
GOOGLE	no	.54
UKA ⋆	yes	.54
UEDIN ⋆	yes	.51
LIU ⋆	yes	.49
RWTH ⋆	yes	.48
STUTTGART	yes	.43

Czech–English
1257–1263 judgments per system

System	C?	≥others
GOOGLE ●	no	.75
UEDIN ⋆	yes	.57
CU-BOJAR ⋆	yes	.51

Spanish–English
613–801 judgments per system

System	C?	≥others
GOOGLE ●	no	.70
TALP-UPC ●⋆	yes	.59
UEDIN ⋆	yes	.56
RBMT1 ●	no	.55
RBMT3 ●	no	.55
RBMT5 ●	no	.55
RBMT4 ●	no	.53
RWTH ⋆	yes	.51
USAAR	no	.51
NICT	yes	.37

English–Spanish
632–746 judgments per system

System	C?	≥others
RBMT3 ●	no	.66
UEDIN ●⋆	yes	.66
GOOGLE ●	no	.65
RBMT5 ●	no	.64
RBMT4	no	.61
NUS ⋆	yes	.59
TALP-UPC	yes	.58
RWTH	yes	.51
RBMT1	no	.25
USAAR	no	.48

English–Czech
4626–4784 judgments per system

System	C?	≥others
PCTRANS ●	no	.67
EUROTRANXP ●	no	.67
GOOGLE	no	.66
CU-BOJAR ⋆	yes	.61
UEDIN	yes	.53
CU-TECTOMT	yes	.48

Systems are listed in the order of how often their translations were ranked higher than or equal to any other system. Ties are broken by direct comparison.

C? indicates constrained condition, meaning only using the supplied training data and possibly standard monolingual linguistic tools (but no additional corpora).

● indicates a **win** in the category, meaning that no other system is statistically significantly better at p-level≤0.1 in pairwise comparison.

⋆ indicates a **constrained win**, no other constrained system is statistically better.

For all pairwise comparisons between systems, please check the appendix.

Table 6: Official results for the WMT09 translation task, based on the human evaluation (ranking translations relative to each other)

Given these low numbers, the numbers presented in Figure 6 should not be read as comparisons between systems, but rather viewed as indicating the state of machine translation for different language pairs.

5 Shared evaluation task overview

In addition to allowing us to analyze the translation quality of different systems, the data gathered during the manual evaluation is useful for validating the automatic evaluation metrics. Last year, NIST began running a similar "Metrics for MAchine TRanslation" challenge (MetricsMATR), and presented their findings at a workshop at AMTA (Przybocki et al., 2008).

In this year's shared task we evaluated a number of different automatic metrics:

- Bleu (Papineni et al., 2002)—Bleu remains 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.

- Meteor (Agarwal and Lavie, 2008)—Meteor measures precision and recall for unigrams and applies a fragmentation penalty. It uses flexible word matching based on stemming and WordNet-synonymy. meteor-ranking is optimized for correlation with ranking judgments.

- 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. Two variants of TER are also included: TERp (Snover et al., 2009), a new version which introduces a number of different features, and $(\text{Bleu} - \text{TER})/2$, a combination of Bleu and Translation Edit Rate.

- MaxSim (Chan and Ng, 2008)—MaxSim calculates a similarity score by comparing items in the translation against the reference. Unlike most metrics which do strict matching, MaxSim computes a similarity score for non-identical items. To find a maximum weight matching that matches each system item to at most one reference item, the items are then modeled as nodes in a bipartite graph.

- wcd6p4er (Leusch and Ney, 2008)—a measure based on cder with word-based substitution costs. Leusch and Ney (2008) also submitted two contrastive metrics: bleusp4114, a modified version of BLEU-S (Lin and Och, 2004), with tuned n-gram weights, and bleusp, with constant weights. wcd6p4er is an error measure and bleusp is a quality score.

- RTE (Pado et al., 2009)—The RTE metric follows a semantic approach which applies recent work in *rich textual entailment* to the problem of MT evaluation. Its predictions are based on a regression model over a feature set adapted from an entailment systems. The features primarily model alignment quality and (mis-)matches of syntactic and semantic structures.

- ULC (Giménez and Màrquez, 2008)—ULC is an arithmetic mean over other automatic metrics. The set of metrics used include Rouge, Meteor, measures of overlap between constituent parses, dependency parses, semantic roles, and discourse representations. The ULC metric had the strongest correlation with human judgments in WMT08 (Callison-Burch et al., 2008).

- wpF and wpBleu (Popovic and Ney, 2009) - These metrics are based on *w*ords and *p*art of speech sequences. wpF is an n-gram based F-measure which takes into account both word n-grams and part of speech n-grams. wpBLEU is a combnination of the normal Blue score and a part of speech-based Bleu score.

- SemPOS (Kos and Bojar, 2009) – the SemPOS metric computes overlapping words, as defined in (Giménez and Màrquez, 2007), with respect to their semantic part of speech. Moreover, it does not use the surface representation of words but their underlying forms obtained from the TectoMT framework.

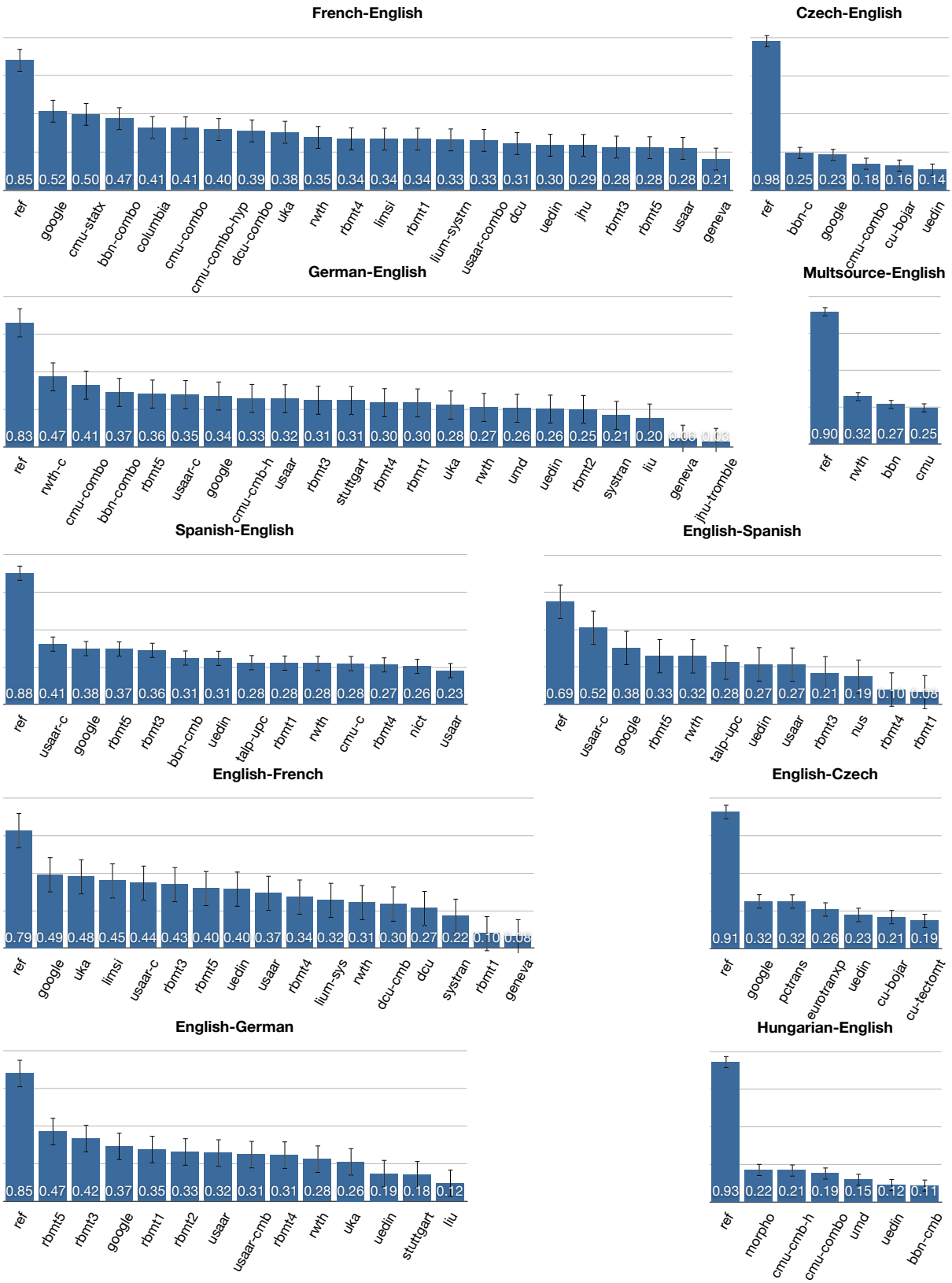

Figure 6: The percent of time that each system's edited output was judged to be an acceptable translation. These numbers also include judgments of the system's output when it was marked either *incomprehensible* or *acceptable* and left unedited. Note that the reference translation was edited alongside the system outputs. Error bars show one positive and one negative standard deviation for the systems in that language pair.

5.1 Measuring system-level correlation

We measured the correlation of the automatic metrics with the human judgments of translation quality at the system-level using Spearman's rank correlation coefficient ρ. We converted the raw scores assigned to each system into ranks. We assigned a human ranking to the systems based on the percent of time that their translations were judged to be better than or equal to the translations of any other system in the manual evaluation.

When there are no ties ρ can be calculated 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 absolute value for ρ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower absolute ρ.

	de-en (21 systems)	fr-en (21 systems)	es-en (13 systems)	cz-en (5 systems)	hu-en (6 systems)	Average
ulc	**.78**	.92	.86	**1**	.6	**.83**
maxsim	.76	.91	**.98**	.7	.66	.8
rte (absolute)	.64	.91	.96	.6	**.83**	.79
meteor-rank	.64	**.93**	.96	.7	.54	.75
rte (pairwise)	.76	.59	.78	.8	**.83**	.75
terp	-.72	-.89	-.94	-.7	-.37	-.72
meteor-0.6	.56	**.93**	.87	.7	.54	.72
meteor-0.7	.55	**.93**	.86	.7	.26	.66
bleu-ter/2	.38	.88	.78	.9	-.03	.58
nist	.41	.87	.75	.9	-.14	.56
wpF	.42	.87	.82	**1**	-.31	.56
ter	-.43	-.83	-.84	-.6	-.01	-.54
nist (cased)	.42	.83	.75	**1**	-.31	.54
bleu	.41	.88	.79	.6	-.14	.51
bleusp	.39	.88	.78	.6	-.09	.51
bleusp4114	.39	.89	.78	.6	-.26	.48
bleu (cased)	.4	.86	.8	.6	-.31	.47
wpbleu	.43	.86	.8	.7	-.49	.46
wcd6p4er	-.41	-.89	-.76	-.6	.43	-.45

Table 7: The system-level correlation of the automatic evaluation metrics with the human judgments for translation into English.

5.2 Measuring sentence-level consistency

Because the sentence-level judgments collected in the manual evaluation are relative judgments rather than absolute judgments, it is not possible for us to measure correlation at the sentence-level in the same way that previous work has done (Kulesza and Shieber, 2004; Albrecht and Hwa, 2007a; Albrecht and Hwa, 2007b).

Rather than calculating a correlation coefficient at the sentence-level we instead ascertained how consistent the automatic metrics were with the human judgments. The way that we calculated consistency was the following: for every pairwise comparison of two systems on a single sentence by a person, we counted the automatic metric as being consistent if the relative scores were the same (i.e. the metric assigned a higher score to the higher ranked system). We divided this by the total number of pairwise comparisons to get a percentage. Because the systems generally assign real numbers as scores, we excluded pairs that the human annotators ranked as ties.

	en-de (13 systems)	en-fr (16 systems)	en-es (11 systems)	en-cz (5 systems)	Average
terp	.03	-.89	-.58	**-.4**	**-.46**
ter	-.03	-.78	-.5	-.1	-.35
bleusp4114	-.3	.88	.51	.1	.3
bleusp	-.3	.87	.51	.1	.29
bleu	-.43	.87	.36	.3	.27
bleu (cased)	-.45	.87	.35	.3	.27
bleu-ter/2	-.37	.87	.44	.1	.26
wcd6p4er	.54	-.89	-.45	-.1	-.22
nist (cased)	-.47	.84	.35	.1	.2
nist	-.52	.87	.23	.1	.17
wpF	-.06	.9	.58	*n/a*	*n/a*
wpbleu	**.07**	**.92**	**.63**	*n/a*	*n/a*

Table 8: The system-level correlation of the automatic evaluation metrics with the human judgments for translation out of English.

SemPOS	**.4**	$BLEU_{tecto}$	.3
Meteor	**.4**	BLEU	.3
$GTM(e=0.5)_{tecto}$	**.4**	$NIST_{lemma}$	.1
$GTM(e=0.5)_{lemma}$	**.4**	NIST	.1
GTM(e=0.5)	**.4**	$BLEU_{lemma}$	.1
WER_{tecto}	.3	WER_{lemma}	-.1
TER_{tecto}	.3	WER	-.1
PER_{tecto}	.3	TER_{lemma}	-.1
$F\text{-measure}_{tecto}$	.3	TER	-.1
$F\text{-measure}_{lemma}$	.3	PER_{lemma}	-.1
F-measure	.3	PER	-.1
		$NIST_{tecto}$	-.3

Table 9: The system-level correlation for automatic metrics ranking five English-Czech systems

6 Evaluation task results

6.1 System-level correlation

Table 7 shows the correlation of automatic metrics when they rank systems that are translating into English. Note that TERp, TER and wcd6p4er are error metrics, so a negative correlation is better for them. The strength of correlation varied for the different language pairs. The automatic metrics were able to rank the French-English systems reasonably well with correlation coefficients in the range of .8 and .9. In comparison, metrics performed worse for Hungarian-English, where half of the systems had negative correlation. The ULC metric once again had strongest correlation with human judgments of translation quality. This was followed closely by MaxSim and RTE, with Meteor and TERp doing respectably well in 4th and 5th place. Notably, Bleu and its variants were the worst performing metrics in this translation direction.

Table 8 shows correlation for metrics which operated on languages other than English. Most of the best performing metrics that operate on English do not work for foreign languages, because they perform some linguistic analysis or rely on a resource like WordNet. For translation into foreign languages TERp was the best system overall. The wpBleu and wpF metrics also did extremely well, performing the best in the language pairs that they were applied to. wpBleu and wpF were not applied to Czech because the authors of the metric did not have a Czech tagger. English-German proved to be the most problematic language pair to automatically evaluate, with all of the metrics having a negative correlation except wpBleu and TER.

Table 9 gives detailed results for how well variations on a number of automatic metrics do for the task of ranking five English-Czech systems.[6] These systems were submitted by Kos and Bojar (2009), and they investigate the effects of using Prague Dependency Treebank annotations during automatic evaluation. They linearizing the Czech trees and evaluated either the lemmatized forms of the Czech (*lemma*) read off the trees or the Tectogrammatical form which retained only lemmatized content words (*tecto*). The table also demonstrates SemPOS, Meteor, and GTM perform better on Czech than many other metrics.

6.2 Sentence-level consistency

Tables 10 and 11 show the percent of times that the metrics' scores were consistent with human rankings of every pair of translated sentences.[7] Since we eliminated sentence pairs that were judged to be equal, the random baseline for this task is 50%. Many metrics failed to reach the baseline (including most metrics in the out-of-English direction). This indicates that sentence-level evaluation of machine translation quality is very difficult. RTE and ULC again do the best overall for the into-English direction. They are followed closely by wpF and wcd6p4er, which considerably improve their performance over their system-level correlations.

We tried a variant on measuring sentence-level consistency. Instead of using the scores assigned to each individual sentence, we used the system-level score and applied it to every sentence that was produced by that system. These can be thought of as a metric's prior expectation about how a system should preform, based on their performance on the whole data set. Tables 12 and 13 show that using the system-level scores in place of the sentence-level scores results in considerably higher consistency with human judgments. This suggests an interesting line of research for improving sentence-level predictions by using the performance on a larger data set as a prior.

7 Summary

As in previous editions of this workshop we carried out an extensive manual and automatic evaluation of machine translation performance for translating from European languages into English,

[6] PCTRANS was excluded from the English-Czech systems because its SGML file was malformed.

[7] Not all metrics entered into the sentence-level task.

	fr-en (6268 pairs)	de-en (6382 pairs)	es-en (4106 pairs)	cz-en (2251 pairs)	hu-en (2193 pairs)	xx-en (1952 pairs)	Overall (23152 pairs)
ulc	**.55**	**.56**	.51	.50	.51	.51	**.54**
rte (absolute)	.54	.56	.51	.50	.55	.51	.53
wpF	.54	.55	.50	*.47*	*.48*	.51	.52
wcd6p4er	.54	.54	*.49*	*.48*	*.48*	.50	.52
maxsim	.53	.55	*.49*	*.47*	.50	*.49*	.52
bleusp	.54	.55	*.49*	*.47*	*.46*	.50	.51
bleusp4114	.53	.55	*.48*	*.47*	*.46*	.50	.51
rte (pairwise)	*.49*	*.48*	**.52**	**.53**	.55	**.52**	.51
terp	.52	.53	*.48*	*.46*	*.45*	*.48*	.50
meteor-0.6	.50	.53	*.46*	*.48*	*.47*	*.47*	*.49*
meteor-rank	.50	.52	*.46*	*.48*	*.47*	*.47*	*.49*
meteor-0.7	.49	.52	*.46*	*.48*	*.47*	*.47*	*.49*
ter	*.48*	*.47*	*.43*	*.41*	*.40*	*.42*	*.45*
wpbleu	*.46*	*.45*	*.46*	*.39*	*.35*	*.45*	*.44*

Table 10: Sentence-level consistency of the automatic metrics with human judgments for translations into English. Italicized numbers fall below the random-choice baseline.

	fr-en (6268 pairs)	de-en (6382 pairs)	es-en (4106 pairs)	cz-en (2251 pairs)	hu-en (2193 pairs)	Overall (21200 pairs)
Oracle	.61	.63	.59	.61	.67	.62
rte (absolute)	.60	.61	**.59**	.57	**.65**	**.61**
ulc	**.61**	**.62**	.58	**.61**	.59	.60
maxsim	**.61**	**.62**	**.59**	.57	.61	.60
meteor-rank	**.61**	.61	**.59**	.57	.61	.60
meteor-0.6	**.61**	.61	.58	.57	.60	.60
rte (pairwise)	.56	.61	.57	.59	.64	.59
terp	.60	.61	**.59**	.57	.56	.59
meteor-0.7	**.61**	.61	.58	.57	.55	.59
ter	.60	.59	.57	.55	.51	.58
wpF	.60	.59	.57	**.61**	*.46*	.58
bleusp	**.61**	.59	.56	.55	*.48*	.57
bleusp4114	**.61**	.59	.56	.55	.46	.57
wcd6p4er	**.61**	.59	.57	.55	*.44*	.57
wpbleu	.60	.59	.57	.57	*.43*	.57

Table 12: Consistency of the automatic metrics when their system-level ranks are treated as sentence-level scores. Oracle shows the consistency of using the system-level human ranks that are given in Table 6.

	en-fr (2967 pairs)	en-de (6563 pairs)	en-es (3249 pairs)	en-cz (11242 pairs)	Overall (24021 pairs)
wcd6p4er	**.57**	*.47*	.52	*.49*	**.50**
bleusp4114	.57	*.46*	.54	*.49*	.50
bleusp	.57	*.46*	.53	*.48*	*.49*
ter	.50	*.41*	*.45*	*.37*	*.41*
terp	.51	*.39*	*.48*	*.27*	*.36*
wpF	.57	*.46*	**.54**	n/a	**.51**
wpbleu	.53	*.37*	*.46*	n/a	*.43*

Table 11: Sentence-level consistency of the automatic metrics with human judgments for translations out of English. Italicized numbers fall below the random-choice baseline.

	en-fr (2967 pairs)	en-de (6563 pairs)	en-es (3249 pairs)	en-cz (11242 pairs)	Overall (24021 pairs)
Oracle	.62	.59	.63	.60	.60
terp	.62	.50	.59	**.53**	**.54**
ter	.61	**.51**	.58	.50	.53
bleusp	.62	*.48*	.59	.50	.52
bleusp4114	.63	*.48*	.59	.50	.52
wcd6p4er	.62	.46	.58	.50	.52
wpbleu	**.63**	**.51**	**.60**	n/a	**.56**
wpF	**.63**	.50	.59	n/a	.55

Table 13: Consistency of the automatic metrics when their system-level ranks are treated as sentence-level scores. Oracle shows the consistency of using the system-level human ranks that are given in Table 6.

and vice versa.

The number of participants remained stable compared to last year's WMT workshop, with 22 groups from 20 institutions participating in WMT09. This year's evaluation also included 7 commercial rule-based MT systems and Google's online statistical machine translation system.

Compared to previous years, we have simplified the evaluation conditions by removing the in-domain vs. out-of-domain distinction focusing on news translations only. The main reason for this was eliminating the advantage statistical systems have with respect to test data that are from the same domain as the training data.

Analogously to previous years, the main focus of comparing the quality of different approaches is on manual evaluation. Here, also, we reduced the number of dimensions with respect to which the different systems are compared, with sentence-level ranking as the primary type of manual evaluation. In addition to the direct quality judgments we also evaluated translation quality by having people edit the output of systems and have assessors judge the correctness of the edited output. The degree to which users were able to edit the translations (without having access to the source sentence or reference translation) served as a measure of the overall comprehensibility of the translation.

Although the inter-annotator agreement in the sentence-ranking evaluation is only fair (as measured by the Kappa score), agreement can be improved by removing the first (up to 50) judgments of each assessor, focusing on the judgments that were made once the assessors are more familiar with the task. Inter-annotator agreement with respect to correctness judgments of the edited translations were higher (moderate), which is probably due to the simplified evaluation criterion (binary judgments versus rankings). Inter-annotator agreement for both conditions can be increased further by removing the judges with the worst agreement. Intra-annotator agreement on the other hand was considerably higher ranging between moderate and substantial.

In addition to the manual evaluation criteria we applied a large number of automated metrics to see how they correlate with the human judgments. There is considerably variation between the different metrics and the language pairs under consideration. As in WMT08, the ULC metric had the highest overall correlation with human judgments when translating into English, with MaxSim and RTE following closely behind. TERp and wpBleu were best when translating into other languages.

Automatically predicting human judgments at the sentence-level proved to be quite challenging with many of the systems performing around chance. We performed an analysis that showed that if metrics' system-level scores are used in place of their scores for individual sentences, that they do quite a lot better. This suggests that prior probabilities ought to be integrated into sentence-level scoring.

All data sets generated by this workshop, including the human judgments, system translations and automatic scores, are publicly available for other researchers to analyze.[8]

Acknowledgments

This work was supported in parts by the EuroMatrix project funded by the European Commission (6th Framework Programme), the GALE program of the US Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022, and the US National Science Foundation under grant IIS-0713448.

We are grateful to Holger Schwenk and Preslav Nakov for pointing out the potential bias in our method for ranking systems when self-judgments are excluded. We analyzed the results and found that this did not hold. We would like to thank Maja Popovic for sharing thoughts about how to improve the manual evaluation. Thanks to Cam Fordyce for helping out with the manual evaluation again this year.

An extremely big thanks to Sebastian Pado for helping us work through the logic of segment-level scoring of automatic evaluation metric.

References

Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio, June. Association for Computational Linguistics.

Joshua Albrecht and Rebecca Hwa. 2007a. A re-examination of machine learning approaches for

[8] http://www.statmt.org/wmt09/results.html

sentence-level MT evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.

Joshua Albrecht and Rebecca Hwa. 2007b. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.

Alexandre Allauzen, Josep Crego, Aurélien Max, and Fran cois Yvon. 2009. LIMSI's statistical translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

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

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT07)*, Prague, Czech Republic.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT08)*, Colmbus, Ohio.

Marine Carpuat. 2009. Toward using morphology in French-English phrase-based SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2008. An automatic metric for machine translation evaluation based on maximum similary. In *In the Metrics-MATR Workshop of AMTA-2008*, Honolulu, Hawaii.

Yu Chen, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, and Hans Uszkoreit. 2009. Combining multi-engine translations with moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Jinhua Du, Yifan He, Sergio Penkale, and Andy Way. 2009. MATREX: The DCU MT system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2009. Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Chris Dyer, Hendra Setiawan, Yuval Marton, and Philip Resnik. 2009. The University of Maryland statistical machine translation system for the fourth workshop on machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Jason Eisner and Roy W. Tromble. 2006. Local search with very large-scale neighborhoods for optimal permutations in machine translation. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2006)*, New York, New York.

Christian Federmann, Silke Theison, Andreas Eisele, Hans Uszkoreit, Yu Chen, Michael Jellinghaus, and Sabine Hunsicker. 2009. Translation combination using factored word substitution. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Alexander Fraser. 2009. Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

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

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198.

Greg Hanneman, Vamshi Ambati, Jonathan H. Clark, Alok Parlikar, and Alon Lavie. 2009. An improved statistical transfer system for French-English machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Kenneth Heafield, Greg Hanneman, and Alon Lavie. 2009. Machine translation system combination with flexible word ordering. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Almut Silja Hildebrand and Stephan Vogel. 2009. CMU system combination for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Maria Holmqvist, Sara Stymne, Jody Foo, and Lars Ahrenberg. 2009. Improving alignment for SMT by reordering and augmenting the training corpus. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Philipp Koehn and Barry Haddow. 2009. Edinburgh's submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

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*, New York, New York.

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. 2007. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University.

Kamil Kos and Ondřej Bojar. 2009. Evaluation of Machine Translation Metrics for Czech as the Target Language. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 4–6.

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

Gregor Leusch and Hermann Ney. 2008. BLEUSP, PINVWER, CDER: Three improved MT evaluation measures. In *In the Metrics-MATR Workshop of AMTA-2008*, Honolulu, Hawaii.

Gregor Leusch, Evgeny Matusov, and Hermann Ney. 2009. The RWTH system combination system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, Barcelona, Spain.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2002)*, Tiburon, California.

Preslav Nakov and Hwee Tou Ng. 2009. NUS at WMT09: Domain adaptation experiments for English-Spanish machine translation of news commentary text. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Jan Niehues, Teresa Herrmann, Muntsin Kolss, and Alex Waibel. 2009. The Universität Karlsruhe translation system for the EACL-WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

NIST. 2008. Evaluation plan for gale go/no-go phase 3 / phase 3.5 translation evaluations. June 18, 2008.

Attila Novák. 2009. Morphologic's submission for the WMT 2009 shared task. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Sebastian Pado, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Machine translation evaluation with textual entailment features. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

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

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2009. NICT@WMT09: Model adaptation and transliteration for Spanish-English SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Maja Popovic and Hermann Ney. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Maja Popovic, David Vilar, Daniel Stein, Evgeny Matusov, and Hermann Ney. 2009. The RWTH machine translation system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Mark Przybocki, Kay Peterson, and Sebastien Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results/.

José A. R. Fonollosa, Maxim Khalilov, Marta R. Costa-jussá, José B. Mariño, Carlos A. Henráquez Q., Adolfo Hernández H., and Rafael E. Banchs. 2009. The TALP-UPC phrase-based translation system for EACL-WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2009. Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Holger Schwenk, Sadaf Abdul Rauf, Loic Barrault, and Jean Senellart. 2009. SMT and SPE machine translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

David Talbot and Miles Osborne. 2007. Smoothed Bloom filter language models: Tera-scale lms on the cheap. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Eric Wehrli, Luka Nerima, and Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March. Association for Computational Linguistics.

A Pairwise system comparisons by human judges

Tables 14–24 show pairwise comparisons between systems for each language pair. The numbers in each of the tables' cells indicate the percentage of times that the system in that column was judged to be better than the system in that row. Bolding indicates the winner of the two systems. The difference between 100 and the sum of the complimentary cells is the percent of time that the two systems were judged to be equal.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied the Sign Test to measure which comparisons indicate genuine differences (rather than differences that are attributable to chance). In the following tables $\star$ indicates statistical significance at $p \leq 0.10$, † indicates statistical significance at $p \leq 0.05$, and ‡ indicates statistical significance at $p \leq 0.01$, according to the Sign Test.

B Automatic scores

Tables 26 and 25 give the automatic scores for each of the systems.

	GENEVA	GOOGLE	JHU-TROMBLE	LIU	RBMT1	RBMT2	RBMT3	RBMT4	RBMT5	RWTH	STUTTGART	SYSTRAN	UEDIN	UKA	UMD	USAAR	BBN-COMBO	CMU-COMBO	CMU-COMBO-HYPOSEL	RWTH-COMBO	USAAR-COMBO
GENEVA		**.76**‡	.08‡	**.63**†	**.54**	**.69**†	**.73**‡	**.83**‡	**.78**‡	**.49***	**.77**‡	**.75**‡	**.74**‡	**.57**†	**.74**‡	**.69**‡	**.75**‡	**.84**‡	**.60**	**.84**‡	**.71**‡
GOOGLE	.15‡		.03‡	.23†	**.50**	.43	.24†	.39	**.42**	.39	.43	.33	.27*	.29*	.38	**.48**	**.57***	**.44**	.32	.35	.36
JHU-TROMBLE	**.75**‡	**.90**‡		**.77**‡	**.81**‡	**.84**‡	**.91**‡	**.94**‡	**.88**‡	**.79**‡	**.83**‡	**.83**‡	**.93**‡	**.89**‡	**.92**‡	**.90**‡	**.94**‡	**.90**‡	**.95**‡	**.91**‡	**.83**‡
LIU	.29†	**.65**†	.12‡		**.49**	**.63**	**.63***	**.57**	**.63***	.41	**.49**	**.46**	**.50**	**.49**	**.50**	.41	**.66**†	**.53**	**.59**‡	**.62**†	**.53**
RBMT1	.32	.43	.11‡	.46		.42	**.46**	**.50**	**.61**†	.34	**.46**	**.58**	**.51**	.42	.42	**.56**	**.47**	**.53**	**.49**	**.58**	**.54**
RBMT2	.25†	**.46**	.09‡	.37	**.45**		.33	.45	.23†	.3	.28	.47	.42	.31*	.34	.39	**.49**	**.61**	.4	.32	.29*
RBMT3	.17‡	**.59**†	.02‡	.26*	.35	**.46**		.27	**.45**	.27	.36	.46	**.42**	.43	.26*	**.49**	.4	.48	**.58**	.29	.31
RBMT4	.12‡	**.47**	.07‡	.37	.4	.45	**.52**		**.60***	.39	**.39**	**.45**	.39	.31*	.29†	.44	**.54**	**.45**	.37	.43	.30
RBMT5	.13‡	.34	.07‡	.30*	.24†	**.57**†	.41	.29*		.31	**.50**	.34	.3	.28†	.43	.30	**.49**	**.57**	.3	**.49**	.21
RWTH	.21*	**.55**	.10‡	.41	**.49**	**.55**	**.46**	**.46**	**.60**		**.44**	**.57**	**.48**	**.51***	**.41**	**.56**	**.64**‡	**.54**	**.56***	**.74**‡	**.59***
STUTTGART	.17‡	.43	.13‡	.39	.43	**.55**	**.39**	.36	.33	.34		.38	**.42**	**.52**	**.42**	**.49**	**.49**	.28	.35	**.56**	**.46**
SYSTRAN	.11‡	**.63**	.06‡	.42	.37	.47	**.50**	.32	**.58**	.34	**.55**		.36	**.44**	.35	.43	**.61**†	**.46**	.41	.33	.44
UEDIN	.10‡	**.50***	.03‡	.35	.49	**.46**	.39	**.52**	**.55**	.29	.39	**.52**		.35	.33	**.42**	**.58***	**.43**	**.56**	**.59**†	**.55**
UKA	.29†	**.58***	.04‡	.32	**.47**	**.63***	**.55**	**.54***	**.64**†	.24*	.28	.39	**.50**		.29	**.50**	**.48**	.36	**.57***	**.45**	**.45**
UMD	.16‡	**.53**	.08‡	.38	**.49**	**.43**	**.63***	**.68**†	**.49**	.38	.39	**.41**	**.50**	**.49**		**.46**	**.54**	**.44**	.38	**.46**	**.50**
USAAR	.19‡	.44	‡	.41	.34	**.49**	.4	.44	**.33**	.36	.33	**.45**	.39	.32	.41		**.46**	**.41**	.31	.42	.11
BBN-COMBO	.14‡	.31*	.06‡	.26†	.44	.44	**.48**	.36	.38	.23‡	.35	.26†	.29*	.34	.36	.37		.32	.23†	**.38**	.32
CMU-COMBO	.10‡	.36	.07‡	.37	.37	.36	.48	.40	.30	.28	**.53**	.41	.4	**.43**	.28	.34	**.50**		.33	**.53**	.44
CMU-COMBO-H	.3	**.46**	‡	.10‡	.39	**.43**	.40	**.48**	**.57**	.27*	**.41**	**.47**	.28	.26*	.38	**.49**	**.65**†	**.46**		.41	**.47**
RWTH-COMBO	.06‡	**.38**	‡	.19‡	.36	**.54**	**.43**	.43	.30	.10‡	.33	**.56**	.22†	.27	.23	.42	.32	.31	.41		.29
USAAR-COMBO	.20‡	**.55**	.17‡	.3	.39	**.57***	**.45**	**.59**	**.32**	.27*	.33	**.47**	.32	.33	.27	**.16**	**.55**	.44	.4	**.50**	
> OTHERS	.22	.51	.06	.38	.44	.52	.49	.49	.50	.33	.44	.48	.44	.42	.41	.47	**.56**	.48	.46	.51	.43
>= OTHERS	.33	.65	.13	.50	.54	.64	.64	.62	**.66**	.50	.61	.60	.59	.58	.56	.65	.68	.63	.62	**.70**	.62

Table 14: Sentence-level ranking for the WMT09 German-English News Task

	GOOGLE	LIU	RBMT1	RBMT2	RBMT3	RBMT4	RBMT5	RWTH	STUTTGART	UEDIN	UKA	USAAR	USAAR-COMBO
GOOGLE		.34†	**.56**	**.51**	**.55**†	.44	**.56**†	.37	.41	.42	.45	.45	**.43**
LIU	**.58**†		**.62**‡	**.55**†	**.55***	**.61**‡	**.59**†	.37	.38	**.47**	**.43**	**.58**†	.44
RBMT1	.39	.33‡		**.56**†	**.44**	**.50***	**.57**†	.41	.32‡	.37*	.35†	.45	.42
RBMT2	.35	.34†	.34†		**.43**	.37*	.40	.25‡	.25‡	.31‡	.36†	.37*	.32†
RBMT3	.31†	.35*	.41	.35		.37*	.41	.24‡	.25‡	.33‡	.43	**.49**	.36*
RBMT4	**.48**	.33‡	.33*	**.56***	**.55***		**.47**	.37	.35†	.34‡	**.45**	**.44**	.38
RBMT5	.36†	.35†	.33†	**.50**	**.53**	.33		.36†	.32‡	.35†	.31‡	.25‡	.32‡
RWTH	**.51**	**.46**	**.50**	**.60**‡	**.65**‡	**.51**	**.60**†		.38	**.47**	**.48**	**.52**	**.54**
STUTTGART	**.50**	**.47**	**.62**‡	**.65**‡	**.64**‡	**.57**†	**.62**‡	**.46**		**.52**†	**.54**†	**.66**‡	**.53**
UEDIN	**.50**	.37	**.53***	**.64**‡	**.62**‡	**.60**‡	**.55**†	.45	.28†		**.41**	**.53**	.35
UKA	**.47**	.42	**.57**†	**.58**†	**.46**	.44	**.62**‡	.35	.32†	.36		**.46**	.41
USAAR	**.46**	.36†	**.46**	**.55***	.42	.42	**.48**‡	.42	.28‡	.39	.44		.41
USAAR-COMBO	.37	**.45**	**.54**	**.55**†	**.55***	**.53**	**.61**‡	.39	.40	**.39**	**.46**	**.52**	
> OTHERS	.44	.38	.48	**.55**	.53	.47	.54	.37	.33	.39	.42	.48	.41
>= OTHERS	.54	.49	.57	**.66**	.64	.58	.64	.48	.43	.51	.54	.58	.52

Table 15: Sentence-level ranking for the WMT09 English-German News Task

	GOOGLE	NICT	RBMT1	RBMT3	RBMT4	RBMT5	RWTH	TALP-UPC	UEDIN	USAAR	BBN-COMBO	CMU-COMBO	USAAR-COMBO
GOOGLE		.21‡	.40	.40	.41	.38	.23‡	.35	.31†	.25‡	**.36**	.14	.21
NICT	**.74**‡		**.52**	**.53**	**.63**‡	**.64**‡	**.55**†	**.61**‡	**.65**‡	**.59**†	**.62**‡	**.78**‡	**.66**‡
RBMT1	**.56**	.40		.34	**.44**	**.46**	.35	**.48**	.42	.42	**.57**†	**.52**	**.54**
RBMT3	.40	.39	**.40**		.34	.36	.42	.4	**.55**	**.50**	**.57***	**.48**	**.62**†
RBMT4	**.55**	.32‡	.41	**.46**		**.47**	.39	**.49**	**.49**	**.48**	**.54**	**.57***	**.54**
RBMT5	**.54**	.30‡	.35	**.44**	.38		.45	**.50**	**.49**	.23	**.51**	**.51**	**.66**‡
RWTH	**.64**‡	.29†	**.50**	**.53**	**.53**	**.49**		**.42**	**.46**	.43	**.44**	**.51**	**.58**‡
TALP-UPC	**.48**	.24‡	.44	**.47**	.41	.36	.39		.36	.32*	**.47**	**.45**	**.50**
UEDIN	**.61**†	.16‡	**.48**	.42	.41	.46	.44	**.43**		.44	**.49**	**.51**	**.41**
USAAR	**.69**‡	.28†	**.47**	.44	.38	**.35**	.43	**.60***	**.48**		**.64**†	**.58**‡	**.56***
BBN-COMBO	.35	.20‡	.32†	.36*	.39	.37	.36	.39	.32	.31†		**.50**	**.40**
CMU-COMBO	**.19**	.15‡	.33	.39	.32*	.37	.36	.31	.37	.21‡	.35		**.31**
USAAR-COMBO	**.23**	.20‡	.42	.31†	.39	.25‡	.27‡	.35	.35	.32*	.36	.29	
> OTHERS	.50	.26	.42	.42	.42	.42	.39	.44	.43	.37	.49	.49	**.50**
>= OTHERS	**.70**	.37	.55	.55	.53	.55	.51	.59	.56	.51	.64	**.70**	.69

Table 16: Sentence-level ranking for the WMT09 Spanish-English News Task

	GOOGLE	NUS	RBMT1	RBMT3	RBMT4	RBMT5	RWTH	TALP-UPC	UEDIN	USAAR	USAAR-COMBO
GOOGLE		.39	.21‡	**.49**	.36	**.48**	.34*	.39	.33	.36*	.21
NUS	**.50**		.11‡	**.62**†	**.51**	**.51**	.35	.25	**.47**	.36	.43
RBMT1	**.76**‡	**.80**‡		**.79**‡	**.79**‡	**.83**‡	**.64**‡	**.76**‡	**.80**‡	**.67**‡	**.64**‡
RBMT3	.42	.31†	.16‡		.30*	.43	.34	.29‡	**.56**	.24‡	.32
RBMT4	**.47**	.32	.11‡	**.52***		**.49**	.38	.36	**.51**	.39	.38
RBMT5	.42	.40	.11‡	**.49**	.35		.31†	.39	**.47**	.18†	.47
RWTH	**.59***	**.52**	.26‡	**.54**	**.51**	**.61**†		**.46**	**.56**†	.39	**.55**†
TALP-UPC	**.49**	**.41**	.17‡	**.63**‡	**.52**	**.51**	.29		**.45***	.39	.41
UEDIN	**.50**	.32	.17‡	.36	.37	.46	.30†	.29*		.32†	.36
USAAR	**.58***	**.56**	.23‡	**.67**‡	**.53**	**.47**†	**.51**	**.49**	**.61**†		**.58***
USAAR-COMBO	**.31**	**.45**	.21‡	**.54**	**.49**	**.50**	.30†	**.43**	**.43**	.33*	
> OTHERS	.50	.45	.17	**.56**	.47	.53	.38	.42	.52	.37	.43
>= OTHERS	.65	.59	.25	**.66**	.61	.64	.51	.58	.66	.48	.61

Table 17: Sentence-level ranking for the WMT09 English-Spanish News Task

	CMU-STATXFER	COLUMBIA	DCU	GENEVA	GOOGLE	JHU	LIMSI	LIUM-SYSTRAN	RBMT1	RBMT3	RBMT4	RBMT5	RWTH	UEDIN	UKA	USAAR	BBN-COMBO	CMU-COMBO	CMU-COMBO-HYPOSEL	DCU-COMBO	USAAR-COMBO
CMU-STATXFER		.37	**.44**	.17‡	**.63**†	**.47**	**.46**	**.58**†	.34	.32	.25†	.42	**.48**	**.46**	.28	.38	**.58**‡	**.47**	**.39**	.41	.35
COLUMBIA	**.56**		**.56**⋆	.37	**.71**‡	**.48**	**.56**‡	**.35**	**.45**	.28⋆	**.38**	.42	**.41**	**.33**	**.58**	**.50**	**.64**†	**.52**	**.64**†	**.71**‡	**.58**†
DCU	.27	.29⋆		.15‡	**.67**‡	**.45**	.33	.34	.29	.31	.29	.27⋆	.24	**.37**	.21†	.39	**.61**‡	.4	.36	.37	.1
GENEVA	**.76**‡	**.54**	**.73**‡		**.71**‡	**.65**‡	**.73**‡	**.62**⋆	**.66**‡	**.76**‡	**.46**	**.79**‡	**.57**	**.74**‡	**.72**‡	**.67**†	**.69**‡	**.52**	**.71**‡	**.67**‡	**.64**†
GOOGLE	.23†	.17‡	.12‡	.13‡		.21‡	.35	.09‡	.20‡	.27†	.31†	**.44**	.16‡	.21‡	.33	.27⋆	**.28**	.30	.34	.37	.16‡
JHU	.40	.26	.38	.22‡	**.60**‡		.31	**.44**	.27	.37	.29†	.41	.33	.37	**.48**	.48	**.53**	**.47**	.31	**.47**	.29
LIMSI	.4	.16‡	**.38**	.19‡	**.56**	**.49**		.29	.37	.27	.20‡	.38	.23⋆	.33	.29	.38	**.61**†	.47	.31	.36	.26⋆
LIUM-SYSTRAN	.23†	.30	**.42**	.33⋆	**.61**‡	.27	**.45**		**.48**	.31	**.41**	**.44**	.32	.35	**.41**	**.39**	**.54**†	**.61**†	.24	**.67**†	.36
RBMT1	**.53**	.23	**.42**	.19‡	**.57**‡	**.46**	**.51**	.45		**.47**	.33	**.46**	.33	.41	.30	**.61**	**.77**‡	**.51**	**.41**	**.50**	.41
RBMT3	**.57**	**.63**⋆	**.55**	.15‡	**.69**†	**.44**	**.57**	**.52**	.41		.22‡	**.38**	**.51**	**.43**	.43	.31	**.57**⋆	**.46**	**.47**	.38	**.55**
RBMT4	**.58**†	.35	**.51**	.36	**.67**†	**.60**†	**.63**‡	.35	**.41**	**.59**‡		**.40**	**.55**	**.50**	**.71**‡	**.52**†	**.63**†	**.65**†	**.65**†	**.66**†	.38
RBMT5	.42	**.49**	**.54**⋆	.09‡	.38	**.49**	**.49**	.37	.27	.29	.34		.38	.39	**.51**	.18	.42	**.58**	**.48**	**.50**	**.60**‡
RWTH	.38	.39	**.45**	.32	**.63**‡	**.46**	**.51**⋆	**.34**	**.56**	.39	.32	**.52**		**.48**	**.46**	**.46**	**.66**‡	**.62**†	**.61**‡	**.66**‡	**.54**⋆
UEDIN	.41	.21	.31	.19‡	**.68**‡	**.46**	**.42**	.35	.41	.38	.31	**.46**	.33		.34	.41	**.41**	.35	**.44**	**.63**‡	.37
UKA	**.40**	.31	**.54**†	.19‡	**.51**	.37	**.44**	.33	**.52**	**.51**	.17‡	.27	.32	**.49**		.34	**.39**	**.53**	**.36**	**.44**	.29
USAAR	**.44**	.43	**.52**	.26†	**.62**⋆	.48	**.46**	.30	.30	**.58**	.17†	**.24**	.44	**.47**	**.41**		**.65**‡	**.52**	**.70**‡	**.55**	.41
BBN-COMBO	.21‡	.21†	.12‡	.23‡	.26	.32	.28†	.23†	.12‡	.26⋆	.22†	**.49**	.09‡	.34	.23	.19‡		**.44**	**.49**†	.28	.21‡
CMU-COMBO	.41	.36	.4	.28	.30	.35	.47	.21†	.29	.42	.23†	.31	.17†	**.49**	.25	.42	.31		**.37**	**.29**	.25
CMU-COMBO-H	.24	.21†	**.38**	.23‡	**.37**	**.39**	.31	.24	.31	.41	.28†	.31	.14‡	.33	.34	.24‡	.18†	.3		.29	.27
DCU-COMBO	.41	.13‡	**.42**	.20‡	.37	.29	**.50**	.19†	.44	**.49**	.23†	.46	.20‡	.21‡	.37	.39	**.31**	.26	**.46**		.19‡
USAAR-COMBO	**.41**	.25†	**.18**	.28†	**.66**‡	**.53**	**.52**⋆	**.48**	.41	.38	**.53**	.17‡	.21⋆	**.42**	**.42**	**.47**	**.58**‡	**.58**	**.47**	**.63**‡	
> OTHERS	.40	.31	.41	.23	**.56**	.43	.46	.36	.37	.41	.30	.40	.33	.41	.40	.40	.50	.47	.46	.49	.36
>= OTHERS	.58	.5	.66	.34	**.76**	.62	.65	.60	.56	.54	.47	.59	.52	.61	.61	.55	**.73**	.66	.71	.67	.57

Table 18: Sentence-level ranking for the WMT09 French-English News Task

	DCU	GENEVA	GOOGLE	LIMSI	LIUM-SYSTRAN	RBMT1	RBMT3	RBMT4	RBMT5	RWTH	SYSTRAN	UEDIN	UKA	USAAR	DCU-COMBO	USAAR-COMBO
DCU		.12‡	.39	**.47**	**.44**	.33	**.44**	.27	**.45**	.24⋆	**.49**	.24	**.46**	.26†	**.39**	.33
GENEVA	**.62**‡		**.73**‡	**.69**‡	**.80**‡	**.50**⋆	**.71**‡	**.50**⋆	**.52**⋆	**.56**†	**.66**‡	**.46**‡	**.56**‡	**.57**	**.74**‡	**.84**‡
GOOGLE	**.46**	.15‡		.28	.42	.26	**.44**	.26†	.34	.29⋆	**.44**	.24	.32	.29	**.36**	.32
LIMSI	.25	.16‡	**.45**		**.48**	.23⋆	**.43**	.30	**.45**	.27	**.42**	**.34**	.4	.36	**.53**†	**.38**
LIUM-SYSTRAN	.24	‡	**.45**	.32		.17†	.29	.17†	.21†	.38	.29	.17‡	.35	.17†	.41	**.41**
RBMT1	**.39**	.25⋆	**.51**	**.51**⋆	**.53**†		**.46**	**.40**	.29	**.52**	.36	**.60**⋆	**.63**‡	**.41**	**.44**	**.60**†
RBMT3	.36	.11‡	.37	.37	**.52**	.24		.25⋆	.27⋆	.31	**.44**	.43	.32	.27⋆	**.53**	**.44**
RBMT4	**.36**	.19⋆	**.58**†	**.37**	**.57**†	.23	**.61**⋆		**.42**	.32	**.50**	.22	.39	**.44**	**.53**	**.56**⋆
RBMT5	.41	.17⋆	**.53**	.39	**.61**†	**.38**	**.58**⋆	.30		**.41**	**.52**⋆	.41	**.48**	.13	**.54**	**.60**
RWTH	**.59**⋆	.21†	**.63**⋆	**.50**	**.47**	.29	**.44**	**.37**	.31		.37	**.35**	**.51**	.16†	**.50**‡	**.57**†
SYSTRAN	.35	.20‡	.33	.39	**.38**	**.40**	.22	.29	.26⋆	**.44**		**.47**	.33	.32	**.60**⋆	**.45**
UEDIN	**.38**	.11‡	**.41**	.28	**.77**‡	.33⋆	**.51**	**.44**	**.49**	.32	.37		**.30**	.31	**.56**	**.56**‡
UKA	.36	.09‡	**.46**	.4	**.45**	.23‡	**.50**	.39	.29	.29	**.47**	.26		.19‡	.41	**.56**†
USAAR	**.66**†	.27	**.52**	**.49**	**.70**†	.31	**.61**⋆	.29	**.32**	**.64**†	**.62**	**.51**	**.61**‡		**.76**‡	**.65**‡
DCU-COMBO	.32	.11‡	.30	.18†	**.45**	.22	.29	.33	.29	.13‡	.27⋆	.26	.41	.12‡		.21
USAAR-COMBO	**.40**	‡	**.39**	.17	.26	.17†	.28	.20⋆	.28	.20†	.39	.04‡	.06†	.08‡	**.39**	
> OTHERS	.41	.15	.47	.39	**.52**	.29	.45	.32	.35	.35	.45	.34	.42	.28	.51	.49
>= OTHERS	.65	.38	.68	.64	**.73**	.54	.65	.59	.57	.58	.65	.60	.66	.48	.74	**.77**

Table 19: Sentence-level ranking for the WMT09 English-French News Task

	CU-BOJAR	GOOGLE	UEDIN	BBN-COMBO	CMU-COMBO
CU-BOJAR		**.54**‡	**.44**	**.45**‡	**.52**‡
GOOGLE	.28‡		.32‡	.18‡	.23
UEDIN	.38	**.51**‡		**.38**	**.45**‡
BBN-COMBO	.31‡	**.39**‡	.32		**.38**‡
CMU-COMBO	.28‡	**.29**	.27‡	.24‡	
> OTHERS	.31	**.43**	.34	.31	.40
>= OTHERS	.51	**.75**	.57	.65	**.73**

Table 20: Sentence-level ranking for the WMT09 Czech-English News Task

	CU-BOJAR	CU-TECTOMT	EUROTRANXP	GOOGLE	PCTRANS	UEDIN
CU-BOJAR		.31‡	**.45**‡	**.43**‡	**.48**‡	.30‡
CU-TECTOMT	**.51**‡		**.54**‡	**.56**‡	**.58**‡	**.42**⋆
EUROTRANXP	.35‡	.26‡		.39	**.38**	.29‡
GOOGLE	.31‡	.30‡	**.42**		**.43**⋆	.26‡
PCTRANS	.33‡	.27‡	.36	.38⋆		.30‡
UEDIN	**.42**‡	.37⋆	**.52**‡	**.50**‡	**.53**‡	
> OTHERS	.38	.30	.46	.45	**.48**	.31
>= OTHERS	.61	.48	.67	.66	**.67**	.53

Table 21: Sentence-level ranking for the WMT09 English-Czech News Task

	MORPHO	UEDIN	UMD	BBN-COMBO	CMU-COMBO	CMU-COMBO-HYPOSEL
MORPHO		.21‡	.28‡	.24‡	.27‡	.28‡
UEDIN	**.70**‡		**.59**‡	**.45**‡	**.55**‡	**.50**‡
UMD	**.61**‡	.26‡		.21‡	.29	**.38**
BBN-COMBO	**.67**‡	.23‡	**.48**‡		**.41**⋆	**.52**‡
CMU-COMBO	**.59**‡	.25‡	**.35**	.29⋆		**.42**
CMU-COMBO-HYPOSEL	**.55**‡	.15‡	.34	.27‡	.34	
> OTHERS	**.62**	.22	.41	.29	.37	.42
>= OTHERS	**.75**	.45	.66	.54	.62	**.68**

Table 22: Sentence-level ranking for the WMT09 Hungarian-English News Task

	GOOGLE$_{CZ}$	GOOGLE$_{ES}$	GOOGLE$_{FR}$	RBMT2$_{DE}$	RBMT3$_{DE}$	RBMT3$_{ES}$	RBMT3$_{FR}$	RBMT5$_{ES}$	RBMT5$_{FR}$	BBN-COMBO$_{CZ}$	BBN-COMBO$_{DE}$	BBN-COMBO$_{ES}$	BBN-COMBO$_{FR}$	BBN-COMBO$_{HU}$	BBN-COMBO$_{XX}$	CMU-COMBO-HYPOSEL$_{DE}$	CMU-COMBO-HYPOSEL$_{HU}$	CMU-COMBO$_{CZ}$	CMU-COMBO$_{HU}$	CMU-COMBO$_{XX}$	DCU-COMBO$_{FR}$	RWTH-COMBO$_{DE}$	RWTH-COMBO$_{XX}$	USAAR-COMBO$_{ES}$
GOOGLE$_{CZ}$		**.61**⋆	**.54**⋆	**.47**	**.52**	**.51**	**.47**	**.61**⋆	.42	.38	**.52**	**.55**	**.54**	.11‡	**.51**	.48	.34	**.49**	.32	**.53**	**.52**	**.50**	**.59**	**.53**
GOOGLE$_{ES}$	.33⋆		.42	.37	.38	.41	.35	**.49**	.45	.11‡	.39	.25	.36	.18‡	.26⋆	.36	.22‡	.32	.18‡	.38	.4	.4	.38	.22
GOOGLE$_{FR}$	.27⋆	.42		.26†	.36	.43	**.47**	.33	.35	.29⋆	.23†	**.50**	.23	.14‡	.29†	.21†	.11‡	.17‡	.22‡	**.39**	**.48**	.32	.36	.27
RBMT2$_{DE}$	.33	**.49**	**.61**†		.41	.43	.25†	**.52**	.38	.33	.41	.4	**.55**	.20‡	**.66**⋆	**.62**⋆	.18‡	**.55**	.35	.35	**.58**	**.54**	**.61**⋆	**.57**†
RBMT3$_{DE}$	.37	**.60**	**.54**	.41		.42	.38	.45	**.61**	**.48**	.39	.40	**.63**‡	.32	.43	.25†	.35	.35	.25†	**.56**	**.69**†	**.46**	**.49**	**.46**
RBMT3$_{ES}$	.34	**.52**	**.46**	**.51**	**.54**		.43	.36	.38	.30⋆	**.54**	.41	.47	.25⋆	**.50**	.42	.26⋆	.43	.27†	**.52**	**.57**	.47	**.46**	.26⋆
RBMT3$_{FR}$	.40	**.58**	.37	**.63**†	**.53**	**.57**		**.54**	**.50**	.36	**.64**⋆	.44	**.55**	.13‡	**.60**	**.64**⋆	.4	**.53**	.31	**.46**	**.48**	.44	**.52**	**.42**
RBMT5$_{ES}$	.29⋆	.41	**.55**	.31	**.48**	.36	.33		.39	.16‡	.44	**.50**	**.68**†	.23†	.35	**.48**	.38	.37	.41	**.60**†	**.51**	**.51**	**.65**⋆	.32
RBMT5$_{FR}$	**.47**	**.52**	**.45**	**.50**	.33	**.51**	.34	**.42**		.29	**.59**	.44	**.49**	‡	**.49**	**.61**⋆	.28⋆	.19‡	.35	**.58**†	**.60**⋆	.27	**.59**	**.57**
BBN-COMBO$_{CZ}$	**.41**	**.74**‡	**.65**⋆	**.55**	.44	**.67**⋆	**.56**	**.80**‡	**.46**		.46	**.58**	**.70**‡	.22†	**.73**‡	**.63**†	.32	**.38**	.48	**.65**⋆	**.72**‡	**.66**‡	**.70**‡	**.58**
BBN-COMBO$_{DE}$	.39	**.54**	**.58**†	.41	**.49**	.44	.31⋆	.44	.28	**.49**		**.49**	**.52**	.16‡	**.52**	.36	.22⋆	.38	.33⋆	.41	**.68**†	.34	**.52**	**.56**
BBN-COMBO$_{ES}$	.38	**.40**	.41	**.43**	**.47**	**.55**	**.46**	.25	**.51**	.31	.43		**.44**	.20†	**.50**	**.42**	.30†	.32	.29⋆	.36	**.62**	.47	.44	**.38**
BBN-COMBO$_{FR}$	.38	**.52**	**.35**	.36	.27‡	**.53**	.40	.26†	.33	.24‡	.44	.36		.12‡	**.47**	**.47**	.32	.44	.27†	**.41**	.42	.33	**.60**‡	.35
BBN-COMBO$_{HU}$	**.84**‡	**.75**‡	**.78**‡	**.60**‡	**.57**	**.70**⋆	**.71**‡	**.62**†	**.84**‡	**.65**†	**.72**‡	**.63**†	**.85**‡		**.78**‡	**.69**†	**.60**‡	**.71**‡	**.50**	**.85**‡	**.78**‡	**.87**‡	**.86**‡	**.75**‡
BBN-COMBO$_{XX}$	.4	**.54**⋆	**.63**†	.34⋆	**.50**	.47	.32	**.45**	.39	.20‡	.39	.45	.41	.14‡		.24‡	.21‡	.3	.21‡	**.46**	**.40**	.47	**.41**	**.41**
CMU-CMB-HYP$_{DE}$	.48	**.43**	**.68**†	.29⋆	**.64**†	**.46**	.31⋆	.30	.30⋆	.23†	**.41**	.39	.32	.19†	**.74**‡		.21‡	.32	.31	**.50**	**.74**‡	.38	**.56**⋆	**.53**
CMU-CMB-HYP$_{HU}$	**.63**	**.75**‡	**.78**‡	**.70**‡	**.55**	**.63**⋆	**.46**	**.58**	**.59**⋆	**.50**	**.61**⋆	**.70**†	**.59**	.13‡	**.68**‡	**.69**‡		**.65**‡	.39	**.75**‡	**.71**‡	**.82**‡	**.80**‡	**.68**†
CMU-COMBO$_{CZ}$	.32	**.59**	**.81**‡	.36	**.50**	**.46**	.41	**.50**	**.60**‡	.28	**.54**	**.52**	**.47**	.20‡	**.55**	**.56**	.26‡		.13‡	**.55**	**.69**†	**.57**	**.66**⋆	**.55**
CMU-COMBO$_{HU}$	**.62**	**.76**‡	**.69**‡	**.58**	**.68**†	**.67**†	**.59**	**.54**	**.54**	.48	**.67**⋆	**.64**⋆	**.70**†	.32	**.74**‡	**.60**	**.50**	**.77**‡		**.66**†	**.72**‡	**.61**	**.82**‡	**.82**‡
CMU-COMBO$_{XX}$	.4	**.50**	.33	**.51**	.37	.43	.44	.29†	.24†	.32⋆	**.56**	**.43**	.39	.13‡	.39	.39	.16‡	.30	.32†		.39	.4	**.46**	.4
DCU-COMBO$_{FR}$	.44	**.57**	.29	.32	.25†	.29	.26	.35	.27⋆	.19‡	.23†	.38	.42	.15‡	.34	.20‡	.12‡	.19†	.17‡	**.50**		**.55**	**.49**	.30⋆
RWTH-COMBO$_{DE}$	.41	**.43**	**.52**	.37	.39	**.53**	**.50**	.35	**.53**	.25‡	**.40**	.47	**.54**	.10‡	.47	**.41**	.07‡	.38	.30	**.53**	.38		**.56**	**.49**
RWTH-COMBO$_{XX}$	.31	.38	**.44**	.26⋆	.41	.39	.31	.26⋆	.32	.18‡	.29	.44	.19‡	.10‡	.36	.25⋆	.11‡	.28⋆	.15‡	.39	.42	.28		.44
USAAR-COMBO$_{ES}$	.37	**.37**	**.54**	.21†	.4	**.58**⋆	.39	**.47**	.31	.32	.34	.28	**.55**	.11‡	.38	.38	.20†	.38	.18‡	**.44**	**.67**⋆	.43	.44	
> OTHERS	.41	.54	.54	.43	.45	.49	.41	.44	.44	.32	.46	.46	.50	.16	.51	.45	.26	.40	.29	.52	**.57**	.48	.55	.47
>= OTHERS	.52	.67	**.70**	.55	.55	.57	.52	.58	.58	.43	.57	.59	.62	.27	.62	.58	.37	.52	.36	.63	.68	.59	**.69**	.62

Table 23: Sentence-level ranking for the WMT09 All-English News Task

	BBN-COMBO	CMU-COMBO	RWTH-COMBO
BBN-COMBO		.37	**.40**‡
CMU-COMBO	**.41**		**.44**‡
RWTH-COMBO	.32‡	.34‡	
> OTHERS	.36	.35	**.42**
>= OTHERS	.62	.58	**.67**

Table 24: Sentence-level ranking for the WMT09 Multisource-English News Task

	RANK	BLEU	BLEU-CASED	BLEU-TER	BLEUSP	BLEUSP4114	MAXSIM	METEOR-0.6	METEOR-0.7	METEOR-RANKING	NIST	NIST-CASED	RTE-ABSOLUTE	RTE-PAIRWISE	TER	TERP	ULC	WCD6P4ER	WPF	WPBLEU
German-English News Task																				
BBN-COMBO	0.68	0.24	0.22	–0.17	0.29	0.31	0.51	0.55	0.6	0.41	7.08	6.78	0.13	0.1	0.54	0.63	0.31	0.45	0.36	0.31
CMU-COMBO	0.63	0.22	0.21	–0.19	0.28	0.29	0.49	0.54	0.58	0.4	6.95	6.71	0.12	0.09	0.56	0.66	0.29	0.47	0.35	0.29
CMU-COMBO-HYPOSEL	0.62	0.23	0.21	–0.19	0.28	0.3	0.49	0.54	0.57	0.4	6.79	6.5	0.11	0.09	0.57	0.66	0.29	0.47	0.35	0.3
GENEVA	0.33	0.1	0.09	–0.33	0.17	0.18	0.38	0.43	0.44	0.30	4.88	4.65	0.03	0.04	0.71	0.86	0.22	0.58	0.25	0.17
GOOGLE	0.65	0.21	0.20	–0.2	0.27	0.28	0.48	0.54	0.57	0.39	6.85	6.65	0.11	0.11	0.56	0.65	0.29	0.48	0.35	0.28
JHU-TROMBLE	0.13	0.07	0.06	–0.38	0.09	0.1	0.34	0.43	0.41	0.29	4.90	4.25	0.02	0.02	0.81	1	0.19	0.61	0.22	0.12
LIU	0.50	0.19	0.18	–0.22	0.25	0.27	0.46	0.51	0.54	0.38	6.35	6.02	0.06	0.05	0.61	0.72	0.27	0.49	0.33	0.26
RBMT1	0.54	0.14	0.13	–0.29	0.20	0.21	0.43	0.50	0.53	0.37	5.30	5.07	0.04	0.04	0.67	0.76	0.26	0.55	0.29	0.22
RBMT2	0.64	0.17	0.16	–0.26	0.23	0.24	0.48	0.52	0.55	0.38	6.06	5.75	0.1	0.12	0.63	0.70	0.29	0.51	0.31	0.24
RBMT3	0.64	0.17	0.16	–0.25	0.23	0.25	0.48	0.52	0.55	0.38	5.98	5.71	0.09	0.09	0.61	0.68	0.29	0.51	0.32	0.25
RBMT4	0.62	0.16	0.14	–0.27	0.21	0.23	0.45	0.5	0.52	0.36	5.65	5.36	0.06	0.07	0.65	0.72	0.27	0.52	0.30	0.23
RBMT5	0.66	0.16	0.15	–0.26	0.22	0.24	0.47	0.51	0.54	0.37	5.76	5.52	0.07	0.06	0.63	0.70	0.28	0.52	0.31	0.24
RWTH	0.50	0.19	0.18	–0.21	0.25	0.26	0.45	0.50	0.53	0.36	6.44	6.24	0.06	0.03	0.60	0.74	0.27	0.49	0.33	0.26
RWTH-COMBO	0.7	0.23	0.22	–0.18	0.29	0.30	0.50	0.55	0.59	0.41	7.06	6.81	0.11	0.07	0.54	0.63	0.30	0.46	0.36	0.31
STUTTGART	0.61	0.2	0.18	–0.22	0.26	0.27	0.48	0.52	0.56	0.38	6.39	6.11	0.1	0.06	0.60	0.69	0.29	0.49	0.33	0.27
SYSTRAN	0.6	0.19	0.17	–0.22	0.24	0.26	0.47	0.52	0.55	0.38	6.40	6.08	0.08	0.07	0.60	0.71	0.28	0.5	0.33	0.26
UEDIN	0.59	0.20	0.19	–0.22	0.26	0.27	0.47	0.52	0.55	0.38	6.47	6.24	0.07	0.04	0.61	0.70	0.27	0.49	0.34	0.27
UKA	0.58	0.21	0.2	–0.20	0.27	0.28	0.47	0.52	0.56	0.38	6.66	6.43	0.08	0.04	0.58	0.69	0.28	0.48	0.34	0.28
UMD	0.56	0.21	0.19	–0.19	0.26	0.28	0.47	0.52	0.56	0.38	6.74	6.42	0.08	0.04	0.56	0.69	0.28	0.48	0.34	0.27
USAAR	0.65	0.17	0.15	–0.26	0.23	0.24	0.47	0.51	0.54	0.38	5.89	5.64	0.06	0.05	0.64	0.71	0.28	0.52	0.31	0.24
USAAR-COMBO	0.62	0.17	0.16	–0.25	0.23	0.24	0.47	0.51	0.55	0.38	5.99	6.85	0.07	0.06	0.64	0.70	0.28	0.51	0.32	0.25
Spanish-English News Task																				
BBN-COMBO	0.64	0.29	0.27	–0.13	0.34	0.35	0.53	0.57	0.62	0.43	7.64	7.35	0.16	0.13	0.51	0.61	0.33	0.42	0.4	0.35
CMU-COMBO	0.7	0.28	0.27	–0.13	0.33	0.35	0.53	0.58	0.62	0.43	7.65	7.46	0.21	0.2	0.51	0.60	0.34	0.42	0.40	0.36
GOOGLE	0.70	0.29	0.28	–0.13	0.34	0.35	0.53	0.58	0.62	0.43	7.68	7.50	0.23	0.22	0.5	0.59	0.34	0.42	0.41	0.36
NICT	0.37	0.22	0.22	–0.19	0.27	0.29	0.48	0.54	0.57	0.39	6.91	6.74	0.1	0.1	0.60	0.71	0.3	0.46	0.36	0.3
RBMT1	0.55	0.19	0.18	–0.24	0.25	0.26	0.49	0.54	0.57	0.40	6.07	5.93	0.11	0.12	0.62	0.69	0.3	0.49	0.34	0.28
RBMT3	0.55	0.20	0.2	–0.22	0.26	0.27	0.50	0.54	0.58	0.41	6.24	6.08	0.13	0.14	0.60	0.65	0.31	0.48	0.36	0.29
RBMT4	0.53	0.2	0.19	–0.22	0.25	0.27	0.48	0.53	0.57	0.4	6.20	6.03	0.10	0.11	0.60	0.67	0.3	0.48	0.35	0.28
RBMT5	0.55	0.20	0.2	–0.22	0.26	0.27	0.5	0.54	0.58	0.40	6.26	6.10	0.12	0.11	0.6	0.65	0.31	0.48	0.36	0.29
RWTH	0.51	0.24	0.23	–0.16	0.3	0.31	0.49	0.54	0.58	0.4	7.12	6.95	0.11	0.08	0.56	0.68	0.31	0.45	0.37	0.32
TALP-UPC	0.59	0.26	0.25	–0.15	0.31	0.33	0.51	0.56	0.6	0.41	7.28	7.02	0.13	0.11	0.54	0.64	0.32	0.44	0.38	0.33
UEDIN	0.56	0.26	0.25	–0.15	0.32	0.33	0.51	0.56	0.60	0.42	7.25	7.04	0.16	0.1	0.55	0.64	0.32	0.43	0.39	0.34
USAAR	0.51	0.2	0.19	–0.22	0.25	0.27	0.48	0.54	0.57	0.4	6.31	6.14	0.11	0.09	0.62	0.67	0.3	0.48	0.34	0.28
USAAR-COMBO	0.69	0.29	0.27	–0.13	0.34	0.35	0.53	0.58	0.62	0.43	7.58	7.25	0.20	0.13	0.51	0.6	0.34	0.42	0.4	0.35
French-English News Task																				
BBN-COMBO	0.73	0.31	0.3	–0.11	0.36	0.38	0.54	0.59	0.64	0.45	7.88	7.58	0.14	0.12	0.2	0.20	0.36	0.40	0.41	0.37
CMU-COMBO	0.66	0.3	0.29	–0.12	0.35	0.36	0.53	0.58	0.63	0.44	7.72	7.57	0.15	0.12	0.24	0.26	0.35	0.41	0.41	0.37
CMU-COMBO-HYPOSEL	0.71	0.28	0.26	–0.14	0.33	0.35	0.53	0.57	0.61	0.43	7.40	7.15	0.1	0.08	0.31	0.33	0.34	0.42	0.4	0.35
CMU-STATXFER	0.58	0.24	0.23	–0.18	0.29	0.31	0.49	0.54	0.58	0.40	6.89	6.75	0.08	0.07	0.38	0.42	0.31	0.46	0.37	0.32
COLUMBIA	0.50	0.23	0.22	–0.18	0.29	0.30	0.49	0.54	0.58	0.40	6.85	6.68	0.07	0.07	0.36	0.39	0.31	0.46	0.36	0.31
DCU	0.66	0.27	0.25	–0.15	0.32	0.34	0.52	0.56	0.61	0.42	7.29	6.94	0.09	0.07	0.32	0.34	0.33	0.43	0.38	0.34
DCU-COMBO	0.67	0.31	0.31	–0.11	0.36	0.37	0.54	0.59	0.64	0.44	7.84	7.69	0.14	0.12	0.21	0.22	0.35	0.41	0.42	0.38
GENEVA	0.34	0.14	0.14	–0.29	0.21	0.22	0.43	0.49	0.52	0.36	5.32	5.15	0.05	0.05	0.54	0.52	0.26	0.53	0.29	0.22
GOOGLE	0.76	0.31	0.30	–0.10	0.36	0.37	0.54	0.58	0.63	0.44	8	7.84	0.17	0.13	0.17	0.2	0.36	0.41	0.42	0.38
JHU	0.62	0.27	0.23	–0.15	0.32	0.33	0.51	0.56	0.6	0.41	7.23	6.68	0.08	0.05	0.33	0.36	0.32	0.43	0.37	0.32
LIMSI	0.65	0.26	0.25	–0.16	0.30	0.32	0.51	0.56	0.60	0.42	7.02	6.87	0.09	0.07	0.35	0.36	0.33	0.44	0.38	0.33
LIUM-SYSTRAN	0.60	0.27	0.26	–0.15	0.32	0.33	0.51	0.56	0.60	0.42	7.26	7.10	0.10	0.06	0.33	0.36	0.33	0.43	0.39	0.35
RBMT1	0.56	0.18	0.18	–0.25	0.24	0.25	0.48	0.53	0.57	0.4	5.89	5.73	0.07	0.06	0.51	0.45	0.3	0.50	0.34	0.26
RBMT3	0.54	0.2	0.19	–0.22	0.25	0.27	0.48	0.53	0.56	0.39	6.12	5.96	0.07	0.06	0.45	0.45	0.30	0.49	0.35	0.28
RBMT4	0.47	0.19	0.18	–0.24	0.24	0.26	0.48	0.52	0.56	0.39	5.97	5.83	0.07	0.06	0.46	0.45	0.3	0.49	0.34	0.27
RBMT5	0.59	0.19	0.19	–0.24	0.25	0.26	0.49	0.54	0.57	0.40	6.03	5.9	0.09	0.07	0.46	0.43	0.31	0.49	0.35	0.28
RWTH	0.52	0.25	0.24	–0.16	0.30	0.32	0.5	0.55	0.59	0.40	7.09	6.94	0.07	0.03	0.35	0.39	0.32	0.44	0.38	0.32
UEDIN	0.61	0.25	0.24	–0.16	0.31	0.32	0.50	0.55	0.59	0.41	7.04	6.85	0.08	0.04	0.35	0.38	0.32	0.44	0.38	0.33
UKA	0.61	0.26	0.25	–0.15	0.31	0.33	0.51	0.55	0.6	0.41	7.17	7.00	0.08	0.04	0.34	0.37	0.32	0.44	0.38	0.34
USAAR	0.55	0.19	0.18	–0.24	0.24	0.26	0.48	0.54	0.57	0.4	6.08	5.92	0.07	0.06	0.46	0.44	0.3	0.49	0.34	0.26
USAAR-COMBO	0.57	0.26	0.25	–0.16	0.31	0.33	0.51	0.55	0.59	0.41	7.13	6.85	0.08	0.02	0.33	0.35	0.32	0.44	0.38	0.33
Czech-English News Task																				
BBN-COMBO	0.65	0.22	0.20	–0.19	0.27	0.29	0.47	0.52	0.56	0.39	6.74	6.45	0.24	0.3	0.52	0.60	0.29	0.47	0.34	0.29
CMU-COMBO	0.73	0.22	0.20	–0.2	0.27	0.29	0.47	0.53	0.57	0.39	6.72	6.46	0.34	0.34	0.53	0.60	0.29	0.47	0.35	0.29
CU-BOJAR	0.51	0.16	0.15	–0.26	0.22	0.24	0.43	0.5	0.52	0.36	5.84	5.54	0.26	0.28	0.61	0.69	0.26	0.52	0.31	0.24
GOOGLE	0.75	0.21	0.20	–0.19	0.26	0.28	0.46	0.52	0.55	0.38	6.82	6.61	0.32	0.33	0.53	0.62	0.29	0.47	0.35	0.28
UEDIN	0.57	0.2	0.19	–0.23	0.25	0.27	0.45	0.50	0.54	0.37	6.2	6	0.22	0.25	0.56	0.63	0.27	0.49	0.33	0.27
Hungarian-English News Task																				
BBN-COMBO	0.54	0.14	0.13	–0.29	0.19	0.21	0.38	0.45	0.46	0.32	5.46	5.2	0.16	0.18	0.71	0.83	0.23	0.55	0.27	0.2
CMU-COMBO	0.62	0.14	0.13	–0.29	0.19	0.21	0.39	0.46	0.47	0.32	5.52	5.24	0.28	0.22	0.71	0.82	0.23	0.55	0.28	0.2
CMU-COMBO-HYPOSEL	0.68	0.14	0.12	–0.29	0.19	0.21	0.39	0.45	0.46	0.32	5.51	5.16	0.25	0.25	0.71	0.82	0.23	0.55	0.27	0.2
MORPHO	0.75	0.1	0.09	–0.36	0.15	0.17	0.39	0.45	0.46	0.32	4.75	4.55	0.34	0.49	0.79	0.83	0.23	0.6	0.26	0.17
UEDIN	0.45	0.12	0.11	–0.32	0.18	0.19	0.37	0.42	0.43	0.30	4.95	4.74	0.12	0.12	0.75	0.87	0.21	0.58	0.27	0.19
UMD	0.66	0.13	0.12	–0.28	0.18	0.2	0.36	0.44	0.45	0.30	5.41	5.12	0.21	0.13	0.68	0.85	0.22	0.55	0.27	0.18

Table 25: Automatic evaluation metric scores for translations into English

	RANK	BLEU	BLEU-CASED	BLEU-TER	BLEUSP	BLEUSP4114	NIST	NIST-CASED	TER	TERP	WCD6P4ER	WPF	WPBLEU
English-German News Task													
GOOGLE	0.54	0.15	0.14	–0.29	0.20	0.22	5.36	5.25	0.62	0.74	0.54	0.3	0.23
LIU	0.49	0.14	0.13	–0.29	0.2	0.21	5.35	5.18	0.65	0.78	0.54	0.3	0.23
RBMT1	0.57	0.11	0.11	–0.32	0.17	0.19	4.69	4.59	0.67	0.81	0.57	0.28	0.21
RBMT2	0.66	0.13	0.13	–0.30	0.19	0.21	5.08	4.99	0.62	0.75	0.55	0.30	0.23
RBMT3	0.64	0.12	0.12	–0.29	0.2	0.21	4.8	4.71	0.62	0.76	0.54	0.31	0.25
RBMT4	0.58	0.11	0.10	–0.33	0.17	0.18	4.66	4.57	0.7	0.84	0.57	0.27	0.2
RBMT5	0.64	0.13	0.12	–0.3	0.19	0.20	5.03	4.94	0.64	0.79	0.55	0.3	0.23
RWTH	0.48	0.14	0.13	–0.28	0.2	0.21	5.51	5.41	0.62	0.78	0.53	0.3	0.23
STUTTGART	0.43	0.12	0.12	–0.31	0.18	0.20	5.06	4.82	0.67	0.79	0.55	0.29	0.21
UEDIN	0.51	0.15	0.15	–0.27	0.21	0.23	5.53	5.42	0.63	0.77	0.53	0.31	0.24
UKA	0.54	0.15	0.15	–0.27	0.21	0.22	5.6	5.48	0.62	0.75	0.52	0.31	0.24
USAAR	0.58	0.12	0.11	–0.33	0.18	0.19	4.83	4.71	0.69	0.8	0.57	0.28	0.21
USAAR-COMBO	0.52	0.16	0.15	–0.27	0.21	0.23	5.6	5.39	0.62	0.75	0.52	0.31	0.24
English-Spanish News Task													
GOOGLE	0.65	0.28	0.27	–0.15	0.33	0.34	7.27	7.07	0.36	0.42	0.42	0.37	0.31
NUS	0.59	0.25	0.23	–0.17	0.30	0.31	6.96	6.67	0.48	0.59	0.44	0.34	0.28
RBMT1	0.25	0.15	0.14	–0.27	0.20	0.22	5.32	5.17	0.55	0.66	0.51	0.24	0.16
RBMT3	0.66	0.18	0.17	–0.18	0.28	0.3	5.79	5.63	0.49	0.59	0.45	0.33	0.27
RBMT4	0.61	0.21	0.2	–0.20	0.26	0.28	6.47	6.28	0.52	0.64	0.47	0.31	0.25
RBMT5	0.64	0.22	0.21	–0.2	0.27	0.29	6.53	6.34	0.52	0.64	0.46	0.32	0.26
RWTH	0.51	0.22	0.21	–0.18	0.27	0.29	6.83	6.63	0.50	0.65	0.46	0.32	0.26
TALP-UPC	0.58	0.25	0.23	–0.17	0.3	0.31	6.96	6.69	0.47	0.58	0.44	0.34	0.28
UEDIN	0.66	0.25	0.24	–0.17	0.30	0.31	6.94	6.73	0.48	0.59	0.44	0.34	0.29
USAAR	0.48	0.20	0.19	–0.21	0.26	0.27	6.36	6.16	0.54	0.66	0.47	0.30	0.24
USAAR-COMBO	0.61	0.28	0.26	–0.14	0.33	0.34	7.36	6.97	0.39	0.48	0.42	0.36	0.31
English-French News Task													
DCU	0.65	0.24	0.22	–0.19	0.29	0.30	6.69	6.39	0.63	0.72	0.47	0.38	0.34
DCU-COMBO	0.74	0.28	0.27	–0.15	0.33	0.34	7.29	7.12	0.58	0.67	0.44	0.42	0.38
GENEVA	0.38	0.15	0.14	–0.27	0.20	0.22	5.59	5.39	0.68	0.82	0.53	0.32	0.25
GOOGLE	0.68	0.25	0.24	–0.17	0.30	0.31	6.90	6.71	0.62	0.7	0.46	0.40	0.36
LIMSI	0.64	0.25	0.24	–0.17	0.3	0.31	6.94	6.77	0.60	0.71	0.46	0.4	0.35
LIUM-SYSTRAN	0.73	0.26	0.24	–0.17	0.31	0.32	7.02	6.83	0.61	0.71	0.45	0.40	0.36
RBMT1	0.54	0.18	0.17	–0.23	0.24	0.26	6.12	5.96	0.65	0.76	0.5	0.35	0.29
RBMT3	0.65	0.22	0.20	–0.20	0.27	0.28	6.48	6.29	0.63	0.72	0.48	0.38	0.33
RBMT4	0.59	0.18	0.17	–0.24	0.24	0.25	6.02	5.86	0.66	0.77	0.50	0.35	0.3
RBMT5	0.57	0.20	0.19	–0.21	0.26	0.27	6.31	6.15	0.63	0.74	0.49	0.36	0.31
RWTH	0.58	0.22	0.21	–0.19	0.27	0.28	6.67	6.51	0.62	0.75	0.48	0.38	0.32
SYSTRAN	0.65	0.23	0.22	–0.19	0.28	0.29	6.7	6.47	0.63	0.74	0.47	0.39	0.34
UEDIN	0.60	0.24	0.23	–0.18	0.29	0.30	6.75	6.57	0.62	0.71	0.47	0.39	0.35
UKA	0.66	0.24	0.23	–0.18	0.29	0.30	6.82	6.65	0.61	0.71	0.46	0.39	0.35
USAAR	0.48	0.19	0.18	–0.23	0.24	0.26	6.16	5.98	0.66	0.76	0.5	0.34	0.29
USAAR-COMBO	0.77	0.27	0.25	–0.15	0.32	0.33	7.24	6.93	0.59	0.69	0.44	0.41	0.37
English-Czech News Task													
CU-BOJAR	0.61	0.14	0.13	–0.28	0.21	0.23	5.18	4.96	0.63	0.82	0.01	n/a	n/a
CU-TECTOMT	0.48	0.07	0.07	–0.35	0.14	0.16	4.17	4.03	0.71	0.96	0.01	n/a	n/a
EUROTRANXP	0.67	0.1	0.09	–0.33	0.16	0.18	4.38	4.26	0.7	0.93	0.01	n/a	n/a
GOOGLE	0.66	0.14	0.13	–0.30	0.20	0.22	4.96	4.84	0.66	0.82	0.01	n/a	n/a
PCTRANS	0.67	0.09	0.09	–0.34	0.17	0.18	4.34	4.19	0.71	0.90	0.01	n/a	n/a
UEDIN	0.53	0.14	0.13	–0.29	0.21	0.22	5.04	4.9	0.64	0.84	0.01	n/a	n/a
English-Hungarian News Task													
MORPHO	0.79	0.08	0.08	–0.37	0.15	0.16	4.04	3.92	0.83	1	0.6	n/a	n/a
UEDIN	0.32	0.1	0.09	–0.33	0.17	0.18	4.48	4.32	0.78	1	0.56	n/a	n/a

Table 26: Automatic evaluation metric scores for translations out of English