Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for 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
c.monz@uva.nl

Kay Peterson and **Mark Przybocki**
National Institute of Standards and Technology
kay.peterson,mark.przybocki@nist.gov

Omar F. Zaidan
Johns Hopkins University
ozaidan@cs.jhu.edu

Abstract

This paper presents the results of the WMT10 and MetricsMATR10 shared tasks,[1] which included a translation task, a system combination task, and an evaluation task. We conducted a large-scale manual evaluation of 104 machine translation systems and 41 system combination entries. We used the ranking of these systems to measure how strongly automatic metrics correlate with human judgments of translation quality for 26 metrics. This year we also investigated increasing the number of human judgments by hiring non-expert annotators through Amazon's Mechanical Turk.

1 Introduction

This paper presents the results of the shared tasks of the joint Workshop on statistical Machine Translation (WMT) and Metrics for MAchine TRanslation (MetricsMATR), which was held at ACL 2010. This builds on four previous WMT workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009), and one previous MetricsMATR meeting (Przybocki et al., 2008). There were three shared tasks this year: a translation task between English and four 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.

[1]The published version of this paper was missing part of the MetricsMATR analysis. This updated version was released on August 23, 2010.

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

- **Non-expert judgments** – In addition to having shared task participants judge translation quality, we also collected judgments from non-expert annotators hired through Amazon's Mechanical Turk. By collecting a large number of judgments we hope to reduce the burden on shared task participants, and to increase the statistical significance of our findings. We discuss the feasibility of using non-experts evaluators, by analyzing the cost, volume and quality of non-expert annotations.
- **Clearer results for system combination** – This year we excluded Google translations from the systems used in system combination. In last year's evaluation, the large margin between Google and many of the other systems meant that it was hard to improve on when combining systems. This year, the system combinations perform better than their component systems more often than last year.
- **Fewer rule-based systems** – This year there were fewer rule-based systems submitted. In past years, University of Saarland compiled a large set of outputs from rule-based machine translation (RBMT) systems. The RBMT systems were not submitted this year. This is unfortunate, because they tended to outperform the statistical systems for German, and they were often difficult to rank properly using automatic evaluation metrics.

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 num-

bers, and to refine evaluation methodologies for machine translation. As with past years, all of the data, translations, and human judgments produced for our workshop are publicly available.[2] 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 four other languages: German, Spanish, French, and Czech. We created a test set for each language pair by translating newspaper articles. We additionally provided training data and two baseline systems.

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 from mid-December 2009. A total of 119 articles were selected, in roughly equal amounts from a variety of Czech, English, French, German and Spanish news sites:[3]

Czech: iDNES.cz (5), iHNed.cz (1), Lidovky (16)

French: Les Echos (25)

Spanish: El Mundo (20), ABC.es (4), Cinco Dias (11)

English: BBC (5), Economist (2), Washington Post (12), Times of London (3)

German: Frankfurter Rundschau (11), Spiegel (4)

The translations were created by the professional translation agency CEET[4]. All of the translations were done directly, and not via an intermediate 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.

[2] http://statmt.org/wmt10/results.html

[3] 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.

[4] http://www.ceet.eu/

2.3 Baseline systems

To lower the barrier of entry for newcomers to the field, we provided two open source toolkits for phrase-based and parsing-based statistical machine translation (Koehn et al., 2007; Li et al., 2009).

2.4 Submitted systems

We received submissions from 33 groups from 29 institutions, as listed in Table 1, a 50% increase over last year's shared task.

We also evaluated 2 commercial off the shelf MT systems, and two online statistical machine translation systems. We note that these companies did not submit entries themselves. The entries for the online systems were done by translating the test data via their web interfaces. The data used to train the online systems is unconstrained. It is possible that part of the reference translations that were taken from online news sites could have been included in the online systems' language models.

2.5 System combination

In total, we received 153 primary system submissions along with 28 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 20 n-best lists accompanying the system submissions.

Table 2 lists the 9 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 that automatic measures are an imperfect substitute for human assessment of translation quality.

Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English	
Sentences	1,650,152		1,683,156		1,540,549	
Words	47,694,560	46,078,122	50,964,362	47,145,288	40,756,801	43,037,967
Distinct words	173,033	95,305	123,639	95,846	316,365	92,464

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	98,598		84,624		100,269		94,742	
Words	2,724,141	2,432,064	2,405,082	2,101,921	2,505,583	2,443,183	2,050,545	2,290,066
Distinct words	69,410	46,918	53,763	43,906	101,529	47,034	125,678	45,306

United Nations Training Corpus

	Spanish ↔ English		French ↔ English	
Sentences	6,222,450		7,230,217	
Words	213,877,170	190,978,737	243,465,100	216,052,412
Distinct words	441,517	361,734	402,491	412,815

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

CzEng Training Corpus

	Czech ↔ English	
Sentences	7,227,409	
Words	72,993,427	84,856,749
Distinct words	1,088,642	522,770

Europarl Language Model Data

	English	Spanish	French	German
Sentence	1,843,035	1,822,021	1,855,589	1,772,039
Words	50,132,615	51,223,902	54,273,514	43,781,217
Distinct words	99,206	178,934	127,689	328,628

News Language Model Data

	English	Spanish	French	German	Czech
Sentence	48,653,884	3,857,414	15,670,745	17,474,133	13,042,040
Words	1,148,480,525	106,716,219	382,563,246	321,165,206	205,614,201
Distinct words	1,451,719	548,169	998,595	1,855,993	1,715,376

News Test Set

	English	Spanish	French	German	Czech
Sentences	2489				
Words	62,988	65,654	68,107	62,390	53,171
Distinct words	9,457	11,409	10,775	12,718	15,825

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

ID	Participant
AALTO	Aalto University, Finland (Virpioja et al., 2010)
CAMBRIDGE	Cambridge University (Pino et al., 2010)
CMU	Carnegie Mellon University's Cunei system (Phillips, 2010)
CMU-STATXFER	Carnegie Mellon University's statistical transfer system (Hanneman et al., 2010)
COLUMBIA	Columbia University
CU-BOJAR	Charles University Bojar (Bojar and Kos, 2010)
CU-TECTO	Charles University Tectogramatical MT (Žabokrtský et al., 2010)
CU-ZEMAN	Charles University Zeman (Zeman, 2010)
DCU	Dublin City University (Penkale et al., 2010)
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz (Federmann et al., 2010)
EU	European Parliament, Luxembourg (Jellinghaus et al., 2010)
EUROTRANS	commercial MT provider from the Czech Republic
FBK	Fondazione Bruno Kessler (Hardmeier et al., 2010)
GENEVA	University of Geneva
HUICONG	Shanghai Jiao Tong University (Cong et al., 2010)
JHU	Johns Hopkins University (Schwartz, 2010)
KIT	Karlsruhe Institute for Technology (Niehues et al., 2010)
KOC	Koc University, Turkey (Bicici and Kozat, 2010; Bicici and Yuret, 2010)
LIG	LIG Lab, University Joseph Fourier, Grenoble (Potet et al., 2010)
LIMSI	LIMSI (Allauzen et al., 2010)
LIU	Linköping University (Stymne et al., 2010)
LIUM	University of Le Mans (Lambert et al., 2010)
NRC	National Research Council Canada (Larkin et al., 2010)
ONLINEA	an online machine translation system
ONLINEB	an online machine translation system
PC-TRANS	commercial MT provider from the Czech Republic
POTSDAM	Potsdam University
RALI	RALI - Université de Montréal (Huet et al., 2010)
RWTH	RWTH Aachen (Heger et al., 2010)
SFU	Simon Fraser University (Sankaran et al., 2010)
UCH-UPV	Universidad CEU-Cardenal Herrera y UPV (Zamora-Martinez and Sanchis-Trilles, 2010)
UEDIN	University of Edinburgh (Koehn et al., 2010)
UMD	University of Maryland (Eidelman et al., 2010)
UPC	Universitat Politècnica de Catalunya (Henríquez Q. et al., 2010)
UPPSALA	Uppsala University (Tiedemann, 2010)
UPV	Universidad Politécnica de Valencia (Sanchis-Trilles et al., 2010)
UU-MS	Uppsala University - Saers (Saers et al., 2010)

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., 2010)
CMU-COMBO-HEAFIELD	CMU system combination (Heafield and Lavie, 2010)
CMU-COMBO-HYPOSEL	CMU system combo with hyp. selection (Hildebrand and Vogel, 2010)
DCU-COMBO	Dublin City University system combination (Du et al., 2010)
JHU-COMBO	Johns Hopkins University system combination (Narsale, 2010)
KOC-COMBO	Koc University, Turkey (Bicici and Kozat, 2010; Bicici and Yuret, 2010)
LIUM-COMBO	University of Le Mans system combination (Barrault, 2010)
RWTH-COMBO	RWTH Aachen system combination (Leusch and Ney, 2010)
UPV-COMBO	Universidad Politécnica de Valencia (González-Rubio et al., 2010)

Table 2: Participants in the system combination task.

Language Pair	Sentence Ranking	Edited Translations	Yes/No Judgments
German-English	5,212	830	824
English-German	6,847	755	751
Spanish-English	5,653	845	845
English-Spanish	2,587	920	690
French-English	4,147	925	921
English-French	3,981	1,325	1,223
Czech-English	2,688	490	488
English-Czech	6,769	1,165	1,163
Totals	**37,884**	**7,255**	**6,905**

Table 3: The number of items that were collected for each task during the manual evaluation. An item is defined to be a rank label in the ranking task, an edited sentence in the editing task, and a yes/no judgment in the judgment task.

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 120 people participated in the manual evaluation[5], with 89 people putting in more than an hour's worth of effort, and 29 putting in more than four hours. A collective total of 337 hours of labor was invested.[6]

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 systems 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).*

Each screen for this task involved judging translations of three consecutive source segments. For each source segment, the annotator was shown the outputs of five submissions. For each of the language pairs, there were more than 5 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 Inter- and Intra-annotator agreement in the ranking task

We were interested in determining the inter- and intra-annotator agreement for the ranking task, since a reasonable degree of agreement must exist to support our process as a valid evaluation setup. To ensure we had enough data to measure agreement, we purposely designed the sampling of source segments shown to annotators so that items were likely to be repeated, both within an annotator's assigned tasks and across annotators. We did so by assigning an annotator a batch of 20 screens (each with three ranking sets; see 3.1) that were to be completed in full before generating new screens for that annotator.

Within each batch, the source segments for nine of the 20 screens (45%) were chosen from a small pool of 60 source segments, instead of being sampled from the larger pool of 1,000 source segments designated for the ranking task.[7] The larger pool was used to choose source segments for nine other screens (also 45%). As for the remaining two screens (10%), they were chosen randomly from the set of eighteen screens already chosen. Furthermore, in the two "local repeat" screens, the system choices were also preserved.

Heavily sampling from a small pool of source segments ensured we had enough data to measure inter-annotator agreement, while purposely making 10% of each annotator's screens repeats of previously seen sets in the same batch ensured we had enough data to measure intra-annotator agreement.

[5] We excluded data from three errant annotators, identified as follows. We considered annotators completing at least 3 screens, whose $P(A)$ with others (see 3.2) is less than 0.33. Out of seven such annotators, four were affiliated with shared task teams. The other three had no apparent affiliation, and so we discarded their data, less than 5% of the total data.

[6] Whenever an annotator appears to have spent more than ten minutes on a single screen, we assume they left their station and left the window open, rather than actually needing more than ten minutes. In those cases, we assume the time spent to be ten minutes.

[7] Each language pair had its own 60-sentence pool, disjoint from other language pairs' pools, but ach of the 60-sentence pools was a subset of the 1,000-sentence pool.

INTER-ANNOTATOR AGREEMENT

	$P(A)$	K
With references	0.658	0.487
Without references	0.626	0.439
WMT '09	0.549	0.323

INTRA-ANNOTATOR AGREEMENT

	$P(A)$	K
With references	0.755	0.633
Without references	0.734	0.601
WMT '09	0.707	0.561

Table 4: Inter- and intra-annotator agreement for the sentence ranking task. In this task, $P(E)$ is 0.333.

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 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 exact interpretation of the kappa coefficient is difficult, 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 is almost perfect.

Based on these interpretations the agreement for sentence-level ranking is *moderate* for inter-annotator agreement and *substantial* for intra-annotator agreement. These levels of agreement are higher than in previous years, partially due to the fact that that year we randomly included the references along the system outputs. In general, judges tend to rank the reference as the best translation, so people have stronger levels of agreement when it is included. That said, even when comparisons involving reference are excluded, we still see an improvement in agreement levels over last year.

3.3 Editing machine translation output

In addition to simply ranking the output of systems, we also had people edit the output of MT systems. We did not show them the reference translation, which makes our edit-based evaluation different from the Human-targeted Translation Edit 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 given to our judges were as follows:

> *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."*

A screenshot is shown in Figure 2. This year, judges were shown the translations of 5 consecutive source sentences, all produced by the same machine translation system. In last year's WMT evaluation they were shown only one sentence at a time, which made the task more difficult because the surrounding context could not be used as an aid to understanding.

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.

3.4 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:

Edit Machine Translation Outputs

Instructions:

- You are shown several **machine translation outputs**.
- Your task is to edit each translation to make it as fluent as possible.
- It is possible that the translation is already fluent. In that case, select **No corrections needed**.
- If you cannot understand the sentence well enough to correct it, select **Unable to correct**.
- The sentences are all from the same article. You can use the earlier and later sentences to help understand a confusing sentence.

Your edited translations	**The machine translations**
The shortage of snow in mountain worries the hoteliers ○ Edited ○ No corrections needed ○ Unable to correct Reset	**The shortage of snow in mountain worries the hoteliers**
The deserted tracks are not putting down problem only at the exploitants of skilift. ○ Edited ○ No corrections needed ○ Unable to correct Reset	**The deserted tracks are not putting down problem only at the exploitants of skilift.**
The lack of snow deters the people to reserving their stays at the ski in the hotels and pension. ○ Edited ○ No corrections needed ○ Unable to correct Reset	**The lack of snow deters the people to reserving their stays at the ski in the hotels and pension.**
Thereby, is always possible to track free bedrooms for all the dates in winter, including Christmas and Nouvel An. ○ Edited ○ No corrections needed ○ Unable to correct Reset	**Thereby, is always possible to track free bedrooms for all the dates in winter, including Christmas and Nouvel An.**

Figure 2: This screenshot shows what an annotator sees when beginning to edit the output of a machine translation system.

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.

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 5 shows the 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 $p \leq 0.1$. Multiple systems are listed as the winners for many language pairs because it was not possible to draw a statistically significant difference between the systems. There is no individual system clearly outperforming all other systems across the different language pairs. With the exception of French-English and English-French one can observe that top-performing constrained systems did as well as the unconstrained system ONLINEB.

Table 6 shows the best combination systems. For all language directions, except Spanish-English, one can see that the system combination runs outperform the individual systems and that in most cases the differences are statistically significant. While this is to be expected, system combination is not guaranteed to improve performance as some of the lower ranked combination runs show, which are outperformed by individual systems. Also note that except for Czech-English translation the online systems ONLINEA and ONLINEB where not included for the system combination runs

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 3 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).

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

- 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 Shared evaluation task overview

In addition to allowing the analysis of subjective translation quality measures for different systems, the judgments gathered during the manual evaluation may be used to evaluate how well the automatic evaluation metrics serve as a surrogate to the manual evaluation processes. NIST began running a "Metrics for MAchine TRanslation" challenge (MetricsMATR), and presented their findings at a workshop at AMTA (Przybocki et al., 2008). This year we conducted a joint MetricsMATR and WMT workshop, with NIST running the shared evaluation task and analyzing the results.

In this year's shared evaluation task 14 different research groups submitted a total of 26 different automatic metrics for evaluation:

Aalto University of Science and Technology (Dobrinkat et al., 2010)

- MT-NCD – A machine translation metric based on normalized compression distance (NCD), a general information-theoretic measure of string similarity. MT-NCD measures the surface level similarity between two strings with a general compression algorithm. More similar strings can be represented with

French-English
551–755 judgments per system

System	C?	≥others
LIUM ●⋆	Y	0.71
ONLINEB ●	N	0.71
NRC ●⋆	Y	0.66
CAMBRIDGE ●⋆	Y +GW	0.66
LIMSI ⋆	Y +GW	0.65
UEDIN	Y	0.65
RALI ●⋆	Y +GW	0.65
JHU	Y	0.59
RWTH ●⋆	Y +GW	0.55
LIG	Y	0.53
ONLINEA	N	0.52
CMU-STATXFER	Y	0.51
HUICONG	Y	0.51
DFKI	N	0.42
GENEVA	Y	0.27
CU-ZEMAN	Y	0.21

English-French
664–879 judgments per system

System	C?	≥others
UEDIN ●⋆	Y	0.70
ONLINEB ●	N	0.68
RALI ●⋆	Y +GW	0.66
LIMSI ●⋆	Y +GW	0.66
RWTH ●⋆	Y +GW	0.63
CAMBRIDGE ⋆	Y +GW	0.63
LIUM	Y	0.63
NRC	Y	0.62
ONLINEA	N	0.55
JHU	Y	0.53
DFKI	N	0.40
GENEVA	Y	0.35
EU	N	0.32
CU-ZEMAN	Y	0.26
KOC	Y	0.26

Czech-English
788–868 judgments per system

System	C?	≥others
ONLINEB ●	N	0.7
UEDIN ⋆	Y	0.61
CMU	Y	0.55
CU-BOJAR	N	0.55
AALTO	Y	0.43
ONLINEA	N	0.37
CU-ZEMAN	Y	0.22

German-English
723–879 judgments per system

System	C?	≥others
ONLINEB ●	N	0.73
KIT ●⋆	Y +GW	0.72
UMD ●⋆	Y	0.68
UEDIN ⋆	Y	0.66
FBK ⋆	Y +GW	0.66
ONLINEA ●	N	0.63
RWTH	Y +GW	0.62
LIU	Y	0.59
UU-MS	Y	0.55
JHU	Y	0.53
LIMSI	Y +GW	0.52
UPPSALA	Y	0.51
DFKI	N	0.50
HUICONG	Y	0.47
CMU	Y	0.46
AALTO	Y	0.42
CU-ZEMAN	Y	0.36
KOC	Y	0.23

English-German
1284–1542 judgments per system

System	C?	≥others
ONLINEB ●	N	0.70
DFKI ●	N	0.62
UEDIN ●⋆	Y	0.62
KIT ⋆	Y	0.60
ONLINEA	N	0.59
FBK ⋆	Y	0.56
LIU	Y	0.55
RWTH	Y	0.51
LIMSI	Y	0.51
UPPSALA	Y	0.47
JHU	Y	0.46
SFU	Y	0.34
KOC	Y	0.30
CU-ZEMAN	Y	0.28

English-Czech
1375–1627 judgments per system

System	C?	≥others
ONLINEB ●	N	0.70
CU-BOJAR ●	N	0.66
PC-TRANS ●	N	0.62
UEDIN ●⋆	Y	0.62
CU-TECTO	Y	0.60
EUROTRANS	N	0.54
CU-ZEMAN	Y	0.50
SFU	Y	0.45
ONLINEA	N	0.44
POTSDAM	Y	0.44
DCU	N	0.38
KOC	Y	0.33

Spanish-English
1448–1577 judgments per system

System	C?	≥others
ONLINEB ●	N	0.70
UEDIN ●⋆	Y	0.69
CAMBRIDGE	Y +GW	0.61
JHU	Y	0.61
ONLINEA	N	0.54
UPC ⋆	Y	0.51
HUICONG	Y	0.50
DFKI	N	0.45
COLUMBIA	Y	0.45
CU-ZEMAN	Y	0.27

English-Spanish
540–722 judgments per system

System	C?	≥others
ONLINEB ●	N	0.71
ONLINEA ●	N	0.69
UEDIN ⋆	Y	0.61
DCU	N	0.61
DFKI ⋆	N	0.55
JHU ⋆	Y	0.55
UPV ⋆	Y	0.55
CAMBRIDGE ⋆	Y +GW	0.54
UHC-UPV ⋆	Y	0.54
SFU	Y	0.40
CU-ZEMAN	Y	0.23
KOC	Y	0.19

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, standard monolingual linguistic tools, and optionally the LDC's GigaWord, which was allowed this year (entries that used the GigaWord are marked +GW).

- ● 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 5: Official results for the WMT10 translation task, based on the human evaluation (ranking translations relative to each other)

French-English
589–716 judgments per combo

System	≥others
RWTH-COMBO ●	0.77
CMU-HYP-COMBO ●	0.77
DCU-COMBO ●	0.72
LIUM ⋆	0.71
CMU-HEA-COMBO ●	0.70
UPV-COMBO ●	0.68
NRC	0.66
CAMBRIDGE	0.66
UEDIN ⋆	0.65
LIMSI ⋆	0.65
JHU-COMBO	0.65
RALI	0.65
LIUM-COMBO	0.64
BBN-COMBO	0.64
RWTH	0.55

English-French
740–829 judgments per combo

System	≥others
RWTH-COMBO ●	0.75
CMU-HEA-COMBO ●	0.74
UEDIN	0.70
KOC-COMBO ●	0.68
UPV-COMBO	0.66
RALI ⋆	0.66
LIMSI	0.66
RWTH	0.63
CAMBRIDGE	0.63

Czech-English
766–843 judgments per combo

System	≥others
CMU-HEA-COMBO ●	0.71
ONLINEB ⋆	0.7
BBN-COMBO ●	0.70
RWTH-COMBO ●	0.65
UPV-COMBO ●	0.63
JHU-COMBO	0.62
UEDIN	0.61

German-English
743–835 judgments per combo

System	≥others
BBN-COMBO ●	0.77
RWTH-COMBO ●	0.75
CMU-HEA-COMBO	0.73
KIT ⋆	0.72
UMD ⋆	0.68
JHU-COMBO	0.67
UEDIN ⋆	0.66
FBK	0.66
CMU-HYP-COMBO	0.65
UPV-COMBO	0.64
RWTH	0.62
KOC-COMBO	0.59

English-German
1340–1469 judgments per combo

System	≥others
RWTH-COMBO ●	0.65
DFKI ⋆	0.62
UEDIN ⋆	0.62
KIT ⋆	0.60
CMU-HEA-COMBO ●	0.59
KOC-COMBO	0.59
FBK ⋆	0.56
UPV-COMBO	0.55

English-Czech
1405–1496 judgments per combo

System	≥others
DCU-COMBO ●	0.75
ONLINEB ⋆	0.70
RWTH-COMBO	0.70
CMU-HEA-COMBO	0.69
UPV-COMBO	0.68
CU-BOJAR	0.66
KOC-COMBO	0.66
PC-TRANS	0.62
UEDIN	0.62

Spanish-English
1385–1535 judgments per combo

System	≥others
UEDIN ⋆	0.69
CMU-HEA-COMBO ●	0.66
UPV-COMBO ●	0.66
BBN-COMBO	0.62
JHU-COMBO	0.55
UPC	0.51

English-Spanish
516–673 judgments per combo

System	≥others
CMU-HEA-COMBO ●	0.68
KOC-COMBO	0.62
UEDIN ⋆	0.61
UPV-COMBO	0.60
RWTH-COMBO	0.59
DFKI ⋆	0.55
JHU	0.55
UPV	0.55
CAMBRIDGE ⋆	0.54
UPV-NNLM ⋆	0.54

System combinations 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. We show the best individual systems alongside the system combinations, since the goal of combination is to produce better quality translation than the component systems.

- ● indicates a **win** for the system combination meaning that no other system or system combination is statistically significantly better at p-level≤0.1 in pairwise comparison.
- ⋆ indicates an **individual system** that none of the system combinations beat by a statistically significant margin at p-level≤0.1.

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

Note: ONLINEA and ONLINEB were not included among the systems being combined in the system combination shared tasks, except in the Czech-English and English-Czech conditions, where ONLINEB was included.

Table 6: Official results for the WMT10 system combination task, based on the human evaluation (ranking translations relative to each other)

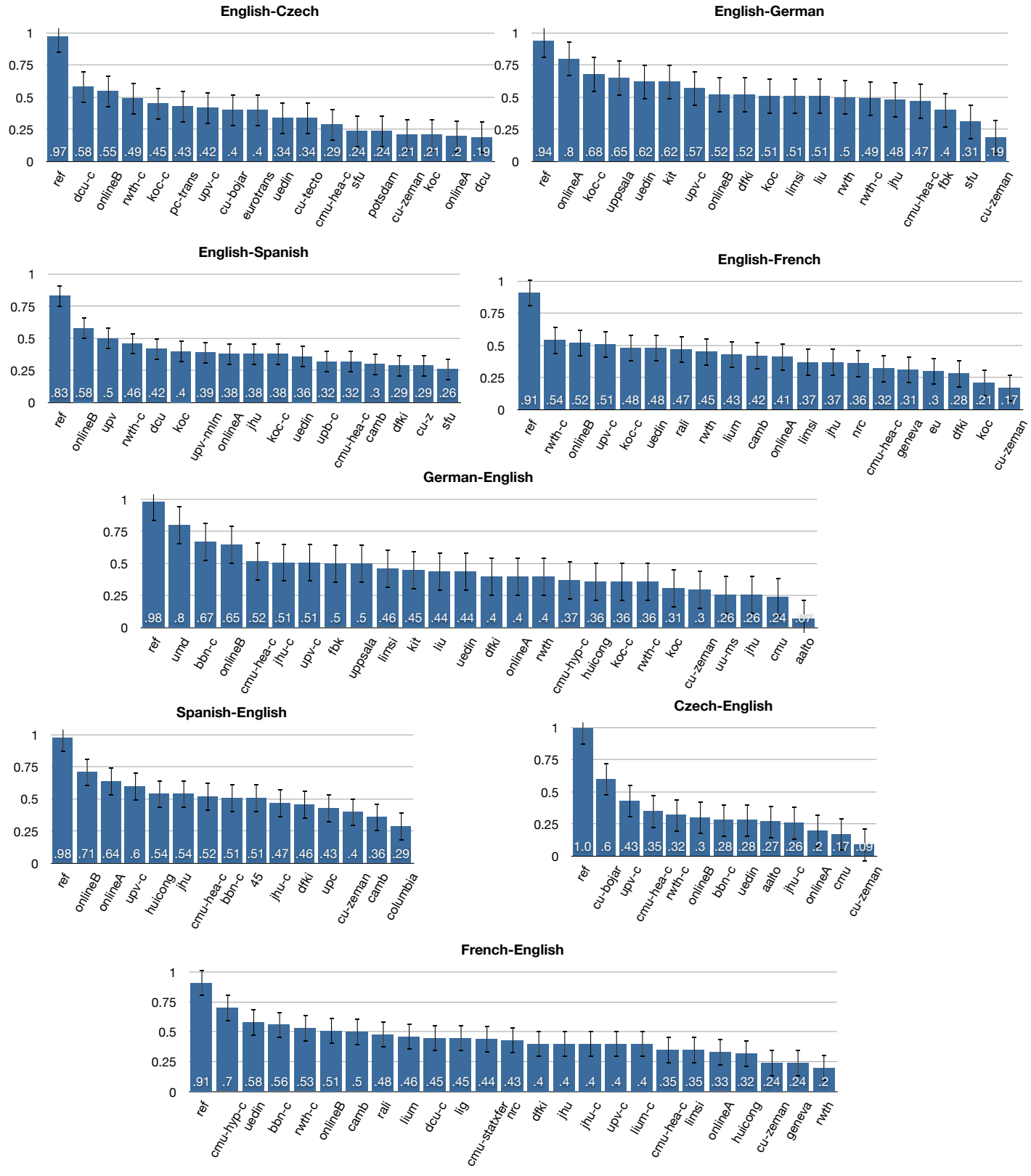

Figure 3: The percent of time that each system's edited output was judged to be an acceptable translation. These numbers also include judgments of the system's output when it was marked either *incomprehensible* or *acceptable* and left unedited. Note that the reference translation was edited alongside the system outputs. Error bars show one positive and one negative standard deviation for the systems in that language pair.

a shorter description when concatenated before compression than when concatenated after compression. MT-NCD does not require any language specific resources.

- MT-mNCD – Enhances MT-NCD with flexible word matching provided by stemming and synonyms. It works analogously to M-BLEU and M-TER and uses METEOR's aligner module to find relaxed word-to-word alignments. MT-mNCD exploits English WordNet data and increases correlation to human judgments for English over MT-NCD.

BabbleQuest International[8]

- Badger 2.0 full – Uses the Smith-Waterman alignment algorithm with Gotoh improvements to measure segment similarity. The full version uses a multilingual knowledge base to assign a substitution cost which supports normalization of word infection and similarity.

- Badger 2.0 lite – The lite version uses default gap, gap extension and substitution costs.

City University of Hong Kong (Wong and Kit, 2010)

- ATEC 2.1 – This version of ATEC extends the measurement of word choice and word order by various means. The former is assessed by matching word forms at linguistic levels, including surface form, stem, sense and semantic similarity, and further by weighting the informativeness of both matched and unmatched words. The latter is quantified in term of the discordance of word position and word sequence between an MT output and its reference.

Due to a version discrepancy of the metric, final scores for ATECD-2.1 differ from those reported here, but only minimally.

Carnegie Mellon University (Denkowski and Lavie, 2010)

- METEOR-NEXT-adq – Evaluates a machine translation hypothesis against one or more reference translations by calculating a similarity score based on an alignment between the hypothesis and reference strings. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases in the strings. Metric parameters are tuned to maximize correlation with human judgments of translation quality (adequacy judgments).

- METEOR-NEXT-hter – METEOR-NEXT tuned to HTER.

- METEOR-NEXT-rank – METEOR-NEXT tuned to human judgments of rank.

Columbia University[9]

- SEPIA – A syntactically-aware machine translation evaluation metric designed with the goal of assigning bigger weight to grammatical structural bigrams with long surface spans that cannot be captured with surface n-gram metrics. SEPIA uses a dependency representation produced for both hypothesis and reference(s). SEPIA is configurable to allow using different combinations of structural n-grams, surface n-grams, POS tags, dependency relations and lemmatization. SEPIA is a precision-based metric and as such employs clipping and length penalty to minimize metric gaming.

Charles University Prague (Bojar and Kos, 2010)

- SemPOS – Computes overlapping of autosemantic (content-bearing) word lemmas in the candidate and reference translations given a fine-grained semantic part of speech (sempos) and outputs average overlapping score over all sempos types. The overlapping is defined as the number of matched lemmas divided by the total number of lemmas in the candidate and reference translations having the same sempos type.

- SemPOS-BLEU – A linear combination of SemPOS and BLEU with equal weights. BLEU is computed on surface forms of autosemantic words that are used by SemPOS, i.e. auxiliary verbs or prepositions are not taken into account.

[8] http://www.babblequest.com/badger2

[9] http://www1.ccls.columbia.edu/~SEPIA/

Dublin City University (He et al., 2010)

- DCU-LFG – A combination of syntactic and lexical information. It measures the similarity of the hypothesis and reference in terms of matches of Lexical Functional Grammar (LFG) dependency triples. The matching module can also access the WordNet synonym dictionary and Snover's paraphrase database[10].

University of Edinburgh (Birch and Osborne, 2010)

- LRKB4 – A novel metric which directly measures reordering success using Kendall's tau permutation distance metrics. The reordering component is combined with a lexical metric, capturing the two most important elements of translation quality. This simple combined metric only has one parameter, which makes its scores easy to interpret. It is also fast to run and language-independent. It uses Kendall's tau permutation.

- LRHB4 – LRKB4, replacing Kendall's tau permutation distance metric with the Hamming distance permutation distance metric.

The scores for these two metrics used in the analysis here are those produced by the developer; due to installation issues, the metrics have not been verified to produce identical scores at NIST.

Harbin Institute of Technology, China

- I-letter-BLEU – Normal BLEU based on letters. Moreover, the maximum length of N-gram is decided by the average length for each sentence, respectively.

- I-letter-recall – A geometric mean of N-gram recall based on letters. Moreover, the maximum length of N-gram is decided by the average length for each sentence, respectively.

- SVM-RANK – Uses support vector machines rank models to predict an ordering over a set of system translations with linear kernel. Features include Meteor-exact, BLEU-cum-1, BLEU-cum-2, BLEU-cum-5, BLEU-ind-1, BLEU-ind-2, ROUGE-L recall, letter-based TER, letter-based BLEU-cum-5, letter-based ROUGE-L recall, and letter-based ROUGE-S recall.

[10]Available at `http://www.umiacs.umd.edu/~snover/terp/`.

National University of Singapore (Liu et al., 2010)

- TESLA-M – Based on matching of bags of unigrams, bigrams, and trigrams, with consideration of WordNet synonyms. The match is done in the framework of real-valued linear programming to enable the discounting of function words.

- TESLA – Built on TESLA-M, this metric also considers bilingual phrase tables to discover phrase-level synonyms. The feature weights are tuned on the development data using SVMrank.

Stanford University

- Stanford – A discriminatively trained string-edit distance metric with various similarity-matching, synonym-matching, and dependency-parse-tree-matching features. The model resembles a Conditional Random Field, but performs regression instead of classification. It is trained on Arabic, Chinese, and Urdu data from the MT-Eval 2008 dataset.

Due to installation issues, the scores included in the analysis here are those submitted by the developer; the metric has not been verified at NIST to produce identical scores on the WMT10 set.

University of Maryland[11]

- TER-plus (TERp) – An extension of the Translation Edit Rate (TER) metric that measures the number of edits between a hypothesized translation and a reference translation. TERp extends TER by using stemming, synonymy, and paraphrases as well as tunable edit costs to better measure the distance between the two translations. This version of TERp improves upon prior versions by adding brevity and length penalties.

WMT10 scores for TERp were not submitted by the developer and therefore could not be verified at NIST; NIST's installation was only verified against developer's scores on a much smaller check set. The WMT10 scores reported in the analysis here were produced using NIST's installation.

[11]`http://www.umiacs.umd.edu/~snover/terp`

University Politècnica de Catalunya/University de Barcelona (Comelles et al., 2010)

- IQmt-DR – An arithmetic mean over a set of three metrics based on discourse representations, respectively computing lexical overlap, morphosyntactic overlap, and semantic tree matching.

- IQmg-DRdoc – Is analogous to DR but, instead of operating at the segment level, it analyzes similarities over whole document discourse representations.

- IQmt-ULCh – An arithmetic mean over a heuristically-defined set of metrics operating at different linguistic levels (ROUGE, METEOR, and measures of overlap between constituent parses, dependency parses, semantic roles, and discourse representations).

The ULCh metric was not verified to produce the exact same scores at NIST as were submitted by the developer. Scores used here are those provided by the developer and differ, but only slightly, from those generated in tests at NIST.

University of Southern California, ISI

- BEwT-E – Basic Elements with Transformations for Evaluation, is a recall-oriented metric that compares basic elements, small portions of contents, between the two translations. The basic elements (BEs) consist of content words and various combinations of syntactically-related words. A variety of transformations are performed to allow flexible matching so that words and syntactic constructions conveying similar content in different manners may be matched. The transformations cover synonymy, preposition vs. noun compounding, differences in tenses, etc. BEwT-E was originally created for summarization evaluation and is English-specific.

- Bkars – Measures overlap between character trigrams in the system and reference translations. It is heavily weighted toward recall and contains a fragmentation penalty. Bkars produces a score both with and without stemming (using the Snowball package of stemmers) and averages the results together. It is not English-specific.

WMT10 scores for BEwT-E were submitted by the developer only for part of the WMT10 data set and therefore were only partially verified at NIST; the scores reported in the analysis here were produced using NIST's installation.

6 Evaluation task results

Metric developers submitted metrics for installation at NIST; they were also asked to submit metric scores on the WMT10 test set along with their metrics. Not all developers submitted scores, and not all metrics were verified to produce the same scores as submitted at NIST in time for publication. Any such caveats are reported with the description of the metrics above.

The results reported here are limited to a comparison of metric scores on the full WMT10 test set with human assessments on the human-assessed subset. An analysis comparing the human assessments with the automatic metrics run only on the human-assessed subset is planned for a later date.

The WMT10 system output used to generate the reported metric scores was found to have improperly escaped characters for a small number of segments. While we plan to regenerate the metric scores with this issue resolved, we do not expect this to significantly alter the results, given the small number of segments affected.

6.1 Metric Scores

System-, document-, and segment-level raw metric scores will be made available via the WMT10 web page.

6.2 System-Level Metric Analysis

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. The reference was not included as an extra translation.

When there are no ties, ρ can be calculated using the simplified equation:

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

	cz-en	fr-en	de-en	es-en	avg
i-letter-BLEU	**.96**	**.93**	.95	.94	**.94**
TESLA	.94	.91	.93	**.98**	**.94**
IQmt-ULCh	.93	.90	.94	.97	.93
ATEC-2.1	.95	.88	.95	.94	.93
TESLA-M	.95	.89	.92	.94	.93
BEwT-E	.91	.91	.95	.93	.92
IQmt-DR	.90	.88	**.96**	.95	.92
Bkars	.87	.88	.95	.97	.92
meteor-next-adq	.91	.88	.94	.95	.92
meteor-next-rank	.91	.88	.94	.93	.92
meteor-next-hter	.95	.86	.92	.94	.92
SemPOS-BLEU	.95	.87	.90	.94	.91
MT-mNCD	.92	.86	.94	.93	.91
MT-NCD	.90	.88	.94	.93	.91
i-letter-recall	.86	.91	.90	.97	.91
DCU-LFG	.91	.88	.91	.93	.91
SVM-rank	.86	.90	.92	.95	.91
IQmt-DRdoc	.88	.87	.92	.96	.91
1-TERp	.90	.86	.93	.93	.90
badger-2.0-full	**.96**	.85	.89	.91	.90
SEPIA	**.96**	.84	.90	.89	.90
badger-2.0-lite	.92	.86	.89	.91	.89
BLEU-4-v13a-c	.92	.84	.89	.90	.89
SemPOS	.93	.82	.83	.93	.88
NIST-c	.89	.83	.90	.87	.87
LRKB4	.73	.82	.85	.85	.81
LRHB4	.77	.80	.86	.82	.81
Stanford	.51	-.08	.47	.43	.33

Table 7: System-level Spearman's rho correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average absolute value.

	en-cz	en-fr	en-de	en-es	avg
TESLA-M	.93	.91	**.94**	.93	**.93**
SVM-rank	**.95**	.89	.91	.93	.92
i-letter-recall	.94	.88	.86	.93	.90
Bkars	.92	**.93**	.78	**.96**	.90
i-letter-BLEU	.92	.90	.85	.91	.90
MT-mNCD	.88	.91	.74	.93	.87
MT-NCD	.88	.90	.74	.92	.86
ATEC-2.1	.89	.91	.72	.90	.85
badger-2.0-lite	.81	.90	.69	.90	.83
badger-2.0-full	.81	.91	.68	.90	.83
meteor-next-rank	.86	.91	.69	.84	.82
1-TERp	.79	.89	.65	.91	.81
LRKB4	.73	.91	.68	.92	.81
NIST-c	.85	.86	.67	.84	.81
BLEU-4-v13a-c	.80	.89	.66	.87	.80
LRHB4	.77	.91	.60	.86	.78
TESLA	.46	.84	.83	.90	.76
Stanford	.56	.27	.41	.62	.46
SemPOS-BLEU	.80	n/a	n/a	n/a	n/a
SemPOS	.76	n/a	n/a	n/a	n/a

Table 8: System-level Spearman's rho correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average absolute value.

	cz-en	fr-en	de-en	es-en	avg
SVM-rank	**.40**	**.35**	**.41**	**.37**	**.38**
Bkars	.36	.34	**.41**	**.37**	.37
i-letter-recall	.37	.33	.39	.32	.36
i-letter-BLEU	.36	.33	.39	.34	.36
TESLA	.34	.34	.38	.34	.35
IQmt-ULCh	.34	.33	.34	.33	.33
Stanford	.34	.29	.37	.32	.33
ATEC-2.1	.33	.27	.37	.32	.32
meteor-next-rank	.33	.27	.36	.33	.32
NIST-c	.33	.27	.32	.31	.31
meteor-next-adq	.30	.26	.36	.31	.31
meteor-next-hter	.29	.27	.34	.31	.30
1-TERp	.29	.28	.35	.27	.30
TESLA-M	.28	.28	.34	.29	.30
SEPIA	.29	.26	.30	.31	.29
MT-NCD	.30	.24	.31	.28	.28
IQmt-DR	.23	.29	.30	.29	.28
MT-mNCD	.30	.25	.30	.27	.28
BLEU-4-v13a-c	.26	.22	.27	.28	.26
IQmt-DRdoc	.28	.23	.27	.23	.25
LRKB4	.27	.22	.25	.25	.25
SemPOS-BLEU	.19	.20	.26	.26	.23
LRHB4	.26	.19	.22	.24	.22
badger-2.0-full	.18	.16	.21	.20	.19
badger-2.0-lite	.17	.16	.21	.19	.18
DCU-LFG	.15	.14	.17	.21	.17
SemPOS	.09	.07	.13	.11	.10
BEwT-E	.05	.00	.12	.05	.05

Table 9: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average absolute value. Number of pairs included in comparison: cz-en 3575, fr-en 5844, de-en 7585, es-en 7911.

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.

The correlations are shown in Table 7 for translations to English, and Table 8 out of English, sorted by average correlation across the four language pairs. The highest correlation for each language pair and the highest overall average are bolded.

6.3 Segment-Level Metric Analysis

To assess the performance of the automatic metrics at the segment level, we correlated the metrics' segment-level scores with the human rankings using Kendall's tau rank correlation coefficient. The reference was not included as an extra

	en-cz	en-fr	en-de	en-es	avg
SVM-rank	**.32**	**.40**	**.29**	**.36**	**.34**
Bkars	.31	**.40**	.25	.34	.33
i-letter-BLEU	.29	.38	.24	.33	.31
i-letter-recall	.28	.37	.25	.32	.30
ATEC-2.1	.25	.37	.17	.29	.27
meteor-next-rank	.22	.37	.17	.31	.27
TESLA-M	.21	.30	.22	.32	.26
NIST-c	.20	.37	.17	.29	.26
TESLA	.17	.33	.20	.29	.25
MT-mNCD	.23	.30	.19	.26	.24
Stanford	.18	.34	.15	.30	.24
BLEU-4-v13a-c	.18	.33	.15	.29	.24
MT-NCD	.23	.29	.20	.23	.24
LRKB4	.14	.27	.14	.26	.20
LRHB4	.15	.25	.10	.26	.19
badger-2.0-full	.10	.28	.07	.23	.17
badger-2.0-lite	.11	.27	.07	.23	.17
1-TERp	.06	.31	.08	.23	.17
SemPOS-BLEU	.21	n/a		n/a	n/a
SemPOS	.09	n/a	n/a	n/a	n/a

Table 10: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average absolute value. Number of pairs included in comparison: en-cz 9613, en-fr 5904, en-de 10892, en-es 3813.

translation.

We calculated Kendall's tau as:

$$\tau = \frac{\text{num concordant pairs - num discordant pairs}}{\text{total pairs}}$$

where a concordant pair is a pair of two translations of the same segment in which the ranks calculated from the same human ranking task and from the corresponding metric scores agree; in a discordant pair, they disagree. In order to account for accuracy- vs. error-based metrics correctly, counts of concordant vs. discordant pairs were calculated specific to these two metric types. The possible values of τ range between 1 (where all pairs are concordant) and -1 (where all pairs are discordant). Thus an automatic evaluation metric with a higher value for τ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower τ.

We did not include cases where the human ranking was tied for two systems. As the metrics produce absolute scores, compared to five relative ranks in the human assessment, it would be potentially unfair to the metric to count a slightly different metric score as discordant with a tie in the relative human rankings. A tie in automatic metric rank for two translations was counted as discordant with two corresponding non-tied human judgments.

The correlations are shown in Table 9 for translations to English, and Table 10 out of English, sorted by average correlation across the four language pairs. The highest correlation for each language pair and the highest overall average are bolded.

7 Feasibility of Using Non-Expert Annotators in Future WMTs

In this section we analyze the data that we collected data by posting the ranking task on Amazon's Mechanical Turk (MTurk). Although we did not use this data when creating the official results, our hope was that it may be useful in future workshops in two ways. First, if we find that it is possible to obtain a sufficient amount of data of good quality, then we might be able to reduce the time commitment expected from the system developers in future evaluations. Second, the additional collected labels might enable us to detect significant differences between systems that would otherwise be insignificantly different using only the data from the volunteers (which we will now refer to as the "expert" data).

7.1 Data collection

To that end, we prepared 600 ranking sets for each of the eight language pairs, with each set containing five MT outputs to be ranked, using the same interface used by the volunteers. We posted the data to MTurk and requested, for each one, five redundant assignments, from different workers. Had all the $5 \times 8 \times 600 = 24{,}000$ assignments been completed, we would have obtained 24,000 $\times$ 5 = 120,000 additional rank labels, compared to the 37,884 labels we collected from the volunteers (Table 3). In actuality, we collected closer to 55,000 rank labels, as we discuss shortly.

To minimize the amount of data that is of poor quality, we placed two requirements that must be satisfied by any worker before completing any of our tasks. First, we required that a worker have an existing approval rating of at least 85%. Second, we required a worker to reside in a country where the target language of the task can be assumed to be the spoken language. Finally, anticipating a large pool of workers located in the United States, we felt it possible for us to add a third restriction for the *-to-English language pairs, which is that a worker must have had at least five tasks previously

INTER-ANNOTATOR AGREEMENT

	$P(A)$	K	K^*
With references	0.466	0.198	0.487
Without references	0.441	0.161	0.439

INTRA-ANNOTATOR AGREEMENT

	$P(A)$	K	K^*
With references	0.539	0.309	0.633
Without references	0.538	0.307	0.601

Table 12: Inter- and intra-annotator agreement for the MTurk workers on the sentence ranking task. (As before, $P(E)$ is 0.333.) For comparison, we repeat here the kappa coefficients of the experts (K^*), taken from Table 4.

approved on MTurk.[12] We organized the ranking sets in groups of 3 per screen, with a monetary reward of $0.05 per screen.

When we created our tasks, we had no expectation that all the assignments would be completed over the tasks' lifetime of 30 days. This was indeed the case (Table 11), especially for language pairs with a non-English target language, due to workers being in short supply outside the US. Overall, we see that the amount of data collected from non-US workers is relatively small (left half of Table 11), whereas the pool of US-based workers is much larger, leading to much higher completion rates for language pairs with English as the target language (right half of Table 11). This is in spite of the additional restriction we placed on US workers.

7.2 Quality of MTurk data

It is encouraging to see that we can collect a large amount of rank labels from MTurk. That said, we still need to guard against data from bad workers, who are either not being faithful and clicking randomly, or who might simply not be competent enough. Case in point, if we examine inter- and intra-annotator agreement on the MTurk data (Table 12), we see that the agreement rates are markedly lower than their expert counterparts.

[12]We suspect that newly registered workers on MTurk already start with an "approval rating" of 100%, and so requiring a high approval rating alone might not guard against new workers. It is not entirely clear if our suspicion is true, but our past experiences with MTurk usually involved a noticeably faster completion rate than what we experienced this time around, indicating our suspicion might very well be correct.

Another indication of the presence of bad workers is a low *reference preference rate* (RPR), which we define as the proportion of time a reference translation wins (or ties in) a comparison when it appears in one. Intuitively, the RPR should be quite high, since it is quite rare that an MT output ought to be judged better than the reference. This rate is 96.5% over the expert data, but only 83.7% over the MTurk data. Compare this to a randomly-clicking RPR of 66.67% (because the two acceptable answers are that the reference is either better than a system's output or tied with it).

Also telling would be the rate at which MTurk workers agree with experts. To ensure that we obtain enough overlapping data to calculate such a rate, we purposely select one-sixth[13] of our ranking sets so that the five-system group is exactly one that has been judged by an expert. This way, at least one-sixth of the comparisons obtained from an MTurk worker's labels are comparisons for which we already have an expert judgment. When we calculate the rate of agreement on this data, we find that MTurk workers agree with the expert workers 53.2% of the time, or K = 0.297, and when references are excluded, the agreement rate is 50.0%, or K = 0.249. Ideally, we would want those values to be in the 0.4–0.5 range, since that is where the inter-annotator kappa coefficient lies for the expert annotators.

7.3 Filtering MTurk data by agreement with experts

We can use the agreement rate with experts to identify MTurk workers who are not performing the task as required. For each worker w of the 669 workers for whom we have such data, we compute the worker's agreement rate with the experts, and from it a kappa coefficient $K_{exp}(w)$ for that worker. (Given that $P(E)$ is 0.333, $K_{exp}(w)$ ranges between -0.5 and +1.0.) We sort the workers based on $K_{exp}(w)$ in ascending order, and examine properties of the MTurk data as we remove the lowest-ranked workers one by one (Figure 4).

We first note that the amount of data we obtained from MTurk is so large, that we could afford to eliminate close to 30% of the labels, and we would still have twice as much data than using the expert data alone. We also note that two

[13]This means that on average Turkers ranked a set of system outputs that had been ranked by experts on every other screen, since each screen's worth of work had three sets.

	en-de	en-es	en-fr	en-cz	de-en	es-en	fr-en	cz-en
Location	DE	ES/MX	FR	CZ	US	US	US	US
Completed 1 time	37%	38%	29%	19%	3.5%	1.5%	14%	2.0%
Completed 2 times	18%	14%	12%	1.5%	6.0%	5.5%	19%	4.5%
Completed 3 times	2.5%	4.5%	0.5%	0.0%	8.5%	11%	20%	10%
Completed 4 times	1.5%	0.5%	0.5%	0.0%	22%	19%	23%	17%
Completed 5 times	0.0%	0.5%	0.0%	0.0%	60%	63%	22%	67%
Completed $\geq$ once	59%	57%	42%	21%	100%	99%	96%	100%
Label count	**2,583**	**2,488**	**1,578**	**627**	**12,570**	**12,870**	**9,197**	**13,169**
(% of expert data)	**(38%)**	**(96%)**	**(40%)**	**(9%)**	**(241%)**	**(228%)**	**(222%)**	**(490%)**

Table 11: Statistics for data collected on MTurk for the ranking task. In total, **55,082** rank labels were collected across the eight language pairs (**145%** of expert data). Each language pair had 600 sets, and we requested each set completed by 5 different workers. Since each set provides 5 labels, we could have potentially obtained 600 $\times$ 5 $\times$ 5 = 15,000 labels for each language pair. The **Label count** row indicates to what extent that potential was met (over the 30-day lifetime of our tasks), and the "Completed..." rows give a breakdown of redundancy. For instance, the right-most column indicates that, in the cz-en group, 2.0% of the 600 sets were completed by only one worker, while 67% of the sets were completed by 5 workers, with 100% of the sets completed at least once. The total cost of this data collection effort was roughly $200.

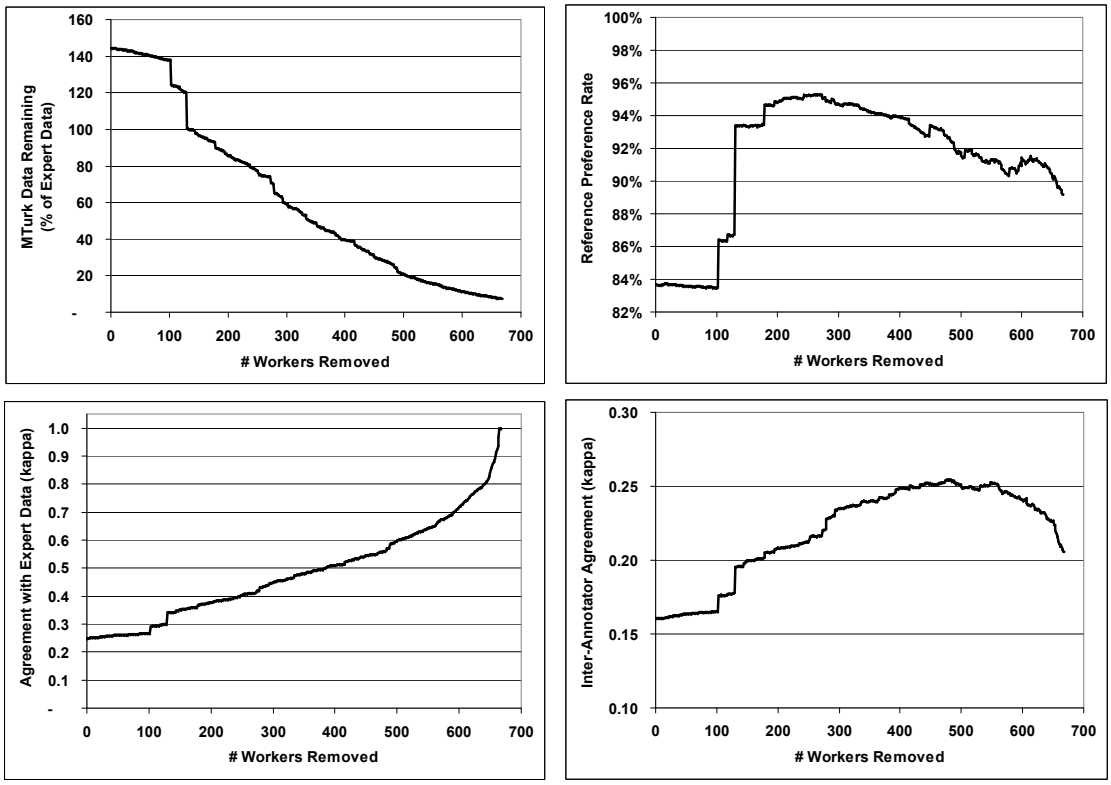

Figure 4: The effect of removing an increasing number of MTurk workers. The order in which workers are removed is by $K_{exp}(w)$, the kappa agreement coefficient with expert data (excluding references).

workers in particular (the 103rd and 130th to be removed) are likely responsible for the majority of the bad data, since removing their data leads to noticeable jumps in the reference preference rate and the inter-annotator agreement rate (right two curves of Figure 4). Indeed, examining the data for those two workers, we find that their RPR values are 55.7% and 51.9%, which is a clear indication of random clicking.[14]

Looking again at those two curves shows degrading values as we continue to remove workers in large droves, indicating a form of "overfitting" to agreement with experts (which, naturally, continues to increase until reaching 1.0; bottom left curve). It is therefore important, if one were to filter out the MTurk data by removing workers this way, to choose a cutoff carefully so that no criterion is degraded dramatically.

In Appendix A, after reporting head-to-head comparisons using only the expert data, we also report head-to-head comparisons using the expert data *combined* with the MTurk data, in order to be able to detect more significant differences between the systems. We choose the 300-worker point as a reasonable cutoff point before combining the MTurk data with the expert data, based on the characteristics of the MTurk data at that point: a high reference preference rate, high inter-annotator agreement, and, critically, a kappa coefficient vs. expert data of 0.449, which is close to the expert inter-annotator kappa coefficient of 0.439.

7.4 Feasibility of using only MTurk data

In the previous subsection, we outlined an approach by which MTurk data can be filtered out using expert data. Since we were to combine the filtered MTurk data with the expert data to obtain more significant differences, it was reasonable to use agreement with experts to quantify the MTurk workers' competency. However, we also would like to know whether it is feasible to use the MTurk data alone. Our aim here is not to boost the differences we see by examining expert data, but to eliminate our reliance on obtaining expert data in the first place.

We briefly examined some simple ways of filtering/combining the MTurk data, and measured the Spearman rank correlations obtained from the MTurk data (alone), as compared to the rankings obtained using the expert data (alone), and report them in Table 13. (These correlations do not include the references.)

We first see that even when using the MTurk data untouched, we already obtain relatively high correlation with expert ranking ("Unfiltered"). This is especially true for the *-to-English language pairs, where we collected much more data than English-to-*. In fact, the relationship between the amount of data and the correlation values is very strong, and it is reasonable to expect the correlation numbers for English-to-* to catch up had more data been collected.

We also measure rank correlations when applying some simple methods of cleaning/weighting MTurk data. The first method ("Voting") is performing a simple vote whenever redundant comparisons (i.e. from different workers) are available. The second method ("K_{exp}-filtered") first removes labels from the 300 worst workers according to agreement with experts. The third method ("RPR-filtered") first removes labels from the 62 worst workers according to their RPR. The numbers 300 and 62 were chosen since those are the points at which the MTurk data reaches the level of expert data in the inter-annotator agreement and RPR of the experts.

The fourth and fifth methods ("Weighted by K_{exp}" and "Weighted by $K(RPR)$") do not remove any data, instead assigning weights to workers based on their agreement with experts and their RPR, respectively. Namely, for each worker, the weight assigned by the fourth method is K_{exp} for that worker, and the weight assigned by the fifth method is $K(RPR)$ for that worker.

Examining the correlation coefficients obtained from those methods (Table 13), we see mixed results, and there is no clear winner among those methods. It is also difficult to draw any conclusion as to which method performs best when. However, it is encouraging to see that the two RPR-based methods perform well. This is noteworthy, since there is no need to use expert data to weight workers, which means that it is possible to evaluate a worker using inherent, 'built-in' properties of that worker's own data, without resorting to making comparisons with other workers or with experts.

[14] In retrospect, we should have performed this type of analysis as the data was being collected, since such workers could have been identified early on and blocked.

	Label count	Unfiltered	Voting	K_{exp}-filtered	RPR-filtered	Weighted by K_{exp}	Weighted by $K(RPR)$
en-de	2,583	0.86	0.78	0.82	0.86	**0.87**	0.86
en-es	2,488	0.76	0.79	0.80	0.80	0.77	**0.81**
en-fr	1,578	0.83	**0.84**	0.79	0.81	0.80	0.81
en-cz	627	0.83	0.82	0.35	0.83	**0.85**	0.83
de-en	12,570	0.91	0.93	0.92	0.93	**0.93**	0.93
es-en	12,870	0.93	0.97	0.97	**0.99**	0.98	**0.99**
fr-en	9,197	0.88	0.87	**0.92**	0.92	0.91	0.92
cz-en	13,169	0.95	0.91	**0.97**	0.94	0.93	0.94

Table 13: Spearman rank coefficients for the MTurk data across the various language pairs, using different methods to clean the data or weight workers. (These correlations were computed after excluding the references.) K_{exp} is the kappa coefficient of the worker's agreement rate with experts, with $P(A) = 0.33$. $K(RPR)$ is the kappa coefficient of the worker's RPR (see 7.2), with $P(A) = 0.66$. In K_{exp}-filtering, 42% of labels remain, after removing 300 workers. In $K(RPR)$-filtering, 69% of labels remain, after removing 62 workers.

8 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, and vice versa.

The number of participants grew substantially compared to previous editions of the WMT workshop, with 33 groups from 29 institutions participating in WMT10. Most groups participated in the translation task only, while the system combination task attracted a somewhat smaller number of participants.

Unfortunately, fewer rule-based systems participated in this year's edition of WMT, compared to previous editions. We hope to attract more rule-based systems in future editions as they increase the variation of translation output and for some language pairs, such as German-English, tend to outperform statistical machine translation systems.

This was the first time that the WMT workshop was held as a joint workshop with NIST's MetricsMATR evaluation initiative. This joint effort was very productive as it allowed us to focus more on the two evaluation dimensions: manual evaluation of MT performance and the correlation between manual metrics and automated metrics.

This year was also the first time we have introduced quality assessments by non-experts. In previous years all assessments were carried out through peer evaluation exclusively consisting of developers of machine translation systems, and thereby people who are used to machine translation output. This year we have facilitated Amazon's Mechanical Turk to investigate two aspects of manual evaluation: How stable are manual assessments across different assessor profiles (experts vs. non-experts) and how reliable are quality judgments of non-expert users? While the intra- and inter-annotator agreements between non-expert assessors are considerably lower than for their expert counterparts, the overall rankings of translation systems exhibit a high degree of correlation between experts and non-experts. This correlation can be further increased by applying various filtering strategies reducing the impact of unreliable non-expert annotators.

As in previous years, all data sets generated by this workshop, including the human judgments, system translations and automatic scores, are publicly available for other researchers to analyze.[15]

Acknowledgments

This work was supported in parts by the EuroMatrixPlus project funded by the European Commission (7th 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.

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

References

Alexandre Allauzen, Josep M. Crego, lknur Durgar El-Kahlout, and Francois Yvon. 2010. Limsi's statistical translation systems for wmt'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 29–34, Uppsala, Sweden, July. Association for Computational Linguistics.

Loïc Barrault. 2010. Many: Open source mt system combination at wmt'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–256, Uppsala, Sweden, July. Association for Computational Linguistics.

Ergun Bicici and S. Serdar Kozat. 2010. Adaptive model weighting and transductive regression for predicting best system combinations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 257–262, Uppsala, Sweden, July. Association for Computational Linguistics.

Ergun Bicici and Deniz Yuret. 2010. L1 regularized regression for reranking and system combination in machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 263–270, Uppsala, Sweden, July. Association for Computational Linguistics.

Alexandra Birch and Miles Osborne. 2010. Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 302–307, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondrej Bojar and Kamil Kos. 2010. 2010 failures in english-czech phrase-based mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 35–41, Uppsala, Sweden, July. Association for Computational Linguistics.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazons mechanical turk. In *Proceedings NAACL-2010 Workshop on Creating Speech and Language Data With Amazons Mechanical Turk*, Los Angeles.

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.

Chris Callison-Burch, , Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*, Athens, Greece.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, Singapore.

Elisabet Comelles, Jesus Gimenez, Lluis Marquez, Irene Castellon, and Victoria Arranz. 2010. Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 308–313, Uppsala, Sweden, July. Association for Computational Linguistics.

Hui Cong, Zhao Hai, Lu Bao-Liang, and Song Yan. 2010. An empirical study on development set selection strategy for machine translation learning. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 42–46, Uppsala, Sweden, July. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2010. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 314–317, Uppsala, Sweden, July. Association for Computational Linguistics.

Marcus Dobrinkat, Tero Tapiovaara, Jaakko Väyrynen, and Kimmo Kettunen. 2010. Normalized compression distance based measures for metricsmatr 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 318–323, Uppsala, Sweden, July. Association for Computational Linguistics.

Jinhua Du, Pavel Pecina, and Andy Way. 2010. An augmented three-pass system combination framework: Dcu combination system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 271–276, Uppsala, Sweden, July. Association for Computational Linguistics.

Vladimir Eidelman, Chris Dyer, and Philip Resnik. 2010. The university of maryland statistical machine translation system for the fifth workshop on machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 47–51, Uppsala, Sweden, July. Association for Computational Linguistics.

Christian Federmann, Andreas Eisele, Yu Chen, Sabine Hunsicker, Jia Xu, and Hans Uszkoreit. 2010. Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on*

Statistical Machine Translation and MetricsMATR, pages 52–56, Uppsala, Sweden, July. Association for Computational Linguistics.

Jesús González-Rubio, Germán Sanchis-Trilles, Joan-Andreu Sánchez, Jesús Andrés-Ferrer, Guillem Gascó, Pascual Martínez-Gómez, Martha-Alicia Rocha, and Francisco Casacuberta. 2010. The upv-prhlt combination system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 277–281, Uppsala, Sweden, July. Association for Computational Linguistics.

Greg Hanneman, Jonathan Clark, and Alon Lavie. 2010. Improved features and grammar selection for syntax-based mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

Christian Hardmeier, Arianna Bisazza, and Marcello Federico. 2010. Fbk at wmt 2010: Word lattices for morphological reduction and chunk-based reordering. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 63–67, Uppsala, Sweden, July. Association for Computational Linguistics.

Yifan He, Jinhua Du, Andy Way, and Josef van Genabith. 2010. The dcu dependency-based metric in wmt-metricsmatr 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 324–328, Uppsala, Sweden, July. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2010. Cmu multi-engine machine translation for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 68–73, Uppsala, Sweden, July. Association for Computational Linguistics.

Carmen Heger, Joern Wuebker, Matthias Huck, Gregor Leusch, Saab Mansour, Daniel Stein, and Hermann Ney. 2010. The rwth aachen machine translation system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 74–78, Uppsala, Sweden, July. Association for Computational Linguistics.

Carlos A. Henríquez Q., Marta Ruiz Costa-jussà, Vidas Daudaravicius, Rafael E. Banchs, and José B. Mariño. 2010. Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 79–83, Uppsala, Sweden, July. Association for Computational Linguistics.

Almut Silja Hildebrand and Stephan Vogel. 2010. Cmu system combination via hypothesis selection for wmt'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 282–285, Uppsala, Sweden, July. Association for Computational Linguistics.

Stéphane Huet, Julien Bourdaillet, Alexandre Patry, and Philippe Langlais. 2010. The rali machine translation system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 84–90, Uppsala, Sweden, July. Association for Computational Linguistics.

Michael Jellinghaus, Alexandros Poulis, and David Kolovratník. 2010. Exodus - exploring smt for eu institutions. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 91–95, Uppsala, Sweden, July. 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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, Prague, Czech Republic.

Philipp Koehn, Barry Haddow, Philip Williams, and Hieu Hoang. 2010. More linguistic annotation for statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 96–101, Uppsala, Sweden, July. Association for Computational Linguistics.

Patrik Lambert, Sadaf Abdul-Rauf, and Holger Schwenk. 2010. Lium smt machine translation system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 102–107, Uppsala, Sweden, July. Association for Computational Linguistics.

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

Samuel Larkin, Boxing Chen, George Foster, Ulrich Germann, Eric Joanis, Howard Johnson, and Roland Kuhn. 2010. Lessons from nrcs portage system at wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 108–113, Uppsala, Sweden, July. Association for Computational Linguistics.

Gregor Leusch and Hermann Ney. 2010. The rwth system combination system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical*

Machine Translation and MetricsMATR, pages 290–295, Uppsala, Sweden, July. 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.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Ziyuan Wang, Jonathan Weese, and Omar Zaidan. 2010. Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 114–118, Uppsala, Sweden, July. Association for Computational Linguistics.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Tesla: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 329–334, Uppsala, Sweden, July. Association for Computational Linguistics.

Sushant Narsale. 2010. Jhu system combination scheme for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 286–289, Uppsala, Sweden, July. Association for Computational Linguistics.

Jan Niehues, Teresa Herrmann, Mohammed Mediani, and Alex Waibel. 2010. The karlsruhe institute for technology translation system for the acl-wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 119–123, Uppsala, Sweden, July. Association for Computational Linguistics.

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

Sergio Penkale, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, and Andy Way. 2010. Matrex: The dcu mt system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 124–129, Uppsala, Sweden, July. Association for Computational Linguistics.

Aaron Phillips. 2010. The cunei machine translation platform for wmt '10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 130–135, Uppsala, Sweden, July. Association for Computational Linguistics.

Juan Pino, Gonzalo Iglesias, Adrià de Gispert, Graeme Blackwood, Jamie Brunning, and William Byrne. 2010. The cued hifst system for the wmt10 translation shared task. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 136–141, Uppsala, Sweden, July. Association for Computational Linguistics.

Marion Potet, Laurent Besacier, and Hervé Blanchon. 2010. The lig machine translation system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 142–147, Uppsala, Sweden, July. Association for Computational Linguistics.

Mark Przybocki, Kay Peterson, and Sebastian Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" challenge (MetricsMATR08). In *AMTA-2008 workshop on Metrics for Machine Translation*, Honolulu, Hawaii.

Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. Bbn system description for wmt10 system combination task. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 296–301, Uppsala, Sweden, July. Association for Computational Linguistics.

Markus Saers, Joakim Nivre, and Dekai Wu. 2010. Linear inversion transduction grammar alignments as a second translation path. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 148–152, Uppsala, Sweden, July. Association for Computational Linguistics.

Germán Sanchis-Trilles, Jesús Andrés-Ferrer, Guillem Gascó, Jesús González-Rubio, Pascual Martínez-Gómez, Martha-Alicia Rocha, Joan-Andreu Sánchez, and Francisco Casacuberta. 2010. Upv-prhlt english–spanish system for wmt10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 153–157, Uppsala, Sweden, July. Association for Computational Linguistics.

Baskaran Sankaran, Ajeet Grewal, and Anoop Sarkar. 2010. Incremental decoding for phrase-based statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 197–204, Uppsala, Sweden, July. Association for Computational Linguistics.

Lane Schwartz. 2010. Reproducible results in parsing-based machine translation: The jhu shared task submission. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 158–163, Uppsala, Sweden, July. Association for Computational Linguistics.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2010. Vs and oovs: Two problems for translation between german and english. In *Proceedings of the*

Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 164–169, Uppsala, Sweden, July. Association for Computational Linguistics.

Jörg Tiedemann. 2010. To cache or not to cache? experiments with adaptive models in statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 170–175, Uppsala, Sweden, July. Association for Computational Linguistics.

Sami Virpioja, Jaakko Väyrynen, Andre Mansikkaniemi, and Mikko Kurimo. 2010. Applying morphological decompositions to statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 176–181, Uppsala, Sweden, July. Association for Computational Linguistics.

Zdeněk Žabokrtský, Martin Popel, and David Mareček. 2010. Maximum entropy translation model in dependency-based mt framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 182–187, Uppsala, Sweden, July. Association for Computational Linguistics.

Billy Wong and Chunyu Kit. 2010. The parameter-optimized atec metric for mt evaluation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 335–339, Uppsala, Sweden, July. Association for Computational Linguistics.

Francisco Zamora-Martinez and Germán Sanchis-Trilles. 2010. Uch-upv english–spanish system for wmt10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 188–192, Uppsala, Sweden, July. Association for Computational Linguistics.

Daniel Zeman. 2010. Hierarchical phrase-based mt at the charles university for the wmt 2010 shared task. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 193–196, Uppsala, Sweden, July. Association for Computational Linguistics.

A Pairwise system comparisons by human judges

Tables 14–21 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 Pairwise system comparisons for combined expert and non-expert data

Tables 22–21 show pairwise comparisons between systems for the into English direction when non-expert judgments have been added.

The number of pairwise comparisons at the $\star$ level of significance increases from 48 to 50, and the number at the † level of significants increases from 79 to 80 (basically same number). However, the ‡ level of significance went up considerably, from 280 to 369. That's a 31% increase. 75 of ‡ are comparisons involving the reference, then the non-reference ‡ count went up from 205 to 294, a 43% increase.

	REF	CAMBRIDGE	CMU-STATXFER	CU-ZEMAN	DFKI	GENEVA	HUICONG	JHU	LIG	LIMSI	LIUM	NRC	ONLINEA	ONLINEB	RALI	RWTH	UEDIN	BBN-COMBO	CMU-HEAFIELD-COMBO	CMU-HYPOSEL-COMBO	DCU-COMBO	JHU-COMBO	LIUM-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.00‡	.00‡	.00‡	.00‡	.00‡	.04‡	.03‡	.00‡	.00‡	.00‡	.00‡	.04‡	.00‡	.04‡	.00‡	.00‡	.00‡	.00‡	.05‡	.06‡	.03‡	.09‡	.04‡	.04‡
CAMBRIDGE	**.79**‡	–	.36	.16‡	.12‡	.23†	.27	**.43**	.26†	.38	.24	.3	.28	**.51**	.34	.23	.37	.24	.32	**.46**	.24	.29	**.45**	**.59**⋆	**.44**
CMU-STATXFER	**.84**‡	**.58**	–	.16‡	**.48**	.14‡	.19	.39	.33	**.54**	**.54**⋆	**.50**†	.36	**.50**	**.70**‡	**.55**⋆	**.50**	.46	**.58**†	**.67**†	**.50**	**.56**†	**.48**	**.58**‡	**.52**†
CU-ZEMAN	**1.00**‡	**.77**‡	**.72**‡	–	**.76**‡	.37	**.73**‡	**.74**‡	**.79**‡	**.77**‡	**.77**‡	**.81**‡	**.75**‡	**.94**‡	**.86**‡	**.77**‡	**.89**‡	**.67**	**.77**‡	**.79**‡	**.81**‡	**.81**‡	**.77**‡	**.96**‡	**.86**‡
DFKI	**1.00**‡	**.72**‡	.45	.12‡	–	.32	**.48**	**.50**	**.52**	**.53**	**.56**	**.65**	**.53**	**.62**	**.55**	.43	**.61**⋆	**.50**	**.68**†	**.73**‡	**.70**†	**.60**	**.59**⋆	**.72**‡	**.71**‡
GENEVA	**1.00**‡	**.69**†	**.76**‡	**.48**	**.56**	–	**.47**	**.71**†	**.79**‡	**.72**†	**.79**‡	**.71**†	**.68**†	**.76**‡	**.83**‡	**.57**	**.86**‡	**.72**‡	**.71**†	**.69**†	**.76**†	**.65**‡	**.88**‡	**.96**‡	**.70**
HUICONG	**.86**‡	**.54**	**.29**	.12‡	.26	.37	–	**.48**	.31	**.43**	**.63**‡	**.62**†	**.53**	**.55**	**.53**‡	.44	**.50**	**.55**	**.52**	**.68**‡	**.52**⋆	**.51**	**.52**⋆	**.57**	**.53**
JHU	**.83**‡	.39	**.42**	.13‡	.33	.19†	.3	–	.3	.36	**.56**†	**.56**⋆	**.47**	**.52**	**.46**	.29	**.36**	**.42**	**.42**	**.59**†	**.50**	.31	.43	**.29**	**.37**
LIG	**.97**‡	**.63**†	**.36**	.15‡	.37	.18‡	**.40**	**.60**	–	**.62**⋆	**.57**‡	**.39**	.35	**.54**†	**.46**	.33	.34	.38	**.54**†	**.48**⋆	**.42**	**.44**	**.50**	**.61**⋆	**.56**
LIMSI	**.96**‡	**.41**	.23	.19‡	.31	.17†	.32	**.50**	.28⋆	–	**.35**	**.42**	.21	**.62**‡	.25	.21	.33	.22	**.42**	.35	**.43**	.32	.26	**.35**	**.41**
LIUM	**.83**‡	**.33**	.21⋆	.13‡	.41	.05‡	.13‡	.15†	.09‡	.3	–	.39	.19	.36	**.43**	.26	.23†	.28	.29	**.45**	.28	.26	.28	.33	.28
NRC	**.96**‡	.3	.10†	.10‡	.32	.24†	.15†	.22⋆	.22	.33	**.43**	–	.26	**.58**	.26	.24	.3	**.50**	**.36**	**.45**	**.47**†	.23	**.38**	**.36**†	**.35**
ONLINEA	**.96**‡	**.55**	**.57**	.14‡	.42	.16†	.42	.4	**.39**	**.53**	**.52**	**.47**	–	**.52**⋆	**.46**	.36	**.64**	**.57**	**.59**	**.50**	**.59**	**.42**	**.46**	.43	**.48**
ONLINEB	**.87**‡	.37	.33	.03‡	.29	.12‡	.31	.26	.16†	.12‡	**.39**	.35	.20⋆	–	.33	.38	.17†	.36	.29	.21	.33	.3	.3	.32	.21‡
RALI	**.89**‡	**.45**	.15‡	.06‡	.35	.04‡	.12‡	.42	.35	**.46**	.32	**.42**	.39	**.52**	–	.32	.31	.26	**.43**	**.41**	.27	**.43**	**.40**	**.63**⋆	.26
RWTH	**.91**‡	**.46**	.21⋆	.05‡	**.51**	.36	.44	**.46**	**.53**	**.39**	**.48**	**.48**	**.39**	**.48**	**.48**	–	**.39**	**.38**	**.39**	**.52**	**.46**	**.53**†	**.52**	**.50**‡	.25
UEDIN	**.96**‡	**.40**	.33	.03‡	.28⋆	.03‡	.28	.29	**.49**	**.38**	**.61**†	.3	.32	**.50**†	**.34**	.24	–	**.42**	**.33**	**.43**	**.48**	.18⋆	.13	**.27**	**.38**
BBN-C	**.90**‡	**.48**	.46	.29	.39	.22‡	.27	.27	**.46**	**.43**	.28	.35	.33	**.39**	**.29**	.34	.26	–	**.28**	**.44**†	**.33**	**.26**	**.62**⋆	**.36**	.28
CMU-HEA-C	**.89**‡	**.50**	.23†	.14‡	.30†	.21†	.26	.25	.17†	.33	**.43**	.16	.36	**.43**	.26	.29	.24	.24	–	**.48**	**.27**	.13	.25	**.30**	.15
CMU-HYP-C	**.81**‡	.17	.19†	.11‡	.19‡	.19†	.14‡	.14†	.19⋆	**.40**	.23	.18	.29	**.46**	.35	.29	.21	.15†	.17	–	.26	.18	.07‡	**.32**	.21
DCU-C	**.88**‡	**.27**	.25	.11‡	.22†	.24†	.20⋆	.28	.21	.35	**.50**	.10†	.31	**.44**	.27	.29	.22	.21	.2	**.30**	–	.12⋆	.26	**.26**	.08
JHU-C	**.86**‡	**.48**	.16†	.16‡	.33	.21‡	.35	**.41**	.32	**.44**	**.39**	**.35**	.39	**.37**	.26	.19†	**.50**⋆	.23	**.32**	**.43**	**.40**⋆	–	.36	**.27**	**.39**
LIUM-C	**.87**‡	.41	.36	.13‡	.31⋆	.08‡	.21⋆	**.48**	.31	**.47**	**.44**	.24	.39	**.52**	.28	.28	**.33**	.27⋆	.25	**.67**‡	.26	**.44**	–	**.54**‡	**.48**
RWTH-C	**.88**‡	.18⋆	.13‡	.04‡	.22‡	.04‡	.14	.24	.25⋆	.3	.33	.05†	.43	**.50**	.30⋆	.13‡	.23	.14	.18	.21	.19	.23	.11‡	–	.24
UPV-C	**.92**‡	.25	.12†	.10‡	.16‡	.3	.25	.34	.29	.31	**.34**	.29	.39	**.65**‡	**.39**	**.36**	.3	**.45**	**.27**	**.36**	**.23**	.16	.24	**.28**	–
> others	.90	.44	.31	.13	.33	.18	.29	.37	.34	.42	.44	.38	.37	**.51**	.41	.31	.38	.35	.38	**.48**	.39	.36	.40	.46	.37
>= others	.98	.66	.51	.21	.42	.27	.51	.59	.53	.65	**.71**	.66	.52	**.71**	.65	.55	.65	.64	.70	**.77**	.72	.65	.64	**.77**	.68

Table 14: Sentence-level ranking for the WMT10 French-English News Task

	REF	CAMBRIDGE	CU-ZEMAN	DFKI	EU	GENEVA	JHU	KOC	LIMSI	LIUM	NRC	ONLINEA	ONLINEB	RALI	RWTH	UEDIN	CMU-HEAFIELD-COMBO	KOC-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.08‡	.02‡	.00‡	.04‡	.08‡	.13‡	.06‡	.09‡	.09‡	.07‡	.16‡	.11‡	.12‡	.12‡	.12‡	.05‡	.07‡	.08‡	.09‡
CAMBRIDGE	**.82**‡	–	.16‡	.24†	.15‡	.07‡	.35	.10‡	**.42**	.36	.43	.27	**.67**‡	**.46**	**.39**	**.44**	**.40**	**.46**	**.48**⋆	**.40**
CU-ZEMAN	**.98**‡	**.82**‡	–	**.47**	**.54**⋆	**.62**‡	**.71**‡	**.41**	**.79**‡	**.82**‡	**.70**‡	**.67**‡	**.85**‡	**.90**‡	**.75**‡	**.72**‡	**.92**‡	**.82**‡	**.88**‡	**.82**‡
DFKI	**.95**‡	**.66**†	.31	–	**.46**	.25⋆	**.78**‡	.36	**.59**	**.62**⋆	**.75**‡	**.65**†	**.45**	**.56**⋆	**.75**‡	**.69**‡	**.71**‡	**.63**⋆	**.57**	**.65**†
EU	**.96**‡	**.78**‡	.30⋆	.41	–	**.55**	**.68**‡	.16‡	**.76**‡	**.72**‡	**.82**‡	**.67**‡	**.63**‡	**.86**‡	**.78**‡	**.78**‡	**.76**‡	**.76**‡	**.75**‡	**.71**‡
GENEVA	**.86**‡	**.81**‡	.23‡	**.55**⋆	.34	–	**.65**‡	.25‡	**.65**†	**.70**‡	**.69**‡	**.66**‡	**.77**‡	**.71**‡	**.70**‡	**.89**‡	**.75**‡	**.63**†	**.84**‡	**.75**‡
JHU	**.77**‡	**.42**	.15‡	.22‡	.22‡	.22‡	–	.06‡	**.58**⋆	**.47**	**.52**†	**.49**	**.70**‡	**.61**†	**.53**	**.64**‡	**.53**⋆	**.65**‡	**.68**‡	**.50**
KOC	**.85**‡	**.67**‡	.4	**.58**	**.55**‡	**.69**‡	**.82**‡	–	**.76**‡	**.85**‡	**.81**‡	**.72**‡	**.86**‡	**.82**‡	**.86**‡	**.85**‡	**.77**‡	**.77**‡	**.74**‡	**.79**‡
LIMSI	**.84**‡	.23	.08‡	.29	.09‡	.30†	.21⋆	.08‡	–	.33	**.37**	.17‡	**.51**	**.40**	.29	**.45**	**.49**	**.40**	**.61**‡	.28
LIUM	**.85**‡	**.39**	.07‡	.32⋆	.11‡	.21‡	.44	.07‡	**.46**	–	.44	.4	.32	.44	.37	**.64**†	.35	**.40**	**.35**	**.42**
NRC	**.91**‡	.43	.15‡	.20‡	.11‡	.25‡	.21†	.09‡	.31	**.45**	–	.32	**.48**	.44	**.49**	**.61**†	**.52**†	**.30**	**.58**⋆	**.40**
ONLINEA	**.80**‡	**.51**	.21‡	.33†	.23‡	.15‡	.41	.14‡	**.60**‡	**.42**	**.54**	–	**.52**⋆	**.56**⋆	.36	**.67**‡	**.61**‡	**.45**	**.50**	**.44**
ONLINEB	**.87**‡	.23‡	.08‡	.43	.23‡	.11‡	.12‡	.08‡	.27	**.36**	.43	.25⋆	–	.38	.31	.33	**.52**	.33⋆	**.46**	.29
RALI	**.83**‡	.38	.05‡	.27⋆	.11‡	.15‡	.22†	.10‡	.36	.44	**.49**	.31⋆	**.50**	–	.38	**.44**	**.42**	.37	**.38**	**.34**
RWTH	**.76**‡	.33	.11‡	.12‡	.15‡	.17‡	.34	.05‡	**.34**	**.44**	.29	**.42**	**.49**	**.40**	–	**.56**	**.48**	**.44**	**.53**‡	**.50**
UEDIN	**.84**‡	.29	.20‡	.17‡	.12‡	.09‡	.19‡	.07‡	.33	.23†	.24†	.24‡	**.56**	.31	.3	–	**.36**⋆	.27	**.51**	.18†
CMU-HEAFIELD-COMBO	**.90**‡	.23	.04‡	.23‡	.18‡	.12‡	.22⋆	.11‡	.32	**.41**	.20†	.23‡	.28	.31	.31	.11⋆	–	.29	.24	.3
KOC-COMBO	**.91**‡	.26	.08‡	.31⋆	.17‡	.28†	.20‡	.07‡	.23	.26	.19	.36	**.57**⋆	.37	.32	**.32**	**.42**	–	**.38**	**.34**
RWTH-COMBO	**.85**‡	.21⋆	.02‡	.36	.16‡	.07‡	.12‡	.07‡	.16‡	.3	.30⋆	.4	.34	.32	.06‡	.26	**.35**	.16	–	.21⋆
UPV-COMBO	**.87**‡	.38	.08‡	.30†	.19‡	.19‡	.37	.11‡	**.39**	.24	.33	.37	**.44**	.27	.34	**.46**†	**.35**	.28	**.50**⋆	–
> others	.87	.43	.15	.30	.22	.25	.38	.13	.44	.45	.46	.41	**.53**	.49	.44	.52	**.53**	.45	**.53**	.45
>= others	.92	.63	.26	.40	.32	.35	.53	.26	.66	.63	.62	.55	**.68**	.66	.63	.70	.74	.68	**.75**	.66

Table 15: Sentence-level ranking for the WMT10 English-French News Task

	REF	AALTO	CMU	CU-ZEMAN	DFKI	FBK	HUICONG	JHU	KIT	KOC	LIMSI	LIU	ONLINEA	ONLINEB	RWTH	UEDIN	UMD	UPPSALA	UU-MS	BBN-COMBO	CMU-HEAFIELD-COMBO	CMU-HYPOSEL-COMBO	JHU-COMBO	KOC-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.00‡	.03‡	.00‡	.06‡	.03‡	.00‡	.00‡	.05‡	.00‡	.00‡	.03‡	.06‡	.09‡	.06‡	.00‡	.09‡	.03‡	.03‡	.14‡	.03‡	.06‡	.03‡	.03‡	.06‡	.00‡
AALTO	**1.00**‡	–	**.50**	.31	**.60**	**.69**‡	**.39**	**.41**	**.71**†	.31	**.45**	**.60**‡	**.59**†	**.65**‡	**.66**‡	**.64**‡	**.81**‡	**.45**	**.41**	**.69**†	**.72**‡	**.75**†	**.55**	**.55**‡	**.76**‡	**.57**†
CMU	**.93**‡	.31	–	.29	**.49**	**.57**‡	**.38**	**.50**	**.74**‡	.13‡	**.44**	**.59**‡	**.57**†	**.59**⋆	**.60**†	**.67**†	**.59**‡	**.41**	**.50**	**.68**‡	**.67**‡	**.46**	**.64**‡	**.55**⋆	**.67**‡	**.54**⋆
CU-ZEMAN	**1.00**‡	**.44**	**.56**	–	**.58**	**.64**‡	.17	**.44**	**.75**‡	.38	**.50**	**.54**†	**.76**†	**.79**‡	**.73**‡	**.72**‡	**.72**‡	**.50**⋆	**.73**‡	**.78**‡	**.80**‡	**.68**‡	**.72**†	**.62**†	**.68**⋆	**.73**‡
DFKI	**.92**‡	.25	.32	.27	–	**.53**	.36	**.46**	**.65**⋆	.07‡	**.50**	**.47**	**.47**	**.69**‡	**.56**	.35	**.55**	**.58**	**.47**	**.67**†	**.61**⋆	**.52**	**.47**	.38	**.67**†	**.51**
FBK	**.97**‡	.20‡	.16‡	.14‡	.38	–	.11‡	.31	**.45**	.10‡	.22⋆	**.36**	**.50**	**.57**†	.37	**.43**	**.40**	.12‡	.17†	**.48**⋆	**.43**	**.35**	**.38**	.22	**.38**	**.39**
HUICONG	**.93**‡	.35	.28	**.46**	**.43**	**.75**‡	–	**.52**	**.69**†	.16†	**.39**	.42	**.64**†	**.79**‡	.31	**.51**†	**.78**‡	**.27**	**.41**	**.49**	**.74**‡	**.68**‡	**.60**⋆	**.37**	**.68**‡	**.56**†
JHU	**.86**‡	.34	.29	.16	.43	.31	.26	–	**.61**‡	.15‡	.35	.36	**.45**	**.69**‡	**.52**⋆	**.56**⋆	**.64**†	.27	**.36**	**.70**‡	**.53**	**.47**	**.66**‡	**.52**	**.68**‡	**.44**
KIT	**.89**‡	.21†	.10‡	.14‡	.29⋆	.33	.19†	.14‡	–	.03‡	.27	.21†	.36	**.46**	.17‡	.29	.24	.25‡	.25‡	**.48**	.23⋆	.31	.38	.2	**.36**	.12‡
KOC	**.96**‡	**.58**	**.77**‡	**.48**	**.70**‡	**.77**‡	**.58**†	**.71**‡	**.97**‡	–	**.77**‡	**.90**‡	**.72**‡	**.82**‡	**.76**‡	**.84**‡	**.81**‡	**.84**‡	**.66**‡	**.83**‡	**.87**‡	**.79**‡	**.77**‡	**.75**‡	**.93**‡	**.71**‡
LIMSI	**1.00**‡	.23	.28	.35	.35	**.53**⋆	.33	**.45**	**.41**	.19‡	–	**.49**	**.48**	**.63**†	**.49**	**.63**‡	**.52**	.36	.29	**.73**‡	**.53**⋆	.45	**.59**‡	.29	**.56**†	**.59**†
LIU	**.88**‡	.12‡	.15‡	.16†	.39	.21	**.46**	.36	**.61**†	.00‡	.27	–	**.44**	**.63**†	**.49**	**.45**	**.53**	.27⋆	.33	**.67**‡	**.55**⋆	**.46**	**.44**	.32	**.37**	**.55**
ONLINEA	**.92**‡	.15†	.23†	.24†	.42	.34	.21†	.35	**.50**	.10‡	.32	.36	–	**.41**	.4	.44	.37	.32	.34	**.36**	.4	**.47**	.3	.26	**.48**	.41
ONLINEB	**.68**‡	.18‡	.29⋆	.17‡	.26‡	.24†	.18‡	.23‡	.33	.18‡	.23†	.27†	.34	–	.3	.15‡	.29	.24†	.15‡	**.44**	.28	.33⋆	.20†	.21‡	.38	.3
RWTH	**.88**‡	.17‡	.20†	.20‡	.37	**.49**	**.41**	.23⋆	**.61**‡	.16‡	.4	.3	**.43**	**.56**	–	.39	**.50**	.26	**.49**	**.37**	.29	**.34**	**.41**	.26	**.44**	.2
UEDIN	**.89**‡	.14‡	.22†	.13‡	**.62**	.34	.18†	.22⋆	**.39**	.03‡	.17‡	.3	.44	**.67**‡	**.42**	–	.39	.15‡	.14‡	**.52**⋆	**.40**	.36	**.43**	.26	**.41**	.38
UMD	**.91**‡	.07‡	.14‡	.08‡	.36	.34	.11‡	.25†	**.48**	.16‡	.24	.34	**.52**	**.56**	.41	**.45**	–	.16‡	.21†	**.41**	.28	.29	**.43**	.29	.25	.23
UPPSALA	**.97**‡	.32	.34	.17⋆	.36	**.54**‡	.23	**.37**	**.70**‡	.00‡	**.41**	**.62**⋆	**.56**	**.68**†	**.57**	**.64**‡	**.59**‡	–	.2	**.63**‡	**.69**‡	**.51**‡	**.60**⋆	**.33**	**.69**‡	**.63**‡
UU-MS	**.82**‡	.22	.43	.14‡	.45	**.51**†	.19	.21	**.68**‡	.14‡	**.39**	**.52**	**.60**	**.64**‡	.44	**.53**‡	**.61**†	**.28**	–	**.36**	**.58**‡	**.52**⋆	**.53**⋆	**.30**	**.64**‡	**.44**
BBN-C	**.86**‡	.25†	.10‡	.07‡	.27†	.17⋆	.23	.18‡	.35	.07‡	.15‡	.12‡	.32	.41	.3	.19⋆	.22	.15‡	.27	–	**.39**	.06†	.23⋆	.11‡	.21	.18†
CMU-HEA-C	**.87**‡	.14‡	.15‡	.08‡	.29⋆	.33	.04‡	.26	**.53**⋆	.00‡	.20⋆	.24⋆	**.44**	**.31**	**.46**	.23	**.53**	.15‡	.13‡	.27	–	**.40**	.2	.14‡	.22	.28
CMU-HYP-C	**.94**‡	.25†	.24	.14‡	.44	.3	.15‡	.26	**.47**	.08‡	.45	.31	.42	**.67**⋆	.24	.36	**.46**	.14‡	.21⋆	**.50**†	.32	–	**.43**	.28	**.51**⋆	**.42**
JHU-C	**.97**‡	.34	.11‡	.20†	.29	.34	.29⋆	.03‡	.38	.12‡	.07‡	.29	**.55**	**.67**†	.34	.32	.23	.24⋆	.24⋆	**.48**⋆	**.40**	.32	–	.27	**.37**	**.31**
KOC-C	**.88**‡	.00‡	.23⋆	.21†	**.53**	**.44**	.29	.22	**.43**	.08‡	**.36**	**.50**	**.53**	**.63**‡	**.39**	**.37**	**.39**	.28	.19	**.64**‡	**.61**‡	**.38**	**.55**	–	**.48**⋆	**.46**
RWTH-C	**.82**‡	.09‡	.06‡	.29⋆	.25†	.25	.18‡	.18‡	.24	.03‡	.19†	.26	.36	**.54**	.25	.26	**.33**	.06‡	.14‡	**.29**	.22	.23⋆	.3	.17⋆	–	.13‡
UPV-C	**.97**‡	.17†	.21⋆	.17‡	.36	.36	.23†	.19	**.67**‡	.20‡	.18†	.29	.41	**.40**	**.40**	.38	**.48**	.17‡	.31	**.50**†	**.43**	.27	.27	.27	**.65**‡	–
> others	.91	.23	.25	.20	.39	.42	.24	.30	.53	.11	.31	.38	.47	**.59**	.42	.43	.48	.27	.30	**.53**	.49	.42	.44	.31	.51	.41
>= others	.96	.42	.46	.36	.50	.66	.47	.53	.72	.23	.52	.59	.63	**.73**	.62	.66	.68	.51	.55	**.77**	.73	.65	.67	.59	.75	.64

Table 16: Sentence-level ranking for the WMT10 German-English News Task

	REF	CU-ZEMAN	DFKI	FBK	JHU	KIT	KOC	LIMSI	LIU	ONLINEA	ONLINEB	RWTH	SFU	UEDIN	UPPSALA	CMU-HEAFIELD-COMBO	KOC-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.03‡	.06‡	.01‡	.02‡	.05‡	.00‡	.00‡	.01‡	.04‡	.03‡	.01‡	.01‡	.01‡	.02‡	.01‡	.01‡	.05‡	.06‡
CU-ZEMAN	**.97**‡	–	**.85**‡	**.67**‡	**.62**‡	**.78**‡	**.58**⋆	**.70**‡	**.64**‡	**.80**‡	**.85**‡	**.64**‡	**.52**	**.80**‡	**.61**†	**.79**‡	**.69**‡	**.76**‡	**.73**‡
DFKI	**.89**‡	.14‡	–	.36†	.24‡	.38	.30‡	.27‡	.36⋆	.36⋆	**.55**	.35†	.21‡	.41	.39	.46	.38⋆	.47	.37⋆
FBK	**.97**‡	.30‡	**.59**†	–	.35†	.42	.12‡	.36	**.48**	**.48**	**.64**‡	.39	.29‡	**.46**	.30†	**.44**	**.46**	**.48**	.38
JHU	**.98**‡	.27‡	**.72**‡	**.57**†	–	**.59**‡	.30‡	**.51**	**.53**	**.56**⋆	**.65**‡	.43	.39	**.66**‡	.45	**.56**	**.61**‡	**.52**	**.47**
KIT	**.92**‡	.18‡	**.55**	.42	.29‡	–	.23‡	.32	.32†	.43	**.53**⋆	.41	.27‡	.43	.23‡	**.41**	**.41**	**.42**	.37
KOC	**1.00**‡	.37⋆	**.64**‡	**.82**‡	**.62**‡	**.70**‡	–	**.74**‡	**.74**‡	**.74**‡	**.82**‡	**.63**‡	.48	**.62**†	**.65**‡	**.73**‡	**.67**‡	**.81**‡	**.71**‡
LIMSI	**.95**‡	.27‡	**.68**‡	**.39**	.45	**.49**	.17‡	–	**.49**	**.74**‡	**.70**‡	**.51**	.28‡	**.58**‡	.32	**.51**	**.53**⋆	**.52**†	.31
LIU	**.95**‡	.32‡	**.59**⋆	.4	.36	**.58**†	.21‡	.37	–	.39	**.74**‡	.33⋆	.23‡	**.55**†	.36⋆	.49	**.42**	**.46**	.38
ONLINEA	**.95**‡	.16‡	**.55**⋆	.4	.36⋆	**.45**	.21‡	.23‡	**.50**	–	**.56**†	.38	.23‡	.41	.23‡	**.48**	.4	**.50**	.33†
ONLINEB	**.92**‡	.12‡	.42	.26‡	.27‡	.33⋆	.14‡	.23‡	.21‡	.32†	–	.24‡	.14‡	.39	.19‡	.29‡	.27‡	.36	.32‡
RWTH	**.98**‡	.33‡	**.61**†	**.51**	**.47**	**.46**	.30‡	.33	**.52**⋆	**.55**	**.71**‡	–	.33†	**.57**⋆	**.45**	**.40**	**.51**†	**.47**	**.46**
SFU	**.98**‡	.42	**.77**‡	**.66**‡	**.51**	**.69**‡	.48	**.68**‡	**.69**‡	**.72**‡	**.77**‡	**.56**†	–	**.82**‡	**.53**	**.65**‡	**.69**‡	**.73**‡	**.62**‡
UEDIN	**.94**‡	.17‡	**.51**	.4	.31‡	**.49**	.34†	.25‡	.30†	**.52**	**.52**	.36⋆	.10‡	–	.33⋆	.31	**.42**	.38	.22‡
UPPSALA	**.97**‡	.36†	**.55**	**.51**†	**.47**	**.70**‡	.25‡	**.46**	**.57**⋆	**.67**‡	**.71**‡	.41	.38	**.54**⋆	–	**.53**†	**.42**	**.58**‡	**.40**
CMU-HEAFIELD-COMBO	**.96**‡	.17‡	**.49**	.36	.36	.37	.21‡	.35	.49	.42	**.64**‡	.38	.28‡	**.48**	.28†	–	.35	**.46**	.35
KOC-COMBO	**.99**‡	.27‡	**.56**⋆	.32	.27‡	.32	.23‡	.32⋆	.41	**.55**	**.64**‡	.30†	.21‡	.37	.36	**.41**	–	.34	.36
RWTH-COMBO	**.92**‡	.17‡	**.50**	.34	.35	.41	.09‡	.25†	.38	.4	**.54**	.38	.20‡	**.42**	.19‡	.28	**.35**	–	.16‡
UPV-COMBO	**.93**‡	.23‡	**.58**⋆	.38	.36	**.51**	.23‡	**.50**	**.49**	**.57**†	**.60**‡	.42	.28‡	**.51**‡	.3	**.38**	**.46**	**.48**‡	–
> others	.95	.24	.57	.44	.37	.48	.24	.39	.45	.51	**.63**	.40	.27	.51	.34	.45	.44	**.49**	.39
>= others	.98	.28	.62	.56	.46	.60	.30	.51	.55	.59	**.70**	.51	.34	.62	.47	.59	.59	**.65**	.55

Table 17: Sentence-level ranking for the WMT10 English-German News Task

	REF	CAMBRIDGE	COLUMBIA	CU-ZEMAN	DFKI	HUICONG	JHU	ONLINEA	ONLINEB	UEDIN	UPC	BBN-COMBO	CMU-HEAFIELD-COMBO	JHU-COMBO	UPV-COMBO
REF	–	.00‡	.01‡	.01‡	.01‡	.00‡	.00‡	.00‡	.00‡	.00‡	.01‡	.02‡	.05‡	.01‡	.04‡
CAMBRIDGE	**.95**‡	–	.23‡	.14‡	.34⋆	.31†	**.41**	.34	**.62**‡	**.45**⋆	.35	**.40**⋆	**.42**	.22†	**.44**
COLUMBIA	**.97**‡	**.58**‡	–	.25‡	**.52**	**.45**	**.59**‡	**.53**⋆	**.65**‡	**.60**‡	**.47**	**.56**‡	**.55**‡	**.45**	**.58**‡
CU-ZEMAN	**.96**‡	**.71**‡	**.59**‡	–	**.60**‡	**.68**‡	**.79**‡	**.66**‡	**.75**‡	**.80**‡	**.66**‡	**.79**‡	**.78**‡	**.69**‡	**.75**‡
DFKI	**.97**‡	**.51**⋆	.37	.23‡	–	**.43**	**.59**‡	**.52**†	**.66**‡	**.62**‡	**.48**	**.53**†	**.55**†	**.55**†	**.64**‡
HUICONG	**.95**‡	**.50**†	.34	.21‡	.41	–	**.45**	**.50**	**.66**‡	**.61**‡	**.39**	**.50**⋆	**.59**‡	**.40**	**.52**‡
JHU	**.98**‡	.39	.22‡	.12‡	.30‡	.33	–	.37	**.56**‡	**.51**‡	.34	**.39**	**.34**†	.22‡	.34
ONLINEA	**.96**‡	**.46**	.37⋆	.23‡	.32†	.38	**.44**	–	**.59**‡	**.53**†	.4	**.50**	.36	.30†	**.54**‡
ONLINEB	**.88**‡	.25‡	.21‡	.16‡	.23‡	.21‡	.27‡	.23‡	–	.35	.24‡	.28‡	.34†	.22‡	.36
UEDIN	**.96**‡	.31⋆	.28‡	.10‡	.25‡	.19‡	.25‡	.31†	**.48**	–	.23‡	.27†	.31	.23‡	.2
UPC	**.94**‡	**.47**	.4	.20‡	.41	.33	**.43**	**.46**	**.66**‡	**.56**‡	–	**.50**⋆	**.52**†	**.48**⋆	**.49**†
BBN-COMBO	**.95**‡	.26⋆	.31‡	.09‡	.32†	.34⋆	.33	.37	**.54**‡	**.44**†	.33⋆	–	**.35**	.24‡	**.34**
CMU-HEAFIELD-COMBO	**.91**‡	.39	.21‡	.08‡	.34†	.22‡	.16†	**.42**	**.57**†	**.45**	.31†	.31	–	.14‡	.27
JHU-COMBO	**.95**‡	**.40**†	.32	.15‡	.36†	.31	**.44**‡	**.50**†	**.66**‡	**.50**‡	.32⋆	**.47**‡	**.43**‡	–	**.43**†
UPV-COMBO	**.92**‡	.35	.28‡	.16‡	.27‡	.23‡	**.38**	.28‡	**.47**	**.30**	.28†	.26	**.35**	.25†	–
> others	.95	.41	.30	.15	.33	.32	.39	.39	**.56**	.48	.34	.41	**.43**	.32	**.43**
>= others	.99	.61	.45	.27	.45	.50	.61	.54	**.70**	.69	.51	.62	**.66**	.55	**.66**

Table 18: Sentence-level ranking for the WMT10 Spanish-English News Task

	REF	CAMBRIDGE	CU-ZEMAN	DCU	DFKI	JHU	KOC	ONLINEA	ONLINEB	SFU	UEDIN	UPV	UCH-UPV	CMU-HEAFIELD-COMBO	KOC-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.00‡	.02‡	.07‡	.15‡	.07‡	.02‡	.11‡	.14‡	.07‡	.07‡	.03‡	.06‡	.09‡	.06‡	.03‡	.07‡
CAMBRIDGE	**.91**‡	–	.28†	**.45**	.38	**.45**	.11‡	**.52**	**.61**†	.21⋆	**.52**	**.47**	.35	**.54**	**.51**	**.39**	**.49**
CU-ZEMAN	**.95**‡	**.70**†	–	**.79**‡	**.75**‡	**.85**‡	**.49**	**.83**‡	**.82**‡	**.74**‡	**.87**‡	**.67**‡	**.85**‡	**.81**‡	**.80**‡	**.70**‡	**.74**‡
DCU	**.93**‡	.32	.21‡	–	.45	.32	.09‡	**.70**†	**.59**	.24‡	**.48**	.38	.29	.32	.36	.24	.14‡
DFKI	**.80**‡	**.41**	.15‡	.45	–	.38	.12‡	**.64**†	**.57**	.4	**.57**	.31	.41	**.59**	**.50**	**.48**	**.47**
JHU	**.90**‡	.37	.10‡	**.52**	**.56**	–	.17‡	**.67**†	**.67**‡	.26†	.34	.3	**.49**	**.54**	**.53**†	**.47**	**.35**
KOC	**.98**‡	**.87**‡	.47	**.88**‡	**.73**‡	**.76**‡	–	**.76**‡	**.87**‡	**.67**‡	**.83**‡	**.86**‡	**.90**‡	**.87**‡	**.90**‡	**.86**‡	**.86**‡
ONLINEA	**.82**‡	.42	.08‡	.30†	.18†	.24†	.20‡	–	**.49**	.36	.25†	.17‡	.25†	.45	.30⋆	.29	.18‡
ONLINEB	**.76**‡	.26†	.10‡	.32	.37	.22‡	.10‡	.34	–	.21‡	.28	.24†	.32	.33	.22‡	.19‡	.27⋆
SFU	**.91**‡	**.54**⋆	.19‡	**.67**‡	**.51**	**.63**†	.27‡	**.64**	**.72**‡	–	**.74**‡	**.57**⋆	**.68**‡	**.77**‡	**.71**‡	**.64**‡	**.46**
UEDIN	**.91**‡	.3	.08‡	.4	.38	.34	.14‡	**.71**†	**.49**	.09‡	–	.34	.4	**.58**	.33	.3	.31
UPV	**.94**‡	.34	.07‡	**.41**	**.53**	**.54**	.07‡	**.73**‡	**.61**†	.27⋆	**.45**	–	.37	**.51**	**.44**	**.38**	**.48**†
UCH-UPV	**.90**‡	**.55**	.07‡	**.58**	**.51**	.41	.08‡	**.69**†	**.52**	.24‡	**.51**	**.46**	–	**.47**	**.41**	**.49**	**.49**
CMU-HEAFIELD-COMBO	**.83**‡	.29	.13‡	**.37**	.38	.35	.07‡	**.48**	**.54**	.08‡	.29	.26	.28	–	.17†	.21⋆	.21
KOC-COMBO	**.88**‡	.27	.15‡	**.40**	.42	.24†	.03‡	**.62**⋆	**.60**‡	.15‡	**.41**	.27	.34	**.53**†	–	.3	**.40**
RWTH-COMBO	**.92**‡	.36	.21‡	**.52**	.33	.31	.10‡	**.55**	**.65**‡	.14‡	**.37**	.22	.41	**.52**⋆	**.48**	–	**.31**
UPV-COMBO	**.91**‡	.32	.13‡	**.69**‡	.4	.32	.09‡	**.76**‡	**.52**⋆	.36	**.38**	.19†	.31	**.45**	.35	.28	–
> others	.89	.39	.15	.48	.44	.41	.14	**.61**	.58	.29	.46	.36	.42	**.51**	.44	.39	.40
>= others	.93	.54	.23	.61	.55	.55	.19	.69	**.71**	.40	.61	.55	.54	**.68**	.62	.59	.60

Table 19: Sentence-level ranking for the WMT10 English-Spanish News Task

	REF	AALTO	CMU	CU-BOJAR	CU-ZEMAN	ONLINEA	ONLINEB	UEDIN	BBN-COMBO	CMU-HEAFIELD-COMBO	JHU-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.04‡	.02‡	.03‡	.00‡	.02‡	.00‡	.03‡	.03‡	.04‡	.01‡	.04‡	.02‡
AALTO	**.88**‡	–	**.49**	**.51**	.22‡	.38	**.64**‡	**.55**†	**.57**⋆	**.71**‡	**.64**‡	**.65**‡	**.59**‡
CMU	**.97**‡	.35	–	.4	.14‡	.18‡	**.59**‡	**.49**†	**.45**†	**.57**‡	**.50**‡	.34	**.43**
CU-BOJAR	**.90**‡	.33	**.43**	–	.12‡	.20‡	**.64**‡	**.45**	**.45**	**.54**‡	**.42**	**.42**	.41
CU-ZEMAN	**.99**‡	**.60**‡	**.77**‡	**.75**‡	–	**.56**†	**.81**‡	**.78**‡	**.88**‡	**.79**‡	**.84**‡	**.84**‡	**.76**‡
ONLINEA	**.92**‡	**.46**	**.68**‡	**.59**‡	.28†	–	**.65**‡	**.54**‡	**.72**‡	**.75**‡	**.58**‡	**.57**‡	**.66**‡
ONLINEB	**.97**‡	.27‡	.28‡	.21‡	.10‡	.17‡	–	.25†	.32	.22	.21†	.32	.28
UEDIN	**.95**‡	.28†	.26†	.38	.07‡	.22‡	**.49**†	–	**.60**‡	**.52**‡	.33	**.31**	.32
BBN-COMBO	**.92**‡	.31⋆	.20†	.39	.08‡	.15‡	**.41**	.16‡	–	.27	.25	.3	.26
CMU-HEAFIELD-COMBO	**.90**‡	.13‡	.23‡	.25‡	.07‡	.15‡	**.31**	.23‡	**.34**	–	.18‡	.35	.28
JHU-COMBO	**.93**‡	.20‡	.19‡	.33	.08‡	.25‡	**.48**†	**.39**	**.38**	**.52**‡	–	**.37**	**.42**
RWTH-COMBO	**.92**‡	.18‡	**.37**	.38	.13‡	.25‡	**.34**	.28	**.43**	**.40**	.26	–	.25
UPV-COMBO	**.96**‡	.25‡	.36	.41	.11‡	.27‡	**.45**	**.35**	**.37**	**.44**	.31	**.34**	–
> others	.93	.28	.36	.38	.11	.23	**.49**	.38	.47	**.48**	.38	.40	.40
>= others	.98	.43	.55	.55	.22	.37	**.70**	.61	.70	**.71**	.62	.65	.63

Table 20: Sentence-level ranking for the WMT10 Czech-English News Task

	REF	CU-BOJAR	CU-TECTO	CU-ZEMAN	DCU	EUROTRANS	KOC	ONLINEA	ONLINEB	PC-TRANS	POTSDAM	SFU	UEDIN	CMU-HEAFIELD-COMBO	DCU-COMBO	KOC-COMBO	RWTH-COMBO	UPV-COMBO
REF	–	.04‡	.04‡	.03‡	.01‡	.05‡	.03‡	.08‡	.04‡	.04‡	.03‡	.02‡	.02‡	.04‡	.08‡	.04‡	.07‡	.04‡
CU-BOJAR	**.87**‡	–	**.46**	.27‡	.12‡	.28‡	.16‡	.17‡	**.44**	.4	.11‡	.27‡	**.41**	.28	**.52**‡	.28	**.42**	**.43**
CU-TECTO	**.88**‡	.36	–	.30†	.23‡	.38	.17‡	.28‡	**.56**†	**.44**	.29†	.27‡	.36	**.45**	**.51**†	.4	**.58**†	.35
CU-ZEMAN	**.91**‡	**.58**‡	**.51**†	–	.38	**.49**	.19‡	.39	**.62**‡	**.63**‡	.36	**.41**	**.48**	**.51**‡	**.58**‡	**.48**†	**.54**†	**.55**‡
DCU	**.98**‡	**.73**‡	**.52**‡	**.43**	–	**.59**‡	.22‡	**.47**	**.74**‡	**.63**‡	**.47**†	**.53**†	**.56**‡	**.77**‡	**.77**‡	**.62**‡	**.76**‡	**.71**‡
EUROTRANS	**.88**‡	**.61**‡	**.47**	.33	.30‡	–	.10‡	.33	**.51**	**.54**†	.25‡	.27‡	**.49**	**.57**‡	**.59**†	**.49**	**.57**‡	**.60**‡
KOC	**.93**‡	**.69**‡	**.67**‡	**.54**‡	**.49**‡	**.77**‡	–	**.54**‡	**.71**‡	**.70**‡	**.51**‡	**.55**‡	**.64**‡	**.72**‡	**.78**‡	**.65**‡	**.76**‡	**.78**‡
ONLINEA	**.91**‡	**.62**‡	**.57**‡	**.51**	.39	**.44**	.24‡	–	**.66**‡	**.62**‡	.39	**.43**	**.55**‡	**.60**‡	**.61**‡	**.59**‡	**.73**‡	**.61**‡
ONLINEB	**.91**‡	.31	.29†	.27‡	.13‡	.33	.14‡	.19‡	–	.44	.22‡	.09‡	**.39**	.19	.34	.24⋆	.22†	**.39**
PC-TRANS	**.88**‡	**.45**	.43	.24‡	.26‡	.29†	.21‡	.24‡	**.49**	–	.22‡	.27‡	.37	**.43**	**.55**†	.33†	**.49**	.41
POTSDAM	**.88**‡	**.60**‡	**.51**†	**.40**	.27†	**.59**‡	.25‡	**.47**	**.63**‡	**.64**‡	–	**.45**	**.52**‡	**.56**‡	**.69**‡	**.61**‡	**.70**‡	**.68**‡
SFU	**.95**‡	**.52**‡	**.56**‡	.4	.30†	**.61**‡	.27‡	.39	**.65**‡	**.64**‡	.29	–	**.55**‡	**.54**‡	**.76**‡	**.53**‡	**.70**‡	**.60**‡
UEDIN	**.94**‡	.39	**.44**	.33	.23‡	.32	.20‡	.26‡	.32	**.49**	.25‡	.26‡	–	**.43**	**.57**‡	.18	**.46**†	**.42**
CMU-HEAFIELD-COMBO	**.91**‡	**.42**	.39	.23‡	.10‡	.27‡	.14‡	.19‡	**.23**	.35	.24‡	.19‡	.28	–	**.48**‡	.28	**.34**	**.29**
DCU-COMBO	**.84**‡	.23‡	.27†	.23‡	.03‡	.31†	.10‡	.21‡	**.42**	.31†	.15‡	.10‡	.16‡	.20‡	–	.18‡	.27⋆	.22‡
KOC-COMBO	**.91**‡	**.37**	**.49**	.25†	.10‡	.39	.17‡	.32‡	**.42**⋆	**.55**†	.17‡	.27‡	**.26**	**.33**	**.41**‡	–	.32	.22
RWTH-COMBO	**.88**‡	.29	.34†	.28†	.05‡	.26‡	.10‡	.17‡	**.48**†	.43	.16‡	.15‡	.24†	.33	**.46**⋆	**.36**	–	.29
UPV-COMBO	**.92**‡	.37	**.52**	.22‡	.09‡	.25‡	.10‡	.19‡	.28	**.47**	.15‡	.25‡	.33	.24	**.49**‡	**.34**	**.39**	–
> others	.91	.45	.44	.32	.20	.39	.16	.29	**.49**	**.49**	.25	.28	.40	.43	**.54**	.39	.50	.45
>= others	.96	.66	.60	.50	.38	.54	.33	.44	**.70**	.62	.44	.45	.62	.69	**.75**	.66	.70	.68

Table 21: Sentence-level ranking for the WMT10 English-Czech News Task

	REF	AALTO	CMU	CU-BOJAR	CU-ZEMAN	ONLINEA	ONLINEB	UEDIN	BBN-C	CMU-HEA-C	JHU-C	RWTH-C	UPV-C
REF	–	.03‡	.02‡	.03‡	.01‡	.03‡	.02‡	.05‡	.02‡	.06‡	.03‡	.05‡	.03‡
AALTO	**.93**‡	–	**.54**‡	**.54**‡	.23‡	.36	**.58**‡	**.56**‡	**.65**‡	**.69**‡	**.64**‡	**.67**‡	**.62**‡
CMU	**.94**‡	.30‡	–	**.47**	.14‡	.22‡	**.52**‡	**.41**	**.50**‡	**.57**‡	**.45**†	**.44**	.38
CU-BOJAR	**.94**‡	.26‡	.38	–	.10‡	.22‡	**.61**‡	**.47**†	**.46**	**.55**‡	**.42**	**.49**‡	**.44**
CU-ZEMAN	**.98**‡	**.58**‡	**.73**‡	**.77**‡	–	**.55**‡	**.79**‡	**.71**‡	**.84**‡	**.80**‡	**.77**‡	**.79**‡	**.75**‡
ONLINEA	**.94**‡	**.41**	**.61**‡	**.57**‡	.23‡	–	**.68**‡	**.63**‡	**.71**‡	**.71**‡	**.63**‡	**.54**‡	**.61**‡
ONLINEB	**.93**‡	.30‡	.31‡	.26‡	.10‡	.17‡	–	.32†	.35	.31	.22‡	.29*	.38
UEDIN	**.91**‡	.27‡	.35	.34†	.11‡	.18‡	**.47**†	–	**.54**‡	**.50**‡	**.35**	.29	**.35**
BBN-C	**.95**‡	.21‡	.22‡	.36	.06‡	.17‡	**.38**	.26‡	–	.32	.24‡	.31*	.26‡
CMU-HEA-C	**.90**‡	.17‡	.19‡	.23‡	.09‡	.18‡	**.32**	.27‡	**.34**	–	.31†	.31*	.30‡
JHU-C	**.93**‡	.19‡	.30†	.35	.09‡	.24‡	**.50**‡	.34	**.47**‡	**.45**†	–	**.41**‡	**.36**
RWTH-C	**.91**‡	.16‡	.35	.29‡	.12‡	.27‡	**.41***	**.37**	**.42***	**.42***	.23‡	–	.24†
UPV-C	**.94**‡	.24‡	**.40**	.36	.09‡	.28‡	**.39**	.32	**.46**‡	**.47**‡	.33	**.36**†	?
$>$ others	.93	.26	.37	.38	.11	.24	.47	.40	.49	.49	.38	.41	.40
$>=$ others	.97	.42	.56	.55	.25	.39	.67	.62	.70	.70	.61	.65	.62

Table 22: Sentence-level ranking for the WMT10 Czech-English News Task **(Combining expert and non-expert Mechanical Turk judgments)**

	REF	AALTO	CMU	CU-ZEMAN	DFKI	FBK	HUICONG	JHU	KIT	KOC	LIMSI	LIU	ONLINEA	ONLINEB	RWTH	UEDIN	UMD	UPPSALA	UU-MS	BBN-C	CMU-HEA-C	CMU-HYPO-C	JHU-C	KOC-C	RWTH-C	UPV-C
REF	–	.00‡	.02‡	.00‡	.07‡	.04‡	.03‡	.00‡	.06‡	.04‡	.00‡	.02‡	.07‡	.07‡	.07‡	.02‡	.09‡	.03‡	.03‡	.10‡	.04‡	.04‡	.03‡	.02‡	.07‡	.06‡
AALTO	**1.00**‡	–	**.43**	.39	**.48**	**.60**‡	**.38**	**.41**	**.74**‡	.18‡	**.42**	**.57**‡	**.50**†	**.63**‡	**.55**‡	**.68**‡	**.79**‡	**.42**	**.33**	**.71**‡	**.61**‡	**.66**‡	**.54**	**.51**‡	**.66**‡	**.56**‡
CMU	**.95**‡	.34	–	.19‡	**.45**	**.52**†	.38	**.50**	**.63**‡	.17‡	**.51**‡	**.55**‡	**.56**†	**.66**‡	**.55**‡	**.60**‡	**.56**‡	.30	.40	**.62**‡	**.64**‡	**.49**‡	**.58**‡	**.46**	**.64**‡	**.46**†
CU-ZEMAN	**1.00**‡	**.44**	**.64**‡	–	**.43**	**.72**‡	.31	**.45**†	**.69**‡	.36	**.55**	**.62**‡	**.75**‡	**.75**‡	**.78**‡	**.75**‡	**.75**‡	**.48**⋆	**.56**†	**.79**‡	**.82**‡	**.72**‡	**.68**‡	**.63**‡	**.67**‡	**.84**‡
DFKI	**.92**‡	.29	.33	.35	–	.37	**.40**	.34	**.59**	.08‡	.42	**.50**	**.49**	**.64**‡	.35	.44	**.44**	**.50**	.41	**.70**‡	**.61**†	**.57**	**.46**	**.47**	**.62**‡	**.44**
FBK	**.93**‡	.26‡	.23†	.17‡	**.49**	–	.12‡	.30	**.52**†	.08‡	.20‡	**.45**⋆	.41	**.62**‡	**.44**	**.44**	**.48**⋆	.18‡	.25†	**.53**‡	**.47**	**.38**	**.38**	.22†	**.41**	**.51**⋆
HUICONG	**.92**‡	.34	**.39**	**.37**	.38	**.71**‡	–	**.53**†	**.67**‡	.18‡	**.51**†	**.47**	**.60**‡	**.65**‡	**.49**⋆	**.55**‡	**.78**‡	**.35**	**.41**	**.56**‡	**.77**‡	**.74**‡	**.58**‡	**.41**	**.65**‡	**.57**‡
JHU	**.92**‡	.35	.30	.17†	**.52**	**.45**	.25†	–	**.58**‡	.16‡	**.43**	.38	**.57**†	**.60**‡	**.54**‡	**.60**‡	**.70**‡	.29	.25	**.65**‡	**.75**‡	**.56**‡	**.62**‡	**.49**⋆	**.66**‡	**.48**†
KIT	**.90**‡	.14‡	.16‡	.14‡	.35	.28†	.19‡	.16‡	–	.03‡	.29⋆	.20‡	.35	**.53**⋆	.21‡	.24†	.30	.20‡	.22‡	**.44**	.29	.38	**.35**	.24	**.40**	.24†
KOC	**.95**‡	**.66**‡	**.71**‡	**.51**	**.75**‡	**.80**‡	**.58**‡	**.68**‡	**.93**‡	–	**.75**‡	**.87**‡	**.72**‡	**.74**‡	**.74**‡	**.81**‡	**.81**‡	**.78**‡	**.66**‡	**.89**‡	**.85**‡	**.80**‡	**.80**‡	**.72**‡	**.91**‡	**.73**‡
LIMSI	**.99**‡	.26	.24‡	.32	**.45**	**.61**‡	.25†	.38	**.50**⋆	.10‡	–	**.50**⋆	**.55**⋆	**.69**‡	**.52**⋆	**.57**‡	**.57**‡	.29†	.22‡	**.60**‡	**.52**†	**.42**	**.47**†	.37	**.60**‡	**.56**‡
LIU	**.87**‡	.17‡	.20‡	.14‡	.34	.22⋆	.31	.38	**.66**‡	.04‡	.27⋆	–	**.51**⋆	**.53**†	**.52**⋆	**.53**⋆	**.51**	.20‡	.33	**.64**‡	**.59**‡	**.48**†	**.48**	**.51**	**.37**	**.53**⋆
ONLINEA	**.90**‡	.25†	.29†	.18‡	.34	**.43**	.23‡	.28†	**.49**	.08‡	.32⋆	.30⋆	–	**.44**	.38	.40	**.42**	.32†	.35⋆	**.39**	**.47**	**.51**	.27‡	.35	**.43**	**.40**
ONLINEB	**.76**‡	.22‡	.24‡	.14‡	.27‡	.27‡	.25‡	.25‡	.32⋆	.22‡	.21‡	.28†	.32	–	.27†	.21‡	.30†	.23‡	.15‡	.41	.31	.40	.23‡	.16‡	.42	.29
RWTH	**.89**‡	.22‡	.23‡	.13‡	**.49**	.35	.29⋆	.21‡	**.62**‡	.15‡	.32⋆	.29⋆	**.46**	**.57**†	–	.39	**.49**	.25	.38	**.41**	.27	**.34**	.36	.27	**.48**⋆	.22‡
UEDIN	**.91**‡	.15‡	.20‡	.12‡	**.49**	.35	.24‡	.22‡	**.49**†	.04‡	.22‡	.30⋆	**.46**	**.62**‡	**.43**	–	.39	.11‡	.15‡	**.45**	.33	**.40**	**.45**	.33	.34	.33
UMD	**.91**‡	.12‡	.23‡	.06‡	.35	.29⋆	.11‡	.16‡	**.47**	.14‡	.23‡	.35	.40	**.55**†	.36	**.47**	–	.16‡	.17‡	**.44**	.29†	.27	**.37**	.26	.27	.24†
UPPSALA	**.94**‡	.30	**.41**	.23⋆	.35	**.53**‡	.26	**.37**	**.66**‡	.03‡	**.54**†	**.71**‡	**.57**†	**.65**‡	**.45**	**.72**‡	**.67**‡	–	.25	**.59**‡	**.69**‡	**.49**‡	**.63**‡	.33	**.60**‡	**.64**‡
UU-MS	**.83**‡	.28	**.42**	.24†	.41	**.49**†	.28	**.42**	**.68**‡	.10‡	**.55**‡	**.48**	**.55**⋆	**.63**‡	**.49**	**.56**‡	**.60**‡	**.32**	–	**.52**†	**.58**‡	**.61**‡	**.64**‡	**.46**‡	**.64**‡	**.50**⋆
BBN-C	**.90**‡	.15‡	.16‡	.10‡	.22‡	.17‡	.22‡	.18‡	.41	.06‡	.16‡	.21‡	.35	**.45**	.30	.26	.34	.13‡	.20†	–	**.42**†	.14†	.27	.11‡	.25	.21†
CMU-HEA-C	**.83**‡	.20‡	.18‡	.07‡	.29†	.32	.06‡	.10‡	**.49**	.05‡	.26†	.21‡	.41	**.33**	**.37**	**.43**	**.58**†	.10‡	.14‡	.18†	–	**.33**	**.32**	.11‡	**.34**	.24⋆
CMU-HYPO-C	**.96**‡	.24‡	.20‡	.07‡	.37	.33	.12‡	.21‡	**.40**	.10‡	.41	.26†	.40	**.54**	.25	.37	**.44**	.13‡	.17‡	**.49**†	.31	–	.34	.23⋆	**.51**†	**.45**
JHU-C	**.97**‡	.33	.22‡	.18‡	.31	.30	.27‡	.18‡	.33	.12‡	.19†	.33	**.59**‡	**.60**‡	**.39**	.32	.30	.19‡	.20‡	**.44**	.29	.34	–	.21⋆	**.36**	.23
KOC-C	**.93**‡	.11‡	.31	.17‡	.41	**.50**†	.25	.27⋆	**.44**	.11‡	**.42**	.36	**.47**	**.68**‡	**.43**	**.41**	**.40**	.33	.18‡	**.59**‡	**.57**‡	**.46**⋆	**.47**⋆	–	**.52**†	**.43**
RWTH-C	**.87**‡	.20‡	.10‡	.21‡	.25‡	.27	.15‡	.23‡	.24	.02‡	.20‡	.30	.34	**.47**	.27⋆	.34	**.36**	.14‡	.20‡	**.33**	.26	.21†	.24	.20†	–	.17‡
UPV-C	**.93**‡	.14‡	.20†	.10‡	.42	.29⋆	.25‡	.25†	**.57**†	.20‡	.22‡	.33⋆	.39	**.45**	**.47**‡	**.40**	**.50**†	.24‡	.28⋆	**.44**†	**.42**⋆	.27	**.34**	.28	**.56**‡	?
> others	.92	.25	.28	.18	.39	.41	.25	.30	.52	.12	.34	.39	.47	.57	.42	.46	.51	.27	.28	.52	.49	.45	.44	.34	.50	.42
>= others	.96	.46	.49	.35	.53	.62	.45	.51	.71	.24	.54	.58	.63	.72	.63	.66	.70	.50	.51	.75	.73	.68	.67	.59	.74	.64

Table 23: Sentence-level ranking for the WMT10 German-English News Task **(Combining expert and non-expert Mechanical Turk judgments)**

	REF	CAMBRIDGE	COLUMBIA	CU-ZEMAN	DFKI	HUICONG	JHU	ONLINEA	ONLINEB	UEDIN	UPC	BBN-C	CMU-HEA-C	JHU-C	UPV-C
REF	–	.05‡	.01‡	.02‡	.03‡	.03‡	.01‡	.02‡	.04‡	.03‡	.04‡	.03‡	.07‡	.05‡	.04‡
CAMBRIDGE	**.90**‡	–	.24‡	.11‡	.35†	.26‡	**.43**	.35	**.50**†	**.45**†	.33⋆	**.40**	**.46**	.28⋆	**.41**
COLUMBIA	**.97**‡	**.61**‡	–	.25‡	**.47**	**.44**	**.61**‡	**.53**‡	**.62**‡	**.59**‡	**.48**†	**.59**‡	**.57**‡	**.45**†	**.57**‡
CU-ZEMAN	**.92**‡	**.73**‡	**.59**‡	–	**.62**‡	**.66**‡	**.71**‡	**.65**‡	**.75**‡	**.79**‡	**.58**‡	**.75**‡	**.78**‡	**.71**‡	**.72**‡
DFKI	**.95**‡	**.50**†	.41	.21‡	–	**.46**	**.56**‡	**.52**‡	**.65**‡	**.62**‡	**.47**	**.52**‡	**.56**‡	**.52**†	**.60**‡
HUICONG	**.93**‡	**.57**‡	.34	.21‡	.36	–	**.47**†	**.43**	**.67**‡	**.58**‡	**.40**	**.51**‡	**.62**‡	**.46**†	**.52**‡
JHU	**.94**‡	.39	.22‡	.16‡	.30‡	.32†	–	.41	**.52**‡	**.47**‡	.37	**.41**	**.33**†	.28	**.35**
ONLINEA	**.92**‡	**.45**	.35‡	.24‡	.34‡	.41	.41	–	**.60**‡	**.58**‡	.38	**.55**‡	**.46**	.36	**.57**‡
ONLINEB	**.87**‡	.34†	.24‡	.15‡	.21‡	.19‡	.33‡	.25‡	–	.34†	.26‡	.34†	.37⋆	.24‡	.40
UEDIN	**.94**‡	.33†	.26‡	.12‡	.24‡	.22‡	.25‡	.25‡	**.50**†	–	.25‡	.28†	.32⋆	.25‡	.26
UPC	**.89**‡	**.45**⋆	.36†	.23‡	.39	.37	**.42**	**.48**	**.62**‡	**.57**‡	–	**.54**‡	**.51**‡	**.50**‡	**.53**‡
BBN-C	**.91**‡	.33	.25‡	.11‡	.32‡	.30‡	.34	.31‡	**.51**†	**.41**†	.30‡	–	**.36**	.26‡	**.31**
CMU-HEA-C	**.89**‡	.37	.20‡	.10‡	.29‡	.23‡	.23†	.35	**.50**⋆	**.44**⋆	.31‡	.34	–	.23‡	**.31**
JHU-C	**.89**‡	**.39**⋆	.31†	.17‡	.37†	.33†	**.38**	**.42**	**.63**‡	**.47**‡	.31‡	**.42**‡	**.42**‡	–	**.37**†
UPV-C	**.91**‡	.35	.30‡	.16‡	.29‡	.26‡	.32	.28‡	**.44**	**.35**	.27‡	.27	.30	.24†	?
> others	.92	.42	.29	.16	.33	.32	.39	.37	.54	.48	.34	.42	.44	.35	.43
>= others	.97	.62	.45	.29	.46	.50	.61	.52	.68	.68	.51	.64	.65	.58	.66

Table 24: Sentence-level ranking for the WMT10 Spanish-English News Task **(Combining expert and non-expert Mechanical Turk judgments)**

	REF	CAMBRIDGE	CMU-STATXFER	CU-ZEMAN	DFKI	GENEVA	HUICONG	JHU	LIG	LIMSI	LIUM	NRC	ONLINEA	ONLINEB	RALI	RWTH	UEDIN	BBN-C	CMU-HEA-C	CMU-HYPO-C	DCU-C	JHU-C	LIUM-C	RWTH-C	UPV-C
REF	–	.02‡	.00‡	.00‡	.00‡	.00‡	.05‡	.02‡	.00‡	.00‡	.00‡	.02‡	.06‡	.02‡	.04‡	.02‡	.04‡	.03‡	.02‡	.05‡	.05‡	.04‡	.05‡	.06‡	.02‡
CAMBRIDGE	**.82**‡	–	.42	.16‡	.12‡	.35	.31	**.45**	.21‡	**.47**	.29	**.38**	.28†	**.54**	**.43**	.33	.38	.28	.39	**.45**†	.24	.25	.34	**.54**†	**.37**
CMU-STATXFER	**.91**‡	**.50**	–	.17‡	.41	.17‡	.28	**.44**	**.36**	**.48**⋆	**.56**‡	**.57**‡	.47	**.56**⋆	**.70**‡	**.49**	**.50**	**.47**	**.61**‡	**.68**‡	**.55**†	**.50**	.42	**.52**†	**.51**†
CU-ZEMAN	**1.00**‡	**.74**‡	**.71**‡	–	**.74**‡	**.46**	**.67**‡	**.73**‡	**.73**‡	**.74**‡	**.75**‡	**.76**‡	**.75**‡	**.89**‡	**.78**‡	**.66**‡	**.83**‡	**.74**‡	**.87**‡	**.73**‡	**.80**‡	**.83**‡	**.77**‡	**.95**‡	**.82**‡
DFKI	**1.00**‡	**.77**‡	**.48**	.17‡	–	.27†	**.49**	**.52**	**.48**	**.64**‡	**.69**‡	**.67**†	**.47**	**.62**⋆	**.53**	.47	**.64**‡	**.60**†	**.73**‡	**.72**‡	**.79**‡	**.58**⋆	**.66**‡	**.73**‡	**.74**‡
GENEVA	**.98**‡	**.58**	**.70**‡	.44	**.59**†	–	**.55**⋆	**.67**‡	**.70**‡	**.70**‡	**.77**‡	**.73**‡	**.63**‡	**.81**‡	**.81**‡	**.69**†	**.77**‡	**.73**‡	**.62**†	**.66**‡	**.75**‡	**.60**‡	**.73**‡	**.88**‡	**.67**†
HUICONG	**.89**‡	**.53**	**.34**	.13‡	.34	.30⋆	–	**.41**	.36	**.43**	**.70**‡	**.56**‡	**.57**	**.59**†	**.56**‡	**.43**	**.55**†	**.45**	**.51**⋆	**.64**‡	**.48**	**.49**	**.49**	**.53**†	**.57**†
JHU	**.88**‡	.36	.38	.11‡	.34	.25‡	.35	–	.33⋆	**.46**	**.49**⋆	**.48**	.40	**.50**	.40	.34	**.36**	.39	**.33**	**.59**‡	**.54**⋆	**.41**	.42	**.40**	**.41**
LIG	**.98**‡	**.65**‡	.34	.18‡	.44	.26‡	**.39**	**.56**⋆	–	**.60**‡	**.55**‡	**.51**‡	**.45**	**.54**†	**.53**	.39	.38	**.52**⋆	**.54**†	**.53**‡	**.51**⋆	**.53**†	**.55**	**.51**	**.58**†
LIMSI	**.98**‡	.40	.24⋆	.23‡	.23‡	.15‡	.29	.38	.25‡	–	.28	.38	.27†	**.64**‡	.35	.30	**.41**	.27	.33	**.49**	**.45**	.37	.28	**.45**	**.39**
LIUM	**.90**‡	**.40**	.19‡	.12‡	.30‡	.11‡	.11‡	.26⋆	.15‡	**.36**	–	.36	.25†	.37	**.39**	.26	.29	.24	.34	**.49**†	.34	.33	.34	.31	.38
NRC	**.93**‡	.31	.06‡	.15‡	.29†	.23‡	.20‡	.32	.16‡	.38	.36	–	.23†	**.53**	.36	.24⋆	.31	**.44**	**.37**	**.47**⋆	**.45**⋆	.29	**.39**	**.38**	**.42**
ONLINEA	**.92**‡	**.60**†	.47	.15‡	.44	.22‡	.32	**.46**	.34	**.57**†	**.52**†	**.60**†	–	**.52**⋆	.34	.44	**.57**†	**.56**	**.51**	**.51**	**.64**†	**.46**	**.51**	.41	**.60**
ONLINEB	**.85**‡	.35	.32⋆	.09‡	.33⋆	.10‡	.29†	.31	.25†	.17‡	**.40**	.34	.24⋆	–	.38	.32⋆	.28	.39	.30	**.42**	.37	**.41**	.35	.32	.22‡
RALI	**.90**‡	.31	.19‡	.10‡	.38	.10‡	.17‡	**.47**	.35	**.38**	.33	**.38**	**.48**	**.48**	–	.29⋆	.31	.29	**.38**	**.40**	**.38**	.34	.31	**.57**†	.21†
RWTH	**.93**‡	**.43**	.33	.12‡	.47	.26†	.39	**.40**	**.47**	**.35**	**.45**	**.49**⋆	.44	**.53**⋆	**.54**⋆	–	**.44**⋆	**.42**	**.48**	**.51**⋆	**.54**⋆	**.48**†	**.49**	**.50**‡	.26
UEDIN	**.92**‡	**.42**	.32	.10‡	.22‡	.10‡	.28†	.30	**.42**	.30	**.55**	**.36**	.23†	**.43**	**.33**	.20⋆	–	**.41**	.24	**.52**†	**.46**	.25	.22	.27	**.37**
BBN-C	**.92**‡	**.49**	.33	.24‡	.28†	.18‡	.40	.39	.28⋆	**.45**	**.27**	.27	.36	.39	**.35**	.35	.31	–	.26	**.45**‡	**.43**	.26	**.58**‡	**.36**	.28
CMU-HEA-C	**.90**‡	**.41**	.21‡	.06‡	.23‡	.29†	.28⋆	.27	.22†	**.39**	**.40**	.22	.39	**.43**	.29	.30	**.40**	**.28**	–	**.43**	**.28**	.15⋆	.25	.26	.16
CMU-HYPO-C	**.84**‡	.18†	.20‡	.14‡	.20‡	.22‡	.21‡	.19‡	.16‡	.31	.22†	.21⋆	.36	.38	.34	.27⋆	.22†	.16‡	.24	–	**.36**	.23	.10‡	**.33**	.24
DCU-C	**.92**‡	**.27**	.24†	.12‡	.17‡	.23‡	.30	.29⋆	.24⋆	.32	**.43**	.22⋆	.28†	**.41**	.23	.27⋆	.28	.22	.23	.25	–	.23	.23	**.24**	.17
JHU-C	**.88**‡	**.47**	.26	.10‡	.33⋆	.24‡	.36	.34	.24†	**.41**	**.39**	**.40**	.42	.39	.34	.25†	**.42**	**.28**	**.37**⋆	**.38**	**.39**	–	.37	**.32**	**.38**⋆
LIUM-C	**.90**‡	**.48**	.42	.13‡	.25‡	.20‡	.33	**.50**	.30	**.44**	**.37**	.34	.37	**.52**	**.43**	.34	**.33**	.22‡	**.34**	**.56**‡	**.33**	**.43**	–	**.49**‡	**.44**
RWTH-C	**.89**‡	.22†	.19†	.03‡	.23‡	.12‡	.19†	.23	.27	.30	**.36**	.19	**.47**	**.54**	.26†	.16‡	.27	.19	.26	.28	.16	.22	.16‡	–	.22
UPV-C	**.89**‡	.27	.15†	.10‡	.16‡	.29†	.30†	.31	.25†	.36	**.42**	.24	.32	**.64**‡	**.46**†	**.34**	.27	**.44**	**.33**	**.44**	**.23**	.17⋆	.31	**.24**	?
> others	.91	.43	.32	.14	.31	.21	.31	.39	.31	.42	.44	.40	.38	.52	.43	.33	.40	.37	.40	.49	.43	.38	.4	.44	.39
>= others	.97	.64	.51	.24	.40	.31	.50	.59	.50	.63	.68	.65	.51	.68	.65	.55	.66	.63	.69	.75	.71	.64	.62	.74	.67

Table 25: Sentence-level ranking for the WMT10 French-English News Task **(Combining expert and non-expert Mechanical Turk judgments)**