Further Meta-Evaluation of Machine Translation

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

Cameron Fordyce
camfordyce@gmail.com

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

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

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

Abstract

This paper analyzes the translation quality of machine translation systems for 10 language pairs translating between Czech, English, French, German, Hungarian, and Spanish. We report the translation quality of over 30 diverse translation systems based on a large-scale manual evaluation involving hundreds of hours of effort. We use the human judgments of the systems to analyze automatic evaluation metrics for translation quality, and we report the strength of the correlation with human judgments at both the system-level and at the sentence-level. We validate our manual evaluation methodology by measuring intra- and inter-annotator agreement, and collecting timing information.

1 Introduction

This paper presents the results the shared tasks of the 2008 ACL Workshop on Statistical Machine Translation, which builds on two past workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007). There were two shared tasks this year: a translation task which evaluated translation between 10 pairs of European languages, and an evaluation task which examines automatic evaluation metrics.

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

- **Test set selection** – Instead of creating our test set by reserving a portion of the training data, we instead hired translators to translate a set of newspaper articles from a number of different sources. This out-of-domain test set contrasts with the in-domain Europarl test set.
- **New language pairs** – We evaluated the quality of Hungarian-English machine translation. Hungarian is a challenging language because it is agglutinative, has many cases and verb conjugations, and has freer word order. German-Spanish was our first language pair that did not include English, but was not manually evaluated since it attracted minimal participation.
- **System combination** – Saarland University entered a system combination over a number of rule-based MT systems, and provided their output, which were also treated as fully fledged entries in the manual evaluation. Three additional groups were invited to apply their system combination algorithms to all systems.
- **Refined manual evaluation** – Because last year's study indicated that fluency and adequacy judgments were slow and unreliable, we dropped them from manual evaluation. We replaced them with yes/no judgments about the acceptability of translations of shorter phrases.
- **Sentence-level correlation** – In addition to measuring the correlation of automatic evaluation metrics with human judgments at the system level, we also measured how consistent they were with the human rankings of individual sentences.

The remainder of this paper is organized as follows: Section 2 gives an overview of the shared

Proceedings of the Third Workshop on Statistical Machine Translation, pages 70–106,
Columbus, June 2008. ©2008 Association for Computational Linguistics

translation task, describing the test sets, the materials that were provided to participants, and a list of the groups who participated. Section 3 describes the manual evaluation of the translations, including information about the different types of judgments that were solicited and how much data was collected. Section 4 presents the results of the manual evaluation. Section 5 gives an overview of the shared evaluation task, describes which automatic metrics were submitted, and tells how they were evaluated. Section 6 presents the results of the evaluation task. Section 7 validates the manual evaluation methodology.

2 Overview of the shared translation task

The shared translation task consisted of 10 language pairs: English to German, German to English, English to Spanish, Spanish to English, English to French, French to English, English to Czech, Czech to English, Hungarian to English, and German to Spanish. Each language pair had two test sets drawn from the proceedings of the European parliament, or from newspaper articles.[1]

2.1 Test data

The test data for this year's task differed from previous years' data. Instead of only reserving a portion of the training data as the test set, we hired people to translate news articles that were drawn from a variety of sources during November and December of 2007. We refer to this as the News test set. A total of 90 articles were selected, 15 each from a variety of Czech-, English-, French-, German-, Hungarian- and Spanish-language news sites:[2]

Hungarian: Napi (3 documents), Index (2), Origo (5), Népszabadság (2), HVG (2), Uniospez (1)

Czech: Aktuálně (1), iHNed (4), Lidovky (7), Novinky (3)

French: Liberation (4), Le Figaro (4), Dernieres Nouvelles (2), Les Echos (3), Canoe (2)

Spanish: Cinco Dias (7), ABC.es (3), El Mundo (5)

English: BBC (3), Scotsman (3), Economist (3), Times (3), New York Times (3)

German: Financial Times Deutschland (3), Süddeutsche Zeitung (3), Welt (3), Frankfurter Allgemeine Zeitung (3), Spiegel (3)

The translations were created by the members of EuroMatrix consortium who hired a mix of professional and non-professional translators. All translators were fluent or native speakers of both languages, and all translations were proofread by a native speaker of the target language. All of the translations were done directly, and not via an intermediate language. So for instance, each of the 15 Hungarian articles were translated into Czech, English, French, German and Spanish. The total cost of creating the 6 test sets consisting of 2,051 sentences in each language was approximately 17,200 euros (around 26,500 dollars at current exchange rates, at slightly more than 10c/word).

Having a test set that is balanced in six different source languages and translated across six languages raises some interesting questions. For instance, is it easier, when the machine translation system translates in the same direction as the human translator? We found no conclusive evidence that shows this. What is striking, however, that the parts differ dramatically in difficulty, based on the original source language. For instance the Edinburgh French-English system has a BLEU score of 26.8 on the part that was originally Spanish, but a score of on 9.7 on the part that was originally Hungarian. For average scores for each original language, see Table 1.

Original source language	avg. BLEU
Hungarian	8.8
German	11.0
Czech	15.2
Spanish	17.3
English	17.7
French	18.6

Table 1: Difficulty of the test set parts based on the original language. For each part, we average BLEU scores from the Edinburgh systems for 12 language pairs of the shared task.

[1] For Czech news editorials replaced the European parliament transcripts as the second test set, and for Hungarian the newspaper articles was the only test set.

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

In order to remain consistent with previous evaluations, we also created a Europarl test set. The Europarl test data was again drawn from the transcripts of EU parliamentary proceedings from the fourth quarter of 2000, which is excluded from the Europarl training data. Our rationale behind investing a considerable sum to create the News test set was that we believe that it more accurately represents the quality of systems' translations than when we simply hold out a portion of the training data as the test set, as with the Europarl set. For instance, statistical systems are heavily optimized to their training data, and do not perform as well on out-of-domain data (Koehn and Schroeder, 2007). Having both the News test set and the Europarl test set allows us to contrast the performance of systems on in-domain and out-of-domain data, and provides a fairer comparison between systems trained on the Europarl corpus and systems that were developed without it.

2.2 Provided materials

To lower the barrier of entry for newcomers to the field, we provided a complete baseline MT system, along with data resources. We provided:

- sentence-aligned training corpora
- language model data
- development and dev-test sets
- Moses open source toolkit for phrase-based statistical translation (Koehn et al., 2007)

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

The training materials are described in Figure 1.

2.3 Submitted systems

We received submissions from 23 groups from 18 institutions, as listed in Table 2. We also evaluated seven additional commercial rule-based MT systems, bringing the total to 30 systems. This is a significant increase over last year's shared task, where there were submissions from 15 groups from 14 institutions. Of the 15 groups that participated in last year's shared task, 11 groups returned this year. One of the goals of the workshop was to attract submissions from newcomers to the field, and we are please to have attracted many smaller groups, some as small as a single graduate student and her adviser.

The 30 submitted systems represent a broad range of approaches to statistical machine translation. These include statistical phrase-based and rule-based (RBMT) systems (which together made up the bulk of the entries), and also hybrid machine translation, and statistical tree-based systems. For most language pairs, we assembled a solid representation of the state of the art in machine translation.

In addition to individual systems being entered, this year we also solicited a number of entries which combined the results of other systems. We invited researchers at BBN, Carnegie Mellon University, and the University of Edinburgh to apply their system combination algorithms to all of the systems submitted to shared translation task. We designated the translations of the Europarl set as the development data for combination techniques which weight each system.[3] CMU combined the French-English systems, BBN combined the French-English and German-English systems, and Edinburgh submitted combinations for the French-English and German-English systems as well as a multi-source system combination which combined all systems which translated from any language pair into English for the News test set. The University of Saarland also produced a system combination over six commercial RBMT systems (Eisele et al., 2008). Saarland graciously provided the output of these systems, which we manually evaluated alongside all other entries.

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

3 Human evaluation

As with last year's workshop, 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. Therefore, rather than select an official automatic evaluation metric like the NIST Machine Translation Workshop does (Przybocki and Peterson, 2008), we define the manual evaluation to be primary, and use

[3]Since the performance of systems varied significantly between the Europarl and News test sets, such weighting might not be optimal. However this was a level playing field, since none of the individual systems had development data for the News set either.

Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		German ↔ Spanish	
Sentences	1,258,778		1,288,074		1,266,520		1,237,537	
Words	36,424,186	35,060,653	38,784,144	36,046,219	33,404,503	35,259,758	32,652,649	35,780,165
Distinct words	149,159	96,746	119,437	97,571	301,006	96,802	298,040	148,206

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		German ↔ Spanish	
Sentences	64,308		55,030		72,291		63,312	
Words	1,759,972	1,544,633	1,528,159	1,329,940	1,784,456	1,718,561	1,597,152	1,751,215
Distinct words	52,832	38,787	42,385	36,032	84,700	40,553	78,658	52,397

Hunglish Training Corpus

	Hungarian ↔ English	
Sentences	1,517,584	
Words	26,082,667	31,458,540
Distinct words	717,198	192,901

CzEng Training Corpus

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

Europarl Language Model Data

	English	Spanish	French	German
Sentence	1,412,546	1,426,427	1,438,435	1,467,291
Words	34,501,453	36,147,902	35,680,827	32,069,151
Distinct words	100,826	155,579	124,149	314,990

Europarl test set

	English	Spanish	French	German
Sentences	2,000			
Words	60,185	61,790	64,378	56,624
Distinct words	6,050	7,814	7,361	8,844

News Commentary test set

	English	Czech
Sentences	2,028	
Words	45,520	39,384
Distinct words	7,163	12,570

News Test Set

	English	Spanish	French	German	Czech	Hungarian
Sentences	2,051					
Words	43,482	47,155	46,183	41,175	36,359	35,513
Distinct words	7,807	8,973	8,898	10,569	12,732	13,144

Figure 1: Properties of the training and test sets used in the shared task. The training data is drawn from the Europarl corpus and from the Project Syndicate, a web site which collects political commentary in multiple languages. For Czech and Hungarian we use other available parallel corpora. Note that the number of words is computed based on the provided tokenizer and that the number of distinct words is the based on lowercased tokens.

ID	Participant
BBN-COMBO	BBN system combination (Rosti et al., 2008)
CMU-COMBO	Carnegie Mellon University system combination (Jayaraman and Lavie, 2005)
CMU-GIMPEL	Carnegie Mellon University Gimpel (Gimpel and Smith, 2008)
CMU-SMT	Carnegie Mellon University SMT (Bach et al., 2008)
CMU-STATXFER	Carnegie Mellon University Stat-XFER (Hanneman et al., 2008)
CU-TECTOMT	Charles University TectoMT (Zabokrtsky et al., 2008)
CU-BOJAR	Charles University Bojar (Bojar and Hajič, 2008)
CUED	Cambridge University (Blackwood et al., 2008)
DCU	Dublin City University (Tinsley et al., 2008)
LIMSI	LIMSI (Déchelotte et al., 2008)
LIU	Linköping University (Stymne et al., 2008)
LIUM-SYSTRAN	LIUM / Systran (Schwenk et al., 2008)
MLOGIC	Morphologic (Novák et al., 2008)
PCT	a commercial MT provider from the Czech Republic
RBMT1–6	Babelfish, Lingenio, Lucy, OpenLogos, ProMT, SDL (ordering anonymized)
SAAR	University of Saarbruecken (Eisele et al., 2008)
SYSTRAN	Systran (Dugast et al., 2008)
UCB	University of California at Berkeley (Nakov, 2008)
UCL	University College London (Wang and Shawe-Taylor, 2008)
UEDIN	University of Edinburgh (Koehn et al., 2008)
UEDIN-COMBO	University of Edinburgh system combination (Josh Schroeder)
UMD	University of Maryland (Dyer, 2007)
UPC	Universitat Politecnica de Catalunya, Barcelona (Khalilov et al., 2008)
UW	University of Washington (Axelrod et al., 2008)
XEROX	Xerox Research Centre Europe (Nikoulina and Dymetman, 2008)

Table 2: Participants in the shared translation task. Not all groups participated in all language pairs.

the human judgments to validate automatic metrics.

Manual evaluation is time consuming, and it requires a monumental 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 100 people participated in the manual evaluation, with 75 people putting in more than an hour's worth of effort, and 25 putting in more than four hours. A collective total of 266 hours of labor was invested.

We wanted to ensure that we were using our annotators' time effectively, so we carefully designed the manual evaluation process. In our analysis of last year's manual evaluation we found that the NIST-style fluency and adequacy scores (LDC, 2005) were overly time consuming and inconsistent.[4] We therefore abandoned this method of evaluating the translations.

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

- Ranking translated sentences relative to each other
- Ranking the translations of syntactic constituents drawn from the source sentence
- Assigning absolute yes or no judgments to the translations of the syntactic constituents.

The manual evaluation software asked for repeated judgments from the same individual, and had multiple people judge the same item, and logged the time it took to complete each judgment. This allowed us to measure intra- and inter-annotator agreement, and to analyze the average amount of time it takes to collect the different kinds of judgments. Our analysis is presented in Section 7.

3.1 Ranking translations of sentences

Ranking translations relative to each other is a relatively intuitive and straightforward task. We therefore kept the instructions simple. The instructions for this task were:

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

Ranking several translations at a time is a variant of force choice judgments where a pair of systems is presented and an annotator is asked "Is A better than B, worse than B, or equal to B." In our experiments, annotators were shown five translations at a time, except for the Hungarian and Czech language pairs where there were fewer than five system submissions. In most cases there were more than 5 systems submissions. We did not attempt to get a complete ordering over the systems, and instead relied on random selection and a reasonably large sample size to make the comparisons fair.

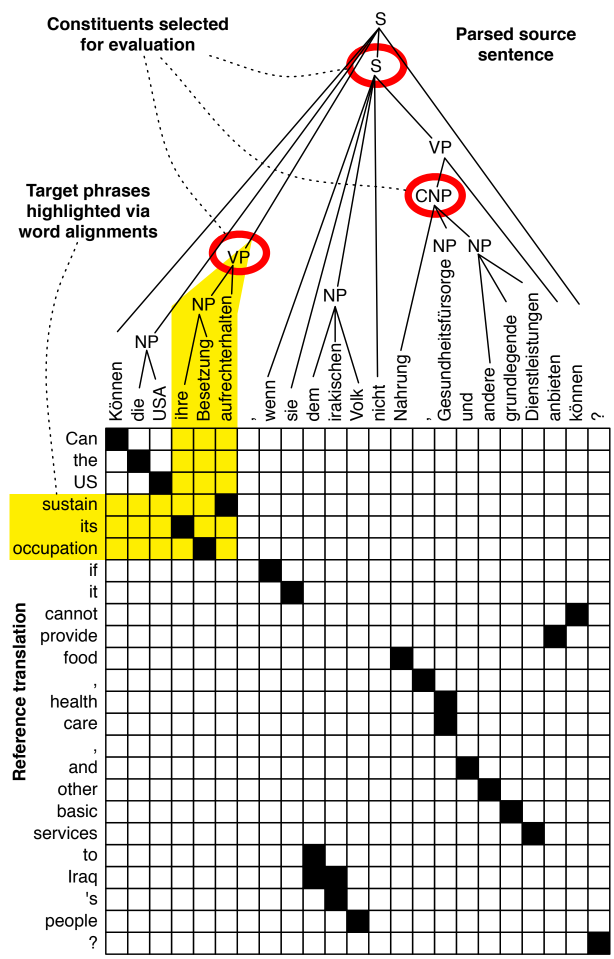

Figure 2: In constituent-based evaluation, the source sentence was parsed, and automatically aligned with the reference translation and systems' translations

[4]It took 26 seconds on average to assign fluency and adequacy scores to a single sentence, and the inter-annotator agreement had a Kappa of between .225–.25, meaning that annotators assigned the same scores to identical sentences less than 40% of the time.

Language Pair	Test Set	Constituent Rank	Yes/No Judgments	Sentence Ranking
English-German	Europarl	2,032	2,034	1,004
	News	2,170	2,221	1,115
German-English	Europarl	1,705	1,674	819
	News	1,938	1,881	1,944
English-Spanish	Europarl	1,200	1,247	615
	News	1,396	1,398	700
Spanish-English	Europarl	1,855	1,921	948
	News	2,063	1,939	1,896
English-French	Europarl	1,248	1,265	674
	News	1,741	1,734	843
French-English	Europarl	1,829	1,841	909
	News	2,467	2,500	2,671
English-Czech	News	2,069	2,070	1,045
	Commentary	1,840	1,815	932
Czech-English	News	0	0	1,400
	Commentary	0	0	1,731
Hungarian-English	News	0	0	937
All-English	News	0	0	4,868
	Totals	**25,553**	**25,540**	**25,051**

Table 3: The number of items that were judged for each task during the manual evaluation. The All-English judgments were reused in the News task for individual language pairs.

3.2 Ranking translations of syntactic constituents

We continued the constituent-based evaluation that we piloted last year, wherein we solicited judgments about the translations of short phrases within sentences rather than whole sentences. We parsed the source language sentence, selected syntactic constituents from the tree, and had people judge the translations of those syntactic phrases. In order to draw judges' attention to these regions, we highlighted the selected source phrases and the corresponding phrases in the translations. The corresponding phrases in the translations were located via automatic word alignments.

Figure 2 illustrates how the source and reference phrases are highlighted via automatic word alignments. The same is done for sentence and each of the system translations. The English, French, German and Spanish test sets were automatically parsed using high quality parsers for those languages (Bikel, 2002; Arun and Keller, 2005; Dubey, 2005; Bick, 2006).

The word alignments were created with Giza++ (Och and Ney, 2003) applied to a parallel corpus containing the complete Europarl training data, plus sets of 4,051 sentence pairs created by pairing the test sentences with the reference translations, and the test sentences paired with each of the system translations. The phrases in the translations were located using standard phrase extraction techniques (Koehn et al., 2003). Because the word-alignments were created automatically, and because the phrase extraction is heuristic, the phrases that were selected may not exactly correspond to the translations of the selected source phrase. We noted this in the instructions to judges:

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

The criteria that we used to select which constituents to evaluate were:

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

The final criterion helped reduce the number of alignment errors, but may have biased the sample to phrases that are more easily aligned.

3.3 Yes/No judgments for the translations of syntactic constituents

This year we introduced a variant on the constituent-based evaluation, where instead of asking judges to rank the translations of phrases relative to each other, we asked them to indicate which phrasal translations were acceptable and which were not.

> Decide if the **highlighted part** of each translation is acceptable, given the reference. This should not be a relative judgment against the other system translations.

The instructions also contained the same caveat about the automatic alignments as above. For each phrase the judges could click on "Yes", "No", or "Not Sure." The number of times people clicked on "Not Sure" varied by language pair and task. It was selected as few as 5% of the time for the English-Spanish News task to as many as 12.5% for the Czech-English News task.

3.4 Collecting judgments

We collected judgments using a web-based tool that presented judges with batches of each type of evaluation. We presented them with five screens of sentence rankings, ten screens of constituent rankings, and ten screen of yes/no judgments. The order of the types of evaluation were randomized.

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

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

In addition to evaluation each language pair individually, we also combined all system translations into English for the News test set, taking advantage of the fact that our test sets were parallel across all languages. This allowed us to gather interesting data about the difficulty of translating from different languages into English.

Table 3 gives a summary of the number of judgments that we collected for translations of individual sentences. We evaluated 14 translation tasks with three different types of judgments for most of them, for a total of 46 different conditions. In total we collected over 75,000 judgments. Despite the large number of conditions we managed to collect between 1,000–2,000 judgments for the constituent-based evaluation, and several hundred to several thousand judgments for the sentence ranking tasks.

4 Translation task results

Tables 4, 5, and 6 summarize the results of the human evaluation of the quality of the machine translation systems. Table 4 gives the results for the manual evaluation which ranked the translations of sentences. It shows the average number of times that systems were judged to be better than or equal to any other system. Table 5 similarly summarizes the results for the manual evaluation which ranked the translations of syntactic constituents. Table 6 shows how many times on average a system's translated constituents were judged to be acceptable in the Yes/No evaluation. The bolded items indicate the system that performed the best for each task under that particular evaluate metric.

Table 7 summaries the results for the All-English task that we introduced this year. Appendix C gives an extremely detailed pairwise comparison between each of the systems, along with an indication of whether the differences are statistically significant.

The highest ranking entry for the All-English task

	BBN-COMBO	CMU-COMBO	CMU-GIMPEL	CMU-SMT	CMU-STATXFER	CU-BOJAR	CU-TECTOMT	CUED	CUED-CONTR	DCU	LIMSI	LIU	LIUM-SYSTRAN	LIUM-SYS-CONTR	MORPHOLOGIC	PC-TRANSLATOR	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-CONTR	SYSTRAN	UCB	UCL	UEDIN	UEDIN-COMBO	UMD	UPC	UW	XEROX
Czech-English Commentary										**.745**														.445			.603		.717			
Czech-English News										**.675**																	.583		.646			
English-Czech Commentary						**.714**	.488									.663											.486					
English-Czech News						.634	.494									**.715**											.502					
English-French Europarl											**.791**		.775					.416	.608	.263	.436	.744	.786			.444	.766					
English-French News											.667		.655					.602	**.780**	.734	.657	.511	.573				.545					.317
English-German Europarl			.612								.584	.581					.615	.583	**.681**	.471	.432	.527				.386	.667					
English-German News											.361	.426					**.787**	.664	.752	.667	.555	.463					.444					
English-Spanish Europarl				.667							**.737**							.554	.560	.413	.436	.717				.500	.714			.593	.735	
English-Spanish News				.494							.537							.683	.674	**.724**	.715	.548			.586		.481			.601		
French-English Europarl					.415			.697		.642	.776		**.792**					.400	.504	.484	.323	.577	.753	.465		.524	.707					
French-English News	.659	.592			.379			.549	.643		.632		.660	**.693**				.581	.575	.654	.565	.540	.639				.614	.608				
German-English Europarl					.364						.485	.614					.627	.596	.610	.543	.537	.677				.416	**.679**					
German-English News	.514				.354						.518	.559					**.742**	.725	.731	.668	.590	.607	.649				.548	.441				
Hungarian-English News															**.853**												.321					
Spanish-English Europarl				.714				.676		.677	**.780**							.427	.488	.350	.470	.671				.425	.660			.687		
Spanish-English News				.567				.563	.674		.583							.667	**.768**	.577	.613	.669			.543		.561			.602		

Table 4: Summary results for the *sentence ranking* judgments. The numbers report the percent of time that each system was judged to be greater than or equal to any other system. Bold indicates the highest score for that task.

was the University of Edinburgh's system combination entry. It uses a technique similar to Rosti et al. (2007) to perform system combination. Like the other system combination entrants, it was tuned on the Europarl test set and tested on the News test set, using systems that submitted entries to both tasks.

The University of Edinburgh's system combination went beyond other approaches by combining output from multiple languages pairs (French-English, German-English and Spanish-English), resulting in 37 component systems. Rather than weighting individual systems, it incorporated weighted features that indicated which language the system was originally translating from. This entry was part of ongoing research in multi-lingual, multi-source translation. Since there was no official multi-lingual system combination track, this entry should be viewed only as a contrastive data point.

We analyzed the All-English judgments to see which source languages were preferred more often, thinking that this might be a good indication of how challenging it is for current MT systems to translate from each of the languages into English. For this analysis we collapsed all of the entries derived from one source language into an equivalence class, and judged them against the others. Therefore, all French systems were judged against all German systems, and so on. We found that French systems were judged to be better than or equal to other systems 69% of the time, Spanish systems 64% of the time, German systems 47% of the time, Czech systems 39% of the time, and Hungarian systems 29% of the time.

We performed a similar analysis by collapsing the RBMT systems into one equivalence class, and the other systems into another. We evaluated how well these two classes did on the sentence ranking task for each language pair and test set, and found that RBMT was a surprisingly good approach in many of the conditions. RBMT generally did better on the News test set and for translations into German, suggesting that SMT's forte is in test sets where it has appropriate tuning data and for language pairs with less reordering than between German and English.

	BBN-COMBO	CMU-COMBO	CMU-GIMPEL	CMU-SMT	CMU-STATXFER	CU-BOJAR	CU-TECTOMT	CUED	CUED-CONTR	DCU	LIMSI	LIU	LIUM-SYSTRAN	LIUM-SYS-CONTR	MORPHOLOGIC	PC-TRANSLATOR	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-CONTR	SYSTRAN	UCB	UCL	UEDIN	UEDIN-COMBO	UMD	UPC	UW	XEROX
Czech-English Commentary																																
Czech-English News																																
English-Czech Commentary						**.594**	.427									.506											.409					
English-Czech News						**.540**	.422									.518											.441					
English-French Europarl											.745		**.843**					.490	.504	.442	.351	.701				.596	.750					
English-French News											.730		**.748**					.589	.593	.640	.576	.591	.586				.633					.302
English-German Europarl			.822								.794	.788					.692	.571	.665	.447	.466	.774				.611	**.849**					
English-German News											.559	.494					.689	.689	**.750**	.553	.598	.544					.518					
English-Spanish Europarl				.804							.872							.582	.598	.635	.600	.806				.714	.888			**.903**	.785	
English-Spanish News				.459							.532							.638	**.759**	.599	.623	.639			.568		.493			.366		
French-English Europarl					.612			.833		.876	.886		.891					.535	.620	.712	.540	.717	.860	.811		.734	**.910**					
French-English News					.554			.736	.788		**.805**		.789	.696				.628	.640	.762	.663	.638	.701				.718					
German-English Europarl					.534						.803	.831					.759	.744	.667	.633	.630	.823				.492	**.856**					
German-English News					.470						.725	.638					.717	.731	**.738**	.589	.684	.669	.716				.632					
Hungarian-English News																																
Spanish-English Europarl				.882				.857		.853	**.902**							.648	.562	.590	.550	.869				.730	.879			.857		
Spanish-English News				.635				.638	.694		.675							.610	.651	.594	.635	.697			.635		.622			**.707**		

Table 6: Summary results for the *Yes/No judgments for constituent translations* judgments. The numbers report the percent of each system's translations that were judged to be acceptable. Bold indicates the highest score for that task.

	BBN-COMBO	CMU-COMBO	CMU-GIMPEL	CMU-SMT	CMU-STATXFER	CU-BOJAR	CU-TECTOMT	CUED	CUED-CONTR	DCU	LIMSI	LIU	LIUM-SYSTRAN	LIUM-SYS-CONTR	MORPHOLOGIC	PC-TRANSLATOR	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-CONTR	SYSTRAN	UCB	UCL	UEDIN	UEDIN-COMBO	UMD	UPC	UW	XEROX
Czech-English Commentary																																
Czech-English News																																
English-Czech Commentary						**.732**	.538									.609											.614					
English-Czech News						.663	.615									**.674**											.610					
English-French Europarl											.876		**.881**					.561	.675	.546	.561	.807				.656	.870					
English-French News											.649		.760					.716	**.768**	.763	.671	.725	.746				.661					.437
English-German Europarl			.774								.750	.812					.577	.585	.582	.508	.518	.770				.690	**.822**					
English-German News											.649	.570					.720	.682	**.748**	.602	.563	.610					.556					
English-Spanish Europarl				.825							**.855**							.561	.592	.458	.573	.849				.592	.818			.775	.790	
English-Spanish News				.721							.694							.694	**.754**	.570	.644	.696			.653		.625			.595		
French-English Europarl					.626			.907		.854	.906		**.917**					.523	.648	.697	.517	.783	.865	.713		.741	.894					
French-English News					.506			.745	.787		**.801**		.765	.780				.652	.655	.726	.615	.640	.735				.773					
German-English Europarl					.554						.752	.795					.580	.640	.643	.579	.587	**.843**				.601	.832					
German-English News					.502						.715	.674					.772	.755	.740	.674	.640	.757	**.775**				.744					
Hungarian-English News																																
Spanish-English Europarl				.847				.846		.868	.854							.455	.561	.469	.567	**.893**				.646	.865			.870		
Spanish-English News				.715				.760	**.818**		.739							.644	.608	.699	.700	.760			.706		.758			.763		

Table 5: Summary results for the *constituent ranking* judgments. The numbers report the percent of time that each system was judged to be greater than or equal to any other system. Bold indicates the highest score for that task.

UEDIN-COMBO$_{xx}$	.717	SAAR$_{fr}$	.584
LIUM-SYSTRAN-C$_{fr}$	.708	SAAR-C$_{de}$	.574
RBMT5$_{fr}$	.706	RBMT4$_{de}$	.573
UEDIN-COMBO$_{fr}$	.704	CUED$_{es}$	.572
LIUM-SYSTRAN$_{fr}$	.702	RBMT3$_{de}$	.552
RBMT4$_{es}$	.699	CMU-SMT$_{es}$	.548
LIMSI$_{fr}$	.699	UCB$_{es}$	.547
BBN-COMBO$_{fr}$	.695	LIMSI$_{es}$	.537
SAAR$_{es}$	.678	RBMT6$_{de}$	.509
CUED-CONTRAST$_{es}$	.674	RBMT5$_{de}$	.493
CMU-COMBO$_{fr}$	.661	LIMSI$_{de}$	.469
UEDIN$_{es}$	.654	LIU$_{de}$	.447
CUED$_{fr}$	.652	SAAR$_{de}$	.445
CUED-CONTRAST$_{fr}$	.638	CMU-STATXFR$_{fr}$	.444
RBMT4$_{fr}$	.637	UMD$_{cz}$	.429
UPC$_{es}$	.633	BBN-COMBO$_{de}$	.407
RBMT3$_{es}$	.628	UEDIN$_{de}$	.402
RBMT2$_{de}$	.627	MORPHOLOGIC$_{hu}$	.387
SAAR-CONTRAST$_{fr}$	.624	DCU$_{cz}$	.380
UEDIN$_{fr}$	.616	UEDIN-COMBO$_{de}$	.327
RBMT6$_{fr}$	.615	UEDIN$_{cz}$	.293
RBMT6$_{es}$	.615	CMU-STATXFER$_{de}$	.280
RBMT3$_{fr}$	.612	UEDIN$_{hu}$	.188

Table 7: The average number of times that each system was judged to be better than or equal to all other systems in the sentence ranking task for the All-English condition. The subscript indicates the source language of the system.

5 Shared evaluation task overview

The manual evaluation data provides a rich source of information beyond simply analyzing the quality of translations produced by different systems. In particular, it is especially useful for validating the automatic metrics which are frequently used by the machine translation research community. We continued the shared task which we debuted last year, by examining how well various automatic metrics correlate with human judgments.

In addition to examining how well the automatic evaluation metrics predict human judgments at the system-level, this year we have also started to measure their ability to predict sentence-level judgments.

The automatic metrics that were evaluated in this year's shared task were the following:

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

- Meteor (Agarwal and Lavie, 2008)—Meteor measures precision and recall for unigrams and applies a fragmentation penalty. It uses flexible word matching based on stemming and WordNet-synonymy. A number of variants are investigated here: meteor-baseline and meteor-ranking are optimized for correlation with adequacy and ranking judgments respectively. mbleu and mter are Bleu and TER computed using the flexible matching used in Meteor.

- Gimenez and Marquez (2008) measure overlapping grammatical dependency relationships (DP), semantic roles (SR), and discourse representations (DR). The authors further investigate combining these with other metrics including TER, Bleu, GTM, Rouge, and Meteor (ULC and ULCh).

- Popovic and Ney (2007) automatically evaluate translation quality by examining sequences of parts of speech, rather than words. They calculate Bleu (posbleu) and F-measure (pos4gramFmeasure) by matching part of speech 4grams in a hypothesis translation against the reference translation.

In addition to the above metrics, which scored the translations on both the system-level[5] and the sentence-level, there were a number of metrics which focused on the sentence-level:

- Albrecht and Hwa (2008) use support vector regression to score translations using past WMT manual assessment data as training examples. The metric uses features derived from target-side language models and machine-generated translations (svm-pseudo-ref) as well as reference human translations (svm-human-ref).

- Duh (2008) similarly used support vector machines to predict an ordering over a set of

[5]We provide the scores assigned to each system by these metrics in Appendix A.

system translations (svm-rank). Features included in Duh (2008)'s training were sentence-level BLEU scores and intra-set ranks computed from the entire set of translations.

- USaar's evaluation metric (alignment-prob) uses Giza++ to align outputs of multiple systems with the corresponding reference translations, with a bias towards identical one-to-one alignments through a suitably augmented corpus. The Model4 log probabilities in both directions are added and normalized to a scale between 0 and 1.

5.1 Measuring system-level correlation

To measure the correlation of the automatic metrics with the human judgments of translation quality at the system-level we used Spearman's rank correlation coefficient ρ. We converted the raw scores assigned each system into ranks. We assigned a ranking to the systems for each of the three types of manual evaluation based on:

- The percent of time that the sentences it produced were judged to be better than or equal to the translations of any other system.
- The percent of time that its constituent translations were judged to be better than or equal to the translations of any other system.
- The percent of time that its constituent translations were judged to be acceptable.

We calculated ρ three times for each automatic metric, comparing it to each type of human evaluation. Since there were no ties ρ can be calculated using the simplified equation:

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

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

	RANK	CONST	YES/NO	OVERALL
meteor-ranking	**.81**	.72	.77	**.76**
ULCh	.68	.79	.82	**.76**
meteor-baseline	.77	.75	.74	.75
posbleu	.77	**.8**	.66	.74
pos4gramFmeasure	.75	.62	.82	.73
ULC	.66	.67	**.84**	.72
DR	.79	.55	.76	.70
SR	.79	.53	.76	.69
DP	.57	.79	.65	.67
mbleu	.61	.77	.56	.65
mter	.47	.72	.68	.62
bleu	.61	.59	.44	.54
svm-rank	.21	.24	.35	.27

Table 8: Average system-level correlations for the automatic evaluation metrics on translations into English

5.2 Measuring consistency at the sentence-level

Measuring sentence-level correlation under our human evaluation framework was made complicated by the fact that we abandoned the fluency and adequacy judgments which are intended to be absolute scales. Some previous work has focused on developing automatic metrics which predict human ranking at the sentence-level (Kulesza and Shieber, 2004; Albrecht and Hwa, 2007a; Albrecht and Hwa, 2007b). Such work generally used the 5-point fluency and adequacy scales to combine the translations of all sentences into a single ranked list. This list could be compared against the scores assigned by automatic metrics and used to calculate correlation coefficients. We did not gather any absolute scores and thus cannot compare translations across different sentences. Given the seemingly unreliable fluency and adequacy assignments that people make even for translations of the same sentences, it may be dubious to assume that their scoring will be reliable across sentences.

The data points that we have available consist of a set of 6,400 human judgments each ranking the output of 5 systems. It's straightforward to construct a ranking of each of those 5 systems using the scores

	RANK	CONST	YES/NO	OVERALL
posbleu	.57	.78	**.80**	**.72**
bleu	.54	**.79**	.6	.64
meteor-ranking	.55	.74	.55	.61
meteor-baseline	.42	.78	.57	.59
pos4gramFmeasure	.37	.49	.79	.55
mter	.54	.50	.55	.53
svm-rank	.55	.56	.46	.52
mbleu	**.63**	.47	.43	.51

Table 9: Average system-level correlations for the automatic evaluation metrics on translations into French, German and Spanish

assigned to their translations of that sentence by the automatic evaluation metrics. When the automatic scores have been retrieved, we have 6,400 pairs of ranked lists containing 5 items. How best to treat these is an open discussion, and certainly warrants further thought. It does not seem like a good idea to calculate ρ for each pair of ranked list, because 5 items is an insufficient number to get a reliable correlation coefficient and its unclear if averaging over all 6,400 lists would make sense. Furthermore, many of the human judgments of 5 contained ties, further complicating matters.

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

6 Evaluation task results

Tables 8 and 9 report the system-level ρ for each automatic evaluation metric, averaged over all trans-

	RANK	CONST	YES/NO
DP	.514	.527	.536
DR	.500	.511	.530
SR	.498	.489	.511
ULC	.559	**.554**	**.561**
ULCh	**.562**	.542	.542
alignment-prob	.517	.538	.535
mbleu	.505	.516	.544
meteor-baseline	.512	.520	.542
meteor-ranking	.512	.517	.539
mter	.436	.471	.480
pos4gramFmeasure	.495	.517	.52
posbleu	.435	.43	.454
svm-human-ref	.542	.541	.552
svm-pseudo-ref	.538	.538	.543
svm-rank	.493	.499	.497

Table 10: The percent of time that each automatic metric was consistent with human judgments for translations into English

lations directions into English and out of English[6] For the into English direction the Meteor score with its parameters tuned on adequacy judgments had the strongest correlation with ranking the translations of whole sentences. It was tied with the combined method of Gimenez and Marquez (2008) for the highest correlation over all three types of human judgments. Bleu was the second to lowest ranked overall, though this may have been due in part to the fact that we were using test sets which had only a single reference translation, since the cost of creating multiple references was prohibitively expensive (see Section 2.1).

In the reverse direction, for translations out of English into the other languages, Bleu does considerably better, placing second overall after the part-of-speech variant on it proposed by Popovic and Ney (2007). Yet another variant of Bleu which utilizes Meteor's flexible matching has the strongest correlation for sentence-level ranking. Appendix B gives a break down of the correlations for each of the lan-

[6]Tables 8 and 9 exclude the Spanish-English News Task, since it had a negative correlation with most of the automatic metrics. See Tables 19 and 20.

	RANK	CONST	YES/NO
mbleu	0.520	0.521	0.52
meteor-baseline	0.514	0.494	0.520
meteor-ranking	**0.522**	0.501	**0.534**
mter	0.454	0.441	0.457
pos4gramFmeasure	0.515	0.525	0.512
posbleu	0.436	0.446	0.416
svm-rank	0.514	**0.531**	0.51

Table 11: The percent of time that each automatic metric was consistent with human judgments for translations into other languages

guage pairs and test sets.

Tables 10 and 11 report the consistency of the automatic evaluation metrics with human judgments on a sentence-by-sentence basis, rather than on the system level. For the translations into English the ULC metric (which itself combines many other metrics) had the strongest correlation with human judgments, correctly predicting the human ranking of a each pair of system translations of a sentence more than half the time. This is dramatically higher than the chance baseline, which is not .5, since it must correctly rank a list of systems rather than a pair. For the reverse direction meteor-ranking performs very strongly. The svn-rank which had the lowest overall correlation at the system level does the best at consistently predicting the translations of syntactic constituents into other languages.

7 Validation and analysis of the manual evaluation

In addition to scoring the shared task entries, we also continued on our campaign for improving the process of manual evaluation.

7.1 Inter- and Intra-annotator agreement

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

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

Evaluation type	$P(A)$	$P(E)$	K
Sentence ranking	.578	.333	.367
Constituent ranking	.671	.333	.506
Constituent (w/identicals)	.678	.333	.517
Yes/No judgments	.821	.5	.642
Yes/No (w/identicals)	.825	.5	.649

Table 12: Kappa coefficient values representing the inter-annotator agreement for the different types of manual evaluation

Evaluation type	$P(A)$	$P(E)$	K
Sentence ranking	.691	.333	.537
Constituent ranking	.825	.333	.737
Constituent (w/identicals)	.832	.333	.748
Yes/No judgments	.928	.5	.855
Yes/No (w/identicals)	.930	.5	.861

Table 13: Kappa coefficient values for intra-annotator agreement for the different types of manual evaluation

where $P(A)$ is the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would agree by chance. We define chance agreement for ranking tasks as $\frac{1}{3}$ since there are three possible outcomes when ranking the output of a pair of systems: $A > B$, $A = B$, $A < B$, and for the Yes/No judgments as $\frac{1}{2}$ since we ignored those items marked "Not Sure".

For inter-annotator agreement we calculated $P(A)$ for the yes/no judgments by examining all items that were annotated by two or more annotators, and calculating the proportion of time they assigned identical scores to the same items. For the ranking tasks we calculated $P(A)$ by examining all pairs of systems which had been judged by two or more judges, and calculated the proportion of time that they agreed that $A > B$, $A = B$, or $A < B$. For intra-annotator agreement we did similarly, but gathered items that were annotated on multiple occasions by a single annotator.

Table 12 gives K values for inter-annotator agreement, and Table 13 gives K values for intra-annotator agreement. These give an indication of how often different judges agree, and how often single judges are consistent for repeated judgments, respectively. The interpretation of Kappa varies, but according to Landis and Koch (1977), $0-.2$ is slight, $.2-.4$ is fair, $.4-.6$ is moderate, $.6-.8$ is substantial and the rest almost perfect. The inter-annotator agreement for the sentence ranking task was fair, for the constituent ranking it was moderate and for the yes/no judgments it was substantial.[7] For the intra-annotator agreement K indicated that people had moderate consistency with their previous judgments on the sentence ranking task, substantial consistency with their previous constituent ranking judgments, and nearly perfect consistency with their previous yes/no judgments.

These K values indicate that people are able to more reliably make simple yes/no judgments about the translations of short phrases than they are to rank phrases or whole sentences. While this is an interesting observation, we do not recommend doing away with the sentence ranking judgments. The higher agreement on the constituent-based evaluation may be influenced based on the selection criteria for which phrases were selected for evaluation (see Section 3.2). Additionally, the judgments of the short phrases are not a great substitute for sentence-level rankings, at least in the way we collected them. The average correlation coefficient between the constituent-based judgments with the sentence ranking judgments is only $\rho = 0.51$. Tables 19 and 20 give a detailed break down of the correlation of the different types of human judgments with each other on each translation task. It may be possible to select phrases in such a way that the constituent-based evaluations are a better substitute for the sentence-based ranking, for instance by selecting more of constituents from each sentence, or attempting to cover most of the words in each sentence in a phrase-by-phrase manner. This warrants further investigation. It might also be worthwhile to refine the instructions given to annotators about how to rank the translations of sentences to try to improve their agreement, which is currently lower than we would like it to be (although it is substantially better than the previous fluency and adequacy scores, which had a $K < .25$ in last year's evaluation).

[7]Note that for the constituent-based evaluations we verified that the high K was not trivially due to identical phrasal translations. We excluded screens where all five phrasal translations presented to the annotator were identical, and report both numbers.

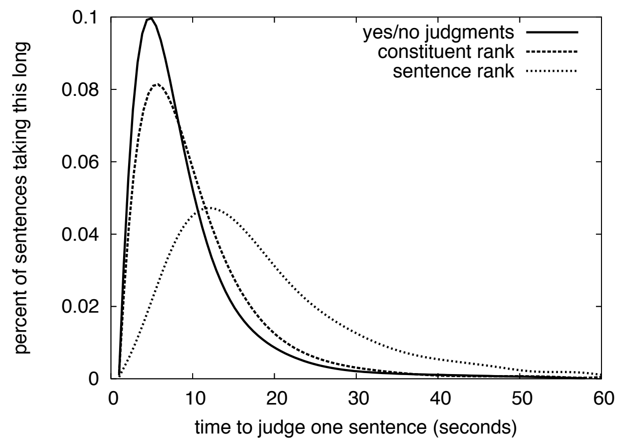

Figure 3: Distributions of the amount of time it took to judge single sentences for the three types of manual evaluation

7.2 Timing

We used the web interface to collect timing information. The server recorded the time when a set of sentences was given to a judge and the time when the judge returned the sentences. It took annotators an average of 18 seconds per sentence to rank a list of sentences.[8] It took an average of 10 seconds per sentence for them to rank constituents, and an average of 8.5 seconds per sentence for them to make yes/no judgments. Figure 3 shows the distribution of times for these tasks.

These timing figures indicate that the tasks which the annotators were the most reliable on (yes/no judgments and constituent ranking) were also much quicker to complete than the ones they were less reliable on (ranking sentences). Given that they are faster at judging short phrases, they can do proportionally more of them. For instance, we could collect 211 yes/no judgments in the same amount of time that it would take us to collect 100 sentence ranking judgments. However, this is partially offset by the fact that many of the translations of shorter phrases are identical, which means that we have to collect more judgments in order to distinguish between two systems.

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

7.3 The potential for re-usability of human judgments

One strong advantage of the yes/no judgments over the ranking judgments is their potential for reuse. We have invested hundreds of hours worth of effort evaluating the output of the translation systems submitted to this year's workshop and last year's workshop. While the judgments that we collected provide a wealth of information for developing automatic evaluation metrics, we cannot not re-use them to evaluate our translation systems after we update their parameters or change their behavior in anyway. The reason for this is that altered systems will produce different translations than the ones that we have judged, so our relative rankings of sentences will no longer be applicable. However, the translations of short phrases are more likely to be repeated than the translations of whole sentences.

Therefore if we collect a large number of yes/no judgments for short phrases, we could build up a database that contains information about what fragmentary translations are acceptable for each sentence in our test corpus. When we change our system and want to evaluate it, we do not need to manually evaluate those segments that match against the database, and could instead have people evaluate only those phrasal translations which are new. Accumulating these judgments over time would give a very reliable idea of what alternative translations were allowable. This would be useful because it could alleviate the problems associated with Bleu failing to recognize allowable variation in translation when multiple reference translations are not available (Callison-Burch et al., 2006). A large database of human judgments might also be useful as an objective function for minimum error rate training (Och, 2003) or in other system development tasks.

8 Conclusions

Similar to 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. One important aspect in which this year's shared task differed from previous years was the introduction of an additional newswire test set that was different in nature to the training data. We also added new language pairs to our evaluation: Hungarian-English and German-Spanish.

As in previous years we were pleased to notice an increase in the number of participants. This year we received submissions from 23 groups from 18 institutions. In addition, we evaluated seven commercial rule-based MT systems.

The goal of this shared-task is two-fold: First we want to compare state-of-the-art machine translation systems, and secondly we aim to measure to what extent different evaluation metrics can be used to assess MT quality.

With respect to MT quality we noticed that the introduction of test sets from a different domain did have an impact on the ranking of systems. We observed that rule-based systems generally did better on the News test set. Overall, it cannot be concluded that one approach clearly outperforms other approaches, as systems performed differently on the various translation tasks. One general observation is that for the tasks where statistical combination approaches participated, they tended to score relatively high, in particular with respect to Bleu.

With respect to measuring the correlation between automated evaluation metrics and human judgments we found that using Meteor and ULCh (which utilizes a variety of metrics, including Meteor) resulted in the highest Spearman correlation scores on average, when translating into English. When translating from English into French, German, and Spanish, Bleu and posbleu resulted in the highest correlations with human judgments.

Finally, we investigated inter- and intra-annotator agreement of human judgments using Kappa coefficients. We noticed that ranking whole sentences results in relatively low Kappa coefficients, meaning that there is only fair agreement between the assessors. Constituent ranking and acceptability judgments on the other hand show moderate and substantial inter-annotator agreement, respectively. Intra-annotator agreement was substantial to almost perfect, except for the sentence ranking assessment where agreement was only moderate. Although it is difficult to draw exact conclusions from this, one might wonder whether the sentence ranking task is simply too complex, involving too many aspects according to which translations can be ranked.

The huge wealth of the data generated by this workshop, including the human judgments, system translations and automatic scores, is available at http://www.statmt.org/wmt08/ for other researchers to analyze.

Acknowledgments

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

We are grateful to Abhaya Agarwal, John Henderson, Rebecca Hwa, Alon Lavie, Mark Przybocki, Stuart Shieber, and David Smith for discussing different possibilities for calculating the sentence-level correlation of automatic evaluation metrics with human judgments in absence of absolute scores. Any errors in design remain the responsibility of the authors.

Thank you to Eckhard Bick for parsing the Spanish test set. See http://beta.visl.sdu.dk for more information about the constraint-based parser. Thanks to Greg Hanneman and Antti-Veikko Rosti for applying their system combination algorithms to our data.

References

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

Joshua Albrecht and Rebecca Hwa. 2007a. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.

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

Joshua Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 187–190, Columbus, Ohio, June. Association for Computational Linguistics.

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

Amittai Axelrod, Mei Yang, Kevin Duh, and Katrin Kirchhoff. 2008. The University of Washington machine translation system for ACL WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 123–126, Columbus, Ohio, June. Association for Computational Linguistics.

Nguyen Bach, Qin Gao, and Stephan Vogel. 2008. Improving word alignment with language model based confidence scores. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 151–154, Columbus, Ohio, June. Association for Computational Linguistics.

Eckhard Bick. 2006. A constraint grammar-based parser for Spanish. In *Proceedings of the 4th Workshop on Information and Human Language Technology (TIL-2006)*, Ribeiro Preto, Brazil.

Dan Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of Second International Conference on Human Language Technology Research (HLT-02)*, San Diego, California.

Graeme Blackwood, Adrià de Gispert, Jamie Brunning, and William Byrne. 2008. European language translation with weighted finite state transducers: The CUED MT system for the 2008 ACL workshop on SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 131–134, Columbus, Ohio, June. Association for Computational Linguistics.

Ondřej Bojar and Jan Hajič. 2008. Phrase-based and deep syntactic English-to-Czech statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

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*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Hélène Bonneau-Maynard, Olivier Galibert, Jean-Luc Gauvain, Philippe Langlais, and François Yvon. 2008. Limsi's statistical translation systems for WMT'08. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 107–110, Columbus, Ohio, June. Association for Computational Linguistics.

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

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2008. Can we relearn an RBMT system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 175–178, Columbus, Ohio, June. Association for Computational Linguistics.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus, Ohio, June. Association for Computational Linguistics.

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

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, June. Association for Computational Linguistics.

Jesus Gimenez and Lluis Marquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June. Association for Computational Linguistics.

Kevin Gimpel and Noah A. Smith. 2008. Rich source-side context for statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 9–17, Columbus, Ohio, June. Association for Computational Linguistics.

Greg Hanneman, Edmund Huber, Abhaya Agarwal, Vamshi Ambati, Alok Parlikar, Erik Peterson, and Alon Lavie. 2008. Statistical transfer systems for French-English and German-English machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 163–166, Columbus, Ohio, June. Association for Computational Linguistics.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, pages 143–152, Budapest, Hungary, May.

Maxim Khalilov, Adolfo Hernández H., Marta R. Costa-jussà, Josep M. Crego, Carlos A. Henríquez Q., Patrik Lambert, José A. R. Fonollosa, José B. Mariño, and Rafael E. Banchs. 2008. The TALP-UPC Ngram-based statistical machine translation system for ACL-WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 127–130, Columbus, Ohio, June. 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 and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the ACL-2007 Workshop on Statistcal Machine Translation (WMT-07)*, Prague, Czech Republic.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, Edmonton, Alberta.

Philipp Koehn, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Alexandra Constantin, Brooke Cowan, Chris Dyer, Marcello Federico, Evan Herbst, Hieu Hoang, Christine Moran, Wade Shen, and Richard Zens. 2007. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University.

Philipp Koehn, Abhishek Arun, and Hieu Hoang. 2008. Towards better machine translation quality for the German-English language pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio, June. Association for Computational Linguistics.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation.

In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 4–6.

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

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

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, Ohio, June. Association for Computational Linguistics.

Vassilina Nikoulina and Marc Dymetman. 2008. Using syntactic coupling features for discriminating phrase-based translations (wmt-08 shared translation task). In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 159–162, Columbus, Ohio, June. Association for Computational Linguistics.

Attila Novák, László Tihanyi, and Gábor Prószéky. 2008. The MetaMorpho translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 111–114, Columbus, Ohio, June. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan.

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

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

Mark Przybocki and Kay Peterson, editors. 2008. *Proceedings of the 2008 NIST Open Machine Translation Evaluation Workshop*. Arlington, Virginia, March.

Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.

Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 183–186, Columbus, Ohio, June. Association for Computational Linguistics.

Holger Schwenk, Jean-Baptiste Fouet, and Jean Senellart. 2008. First steps towards a general purpose French/English statistical machine translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 119–122, Columbus, Ohio, June. Association for Computational Linguistics.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138, Columbus, Ohio, June. Association for Computational Linguistics.

John Tinsley, Yanjun Ma, Sylwia Ozdowska, and Andy Way. 2008. MaTrEx: The DCU MT system for WMT 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 171–174, Columbus, Ohio, June. Association for Computational Linguistics.

Zhuoran Wang and John Shawe-Taylor. 2008. Kernel regression framework for machine translation: UCL system description for WMT 2008 shared translation task. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 155–158, Columbus, Ohio, June. Association for Computational Linguistics.

Zdenek Zabokrtsky, Jan Ptacek, and Petr Pajas. 2008. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, June. Association for Computational Linguistics.

A Automatic scores for each system

	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	SVM-RANK
English-Czech News Commentary Task						
CU-BOJAR	**0.15**	**0.21**	**0.43**	**0.35**	**0.28**	**4.57**
CU-BOJAR-CONTRAST-1	0.04	0.11	0.32	0.25	0.18	0.90
CU-BOJAR-CONTRAST-2	0.14	0.2	0.42	0.34	0.27	2.86
CU-TECTOMT	0.09	0.15	0.37	0.29	0.23	2.13
PC-TRANSLATOR	0.08	0.14	0.35	0.28	0.19	2.09
UEDIN	0.12	0.18	0.4	0.32	0.25	2.28
English-Czech News Task						
CU-BOJAR	**0.11**	**0.18**	**0.37**	**0.3**	**0.18**	**4.72**
CU-BOJAR-CONTRAST-1	0.02	0.10	0.26	0.2	0.12	0.80
CU-BOJAR-CONTRAST-2	0.09	0.16	0.35	0.28	0.15	2.65
CU-TECTOMT	0.06	0.13	0.32	0.25	0.16	2.14
PC-TRANSLATOR	0.08	0.14	0.33	0.26	0.14	2.40
UEDIN	0.08	0.15	0.34	0.27	0.15	2.13

Table 14: Automatic evaluation metric for translations into Czech

	BLEU	MBLEU	METEOR-B	METEOR-R	MTER	POSF4G-AM	POSF4G-GM	POSBLEU	SVM-RANK
English-French News Task									
LIMSI	0.2	0.26	0.16	0.34	0.33	0.48	0.44	0.43	**9.74**
LIUM-SYSTRAN	**0.20**	**0.26**	**0.16**	0.35	**0.34**	**0.49**	**0.44**	**0.44**	7.38
LIUM-SYSTRAN-CONTRAST	0.20	0.26	0.16	**0.35**	0.34	0.48	0.44	0.44	7.02
RBMT1	0.13	0.19	0.12	0.28	0.24	0.42	0.37	0.35	5.46
RBMT3	0.17	0.23	0.14	0.31	0.31	0.45	0.4	0.40	5.60
RBMT4	0.19	0.24	0.15	0.33	0.32	0.48	0.43	0.43	6.80
RBMT5	0.17	0.23	0.14	0.32	0.31	0.47	0.42	0.42	6.15
RBMT6	0.16	0.22	0.13	0.32	0.3	0.46	0.40	0.41	5.60
SAAR	0.15	0.22	0.15	0.33	0.28	0.46	0.41	0.42	6.12
SAAR-CONTRAST	0.17	0.23	0.15	0.33	0.30	0.47	0.42	0.41	5.50
UEDIN	0.16	0.23	0.14	0.32	0.32	0.44	0.39	0.38	4.79
XEROX	0.13	0.2	0.12	0.29	0.29	0.41	0.34	0.34	3.91
XEROX-CONTRAST	0.13	0.2	0.12	0.29	0.29	0.41	0.35	0.35	3.86
English-French Europarl Task									
LIMSI	0.32	0.36	0.24	0.42	0.44	0.56	0.53	0.53	**8.84**
LIUM-SYSTRAN	**0.32**	**0.36**	**0.24**	**0.42**	**0.45**	**0.56**	**0.53**	**0.53**	7.46
LIUM-SYSTRAN-CONTRAST	0.31	0.36	0.23	0.42	0.44	0.56	0.52	0.53	6.69
RBMT1	0.15	0.20	0.13	0.29	0.26	0.44	0.4	0.37	3.89
RBMT3	0.18	0.24	0.15	0.34	0.33	0.47	0.42	0.43	4.13
RBMT4	0.2	0.25	0.17	0.35	0.35	0.5	0.45	0.45	4.70
RBMT5	0.12	0.16	0.09	0.22	0.06	0.37	0.32	0.32	3.01
RBMT6	0.17	0.23	0.14	0.33	0.32	0.47	0.42	0.42	3.93
SAAR	0.26	0.29	0.21	0.41	0.34	0.53	0.49	0.48	7.75
SAAR-CONTRAST	0.28	0.32	0.23	0.41	0.39	0.55	0.51	0.52	6.45
UCL	0.24	0.28	0.19	0.37	0.41	0.49	0.44	0.42	4.16
UEDIN	0.30	0.35	0.23	0.42	0.43	0.54	0.51	0.51	6.56

Table 15: Automatic evaluation metric for translations into French

	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
English-German News Task									
LIMSI	0.11	0.18	0.19	0.45	0.22	0.36	0.29	0.28	**7.83**
LIU	0.10	0.17	0.18	0.44	0.24	0.36	0.28	0.27	4.03
RBMT1	0.12	0.18	0.18	0.44	0.22	0.39	0.33	0.32	5.42
RBMT2	0.13	0.19	**0.20**	**0.46**	0.24	0.4	0.33	0.33	5.76
RBMT3	0.12	0.18	0.19	0.44	0.24	0.39	0.32	0.32	4.70
RBMT4	**0.14**	**0.19**	0.2	0.46	0.25	**0.41**	**0.35**	**0.34**	5.58
RBMT5	0.11	0.17	0.17	0.43	0.21	0.38	0.31	0.31	4.49
RBMT6	0.10	0.16	0.17	0.43	0.2	0.37	0.3	0.29	4.81
SAAR	0.13	0.19	0.19	0.44	**0.27**	0.38	0.31	0.3	4.04
SAAR-CONTRAST	0.12	0.18	0.18	0.43	0.26	0.37	0.3	0.28	3.71
UEDIN	0.12	0.17	0.18	0.45	0.23	0.37	0.30	0.29	4.37
English-German Europarl Task									
CMU-GIMPEL	0.20	0.24	0.27	**0.54**	0.32	0.43	0.37	0.37	**9.54**
LIMSI	0.20	0.24	0.27	0.53	0.32	0.43	0.37	0.37	6.97
LIU	0.2	0.24	0.27	0.53	**0.32**	0.43	0.38	0.37	6.95
RBMT1	0.11	0.16	0.16	0.42	0.19	0.38	0.32	0.32	5.01
RBMT2	0.12	0.17	0.19	0.46	0.21	0.39	0.32	0.31	5.93
RBMT3	0.11	0.16	0.17	0.43	0.21	0.38	0.31	0.30	4.75
RBMT4	0.12	0.17	0.18	0.45	0.22	0.41	0.34	0.33	5.42
RBMT5	0.1	0.14	0.16	0.42	0.19	0.39	0.32	0.31	4.42
RBMT6	0.09	0.14	0.15	0.42	0.18	0.38	0.30	0.29	4.40
SAAR	0.20	0.25	0.26	0.53	0.32	0.43	0.38	0.37	6.67
SAAR-CONTRAST	0.2	0.24	0.26	0.52	0.31	0.43	0.37	0.37	6.35
UCL	0.16	0.20	0.23	0.49	0.31	0.4	0.33	0.31	5.12
UEDIN	**0.21**	**0.25**	**0.27**	0.54	0.32	**0.44**	**0.38**	**0.38**	7.02
English-Spanish News Task									
CMU-SMT	0.19	0.24	0.25	0.34	0.32	0.32	0.25	0.26	**8.34**
LIMSI	0.19	0.25	0.26	0.34	0.34	0.33	0.26	0.26	5.92
RBMT1	0.16	0.22	0.23	0.32	0.30	0.31	0.23	0.23	5.36
RBMT3	0.19	0.24	0.25	0.33	0.34	0.33	0.26	0.26	5.42
RBMT4	**0.21**	0.26	**0.26**	0.34	0.35	0.34	0.28	**0.28**	6.36
RBMT5	0.18	0.24	0.25	0.33	0.32	0.33	0.26	0.26	5.84
RBMT6	0.19	0.24	0.24	0.33	0.33	0.32	0.25	0.26	5.42
SAAR	0.20	**0.27**	0.26	0.34	**0.37**	**0.34**	**0.28**	0.28	5.04
SAAR-CONTRAST	0.2	0.26	0.25	0.34	0.37	0.34	0.27	0.27	4.86
UCB	0.20	0.26	0.26	**0.34**	0.34	0.33	0.26	0.27	5.70
UEDIN	0.18	0.25	0.25	0.33	0.35	0.33	0.26	0.26	4.30
UPC	0.18	0.23	0.24	0.32	0.35	0.32	0.25	0.24	3.97
English-Spanish Europarl Task									
CMU-SMT	**0.32**	0.36	0.33	0.42	0.45	0.40	0.35	0.36	**0.10**
LIMSI	0.31	0.36	0.33	0.42	0.45	0.4	0.35	0.35	7.80
RBMT1	0.16	0.22	0.24	0.32	0.31	0.32	0.25	0.25	4.47
RBMT3	0.20	0.25	0.25	0.34	0.35	0.33	0.27	0.27	4.66
RBMT4	0.21	0.25	0.26	0.34	0.36	0.34	0.28	0.28	4.85
RBMT5	0.18	0.24	0.25	0.34	0.33	0.34	0.27	0.27	5.03
RBMT6	0.18	0.23	0.25	0.33	0.33	0.33	0.26	0.26	4.57
SAAR	0.31	0.35	0.33	0.41	0.44	0.40	0.35	0.35	7.59
SAAR-CONTRAST	0.30	0.34	0.33	0.41	0.44	0.4	0.34	0.35	7.42
UCL	0.25	0.29	0.29	0.37	0.43	0.36	0.29	0.29	4.67
UEDIN	0.32	**0.36**	0.33	0.42	0.45	0.40	0.35	0.35	7.25
UPC	0.30	0.34	0.32	0.40	**0.46**	0.4	0.35	0.34	6.18
UW	0.32	0.36	**0.33**	**0.42**	0.45	0.40	0.35	0.35	7.36
UW-CONTRAST	0.32	0.35	0.33	0.42	0.45	**0.40**	**0.35**	**0.36**	7.21

Table 16: Automatic evaluation metric for translations into German and Spanish

	DP	DR	SR	ULC	ULCH	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
Spanish-English Europarl Task														
CMU-SMT	0.34	**0.44**	0.26	0.29	0.33	**0.33**	0.39	0.59	0.48	0.51	0.52	0.48	0.48	**9.72**
CUED	0.33	0.43	0.25	0.29	0.33	0.32	0.38	0.59	0.48	0.50	0.51	0.47	0.47	7.41
CUED-CONTRAST	0.34	0.44	0.26	0.29	0.33	0.33	0.39	0.59	0.48	0.51	0.52	0.48	0.48	7.00
DCU	0.34	0.43	0.25	0.29	0.33	0.32	0.38	0.59	0.48	0.50	0.51	0.47	0.48	6.78
LIMSI	**0.34**	0.44	0.26	**0.29**	**0.33**	0.33	**0.39**	**0.59**	**0.48**	**0.51**	0.52	0.48	**0.48**	6.73
RBMT3	0.26	0.37	0.19	0.22	0.27	0.19	0.26	0.51	0.41	0.36	0.45	0.4	0.39	5.46
RBMT4	0.26	0.37	0.19	0.22	0.27	0.18	0.26	0.52	0.42	0.36	0.45	0.39	0.38	5.57
RBMT5	0.25	0.36	0.18	0.22	0.27	0.18	0.25	0.51	0.41	0.36	0.44	0.39	0.38	4.74
RBMT6	0.24	0.34	0.18	0.21	0.26	0.17	0.25	0.51	0.41	0.36	0.44	0.38	0.37	4.71
SAAR	0.34	0.44	**0.26**	0.29	0.33	0.32	0.39	0.59	0.48	0.51	**0.52**	**0.49**	0.48	6.30
SAAR-CONTRAST	0.33	0.43	0.25	0.28	0.33	0.30	0.37	0.59	0.48	0.47	0.51	0.47	0.46	7.33
UCL	0.29	0.4	0.21	0.25	0.29	0.25	0.32	0.55	0.43	0.47	0.47	0.42	0.4	4.02
UEDIN	0.34	0.44	0.26	0.29	0.33	0.33	0.39	0.59	0.48	0.50	0.52	0.48	0.48	6.61
UPC	0.33	0.43	0.25	0.28	0.33	0.32	0.38	0.59	0.48	0.5	0.52	0.48	0.48	6.82
French-English News Task														
BBN-COMBO	**0.27**	0.37	**0.2**	**0.23**	**0.28**	0.21	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-COMBO	0.26	0.36	0.18	0.22	0.27	0.19	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-COMBO-CONTRAST	n/a	n/a	n/a	n/a	n/a	0.19	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-STATXFER	0.21	0.32	0.14	0.19	0.23	0.14	0.22	0.48	0.39	0.28	0.38	0.32	0.30	**9.91**
CMU-STATXFER-CONTRAST	0.21	0.30	0.14	0.18	0.23	0.14	0.21	0.47	0.38	0.26	0.38	0.31	0.29	6.47
CUED	0.25	0.35	0.17	0.21	0.26	0.18	0.27	0.51	0.41	0.37	0.41	0.35	0.34	6.34
CUED-CONTRAST	0.26	0.37	0.18	0.22	0.27	0.19	0.28	**0.52**	**0.42**	0.38	0.42	0.37	0.36	6.29
LIMSI	0.26	0.37	0.18	0.22	0.27	0.20	0.28	0.51	0.40	0.40	0.43	0.38	0.37	5.75
LIUM-SYSTRAN	0.27	**0.38**	0.19	0.23	0.27	0.21	0.29	0.51	0.41	0.41	**0.44**	**0.39**	**0.38**	6.32
LIUM-SYSTRAN-CONTRAST	0.27	0.38	0.19	0.23	0.28	**0.21**	**0.29**	0.51	0.41	**0.41**	0.44	0.39	0.38	5.93
RBMT3	0.24	0.36	0.17	0.21	0.26	0.16	0.24	0.49	0.40	0.29	0.42	0.36	0.34	7.61
RBMT4	0.25	0.37	0.17	0.21	0.26	0.17	0.25	0.49	0.4	0.33	0.42	0.36	0.35	6.17
RBMT5	0.25	0.37	0.18	0.22	0.27	0.18	0.25	0.51	0.41	0.33	0.43	0.37	0.36	6.97
RBMT6	0.24	0.36	0.17	0.21	0.26	0.16	0.24	0.49	0.39	0.30	0.41	0.35	0.34	6.51
SAAR	0.24	0.14	0.17	0.19	0.22	0.15	0.24	0.47	0.37	0.39	0.39	0.32	0.31	3.22
SAAR-CONTRAST	0.26	0.36	0.18	0.22	0.27	0.17	0.27	0.51	0.41	0.36	0.41	0.35	0.35	6.01
UEDIN	0.25	0.36	0.17	0.21	0.26	0.18	0.26	0.51	0.41	0.35	0.42	0.36	0.35	5.97
UEDIN-COMBO	0.26	0.36	0.18	0.23	0.27	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
French-English Europarl Task														
CMU-STATXFER	0.24	0.34	0.18	0.22	0.26	0.2	0.26	0.52	0.42	0.37	0.42	0.36	0.35	9.85
CMU-STATXFER-CONTRAST	0.25	0.34	0.19	0.22	0.26	0.2	0.26	0.53	0.42	0.38	0.42	0.36	0.35	7.10
CUED	0.34	0.44	0.26	0.29	0.33	0.32	0.38	0.59	0.48	0.50	0.51	0.47	0.47	**0.11**
CUED-CONTRAST	0.34	0.44	0.26	0.29	0.34	0.32	0.39	0.59	0.48	0.51	0.51	0.47	0.47	9.34
DCU	0.33	0.43	0.25	0.28	0.33	0.31	0.37	0.58	0.47	0.49	0.50	0.46	0.46	9.16
LIMSI	0.34	0.44	0.26	0.29	0.34	0.33	0.39	0.59	0.48	0.51	0.52	0.48	0.48	9.59
LIUM-SYSTRAN	**0.35**	**0.45**	**0.27**	**0.3**	**0.34**	**0.33**	**0.39**	**0.59**	**0.48**	**0.51**	**0.52**	**0.48**	**0.49**	9.75
LIUM-SYSTRAN-CONTRAST	0.34	0.44	0.26	0.29	0.34	0.33	0.39	0.59	0.48	0.50	0.52	0.48	0.48	9.23
RBMT3	0.25	0.36	0.10	0.20	0.24	0.17	0.25	0.51	0.41	0.35	0.43	0.37	0.36	7.36
RBMT4	0.27	0.36	0.19	0.22	0.27	0.18	0.26	0.51	0.41	0.37	0.43	0.38	0.37	5.92
RBMT5	0.27	0.38	0.21	0.23	0.28	0.20	0.28	0.53	0.43	0.4	0.45	0.4	0.39	7.20
RBMT6	0.24	0.35	0.18	0.21	0.26	0.16	0.24	0.5	0.40	0.35	0.42	0.36	0.35	5.96
SAAR	0.32	0.41	0.23	0.27	0.31	0.27	0.33	0.54	0.43	0.49	0.49	0.44	0.41	4.76
SAAR-CONTRAST	0.33	0.43	0.25	0.28	0.33	0.3	0.36	0.58	0.48	0.47	0.51	0.47	0.46	0.10
SYSTRAN	0.3	0.4	0.23	0.26	0.30	0.26	0.34	0.55	0.45	0.46	0.48	0.43	0.43	7.01
UCL	0.3	0.40	0.22	0.26	0.3	0.26	0.32	0.55	0.44	0.47	0.47	0.42	0.41	6.35
UEDIN	0.34	0.44	0.26	0.29	0.33	0.33	0.39	0.59	0.48	0.50	0.52	0.48	0.48	9.41

Table 17: Automatic evaluation metric for translations into English

	DP	DR	SR	ULC	ULCH	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
Czech-English News Commentary Task														
DCU	0.25	**0.34**	0.18	0.22	0.27	0.21	0.29	0.54	0.44	0.42	0.42	0.36	0.36	**2.45**
SYSTRAN	0.19	0.28	0.12	0.17	0.21	0.15	0.23	0.45	0.36	0.34	0.36	0.29	0.29	0.76
UEDIN	0.24	0.31	0.16	0.21	0.25	0.22	0.30	0.54	0.44	0.43	0.41	0.35	0.35	1.37
UMD	**0.26**	**0.34**	**0.19**	**0.23**	**0.28**	**0.24**	**0.33**	**0.56**	**0.45**	**0.49**	**0.44**	**0.39**	**0.38**	1.41
Czech-English News Task														
DCU	0.19	**0.30**	**0.13**	0.17	**0.22**	0.12	**0.22**	**0.45**	**0.35**	0.32	0.36	0.28	**0.28**	**1.78**
UEDIN	0.19	0.28	0.12	0.17	0.21	0.12	0.21	0.44	0.34	0.32	0.35	0.27	0.27	0.65
UMD	**0.2**	0.29	0.12	**0.18**	0.22	**0.13**	0.22	0.44	0.34	**0.36**	**0.36**	**0.29**	0.27	0.52
German-English News Task														
BBN-COMBO	0.23	0.34	0.14	0.21	0.25	**0.18**	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
CMU-STATXFER	0.16	0.27	0.09	0.15	0.19	0.11	0.18	0.43	0.34	0.25	0.33	0.25	0.24	**7.84**
LIMSI	0.22	0.33	0.13	0.19	0.23	0.17	0.25	0.47	0.37	**0.36**	0.4	0.33	0.32	5.58
LIU	0.21	0.32	0.06	0.18	0.22	0.15	0.24	0.48	0.38	0.33	0.38	0.31	0.31	5.51
RBMT1	0.22	0.33	0.14	0.19	0.23	0.14	0.22	0.44	0.35	0.28	0.37	0.31	0.30	6.13
RBMT2	0.24	0.37	0.17	0.21	0.26	0.15	0.24	0.5	0.40	0.31	0.4	0.33	0.32	7.14
RBMT3	0.24	0.37	0.16	0.21	0.26	0.16	0.24	0.49	0.4	0.32	0.41	0.34	0.34	6.97
RBMT4	**0.25**	**0.38**	**0.17**	**0.21**	**0.27**	0.16	0.25	**0.50**	**0.40**	0.34	**0.41**	**0.35**	**0.34**	7.03
RBMT5	0.23	0.36	0.15	0.20	0.25	0.15	0.23	0.48	0.39	0.32	0.4	0.33	0.32	5.94
RBMT6	0.22	0.34	0.14	0.19	0.24	0.14	0.22	0.47	0.38	0.31	0.39	0.32	0.31	5.65
SAAR	0.22	0.33	0.14	0.2	0.24	0.15	0.24	0.47	0.37	0.36	0.39	0.32	0.31	4.67
SAAR-CONTRAST	0.24	0.35	0.16	0.21	0.25	0.17	**0.26**	0.5	0.4	0.36	0.4	0.33	0.33	5.80
SAAR-CONTRAST-2	0.21	0.33	0.14	0.19	0.23	0.15	0.24	0.47	0.37	0.36	0.39	0.32	0.31	4.80
UEDIN	0.23	0.34	0.09	0.19	0.23	0.16	0.25	0.48	0.39	0.35	0.4	0.33	0.33	5.72
German-English Europarl Task														
CMU-STATXFER	0.2	0.31	0.12	0.19	0.22	0.17	0.23	0.49	0.39	0.34	0.39	0.32	0.31	7.11
LIMSI	0.28	0.38	0.18	0.24	0.28	0.27	0.33	0.55	0.44	0.43	0.47	0.42	0.42	**8.04**
LIU	0.28	0.39	0.09	0.23	0.26	0.27	0.33	0.55	0.44	0.44	0.47	0.43	0.43	7.46
RBMT1	0.21	0.3	0.14	0.18	0.22	0.12	0.19	0.42	0.33	0.27	0.36	0.30	0.28	4.61
RBMT2	0.24	0.35	0.16	0.20	0.25	0.14	0.23	0.49	0.39	0.32	0.39	0.33	0.32	5.42
RBMT3	0.24	0.35	0.16	0.20	0.25	0.15	0.23	0.48	0.39	0.32	0.40	0.34	0.33	5.43
RBMT4	0.24	0.36	0.15	0.20	0.25	0.14	0.23	0.49	0.39	0.34	0.41	0.34	0.34	5.11
RBMT5	0.23	0.34	0.15	0.2	0.24	0.14	0.22	0.48	0.38	0.33	0.4	0.33	0.32	4.55
RBMT6	0.22	0.33	0.13	0.18	0.23	0.13	0.21	0.47	0.37	0.31	0.38	0.31	0.31	4.08
SAAR	0.29	0.39	0.19	0.25	0.28	0.27	0.33	0.55	0.44	0.43	0.47	0.42	0.42	7.32
SAAR-CONTRAST	0.28	0.37	0.18	0.24	0.28	0.26	0.32	0.54	0.43	0.43	0.47	0.42	0.42	6.77
UCL	0.24	0.36	0.16	0.22	0.25	0.2	0.25	0.49	0.39	0.41	0.42	0.35	0.32	4.26
UEDIN	**0.30**	**0.41**	**0.20**	**0.26**	**0.3**	**0.28**	**0.34**	**0.56**	**0.45**	**0.45**	**0.48**	**0.44**	**0.44**	7.96
Spanish-English News Task														
CMU-SMT	0.24	0.35	0.17	0.21	0.25	0.18	0.26	0.48	0.38	0.39	0.41	0.35	0.34	**8.00**
CUED	0.25	0.36	0.17	0.21	0.26	0.19	0.28	0.50	0.40	0.38	0.42	0.36	0.36	6.03
CUED-CONTRAST	0.26	0.37	0.18	0.22	0.27	**0.21**	**0.3**	**0.52**	0.42	0.39	0.44	0.38	**0.38**	6.27
LIMSI	0.26	0.37	0.18	0.22	0.27	0.20	0.28	0.50	0.4	**0.41**	0.43	0.38	0.37	4.93
RBMT3	0.25	0.38	0.17	0.22	0.27	0.18	0.26	0.50	0.41	0.32	0.43	0.38	0.36	7.54
RBMT4	0.26	**0.38**	0.18	0.22	0.27	0.18	0.26	0.51	**0.42**	0.32	**0.44**	**0.39**	0.37	7.81
RBMT5	0.26	0.38	0.08	0.20	0.25	0.2	0.27	0.51	0.42	0.33	0.44	0.38	0.37	6.89
RBMT6	0.25	0.36	0.17	0.21	0.26	0.18	0.25	0.51	0.41	0.33	0.43	0.37	0.36	6.83
SAAR	0.26	0.37	0.19	0.22	0.27	0.19	0.29	0.51	0.41	0.39	0.43	0.37	0.37	5.23
SAAR-CONTRAST	0.26	0.37	0.18	0.22	0.27	0.19	0.28	0.51	0.41	0.37	0.42	0.37	0.36	5.95
UCB	0.25	0.35	0.17	0.21	0.26	0.19	0.27	0.5	0.39	0.39	0.42	0.36	0.35	4.40
UEDIN	0.24	0.35	0.17	0.21	0.26	0.18	0.27	0.50	0.40	0.36	0.41	0.35	0.34	5.07
UEDIN-COMBO	**0.27**	0.36	**0.19**	**0.23**	**0.27**	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
UPC	0.25	0.36	0.17	0.21	0.26	0.19	0.26	0.49	0.39	0.4	0.43	0.37	0.36	4.38

Table 18: Automatic evaluation metric for translations into English

B Break down of correlation for each task

	RANK	CONST	YES/NO	DP	DR	SR	ULC	ULCH	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
All-English News Task																	
RANK	1	n/a	n/a	0.83	0.73	0.83	0.83	**0.87**	0.71	0.7	0.82	0.79	0.41	0.79	0.8	0.80	0.25
French-English News Task																	
RANK	1	0.69	0.63	**0.92**	0.83	0.89	0.90	0.90	0.81	0.80	0.88	0.80	0.57	0.87	0.9	0.9	*–0.21*
CONST	—	1	0.81	0.83	0.52	0.81	0.86	0.81	**0.93**	0.9	0.76	0.64	0.73	0.69	0.72	0.85	*–0.52*
YES/NO	—	—	1	0.71	0.57	0.76	0.77	0.74	**0.79**	0.75	0.67	0.59	0.62	0.66	0.67	0.79	*–0.26*
French-English Europarl Task																	
RANK	1	0.95	0.9	0.94	0.95	0.93	**0.95**	0.93	0.92	0.90	0.88	0.87	0.92	0.94	0.94	0.91	0.50
CONST	—	1	0.91	0.97	0.97	**0.98**	0.98	0.97	0.97	0.96	0.97	0.95	0.96	0.97	0.97	0.96	0.56
YES/NO	—	—	1	0.94	0.94	0.94	0.96	0.96	0.96	**0.97**	0.92	0.93	0.92	0.95	0.95	0.97	0.47
German-English News Task																	
RANK	1	0.56	0.56	0.85	0.93	0.92	0.85	**0.95**	0.12	0.09	0.83	0.89	*–0.11*	0.63	0.60	0.58	0.36
CONST	—	1	0.48	0.54	0.48	0.59	**0.66**	0.57	0.64	0.65	0.61	0.55	0.51	0.57	0.63	0.56	*–0.02*
YES/NO	—	—	1	0.68	0.61	0.69	0.73	0.67	0.60	0.41	0.54	0.56	0.33	0.79	**0.83**	0.70	0.08
German-English Europarl Task																	
RANK	1	0.63	0.81	**0.76**	0.59	0.46	0.57	0.60	0.30	0.39	0.40	0.66	0.25	0.53	0.53	0.64	0.35
CONST	—	1	0.78	0.87	**0.92**	0.51	0.83	0.86	0.69	0.69	0.76	0.80	0.69	0.88	0.88	0.88	0.61
YES/NO	—	—	1	**0.88**	0.77	0.48	0.77	0.78	0.66	0.67	0.64	0.86	0.58	0.74	0.74	0.85	0.78
Spanish-English News Task																	
RANK	1	*–0.07*	0.44	0.75	0.76	0.68	0.71	**0.81**	0.19	0.01	0.66	0.63	*–0.12*	0.73	0.76	0.66	0.36
CONST	—	1	0.66	*–0.03*	*–0.44*	0.29	0.29	0.14	0.45	0.66	*–0.11*	*–0.33*	**0.77**	*–0.37*	*–0.34*	0.16	*–0.58*
YES/NO	—	—	1	0.29	0.05	**0.73**	0.64	0.55	0.48	0.47	0.09	*–0.11*	0.71	0.06	0.1	0.39	*–0.43*
Spanish-English Europarl Task																	
RANK	1	0.69	0.76	0.78	0.73	0.73	0.8	0.77	0.78	0.79	0.83	0.84	0.77	0.73	0.73	0.80	**0.87**
CONST	—	1	0.68	0.76	0.77	0.75	0.69	0.73	0.64	0.67	0.64	0.68	0.73	**0.78**	**0.78**	0.73	0.56
YES/NO	—	—	1	0.94	0.93	0.95	0.96	0.95	**0.98**	0.97	0.91	0.91	0.95	0.94	0.94	**0.98**	0.69

Table 19: Correlation of automatic evaluation metrics with the three types of human judgments for translation into English

	RANK	CONST	YES/NO	BLEU	MBLEU	METEOR-BASELINE	METEOR-RANKING	MTER	POSF4GRAM-AM	POSF4GRAM-GM	POSBLEU	SVM-RANK
English-French News Task												
RANK	1	0.55	0.48	**0.73**	0.62	0.3	0.47	0.56	0.69	0.69	0.66	0.72
CONST	—	1	0.35	0.49	0.47	0.39	0.49	0.24	**0.59**	**0.59**	0.58	0.45
YES/NO	—	—	1	0.81	**0.92**	0.71	0.73	0.78	0.73	0.73	0.76	0.76
English-French Europarl Task												
RANK	1	0.98	0.88	0.95	0.95	0.95	0.95	0.90	**0.97**	**0.97**	0.93	0.93
CONST	—	1	0.94	0.98	0.98	0.98	0.98	0.93	**1**	**1**	0.97	0.91
YES/NO	—	—	1	**0.97**	**0.97**	**0.97**	**0.97**	0.92	0.95	0.95	0.92	0.83
English-German News Task												
RANK	1	0.57	0.71	0.58	0.42	0.43	0.13	0.25	**0.90**	**0.90**	**0.90**	0.32
CONST	—	1	0.78	0.75	**0.83**	0.82	0.55	0.60	0.72	0.72	0.72	0.58
YES/NO	—	—	1	0.62	0.54	0.51	0.36	0.23	0.75	0.75	0.75	**0.76**
English-German Europarl Task												
RANK	1	0.28	0.57	0.36	0.36	0.42	0.39	0.26	0.38	0.38	0.50	**0.56**
CONST	—	1	0.87	0.88	0.88	0.91	0.90	**0.93**	0.88	0.88	0.80	0.85
YES/NO	—	—	1	0.89	0.89	0.96	0.96	0.84	0.86	0.86	0.87	**0.98**
English-Spanish News Task												
RANK	1	–	0.49	–	–	–	–	–	–	–	–	**0.02**
		0.30		*0.04*	*0.47*	*0.25*	*0.29*	*0.33*	*0.19*	*0.19*	*0.07*	
CONST	—	1	0.43	**0.79**	0.61	0.64	0.56	0.2	0.59	0.59	0.55	0.56
YES/NO	—	—	1	0.55	0.41	0.43	0.31	0.13	0.65	0.65	**0.72**	0.16
English-Spanish Europarl Task												
RANK	1	0.90	0.63	0.8	0.83	**0.84**	0.83	0.73	0.79	0.79	0.76	0.80
CONST	—	1	0.73	0.84	**0.86**	0.81	0.8	0.74	0.84	0.83	0.84	**0.86**
YES/NO	—	—	1	0.68	0.75	0.66	0.67	**0.90**	0.67	0.66	0.73	0.68

Table 20: Correlation of automatic evaluation metrics with the three types of human judgments for translation into other languages

C Pairwise system comparisons by human judges

The following tables show pairwise comparisons between systems for each language pair, test set, and manual evaluation type. The numbers in each of the tables' cells indicate the percent of 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 $p <= 0.05$ and † indicates statistical significance at $p <= 0.01$, according to the Sign Test.

	BBN-CMB	CMU-CMB	CMU-XFR	CUED	CUED-C	LIMSI	LIUM-SYS	LIUM-SYS-C	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	UEDIN-CMB
BBN-CMB		0.32	0.18†	0.21	**0.42**	**0.37**	0.29	0.24	0.33	**0.48**	**0.48**	0.32	0.29	**0.44**	**0.48**	0.21
CMU-CMB	**0.50**		0.26	0.29	**0.42**	0.4	**0.44**	**0.48**	**0.49**	**0.38**	**0.45**	**0.55**	0.32	0.34	0.34	**0.46**
CMU-XFR	**0.67**†	**0.44**		**0.60***	**0.75**†	**0.58**	**0.73**†	**0.62**	**0.59**	**0.54**	**0.77**†	**0.48**	**0.54**	**0.65***	**0.71**†	**0.58**
CUED	**0.46**	**0.41**	0.20*		**0.47**	**0.56**	**0.47**	**0.51***	**0.41**	**0.54**	**0.57**	**0.37**	**0.43**	**0.61**	0.39	0.15
CUED-C	0.27	0.22	0.08†	0.20		0.31	**0.54**	**0.52***	0.32	**0.52**	**0.50**	0.31	0.40	**0.38**	0.30	**0.52**
LIMSI	0.34	0.4	0.29	0.31	**0.41**		0.23*	**0.52**	0.38	**0.50**	**0.39**	**0.49**	0.42	0.32	0.26	0.30
LIUM-SYS	**0.37**	0.32	0.13†	0.39	0.27	**0.60***		0.24	**0.44**	0.46	**0.46**	0.33	0.24*	0.25	0.30	0.19
LI-SYS-C	**0.40**	0.26	0.24	0.20*	0.13*	0.30	0.24		**0.44**	0.42	0.43	0.35	0.21*	0.30	0.30	0.31
RBMT3	**0.46**	0.43	0.26	0.38	**0.46**	**0.48**	0.39	0.39		0.41	**0.44**	0.26	0.36	**0.50**	**0.68***	**0.44**
RBMT4	0.36	0.33	0.31	0.36	0.39	0.35	**0.50**	**0.45**	**0.45**		**0.49**	**0.40**	0.35	**0.57**	**0.51**	**0.53**
RBMT5	0.37	0.33	0.12†	0.32	0.33	0.33	0.39	**0.46**	0.25	0.22		0.21	0.37	**0.44**	**0.49**	**0.57**
RBMT6	**0.50**	0.33	0.37	0.34	**0.50**	0.39	**0.44**	**0.50**	**0.48**	0.37	**0.55**		**0.42**	**0.48**	0.41	0.41
SAAR	**0.50**	**0.46**	0.37	0.38	**0.44**	**0.52**	**0.6***	**0.54***	**0.44**	**0.53**	**0.44**	0.29		**0.34**	**0.52**	**0.50**
SAAR-C	0.31	**0.47**	0.23*	0.30	0.24	**0.51**	**0.50**	**0.47**	0.25	0.31	0.33	0.35	0.26		0.47	0.38
UED	0.35	**0.37**	0.13†	0.39	**0.55**	**0.50**	**0.50**	**0.43**	0.24*	0.37	0.36	0.41	0.31	0.47		0.36
UED-CMB	**0.57**	0.36	0.16	**0.46**	0.38	0.30	**0.63**	**0.39**	0.39	0.37	0.35	**0.53**	0.27	**0.48**	0.36	
$>$ OTHERS	0.43	0.37	0.22	0.34	0.41	0.44	0.45	0.45	0.4	0.42	**0.47**	0.37	0.34	0.43	0.44	0.42
$\geq$ OTHERS	0.66	0.59	0.38	0.55	0.64	0.63	0.66	**0.69**	0.58	0.58	0.65	0.57	0.54	0.64	0.61	0.61

Table 21: Sentence-level ranking for the French-English News Task.

	CMU-XFR	CUED	DCU	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	SYSTRAN	UCL	UEDIN
CMU-XFR		**0.53**	**0.50**	**0.74**†	**0.79***	**0.55**	0.46	**0.50**	0.36	**0.73**	**0.92**†	0.36	**0.44**	**0.77***
CUED	0.29		**0.42**	**0.29**	**0.48**	0.16†	**0.53**	0.16†	0.18†	0.18*	**0.55**	0.06†	0.21	0.38
DCU	0.46	0.29		0.38	**0.47**	0.37	0.27	0.24	0.29	0.35	**0.55**	0.18	0.25	**0.50**
LIMSI	0.11†	0.21	**0.44**		0.11	0.12†	0.17	0.29	0.05†	0.30	0.32	0.19	0.29	0.33
LIUM-SYS	0.14*	0.16	0.24	**0.32**		0.06†	0.13	0.22	0.12†	0.14†	0.33	0.20*	0.26	0.32
RBMT3	0.36	**0.79**†	**0.58**	**0.88**†	**0.72**†		0.40	**0.57**	0.21	**0.67**	**0.72**†	**0.50**	**0.54**	**0.67**
RBMT4	**0.50**	0.40	**0.64**	**0.67**	**0.56**	0.40		0.42	0.21†	**0.52**	**0.67**	0.33	**0.47**	**0.75**
RBMT5	0.38	**0.79**†	**0.60**	**0.57**	**0.56**	0.24	0.42		0.26	0.48	**0.72**†	**0.50**	**0.46**	**0.60**
RBMT6	**0.54**	**0.79**†	**0.67**	**0.77**†	**0.82**†	**0.47**	**0.79**†	**0.53**		**0.71***	**0.83**†	**0.56**	**0.47**	**0.77**†
SAAR	0.27	**0.59***	**0.57**	**0.47**	**0.71**†	0.22	0.29	0.48	0.18*		**0.50**	0.35	0.23	**0.50**
SAAR-C	0.04†	0.15	0.31	**0.39**	**0.48**	0.14†	0.24	0.21†	0.08†	0.21		0.17†	0.20	**0.57**
SYSTRAN	**0.50**	**0.81**†	**0.65**	**0.52**	**0.64***	0.38	**0.62**	0.33	0.32	**0.41**	**0.71**†		**0.56**	**0.55**
UCL	0.31	**0.64**	**0.56**	**0.57**	**0.47**	0.46	0.40	0.39	0.27	**0.55**	**0.60**	0.44		**0.47**
UED	0.24*	**0.43**	0.35	0.33	**0.42**	0.28	0.25	0.33	0.15†	0.29	0.26	0.25	0.27	
$>$ OTHERS	0.32	0.50	0.5	0.54	0.55	0.28	0.4	0.35	0.21	0.41	**0.59**	0.32	0.35	0.55
$\geq$ OTHERS	0.42	0.7	0.64	0.78	**0.79**	0.40	0.50	0.48	0.32	0.58	0.75	0.47	0.52	0.71

Table 22: Sentence-level ranking for the French-English Europarl Task.

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	XEROX
LIMSI		0.29	0.25	**0.60**†	**0.52**	**0.48**	0.13	0.30	0.13*	0.17*
LIUM-SYSTRAN	**0.36**		**0.41**	**0.51**	**0.41**	**0.53**	0.22	0.26	0.27	0.04†
RBMT3	**0.56**	0.34		**0.48**	**0.52**	**0.40**	0.31	**0.53**	0.37	0.11†
RBMT4	0.13†	0.36	0.31		0.29	0.19*	0.26	0.15†	0.17†	0.09†
RBMT5	0.33	0.35	0.29	**0.42**		0.26	0.17†	0.32	0.17†	0.12†
RBMT6	0.42	0.38	0.37	**0.43***	**0.44**		0.32	0.32	0.28	0.11†
SAAR	**0.56**	**0.52**	**0.51**	**0.56**	**0.69**†	**0.41**		**0.33**	**0.46**	0.3
SAAR-CONTRAST	**0.55**	**0.44**	0.33	**0.63**†	**0.56**	**0.46**	0.21		0.41	0.22*
UEDIN	**0.48***	**0.48**	0.41	**0.60**†	**0.65**†	**0.53**	0.41	**0.43**		0.09†
XEROX	**0.63***	**0.74**†	**0.78**†	**0.74**†	**0.71**†	**0.75**†	**0.44**	**0.64***	**0.63**†	
$>$ OTHERS	0.44	0.43	0.41	**0.54**	0.53	0.43	0.28	0.37	0.32	0.13
$\geq$ OTHERS	0.67	0.66	0.60	**0.78**	0.73	0.66	0.51	0.57	0.55	0.32

Table 23: Sentence-level ranking for the English-French News Task.

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UCL	UEDIN
LIMSI		**0.23**	0.21*	0.32	0.10†	0.15†	0.35	0.27	0.15†	0.17
LIUM-SYSTRAN			0.28	**0.39**	0.11†	0.21*	0.22	0.40	0.19†	0.15
RBMT3	**0.75***	**0.59**		**0.38**	0.39	**0.49**	**0.70**†	**0.81**†	**0.47**	**0.81**†
RBMT4	**0.64**	0.36	0.28		0.24*	0.18	**0.61**	**0.48**	0.42	**0.50**
RBMT5	**0.85**†	**0.89**†	**0.49**	**0.62***		**0.67***	**0.78**†	**0.91**†	**0.63***	**0.93**†
RBMT6	**0.85**†	**0.62***	0.26	**0.42**	0.24*		**0.83**†	**0.82**†	**0.47**	**0.68**†
SAAR	**0.41**	**0.52**	0.17†	0.30	0.11†	0.06†		**0.41**	0.11†	**0.41**
SAAR-CONTRAST	**0.47**	0.40	0.11†	0.26	0.03†	0.06†	0.32		0.27	0.26
UCL	**0.80**†	**0.70**†	0.42	**0.47**	0.22*	0.44	**0.71**†	**0.61**		**0.78**†
UEDIN	**0.46**	**0.41**	0.11†	0.33	0.04†	0.15†	0.32	**0.36**	0.03†	
$>$ OTHERS	**0.62**	0.54	0.26	0.4	0.17	0.27	0.56	0.6	0.32	0.54
$\geq$ OTHERS	**0.79**	0.78	0.42	0.61	0.26	0.44	0.74	0.79	0.44	0.77

Table 24: Sentence-level ranking for the English-French Europarl Task.

	BBN-CMB	CMU-XFR	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	UEDIN-CMB
BBN-COMBO		0.1†	0.22	0.37	**0.62***	**0.69**†	**0.74***	**0.66**†	0.41	**0.63***	**0.60***	**0.35**	**0.40**
CMU-STATXFER	**0.71**†		**0.44**	**0.54**	**0.76**†	**0.79**†	**0.73**†	**0.74**†	**0.80**†	**0.62**†	**0.65**†	**0.54**†	0.37
LIMSI	**0.44**	0.24		**0.41**	**0.67***	**0.65**†	**0.69**†	**0.54**	**0.50**	**0.50**	**0.63**	0.38	0.22
LIU	0.37	0.27	0.34		**0.55***	**0.56**	**0.61**†	**0.50**	**0.45**	**0.48**	**0.56**	**0.32**	0.34
RBMT2	0.21*	0.14†	0.31*	0.20*		0.27	0.43	0.29	0.34	0.30	0.13†	0.25†	0.24*
RBMT3	0.18†	0.13†	0.19†	0.27	**0.56**		0.37	0.33	0.32	0.29	0.29	0.19†	0.17†
RBMT4	0.22*	0.12†	0.17†	0.18†	**0.46**	**0.51**		0.3	0.31	0.18†	0.26*	0.28	0.17†
RBMT5	0.22†	0.12†	0.32	0.36	**0.58**	**0.51**	**0.40**		0.29	0.23*	0.37	0.3	0.28
RBMT6	**0.55**	0.08†	0.40	0.4	**0.51**	**0.51**	**0.47**	**0.51**		**0.49**	**0.52**	0.22*	0.43
SAAR	0.23*	0.21†	0.40	0.39	**0.52**	**0.50**	**0.61**†	**0.53***	0.38		**0.50***	0.26*	0.13*
SAAR-CONTRAST	0.23*	0.19†	0.3	0.37	**0.71**†	**0.37**	**0.60***	0.37	0.33	0.17*		**0.48**	0.13*
UEDIN	0.23	0.13†	0.38	0.3	**0.68**†	**0.65**†	**0.55**	**0.59**	**0.64***	**0.67***	0.38		**0.42**
UEDIN-COMBO	0.35	**0.41**	**0.59**	**0.50**	**0.72***	**0.66**†	**0.83**†	**0.56**	**0.52**	**0.50***	**0.67***	0.38	
$>$ OTHERS	0.32	0.17	0.34	0.35	**0.61**	0.56	0.57	0.49	0.45	0.41	0.46	0.33	0.28
$\geq$ OTHERS	0.51	0.35	0.52	0.56	**0.74**	0.73	0.73	0.67	0.59	0.61	0.65	0.55	0.44

Table 25: Sentence-level ranking for the German-English News Task.

	CMU-XFR	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-STATXFER		**0.57***	**0.77**†	**0.53**	**0.71**†	**0.69**†	**0.50**	**0.58**	**0.82**†	**0.46**	**0.75**†
LIMSI	0.17*		0.35	**0.71***	**0.63**	**0.76***	**0.50**	**0.59**	**0.52**	0.23	**0.67**†
LIU	0.14†	0.35		**0.50**	0.29	**0.67**	0.3	0.42	**0.35**	0.27	**0.57**
RBMT2	0.27	0.24*	0.46		0.39	0.33	0.36	0.42	**0.50**	0.33	**0.46**
RBMT3	0.23†	0.3	**0.57**	**0.45**		0.40	0.31	0.38	**0.56**	0.32	**0.55**
RBMT4	0.22†	0.19*	0.29	**0.50**	**0.48**		0.39	**0.48**	0.41	0.32	**0.61**
RBMT5	0.40	0.40	**0.56**	**0.54**	**0.57**	**0.52**		0.3	**0.48**	0.29*	**0.54**
RBMT6	0.27	0.32	**0.48**	**0.46**	**0.53**	0.44	**0.51**		**0.55**	0.36	**0.61**
SAAR	0.12†	0.19	0.30	0.44	0.41	**0.48**	0.32	0.42		0.20†	0.40
UCL	0.35	**0.54**	**0.46**	**0.63**	**0.61**	**0.68**	**0.68***	**0.61**	**0.63**†		**0.65**†
UEDIN	0.22†	0.17†	0.32	0.42	0.42	0.36	0.41	0.27	0.40	0.23†	
$>$ OTHERS	0.24	0.32	0.46	0.51	0.51	0.53	0.43	0.43	0.53	0.30	**0.58**
$\geq$ OTHERS	0.36	0.49	0.61	0.63	0.6	0.61	0.54	0.54	0.68	0.42	**0.68**

Table 26: Sentence-level ranking for the German-English Europarl Task.

	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UEDIN
LIMSI		**0.44**	**0.8**†	**0.67**†	**0.81**†	**0.76**†	**0.63**†	**0.53**	**0.47***
LIU	0.29		**0.80**†	**0.68**†	**0.81**†	**0.62**†	**0.63**†	0.25	0.31
RBMT2	0.13†	0.07†		0.35	0.33	0.32*	0.20†	0.17†	0.09†
RBMT3	0.18†	0.27†	**0.50**		**0.52**	**0.45**	0.29†	0.26	0.21†
RBMT4	0.09†	0.12†	**0.47**	0.30		**0.42**	0.22†	0.15†	0.17†
RBMT5	0.12†	0.26†	**0.59***	0.42	0.40		0.33	0.28	0.24†
RBMT6	0.25†	0.22†	**0.6**†	**0.61**†	**0.63**†	**0.50**		0.36	0.33
SAAR	0.28	**0.63**	**0.66**†	**0.56**	**0.7**†	**0.62**	**0.46**		**0.45**
UEDIN	0.24*	**0.42**	**0.75**†	**0.66**†	**0.73**†	**0.68**†	**0.51**	0.36	
$>$ OTHERS	0.19	0.28	**0.64**	0.54	0.61	0.54	0.40	0.3	0.27
$\geq$ OTHERS	0.36	0.43	**0.79**	0.66	0.75	0.67	0.56	0.46	0.44

Table 27: Sentence-level ranking for the English-German News Task.

	CMU-GIMPEL	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-GIMPEL		0.29	0.28	0.41	**0.49**	**0.56**	0.44	0.24†	0.09*	0.24*	**0.52**
LIMSI	**0.45**		0.31	0.48	0.45	**0.54**	0.40	0.35	0.40	0.29*	**0.47**
LIU	**0.34**	**0.47**		**0.56**	0.44	**0.65***	0.37	0.30	0.31	0.19†	**0.50**
RBMT2	**0.51**	0.48	0.41		0.41	**0.48**	0.22†	0.24*	**0.62**	0.26*	0.43
RBMT3	0.40	**0.50**	**0.47**	**0.47**		**0.60**	0.33	0.3*	0.11	0.26*	**0.50**
RBMT4	0.39	0.37	0.27*	0.41	0.35		0.22†	0.14†	0.25	0.33	0.46
RBMT5	**0.49**	**0.47**	**0.54**	**0.64**†	**0.60**	**0.64**†		0.32	0.47	0.45	**0.64**†
RBMT6	**0.71**†	**0.50**	**0.58**	**0.57***	**0.65***	**0.74**†	**0.46**		**0.41**	0.36	**0.60**
SAAR	**0.73***	0.40	**0.39**	0.39	**0.78**	**0.58**	0.47	0.35		0.31	**0.50**
UCL	**0.61***	**0.6***	**0.67**†	**0.59***	**0.68***	**0.64**	**0.53**	**0.51**	**0.62**		**0.70**†
UEDIN	0.25	0.27	0.30	**0.52**	0.41	**0.49**	0.26†	0.31	0.25	0.23†	
$>$ OTHERS	0.47	0.43	0.43	0.51	0.51	**0.59**	0.36	0.3	0.37	0.3	0.54
$\geq$ OTHERS	0.61	0.58	0.58	0.62	0.58	**0.68**	0.47	0.43	0.53	0.39	0.67

Table 28: Sentence-level ranking for the English-German Europarl Task.

	CMU-SMT	CUED	CUED-C	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		**0.41**	**0.62***	0.33	**0.54***	**0.57**†	**0.42**	**0.46**	**0.46**	0.29	0.34	**0.37**
CUED	0.29		0.24	0.27	**0.54***	**0.76**†	**0.61***	**0.50**	**0.39**	**0.46**	0.26	**0.42**
CUED-CONTRAST	0.19*	0.24		0.23	**0.47**	**0.48**	0.28	**0.41**	0.37	0.26	0.26	0.33
LIMSI	0.33	**0.30**	**0.51**		0.41	**0.56**†	**0.47**	**0.41**	**0.46**	0.33	**0.37**	**0.43**
RBMT3	0.19*	0.23*	0.37	**0.43**		**0.39**	0.28	0.3	0.33	0.39	0.30	**0.49**
RBMT4	0.19†	0.14†	0.27	0.21†	0.27		0.21†	0.30	0.27	0.17†	0.29*	0.23*
RBMT5	0.37	0.19*	**0.56**	0.35	**0.47**	**0.57**†		**0.56**	**0.43**	0.24*	0.35	**0.52**
RBMT6	0.41	0.30	0.29	0.39	**0.43**	**0.50**	0.25		**0.46**	0.34	**0.44**	**0.46**
SAAR	0.29	0.25	**0.43**	0.32	**0.50**	**0.42**	0.33	0.31		0.2*	0.26	0.3
UCB	0.29	0.36	**0.52**	**0.49**	**0.46**	**0.61**†	**0.6***	**0.41**	**0.56***		**0.39**	0.28
UEDIN	**0.39**	**0.37**	**0.52**	0.30	**0.50**	**0.61***	**0.58**	0.39	**0.46**	0.24		**0.44**
UPC	0.26	0.36	**0.47**	0.35	0.40	**0.59***	0.32	0.42	**0.46**	**0.33**	0.41	
$>$ OTHERS	0.29	0.28	0.43	0.34	0.45	**0.55**	0.39	0.40	0.42	0.29	0.34	0.39
$\geq$ OTHERS	0.57	0.56	0.67	0.58	0.67	**0.77**	0.58	0.61	0.67	0.54	0.56	0.60

Table 29: Sentence-level ranking for the Spanish-English News Task.

	CMU-SMT	CUED	DCU	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC
CMU-SMT		0.36	**0.38**	**0.37**	0.10†	0.20†	0.14†	0.32	0.39	0.22	0.25	0.38
CUED	**0.40**		0.38	**0.53**	0.33	0.30	0.30	0.20†	0.32	0.08†	0.36	0.29
DCU	0.34	0.38		**0.46**	0.32	0.19*	0.26*	0.21*	0.32	0.33	0.25	**0.46**
LIMSI	0.31	0.30	0.21		0.05†	0.09†	0.15†	0.18*	0.24	0.10†	0.19	**0.48**
RBMT3	**0.83**†	**0.62**	**0.58**	**0.73**†		**0.56**	0.25	0.37	**0.60**†	0.31	**0.66***	**0.78**†
RBMT4	**0.73**†	**0.54**	**0.76***	**0.74**†	0.28		0.38	0.24	**0.53**	0.29	**0.56**	**0.65***
RBMT5	**0.79**†	**0.55**	**0.67***	**0.75**†	**0.58**	**0.57**		**0.59***	**0.70**†	**0.44**	**0.71***	**0.67**
RBMT6	**0.52**	**0.77**†	**0.66***	**0.68***	**0.42**	**0.49**	0.18*		**0.55**	0.41	**0.54**	**0.71**
SAAR	**0.43**	**0.42**	**0.41**	**0.47**	0.20†	0.32	0.17†	0.30		0.22*	0.35	**0.32**
UCL	**0.56**	**0.71**†	**0.56**	**0.70**†	**0.42**	**0.57**	0.33	**0.44**	**0.59***		**0.81**†	**0.67**
UEDIN	**0.28**	**0.46**	**0.39**	**0.31**	0.29*	0.42	0.25*	0.39	0.35	0.15†		0.40
UPC	**0.44**	**0.39**	0.43	0.36	0.07†	0.23*	0.24	0.29	0.27	0.20	0.40	
$>$ OTHERS	0.50	0.5	0.49	**0.53**	0.28	0.36	0.24	0.32	0.44	0.26	0.45	0.51
$\geq$ OTHERS	0.71	0.68	0.68	**0.78**	0.43	0.49	0.35	0.47	0.67	0.43	0.66	0.69

Table 30: Sentence-level ranking for the Spanish-English Europarl Task.

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		0.39	**0.57**	**0.52***	**0.62**†	**0.56***	**0.50**	**0.41**	**0.42**	**0.56**†
LIMSI	**0.42**		**0.56**	**0.53**	**0.63***	**0.58**	0.32	0.39	0.35	0.35
RBMT3	0.23	0.3		0.34	**0.46**	**0.50**	0.39	0.17	0.21†	0.06*
RBMT4	0.25*	0.30	**0.47**		0.31	0.35	0.38	0.36	0.32	0.19
RBMT5	0.21†	0.20*	0.28	**0.42**		0.42	0.29*	0.24	0.17†	0.23
RBMT6	0.23*	0.23	0.31	**0.41**	0.42		0.23*	0.19	0.24*	0.24
SAAR	0.36	**0.52**	0.39	**0.43**	**0.67***	**0.54***		0.36	0.29	0.42
UCB	0.37	0.39	**0.52**	0.39	**0.49**	**0.52**	**0.46**		0.27	0.25
UEDIN	0.35	**0.48**	**0.62**†	**0.48**	**0.64**†	**0.61***	**0.50**	**0.47**		**0.53***
UPC	0.11†	**0.41**	**0.63***	**0.48**	**0.50**	**0.57**	0.42	**0.63**	0.06*	
$>$ OTHERS	0.28	0.36	0.47	0.45	**0.52**	0.51	0.38	0.34	0.27	0.33
$\geq$ OTHERS	0.49	0.54	0.68	0.67	**0.72**	0.72	0.55	0.59	0.48	0.60

Table 31: Sentence-level ranking for the English-Spanish News Task.

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC	UW
CMU-SMT		0.28	**0.47**	0.33	0.17†	0.26	**0.50**	0.25	**0.48***	**0.44**	0.28
LIMSI	**0.38**		0.19*	0.33	0.16*	0.23	0.33	0.14†	0.14	0.35	0.32
RBMT3	0.42	**0.62***		0.42	0.36	0.29	**0.54**	0.28	0.39	0.50	**0.75**†
RBMT4	**0.46**	**0.47**	0.42		0.19	0.31	**0.61**	**0.50**	0.40	**0.50**	**0.57**
RBMT5	**0.70**†	**0.64***	**0.59**	**0.48**		0.35	**0.65***	**0.52**	**0.64**	**0.61**	**0.63***
RBMT6	**0.63**	**0.58**	**0.47**	**0.56**	**0.50**		**0.78**†	0.32	**0.58**	0.33	**0.71***
SAAR	0.33	**0.40**	0.33	0.30	0.23*	0.19†		0.20	0.27	0.24	0.33
UCL	**0.46**	**0.64**†	**0.41**	0.46	0.36	**0.41**	**0.60**		**0.65***	**0.42**	**0.57***
UEDIN	0.09*	**0.29**	**0.48**	**0.45**	0.28	0.27	**0.41**	0.19*		0.25	0.17
UPC	0.22	**0.40**	0.50	0.43	0.28	**0.40**	**0.52**	0.26	**0.56**		**0.58**
UW	**0.44**	0.32	0.06†	0.29	0.17*	0.21*	0.33	0.14*	**0.33**	0.33	
$>$ OTHERS	0.43	0.46	0.4	0.4	0.26	0.28	**0.53**	0.28	0.46	0.4	0.49
$\geq$ OTHERS	0.67	**0.74**	0.55	0.56	0.41	0.44	0.72	0.50	0.71	0.59	0.74

Table 32: Sentence-level ranking for the English-Spanish Europarl Task.

	DCU	UEDIN	UMD
DCU		0.26†	0.4
UEDIN	**0.37**†		**0.46**†
UMD	0.4	0.31†	
$>$ OTHERS	0.38	0.28	**0.43**
$\geq$ OTHERS	**0.68**	0.58	0.65

Table 33: Sentence-level ranking for the Czech-English News Task.

	DCU	SYSTRAN	UEDIN	UMD
DCU		0.21†	0.19†	0.37
SYSTRAN	**0.59**†		**0.47**†	**0.61**†
UEDIN	**0.42**†	0.27†		**0.50**†
UMD	**0.38**	0.18†	0.29†	
$>$ OTHERS	0.46	0.22	0.31	**0.49**
$\geq$ OTHERS	**0.75**	0.45	0.60	0.72

Table 34: Sentence-level ranking for the Czech-English Commentary Task.

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.32†	**0.51**†	0.27†
CU-TECTOMT	**0.52**†		**0.58**†	**0.42**
PC-TRANSLATOR	0.35†	0.25†		0.26†
UEDIN	**0.5**†	0.40	**0.59**†	
$>$ OTHERS	0.45	0.32	**0.56**	0.32
$\geq$ OTHERS	0.63	0.49	**0.72**	0.50

Table 35: Sentence-level ranking for the English-Czech News Task.

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.28†	0.38	0.19†
CU-TECTOMT	**0.58**†		**0.53**†	**0.43**
PC-TRANSLATOR	**0.45**	0.3†		0.26†
UEDIN	**0.60**†	0.37	**0.56**†	
$>$ OTHERS	**0.54**	0.32	0.49	0.29
$\geq$ OTHERS	**0.71**	0.49	0.66	0.49

Table 36: Sentence-level ranking for the English-Czech Commentary Task.

	MLOGIC	UEDIN
MORPHOLOGIC		0.15†
UEDIN	**0.68**†	
$>$ OTHERS	**0.68**	0.15
$\geq$ OTHERS	**0.85**	0.32

Table 37: Sentence-level ranking for the Hungarian-English News Task.

	CMU-XFR	CUED	CUED-C	LIMSI	LIUM-SYS	LIUM-SYS-C	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN
CMU-XFR		**0.37**	**0.49**†	**0.62**†	**0.57**†	**0.61**†	**0.49**	**0.49**	**0.48***	**0.41**	**0.56**†	**0.39**	**0.46***
CUED	0.28		**0.21**	**0.30**	**0.30**	0.13	0.28	0.18	**0.27**	0.28	0.31	**0.34**	**0.18**
CUED-C	0.2†	0.11		**0.30***	0.19	**0.33**	0.18†	0.21	0.24	0.2†	0.2*	0.17*	**0.24**
LIMSI	0.13†	0.20	0.13*		**0.27**	0.22	0.23	0.24	0.2	0.20*	0.16*	0.23	0.22
LIUM-SYS	0.18†	0.17	**0.27**	0.17		**0.20**	0.18*	**0.41**	**0.29**	0.24	0.26	0.22	0.26
LI-SYS-C	0.18†	**0.28**	0.24	**0.25**	0.07		0.33	0.2*	0.27	0.18†	0.23	0.25	0.19
RBMT3	0.28	**0.34**	**0.52**†	**0.28**	**0.40***	**0.37**		0.27	**0.46**†	0.27	0.30	**0.39**	**0.34**
RBMT4	0.29	**0.40**	**0.34**	**0.31**	0.39	**0.43***	**0.33**		**0.34**	**0.34**	0.27	**0.41**	**0.31**
RBMT5	0.22*	0.24	**0.34**	**0.3**	0.27	**0.43**	0.14†	0.24		0.13*	0.32	**0.32**	**0.32**
RBMT6	0.3	**0.41**	**0.50**†	**0.39***	**0.33**	**0.58**†	**0.3**	0.33	**0.37***		0.33	**0.52***	**0.37**
SAAR	0.27†	**0.33**	**0.43***	**0.37***	**0.4**	**0.42**	**0.41**	**0.36**	0.32	**0.41**		**0.23**	**0.41**
SAAR-C	0.28	0.32	**0.38***	**0.27**	**0.27**	**0.45**	0.23	0.21	0.20	0.23*	0.18		0.19
UED	0.19*	0.15	0.20	**0.25**	**0.29**	0.19	0.28	0.27	0.19	0.24	0.21	**0.26**	
$>$ OTHERS	0.24	0.27	0.33	0.32	0.32	**0.37**	0.29	0.28	0.30	0.27	0.29	0.31	0.29
$\geq$ OTHERS	0.51	0.75	0.79	**0.80**	0.77	0.78	0.65	0.66	0.73	0.62	0.64	0.74	0.77

Table 38: Constituent ranking for the French-English News Task

	CMU-XFR	CUED	DCU	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	SYSTRAN	UCL	UEDIN
CMU-XFR		**0.42**†	**0.4**†	**0.37***	**0.54**†	0.16*	0.21	**0.41**	0.23	**0.49**†	**0.42**†	0.34	**0.45**	**0.50**†
CUED	0.03†		**0.13**	**0.08**	0.14	0.13†	0.13†	0.08†	0.05†	0.08	0.04	0.15	0.11	0.07
DCU	0.09†	0.08		0.10	0.12	0.06†	0.20	**0.31**	0.16†	0.14	**0.22**	0.13	0.10	0.16
LIMSI	0.1*	0.05	**0.19**		0.05	0.04†	0.08†	0.19	0.11†	0.18	0.09	0.05†	0.05†	
LIUM-SYS	0.03†	0.14	**0.19**	**0.07**		0	0.08*	0.03†	0.05†	0.03†	0.09	0.15	0.14	0.08
RBMT3	**0.44***	**0.61**†	**0.50**†	**0.58**†	**0.56**†		**0.41***	**0.38**	**0.32**	**0.37**	**0.53**†	**0.44**	**0.50***	**0.58**†
RBMT4	**0.39**	**0.44**†	**0.43**	**0.45**†	**0.35***	0.12*		**0.31**	0.23	**0.42**	**0.39**	0.33	**0.32**	**0.35**
RBMT5	0.19	**0.47**†	0.29	**0.35**	**0.37**†	0.18	0.17		0.23	**0.35**	**0.33**	0.19	**0.46**	**0.40**
RBMT6	**0.36**	**0.65**†	**0.54**†	**0.48**†	**0.55**†	0.26	**0.40**	**0.50**		**0.50**†	**0.52**†	**0.47***	**0.60**†	**0.44**
SAAR	0.07†	**0.25**	**0.24**	0.18	**0.37**†	0.23	0.36	0.23	0.12†		0.12	0.23	0.13	**0.37***
SAAR-C	0.09†	**0.18**	0.12	**0.16**	**0.16**	0.09†	0.18	0.2	0.06†	0.12		0.09	0.14	**0.15**
SYSTRAN	0.34	**0.40**	**0.21**	**0.38**†	**0.23**	0.25	**0.36**	**0.22**	0.15*	0.23	**0.28**		0.31	**0.30***
UCL	0.25	**0.34**	**0.28**	**0.31**†	**0.19**	0.11*	0.24	0.23	0.11†	**0.24**	**0.31**	**0.34**		**0.37***
UED	0.10†	**0.10**	0.16	**0.05**	0.08	0.03†	0.15	0.14	0.18	0.07*	0.13	0.07*	0.11*	
$>$ OTHERS	0.2	**0.32**	0.27	0.28	0.28	0.12	0.22	0.25	0.15	0.26	0.27	0.22	0.25	0.28
$\geq$ OTHERS	0.63	0.91	0.85	0.91	**0.92**	0.52	0.65	0.7	0.52	0.78	0.87	0.71	0.74	0.89

Table 39: Constituent ranking for the French-English Europarl Task

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN	XEROX
LIMSI		**0.27**	**0.43**	**0.43**	0.29	**0.53⋆**	0.32	**0.37**	**0.30**	0.14†
LIUM-SYSTRAN	0.09		0.33	**0.36**	0.18	**0.35**	0.16⋆	0.25	0.22	0.13†
RBMT3	0.36	0.33		0.22	**0.31**	0.28	**0.4**	0.26	0.26⋆	0.20†
RBMT4	0.25	0.26	**0.30**		0.23	0.16†	0.28	0.26	0.24	0.13†
RBMT5	**0.31**	**0.33**	0.22	**0.28**		0.17	0.27	0.25	0.23	0.13†
RBMT6	0.26⋆	0.30	**0.31**	**0.38†**	**0.32**		**0.33**	0.36	0.39	0.25⋆
SAAR	0.32	**0.41⋆**	0.35	**0.38**	**0.32**	0.28		**0.14**	0.23	0.11†
SAAR-CONTRAST	0.25	**0.26**	**0.36**	**0.30**	**0.33**	0.36	0.05		0.22	0.13†
UEDIN	0.29	**0.34**	**0.45⋆**	**0.4**	**0.33**	**0.40**	**0.31**	**0.35**		0.13†
XEROX	**0.66†**	**0.55†**	**0.61†**	**0.65†**	**0.58†**	**0.51⋆**	**0.53†**	**0.57†**	**0.45†**	
> OTHERS	0.31	0.34	0.38	**0.38**	0.33	0.33	0.3	0.31	0.29	0.15
≥ OTHERS	0.65	0.76	0.72	**0.77**	0.76	0.67	0.73	0.75	0.66	0.44

Table 40: Constituent ranking for the English-French News Task

	LIMSI	LIUM-SYS	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
LIMSI		**0.14**	0.09†	0.10†	0.24	0.11†	0.13	0.08†	0.12
LIUM-SYSTRAN			0.19†	0.19⋆	0.15	0.12†	0.06	0.06†	0.09
RBMT3	**0.65†**	**0.59†**		**0.33**	**0.43**	0.32	**0.50⋆**	**0.39**	**0.46†**
RBMT4	**0.53†**	**0.47⋆**	0.19		0.27	0.18⋆	**0.33**	**0.38**	**0.39**
RBMT5	**0.48**	**0.38**	0.32	**0.48**		**0.47**	**0.55†**	**0.44**	**0.51†**
RBMT6	**0.54†**	**0.49†**	0.32	**0.41⋆**	0.26		**0.52†**	**0.45**	**0.58†**
SAAR	**0.21**	**0.17**	0.23⋆	0.25	0.21†	0.17†		0.19	**0.13**
UCL	**0.37†**	**0.33†**	0.38	0.35	0.36	0.32	**0.34**		**0.31†**
UEDIN	0.12	**0.11**	0.17†	0.23	0.13†	0.13†	0.07	0.07†	
> OTHERS	**0.38**	0.36	0.25	0.30	0.26	0.24	0.33	0.27	0.34
≥ OTHERS	0.88	**0.88**	0.56	0.68	0.55	0.56	0.81	0.66	0.87

Table 41: Constituent ranking for the English-French Europarl Task

	CMU-XFER	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	SAAR-C	UEDIN
CMU-STATXFER		**0.47†**	**0.44**	**0.52†**	**0.53†**	**0.57†**	**0.49⋆**	**0.41**	**0.49**	**0.58†**	**0.49†**
LIMSI	0.17†		0.18	**0.35**	0.34	**0.40**	**0.33**	**0.43**	0.19	0.28	0.19
LIU	0.25	**0.3**		**0.37**	**0.35**	**0.44**	0.28	**0.40**	0.21	**0.33**	**0.32⋆**
RBMT2	0.19†	0.26	0.30		0.19	**0.32**	0.16⋆	0.20	0.26	**0.23**	0.21
RBMT3	0.22†	**0.36**	0.26	**0.23**		0.24	0.23	0.14†	**0.15**	**0.28**	0.29
RBMT4	0.20†	0.35	0.23	0.21	0.24		0.22	0.19⋆	**0.36**	**0.32**	**0.31**
RBMT5	0.26⋆	0.28	**0.38**	**0.34⋆**	**0.31**	**0.35**		0.26	**0.3**	**0.43†**	**0.35**
RBMT6	0.38	0.37	0.39	**0.34**	**0.44†**	**0.4⋆**	**0.30**		**0.28**	0.26	**0.38**
SAAR	0.29	**0.22**	**0.37**	**0.29**	0.10	0.28	0.19	0.22		**0.26**	0.18
SAAR-CONTRAST	0.18†	**0.33**	0.29	0.19	0.22	0.24	0.15†	0.26	0.18		0.23
UEDIN	0.11†	**0.3**	0.13⋆	**0.23**	**0.35**	0.3	0.2	0.37	**0.30**	**0.31**	
> OTHERS	0.22	0.33	0.3	0.31	0.32	**0.35**	0.25	0.29	0.28	0.33	0.30
≥ OTHERS	0.50	0.72	0.67	0.77	0.76	0.74	0.67	0.64	0.76	**0.78**	0.74

Table 42: Constituent ranking for the German-English News Task

	CMU-XFR	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-STATXFER		**0.51**†	**0.51**†	0.38	**0.38**	**0.41**	0.37	**0.44**	**0.48**†	**0.39**	**0.6**†
LIMSI	0.18†		0.22	0.3	0.30	0.23	0.22†	0.32	**0.27**	0.18*	**0.29**
LIU	0.14†	0.22		0.26*	0.32	0.22*	0.16†	0.31	**0.20**	0.08†	0.12
RBMT2	0.38	**0.51**	**0.52***		**0.40**	**0.32**	0.25	0.31	**0.51**	0.40	**0.7**†
RBMT3	0.32	**0.42**	**0.45**	0.28		**0.46**	0.16	0.20*	**0.56**†	0.38	**0.43**
RBMT4	0.32	**0.45**	**0.52***	0.31	0.24		0.13†	0.30	**0.49**†	**0.44**	**0.48***
RBMT5	**0.44**	**0.57**†	**0.53**†	**0.34**	**0.31**	**0.43**†		0.19	**0.54**†	0.39	**0.54**†
RBMT6	0.33	**0.51**	**0.48**	**0.33**	**0.47***	**0.33**	**0.33**		**0.47***	**0.42**	**0.51***
SAAR	0.12†	0.1	0.15	0.26	0.09†	0.19†	0.17†	0.23*		0.11†	0.14
UCL	0.30	**0.43***	**0.49**†	0.40	**0.40**	0.30	**0.41**	0.39	**0.38**†		**0.51**†
UEDIN	0.11†	0.16	0.12	0.18†	0.25	0.2*	0.18†	0.23*	0.14	0.12†	
$>$ OTHERS	0.27	0.40	0.41	0.31	0.32	0.32	0.25	0.3	0.41	0.30	**0.44**
$\geq$ OTHERS	0.55	0.75	0.8	0.58	0.64	0.64	0.58	0.59	**0.84**	0.60	0.83

Table 43: Constituent ranking for the German-English Europarl Task

	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UEDIN
LIMSI		0.29	**0.46**	**0.45**	**0.37**	0.36	0.29†	0.33	0.22
LIU	**0.32**		**0.53**†	**0.45***	**0.51**†	**0.5***	0.38	**0.31**	**0.36**
RBMT2	0.33	0.32†		0.29	0.29	0.20†	0.25†	0.28	0.28†
RBMT3	0.34	0.3*	**0.4**		**0.33**	0.3*	0.34	0.20*	0.27†
RBMT4	0.26	0.25†	**0.31**	0.3		0.23*	0.23†	0.20*	0.21†
RBMT5	**0.46**	0.33*	**0.55**†	**0.46***	**0.40***		0.32	0.32	0.29†
RBMT6	**0.52**†	**0.40**	**0.47**†	**0.44**	**0.53**†	**0.40**		0.27	0.37
SAAR	**0.38**	0.3	**0.39**	**0.42***	**0.44***	**0.40**	**0.44**		0.34
UEDIN	**0.30**	0.24	**0.53**†	**0.52**†	**0.51**†	**0.56**†	**0.45**	**0.36**	
$>$ OTHERS	0.36	0.31	**0.46**	0.41	0.42	0.37	0.33	0.28	0.29
$\geq$ OTHERS	0.65	0.57	0.72	0.68	**0.75**	0.60	0.56	0.61	0.56

Table 44: Constituent ranking for the English-German News Task

	CMU-GIMPEL	LIMSI	LIU	RBMT2	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN
CMU-GIMPEL		0.12	**0.27**	0.21†	0.30	0.21†	0.27*	0.21†	0.22	0.22	**0.23**
LIMSI	**0.22**		**0.22**	0.34	0.29*	0.29†	0.23†	0.29†	0.2	0.21	**0.19**
LIU	0.18	0.2		0.20†	0.25*	0.17†	0.16†	0.12†	**0.28**	0.21	0.18
RBMT2	**0.54**†	**0.41**	**0.62**†		0.28	0.33	0.35	0.28	**0.61***	**0.43**	**0.47**†
RBMT3	**0.47**	**0.47***	**0.47***	**0.4**		**0.33**	0.32	0.28	**0.56***	**0.47**	**0.48**†
RBMT4	**0.52**†	**0.57**†	**0.52**†	**0.42**	0.32		0.27*	0.28	**0.47**	**0.45**	**0.39**
RBMT5	**0.49***	**0.57**†	**0.65**†	**0.42**	**0.38**	**0.48***		0.31	**0.76**†	**0.51**	**0.52**†
RBMT6	**0.51**†	**0.54**†	**0.60**†	**0.41**	**0.39**	**0.40**	**0.41**		**0.51***	**0.53***	**0.51**†
SAAR	**0.24**	**0.29**	0.17	0.26*	0.22*	0.25	0.20†	0.21*		**0.31**	0.12
UCL	**0.28**	**0.32**	**0.29**	0.33	0.38	0.32	0.32	0.29*	0.19		**0.30**
UEDIN	0.1	0.13	**0.22**	0.2†	0.18†	0.22	0.21†	0.18†	**0.15**	0.17	
$>$ OTHERS	0.37	0.37	**0.42**	0.32	0.30	0.31	0.28	0.25	0.39	0.35	0.35
$\geq$ OTHERS	0.77	0.75	0.81	0.58	0.59	0.58	0.51	0.52	0.77	0.69	**0.82**

Table 45: Constituent ranking for the English-German Europarl Task

	CMU-SMT	CUED	CUED-C	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		0.19	0.17	**0.26**	**0.38**	0.27	**0.45**	0.32	**0.35**	**0.27**	**0.26**	**0.2**
CUED	**0.21**		**0.21**	0.24	0.24	0.2	**0.34**	0.25	0.27	0.18	**0.26**	0.21
CUED-CONTRAST	0.17	0.08		0.12	0.24	0.23*	0.27	0.25	0.21	0.12	0.11	**0.26**
LIMSI	0.17	**0.25**	**0.26**		0.34	0.18†	0.33	0.33	**0.31**	0.17	**0.26**	**0.23**
RBMT3	0.29	**0.31**	**0.35**	**0.37**		0.21	**0.4**	**0.31**	**0.32**	**0.43**	**0.42**	**0.52***
RBMT4	**0.38**	**0.34**	**0.54***	**0.47**†	**0.35**		**0.24**	**0.32**	**0.46**†	**0.37**	**0.40**	**0.53**
RBMT5	0.24	0.31	**0.40**	0.33	0.25	0.18		**0.31**	**0.33**	0.32	0.28	0.38
RBMT6	**0.33**	**0.29**	**0.28**	0.33	0.26	0.27	0.16		**0.26**	0.3	**0.39**	**0.41**
SAAR	0.26	0.27	**0.33**	0.26	0.21	0.12†	0.25	0.24		0.20	**0.28**	0.20
UCB	0.25	**0.30**	**0.23**	**0.27**	0.31	0.27	**0.40**	**0.34**	**0.28**		**0.32**	0.26
UEDIN	0.19	0.20	**0.19**	0.24	0.27	0.33	**0.31**	0.27	0.21	0.21		**0.25**
UPC	0.1	0.21	0.17	0.2	0.22*	0.28	**0.4**	0.24	**0.29**	**0.30**	0.2	
$>$ OTHERS	0.24	0.25	0.28	0.28	0.28	0.23	**0.33**	0.29	0.3	0.26	0.3	0.32
$\geq$ OTHERS	0.72	0.76	**0.82**	0.74	0.64	0.61	0.7	0.70	0.76	0.71	0.76	0.76

Table 46: Constituent ranking for the Spanish-English News Task

	CMU-SMT	CUED	DCU	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC
CMU-SMT		**0.2**	**0.20**	**0.1**	0.1†	0.18†	0.04†	0.18†	**0.16**	0.17	**0.19**	0.19
CUED	0.18		0.13	**0.19**	0.14†	0.12†	0.1†	0.2*	**0.13**	0.12*	**0.22**	0.12
DCU	0.15	0.13		0.11	0.09†	0.10†	0.13†	0.09†	**0.19**	0.15*	0.14	0.15
LIMSI	0.03	0.15	**0.16**		0.19†	0.18†	0.15†	0.19†	**0.19**	0.08†	0.07	**0.22**
RBMT3	**0.7**†	**0.73**†	**0.59**†	**0.49**†		0.19	**0.36**	0.22	**0.62**†	**0.55***	**0.68**†	**0.73**†
RBMT4	**0.55**†	**0.62**†	**0.51**†	**0.55**†	**0.23**		0.22	0.17	**0.56**†	**0.43**	**0.56**†	**0.44***
RBMT5	**0.60**†	**0.61**†	**0.53**†	**0.61**†	0.32	**0.38**		**0.28**	**0.63**†	**0.53**	**0.7**†	**0.59**†
RBMT6	**0.52**†	**0.48***	**0.51**†	**0.49**†	**0.23**	**0.26**	0.19		**0.49**†	**0.53***	**0.52**†	**0.50**†
SAAR	0.14	0.10	0.12	0.15	0.10†	0.12†	0.05†	0.07†		0.14*	0.05	**0.18**
UCL	**0.38**	**0.37***	**0.46***	**0.45**†	0.28*	0.32	0.29	0.24*	**0.38***		**0.38***	**0.36**
UEDIN	0.06	0.14	0.14	**0.18**	0.15†	0.16†	0.05†	0.16†	**0.15**	0.10*		**0.21**
UPC	0.19	0.12	**0.20**	0.12	0.07†	0.17*	0.09†	0.14†	0.04	0.17	0.14	
$>$ OTHERS	0.32	0.33	0.32	0.32	0.17	0.2	0.15	0.17	0.33	0.28	0.34	**0.35**
$\geq$ OTHERS	0.85	0.85	0.87	0.85	0.46	0.56	0.47	0.57	**0.89**	0.65	0.87	0.87

Table 47: Constituent ranking for the Spanish-English Europarl Task

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCB	UEDIN	UPC
CMU-SMT		0.20	**0.36**	**0.37**	0.24†	0.36	0.32	0.21	0.17	0.27
LIMSI	**0.23**		**0.4**	**0.46***	0.33	**0.39**	0.31	0.23	0.17	0.18
RBMT3	0.33	0.35		0.22	0.19†	0.3	0.31	**0.49**	0.34	0.22
RBMT4	0.30	0.25*	**0.25**		0.17*	0.17*	0.24	0.19†	0.34	0.30
RBMT5	**0.53**†	**0.42**	**0.50**†	**0.41***		**0.35**	**0.50***	**0.44**	0.37	0.29
RBMT6	0.36	0.35	**0.34**	**0.39***	0.32		0.35	0.36	0.37	0.38
SAAR	**0.33**	**0.36**	**0.38**	**0.28**	0.24*	**0.38**		0.29	0.22*	0.24
UCB	**0.32**	**0.29**	0.35	**0.54**†	0.33	**0.45**	**0.31**		0.19	**0.29**
UEDIN	**0.29**	**0.33**	**0.36**	**0.42**	**0.42**	**0.39**	**0.45***	**0.30**		**0.44**
UPC	**0.36**	**0.42**	**0.50**	**0.49**	**0.42**	**0.44**	**0.51**	0.21	0.26	
$>$ OTHERS	0.34	0.33	0.38	**0.39**	0.29	0.35	0.36	0.31	0.27	0.29
$\geq$ OTHERS	0.72	0.69	0.69	**0.75**	0.57	0.64	0.7	0.65	0.63	0.6

Table 48: Constituent ranking for the English-Spanish News Task

	CMU-SMT	LIMSI	RBMT3	RBMT4	RBMT5	RBMT6	SAAR	UCL	UEDIN	UPC	UW
CMU-SMT		0.13	0.10†	0.21*	0.2†	0.2†	**0.26**	0.22	0.13	0.16	0.14
LIMSI	**0.17**		0.24	0.16†	0.20†	0.13†	**0.21**	0.06†	0.09	**0.14**	0.08
RBMT3	**0.64**†	**0.45**		0.24	0.30	0.21	**0.57**†	**0.56**	**0.58***	0.32	**0.58**†
RBMT4	**0.54***	**0.52**†	**0.42**		0.26	0.24	**0.50***	0.35	**0.43**	**0.47**	**0.44**
RBMT5	**0.61**†	**0.68**†	**0.46**	**0.44**		**0.37**	**0.64**†	**0.50**	**0.63**†	**0.62**†	**0.54**
RBMT6	**0.57**†	**0.48**†	**0.39**	**0.33**	0.25		**0.52**†	0.33	**0.54**†	**0.46**	**0.46**
SAAR	0.19	0.14	0.07†	0.19*	0.09†	0.14†		0.13†	0.17	0.26	0.18
UCL	**0.43**	**0.46**†	0.29	**0.37**	0.38	**0.42**	**0.49**†		**0.37***	**0.48**	**0.40**
UEDIN	**0.15**	**0.11**	0.24*	0.20	0.13†	0.17†	**0.30**	0.14*		0.20	0.20
UPC	**0.26**	0.05	**0.35**	0.25	0.16†	0.23	**0.34**	0.21	**0.23**		0.10
UW	0.14	**0.14**	0.17†	0.22	0.23	0.2	**0.32**	0.20	0.20	**0.35**	
$>$ OTHERS	0.37	0.32	0.28	0.26	0.22	0.23	**0.42**	0.27	0.35	0.35	0.33
$\geq$ OTHERS	0.83	**0.86**	0.56	0.59	0.46	0.57	0.85	0.59	0.82	0.78	0.79

Table 49: Constituent ranking for the English-Spanish Europarl Task

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.33	**0.41**	0.28*
CU-TECTOMT	**0.37**		**0.42**†	0.36
PC-TRANSLATOR	0.34	0.31†		0.32†
UEDIN	**0.37***	**0.37**	**0.43**†	
$>$ OTHERS	0.36	0.34	**0.42**	0.32
$\geq$ OTHERS	0.66	0.62	**0.67**	0.61

Table 50: Constituent ranking for the English-Czech News Task

	CU-BOJAR	CU-TECTOMT	PC-TRANSLATOR	UEDIN
CU-BOJAR		0.25†	0.33†	0.22†
CU-TECTOMT	**0.50**†		**0.44**†	**0.45**
PC-TRANSLATOR	**0.47**†	0.3†		**0.40**
UEDIN	**0.39**†	0.37	0.39	
$>$ OTHERS	**0.45**	0.31	0.39	0.36
$\geq$ OTHERS	**0.73**	0.54	0.61	0.61

Table 51: Constituent ranking for the English-Czech Commentary Task

French–English

Europarl	Yes	No
CMU-XFR	0.61	0.39
CUED	0.83	0.17
DCU	0.88	0.12
LIMSI	0.89	0.11
LIUM-SYS	0.89	0.11
RBMT3	0.54	0.47
RBMT4	0.62	0.38
RBMT5	0.71	0.29
RBMT6	0.54	0.46
SAAR	0.72	0.28
SAAR-C	0.86	0.14
SYSTRAN	0.81	0.19
UCL	0.73	0.27
UEDIN	**0.91**	**0.09**

News	Yes	No
CMU-XFR	0.55	0.45
CUED	0.74	0.26
CUED-C	0.79	0.21
LIMSI	**0.81**	**0.2**
LIUM-SYS	0.79	0.21
LI-SYS-C	0.7	0.30
RBMT3	0.63	0.37
RBMT4	0.64	0.36
RBMT5	0.76	0.24
RBMT6	0.66	0.34
SAAR	0.64	0.36
SAAR-C	0.70	0.3
UEDIN	0.72	0.28

English–French

Europarl	Yes	No
LIMSI	0.75	0.26
LIUM-SYS	**0.84**	**0.16**
RBMT3	0.49	0.51
RBMT4	0.50	0.5
RBMT5	0.44	0.56
RBMT6	0.35	0.65
SAAR	0.70	0.3
UCL	0.6	0.40
UEDIN	0.75	0.25

News	Yes	No
LIMSI	0.73	0.27
LIUM-SYS	**0.75**	**0.25**
RBMT3	0.59	0.41
RBMT4	0.59	0.41
RBMT5	0.64	0.36
RBMT6	0.58	0.42
SAAR	0.59	0.41
SAAR-C	0.59	0.41
UEDIN	0.63	0.37
XEROX	0.30	0.7

German–English

Europarl	Yes	No
CMU-XFER	0.53	0.47
LIMSI	0.80	0.2
LIU	0.83	0.17
RBMT2	0.76	0.24
RBMT3	0.74	0.26
RBMT4	0.67	0.33
RBMT5	0.63	0.37
RBMT6	0.63	0.37
SAAR	0.82	0.18
UCL	0.49	0.51
UEDIN	**0.86**	**0.14**

News	Yes	No
CMU-XFER	0.47	0.53
LIMSI	0.73	0.28
LIU	0.64	0.36
RBMT2	0.72	0.28
RBMT3	0.73	0.27
RBMT4	**0.74**	**0.26**
RBMT5	0.59	0.41
RBMT6	0.68	0.32
SAAR	0.67	0.33
SAAR-C	0.72	0.28
UEDIN	0.63	0.37

English–German

Europarl	Yes	No
CMU-GIMPEL	0.82†	0.18
LIMSI	0.79†	0.21
LIU	0.79†	0.21
RBMT2	0.69†	0.31
RBMT3	0.57	0.43
RBMT4	0.67†	0.34
RBMT5	0.45	0.55
RBMT6	0.47	0.53
SAAR	0.77†	0.23
UCL	0.61†	0.39
UEDIN	**0.85**†	**0.15**

News	Yes	No
LIMSI	0.56	0.44
LIU	0.49	0.51
RBMT2	0.69	0.31
RBMT3	0.69	0.31
RBMT4	**0.75**	**0.25**
RBMT5	0.55	0.45
RBMT6	0.6	0.40
SAAR	0.54	0.46
UEDIN	0.52	0.48

Spanish–English

Europarl	Yes	No
CMU-SMT	0.88	0.12
CUED	0.86	0.14
DCU	0.85	0.15
LIMSI	**0.90**	**0.1**
RBMT3	0.65	0.35
RBMT4	0.56	0.44
RBMT5	0.59	0.41
RBMT6	0.55	0.45
SAAR	0.87	0.13
UCL	0.73	0.27
UEDIN	0.88	0.12
UPC	0.86	0.14

News	Yes	No
CMU-SMT	0.64	0.37
CUED	0.64	0.36
CUED-C	0.69	0.31
LIMSI	0.68	0.33
RBMT3	0.61	0.39
RBMT4	0.65	0.35
RBMT5	0.59	0.41
RBMT6	0.64	0.37
SAAR	0.7	0.30
UCB	0.64	0.37
UEDIN	0.62	0.38
UPC	**0.71**	**0.29**

English–Spanish

Europarl	Yes	No
CMU-SMT	0.80	0.2
LIMSI	0.87	0.13
RBMT3	0.58	0.42
RBMT4	0.6	0.40
RBMT5	0.64	0.37
RBMT6	0.60	0.40
SAAR	0.81	0.19
UCL	0.71	0.29
UEDIN	0.89	0.11
UPC	**0.90**	**0.1**
UW	0.79	0.22

News	Yes	No
CMU-SMT	0.46	0.54
LIMSI	0.53	0.47
RBMT3	0.64	0.36
RBMT4	**0.76**	**0.24**
RBMT5	0.6	0.40
RBMT6	0.62	0.38
SAAR	0.64	0.36
UCB	0.57	0.43
UEDIN	0.49	0.51
UPC	0.37	0.63

English–Czech

Commentary	Yes	No
CU-BOJAR	**0.59**	**0.41**
CU-TECTO	0.43	0.57
PC-TRANS	0.51	0.49
UEDIN	0.41	0.59

News	Yes	No
CU-BOJAR	**0.54**	**0.46**
CU-TECTO	0.42	0.58
PC-TRANS	0.52	0.48
UEDIN	0.44	0.56

Table 52: Yes/No Acceptability of Constituents

	BBN-CMB-DE	BBN-CMB-FR	CMU-CMB-FR	CMU-SMT-ES	CMU-XFR-DE	CMU-XFR-FR	CUED-C-ES	CUED-C-FR	CUED-ES	CUED-FR	DCU-CZ	LIMSI-DE	LIMSI-ES	LIMSI-FR	LIU-DE	LIUM-S-C-FR	LIUM-SYS-FR	MLOGIC-HU	RBMT2-DE	RBMT3-DE	RBMT3-ES	RBMT3-FR	RBMT4-DE	RBMT4-ES	RBMT4-FR	RBMT5-DE	RBMT5-ES	RBMT5-FR	RBMT6-DE	RBMT6-ES	RBMT6-FR	SAAR-C-DE	SAAR-C-FR	SAAR-DE	SAAR-ES	SAAR-FR	UCB-ES	UED-CMB-DE	UED-CMB-FR	UED-CMB-XX	UED-CZ	UED-DE	UED-ES	UED-FR	UED-HU	UMD-CZ	UPC-ES
BBN-CMB-DE		**.50**	.40	**1**	.20	.50	**1**⋆	**.64**	**1**	**.73**	.31		**.69**⋆	**.71**	.38	**.70**	**.60**	**.60**	**.80**	**.77**⋆	**.60**	**.63**	**.89**⋆	**1**	**.57**	**.62**	**.83**	**.60**	.17	**.57**	**.55**	.41	**.70**	**.58**	**.71**	**.82**	**.75**	**.40**	.33	**1**	.25	**.36**	**.85**⋆	.50		.40	**.60**
BBN-CMB-FR	.38		.14	.38	.09⋆	.13⋆	.33	**.63**	.20	.25	.13	.13	**.60**	.31	.46	**.43**	.27	.13	**.67**	.25	.46	.33	.38	.22	**.43**	.07⋆	.33	.42	**.50**	.36	.25	.46	.40	.06†	.30	.33	**.50**			**.80**	.14⋆	.20	**.67**	.33	.25	.13	.42
CMU-CMB-FR	**.60**	**.71**		**.54**	.09⋆	**.60**	.29	.13	**.57**	**.33**	.23	.33	.33	**.46**	**.44**	**.58**	**.40**	.20	**.54**	.27	.50	**.67**	.11	.14	**.44**	.11	.25	**.60**	.09⋆	.40	.29		.29	.25	**.56**	.20	**.56**	.25	.14	.38	.11⋆	.11	.22			.36	.44
CMU-SMT-ES		**.50**	.31		.50	.17	**.75**	**.46**	**.64**	.43	.25	**.54**	**.60**	**.83**⋆	.40	.50	.17	.14	.46	.50	**.64**	**.73**	**.80**	**.67**	**.64**	.33	**.33**	**.67**	**.46**	.50	**.57**	**.50**	.39	.36	**.64**	**.70**	.17		.50	.33	.14		.25	.33	.13	.38	**.43**
CMU-XFR-DE	**.60**	**.82**⋆	**.91**⋆	.50		**.78**	**.56**	**.89**⋆	**.42**	**.73**⋆	**.55**	.27	**.33**	**.88**⋆	**.57**	**.73**†	**.92**†	**.75**	**.80**†	**.82**	**.75**⋆	**.67**	**.75**	**.86**	**.78**	**.91**⋆	**.89**⋆	**.79**⋆	**.81**†	**.80**	**.80**⋆	**.67**	**.90**†	**.64**	**.73**†	**.80**	**.64**	.33	**1**	**.83**	.11	**.20**	**1**†	**.90**†	.33	.50	**.85**⋆
CMU-XFR-FR	.50	**.75**⋆		**.67**	.11		**.70**	**.80**⋆	**.88**⋆	**.71**	.50		**.75**	**.50**	**.60**	**.71**	**.67**	.50	**.67**	**.60**	.40	**.43**	**.60**	**.67**	.29	.25	**.64**	**.75**	.38	**.75**	.38	.50	**.67**	.18	**.57**	.44	**.73**	.33	**.50**	**.75**		**.80**	**.69**	**.64**	.50	.33	**1**
CUED-C-ES	0	**.56**	**.59**		.22	.20		.18	.21	.19	0	.29	.15	**.47**	.14	.39	.50	.25	.39	.36	**.43**	**.46**	.33	.31	**.56**	**.50**	.07†	**.73**⋆	.31	**.42**	.42	.43	.50	.42	.40	.27	.18		.50	**.38**	.29	.10⋆	.22	.33		**.43**	.33
CUED-C-FR	.29	.13	**.38**	.39	.11⋆	.10⋆	**.73**		**.50**	.25	.36	**.57**	.40	.36	.11⋆	**.70**⋆	**.60**	.40	**.58**	.36	**.56**	.20	.39	**.50**	**.60**	.10⋆	.50	**.50**	.30	**.55**		.46	.33	0	.17	.20	.39	.13	**.88**		0	.29	.39	.36	0	.40	**.50**
CUED-ES		**.80**	.29	.18	.25	0	**.29**	.38		**.64**	.25	.20	.14	**.78**	.25	.36	**.88**	.25	.36	**.39**	**.69**	**.71**	**.58**	**.83**⋆	**.67**	.30	**.50**	**.60**	.47	**.67**	.43	.40	.20	.38	**.50**	.50	**.57**	.25	.50	.50	.08†	0		**.57**	.20	.50	**.67**
CUED-FR	.18	.25	.22	.43	.09⋆	.29	**.69**	**.38**	.27			.11	.10⋆	**.47**	.33	**.64**	.15	.50	.38	**.57**	.50	**.42**	**.43**	.33	**.50**	.22	.46	**.46**	.33	**.58**	**.43**	**.50**	**.56**	.18	.44	.25	.38		.20	.25	0	.33	.13	.44	0	.10⋆	**.50**
DCU-CZ	**.39**	**.75**	**.69**	**.67**	.36	.50	**.90**†	**.46**	**.58**	**1**		**.44**	.22	**.91**⋆	**.56**	**.60**	**.85**⋆	**1**	**.77**⋆	**.78**	**.86**	**.75**	**.62**	**.57**	**.83**⋆	.30	**.55**	**.80**	**.67**	**.77**⋆	**.80**⋆	**.79**⋆	**.50**	.33	**.80**⋆	**.89**†	**.73**	.17	**.50**	**.60**	**.50**	**.54**	**.78**	**.80**	.13	**.38**	.39
LIMSI-DE	**.17**	**.63**	**.67**	.39	.27	**1**	**.71**	.43	**.80**	**.78**	.38		.33	**.57**	**.50**	**.77**⋆	**1**⋆	.29	.50	**.78**	**.50**	**.67**	**.71**	**.88**	**.71**⋆	.33	**.57**	**.89**⋆	.30	**.60**	**.80**	.43	**.78**⋆	.27	.36	.17	.44		**1**	**1**	.13	**.50**	**.67**	**.50**	.40	.30	.43
LIMSI-ES	.08⋆	.40	**.67**	.30	.17	.25	**.54**	**.60**	**.57**	**.80**⋆	**.56**	**.50**		**.50**	.43	**.55**	**.50**	.33	**.43**	.10⋆	**.67**	**.50**	**.63**	**.39**	**.69**	.29	**.75**	**.50**	.29	**.60**	**.82**	**.63**	.20	.22	**.55**	.33	.29	.25	**1**	**.75**	.17	.25	**.57**	**.50**		**.64**	**.50**
LIMSI-FR	.14	**.38**	.18	.08⋆	0	.40	.27	.36	.11	.35	.09⋆	.29	.25		.23	**.63**	.30	.25	.38	**.56**	.36	.44	.22	.25	**.50**	.10⋆	.31	.20	.20	**.56**	.40	.17	.20	.38	.50	.36	.33	**1**	**.33**	**.50**	.25	.08†	.27	.24	.25	.29	.50
LIU-DE	**.50**	**.55**	.33	**.60**		.40	**.86**	**.89**⋆	**.38**	**.44**	.22	.25	**.57**	**.54**		**.73**⋆	**.80**	**1**	.22	**.55**	**.86**	**.83**	**.67**	**.67**	**.67**	.25	**.89**⋆	**.71**	**.50**	**.60**	**.58**	**.60**	**.60**	**.60**	**.75**	**.71**	**.67**	.33	**1**	**1**	.22	.22	**.43**	**.67**		.40	**.69**
LIUM-S-C-FR	.20	.29	.17	.50	0	.14	**.46**	0	**.43**	.21	.20	.08⋆	.18	.13	.09⋆		.25	.11⋆	.18	.39	.50	.27	.27	.46	**.50**	.13⋆	.31	**.55**	.33	.46	**.50**	.42	.25	.33	**.59**⋆	.33	.33	.33	.50			.25	.22	.18	0	.44	**.60**
LIUM-S-FR	.20	**.36**	.20	**.67**	.08†	.11	.50	.20	.13	**.62**	.15⋆	0	.38	**.50**	.20	.25		.14	.42	.36	.17	.43	.13	**.60**	.30	.25	.33	**.52**	.25	.18	.43	.39	.29	.20	.44	.16⋆	.44		**.50**	**.60**	.08⋆	.17⋆	.23	.17	0	.16†	.46
MLOGIC-HU	.40	**.75**	**.60**	**.71**	.25	.50	**.63**	**.60**	**.75**	.50		**.43**	**.67**	**.50**		**.89**⋆	**.71**		**.88**	**.67**	.44	**.86**⋆		**1**⋆	.50	**.75**	**.67**	**.83**	**.63**	**.63**	**.54**	**.63**	**.67**	**.50**	**.86**	.33	**.63**	.33	**.75**	**1**	.40	**1**	**1**	.44	.25		**.80**
RBMT2-DE		.33	.39	.46	.07†	.33	**.46**	.33	**.64**	.38	.08⋆	.50	.36	**.50**	**.33**	**.55**	**.58**	.13		.17	**.67**	.38	.38	**.70**	**.55**	.22	**.46**	**.46**	**.46**	**.43**	.17	.10	.42	.43	**.67**	.29	.33	.40	.40	**1**	.10⋆	.31	**.54**	.36	.14	.07†	**.56**
RBMT3-DE	.08⋆	**.75**	**.64**	.50	.18	.20	**.64**	**.64**	.31	.43	.22	.11	**.80**⋆	.44	.36	**.62**	**.64**	.33	**.67**		**.55**	**.46**	.35	**.90**⋆	**.40**	.14	**.80**⋆	.40	.38	.38	**.60**	.25	**.53**	.44	.31	**.56**	**.63**	.17	**.80**	**.60**		.22	**.55**	**.60**	.17	.20	**.67**
RBMT3-ES	.40	**.55**	.50	.18	.13⋆	**.50**	.36	.22	.23	.50	.14	.42	.33	**.50**	.14	.50	**.83**	.44	.33	.27		.39	**.64**	.50	.50		.36	.33	.31	.27	**.46**	.33	.09	**.43**	.23	**.50**	.46	.29	**.75**	**.75**	.25	.25	.36	.50	0	.20	**.78**
RBMT3-FR	.25	**.58**	.22	.27	.33	.29	.27	**.40**	.29	.33	.13	.33	.43	**.50**	.17	**.55**	**.50**	0	**.56**	.31	**.62**		**.75**	**.63**	.33	.13	.44	.38	.27	**.60**	.09⋆	**.57**	**.63**	.17	**.50**	.38	.40			**1**	.11	.25	**.73**	**.50**		.25	**.53**
RBMT4-DE	.11⋆	**.50**	**.78**	.20		.40	**.67**	**.62**	.25	.14	.15	.29	.13	**.78**	.25	**.73**	**.75**		**.50**	**.55**	.18			.22	.43	**.50**	**.57**	.29	.25	**.50**	**.63**	.42	.33	.25	**.67**	**.58**	.50		**.60**	**.67**	.14	.38	.50	**.75**	0	.22	**.63**
RBMT4-ES		**.44**	**.57**	.22	.14	.17	**.38**	.38	.08⋆	**.56**	.14	.13	.31	**.50**	.33	.46	.30	0	.10	.10⋆	.50	.38	**.56**		.29	.43	.25	.33	.38	.25	.25	.46	.33	.14	.09	.25	.33	.11	**1**	**1**	0	.09⋆	.39	**.50**	0	.38	.38
RBMT4-FR	.43	.29	.22	.27	.11	**.57**	.22	.27	.33	.33	.08⋆	.14⋆	.23	.33	.22	.25	**.60**	.50	.46	.30	.50	**.47**	**.57**	**.41**		**.67**	.46	**.58**	**.54**	.43	.38	.36	**.63**	.14	**.86**	.36			.25	**.60**	.39	**.50**	**.50**	.33	.20		.38
RBMT5-DE	.23	**.71**⋆	**.67**	**.50**	.09⋆	**.50**	.38	**.80**⋆	**.40**	**.56**	**.60**	**.44**	**.57**	**.80**⋆	**.63**	**.75**⋆	**.75**	.25	**.67**	**.43**	**.83**	**.75**	.29	**.57**	.17		**.50**	**1**†	.13	**.67**	**.53**	**.40**	**.71**	.20	**.67**	**.47**	.38		**.86**	.25	.27	.25	**.70**	**.67**	.20	.38	**.50**
RBMT5-ES	.17	**.67**	**.75**	.27	.11⋆	.27	**.73**†	.50	.13	.46	.27	.43		**.62**	.11⋆	**.62**	**.50**	.22	.36	.10⋆	**.43**	.44	.43	**.63**	.46	.36		**.50**	.29	**.44**	.17	**.57**	.27	.29	.25	**.60**	.11	.33	**.60**	**.67**	.44	0	.30	.43	.17	.25	**.58**
RBMT5-FR	.30	.42	.10	.11	.14⋆	.17	.09⋆	.42	.30	.39	.20	.11⋆	.33	**.50**	.21	.27	.29		.15	.40	**.67**	.38	**.71**	.33	.17	0	.40		.50	.40	.20	.25	**.56**	.07†	**.50**	.31	.14		**.50**		.22	.40	**.57**	**.54**			.29
RBMT6-DE	**.67**	.25	**.91**⋆	.36	.06†	**.50**	**.54**	**.70**	.47	**.53**	.33	**.50**	**.57**	**.70**	.40	**.67**	**.58**	.25	.39	.38	**.69**	**.73**	**.50**	**.50**	.46	**.50**	**.43**	.50		.13	**.50**	**.46**	**.62**	**.50**	**.46**	.50	**.46**	.40	.33	**.80**	.29	.20	**.57**	**.50**		.29	**.86**
RBMT6-ES	.29	**.55**	**.60**	.50		.25	.25	.46	.22	.33	.08⋆	.40	.20	.44		**.55**	**.64**	.38	.29	**.63**	**.40**	.30	.30	.25	**.57**	.33	.22	**.60**	**.63**		**.67**	**.64**	.42	.38	**.67**	**.71**	.46	0		.50	.22	.25	.50	.33	0	.13	**.67**
RBMT6-FR	.36	**.63**	**.57**	.43	.20⋆	**.63**	**.50**	**.83**	.43	.36	.10⋆		.18	.40	.17	.43	**.50**	.39	**.67**	.30	.39	**.73**⋆	.38	**.63**	.38	.27	**.83**	**.50**	.38	.17		.29	**.67**	.22	.40	.33	.38		.33	.50	.13	.25	**.63**	.33		.14	.25
SAAR-C-DE	.41	**.55**	**.67**	.30	.25	.50	**.50**	**.54**	**.60**	.40	.14⋆	**.57**	.25	**.83**	.30	**.50**	**.62**	.25	**.70**	**.50**	**.47**	.36	**.50**	.46	**.55**	.33	.14	**.75**	.36	.27	**.59**		.21	.13	**.50**	.36	.42	.33	**1**	.20	.33	**.54**	**.88**	**.69**		0	.44
SAAR-C-FR	.20	.40	**.50**	**.54**	0	.33	.50	.33	**.60**	.44	.38	0	**.60**	**.60**	.40	**.42**	**.43**	.17	**.50**	.33	**.64**	.13	**.67**	**.67**	.25		**.46**	.22	.31	.42	.27	**.79**		.18	**.56**	.18	.17		**.60**	.25	.09		**.86**	**.50**	.18	.44	.44
SAAR-DE	.33	**.77**†	**.63**	**.43**	.14	**.64**	**.50**	**.91**†	**.44**	**.82**	**.58**	**.73**	**.67**	**.50**	.40	**.67**	**.70**	.33	.43	.44	.36	**.67**	**.67**	**.86**	**.86**	**.60**	**.50**	**.86**†	.29	**.56**	**.44**	**.53**	**.73**		.33	**.83**⋆	.31		**.67**	**.86**⋆	.43	.33	**.83**⋆	**.56**	.17	**.78**	**.55**
SAAR-ES	.29	**.60**	.44	.18	.07†	.29	.40	**.67**	.17	.44	.10⋆	**.46**	.27	.50	.13	.12⋆	**.50**	.14	.33	**.46**	**.62**	.36	.25	**.55**	.14	.17	.25	.40	.39	.17	**.50**	.42	.44	**.47**		**.50**	.11		.33		.11†	.33	**.39**	.18		.08	.33
SAAR-FR	.18	**.44**	**.60**	.30	.20	**.56**	**.73**	**.60**	.50	**.58**	0	**.50**	.33	**.64**	.29	**.42**	**.63**⋆	**.67**	**.57**	.33	.25	.38	.42	**.58**	**.43**	.41	.30	**.54**	.50	.29	.33	.36	**.36**	.17⋆	.43		.33	.40	**.60**	.33	.27	.20	**.64**⋆	.31		.25	**.58**
UCB-ES		.33	.44	**.33**	.18	.18	**.55**	**.62**	.14	**.63**	.18	.44	**.57**	**.44**	.33	**.67**	.44	.25	**.56**	.25	.46	**.60**	.50	**.56**	**.83**	**.63**	**.56**	**.57**	.36	.46	**.50**	**.50**	**.75**	**.50**	**.56**	**.58**		**.88**	**.71**	**.57**	.14		**.42**	**.67**		.44	.31
UED-CMB-DE	.20	**.67**	**.75**	**1**	**.67**	.33	**1**	**.88**	**.75**	**1**	**.83**	**.71**	**.75**		**.67**	**.67**	**1**	**.44**	**.60**	**.67**	**.57**	**1**	**1**	**.78**	**1**	**.67**	**.44**	**1**	**.60**	**1**⋆	**1**	**.67**	**1**	**.83**	**1**	.40	.13		**.67**	**.75**⋆	.25	.50	.40	**1**	.50	**.75**	.50
UED-CMB-FR	**.67**	**1**	**.57**	.50			.50	.13	.50	**.40**	.33					.50		.25	.40	.20	.25	**.50**	.40		**.50**	.14	.20	.33	.33	**1**	**.67**		.40	.33	.33		.29			.38					.33	.33	.50
UED-CMB-XX			**.50**	**.67**		.25	.13	**1**	.50	**.50**	.40		.25				.20			.40	.13		.33		.40	**.75**	.33	**1**		.50	.50	**.60**	**.50**	0	**.67**	**.56**	.14	0	.38		0				.33		.25
UED-CZ	**.75**	**.79**⋆	**.89**⋆	**.86**	**.56**	**.83**	**.71**	**.77**†	**.92**†	**1**†	.20	**.88**	**.67**	**.63**	**.67**	**1**	**.69**⋆	**.60**	**.80**⋆	**.80**	**.75**	**.67**	**.86**	**.75**⋆	**.62**	**.46**	**.56**	**.67**	**.71**	**.78**	**.88**	**.67**	**.64**	.43	**.78**†	**.73**	**.86**	**.50**	**1**	**1**⋆		**.55**	**1**†	**.80**⋆	0	**.46**	**.90**⋆
UED-DE	.27	**.80**	**.78**	**1**		.20	**.80**⋆	**.71**	**1**†	**.67**	.31	.25	**.75**	**.92**†	**.56**	**.67**	**.83**⋆		**.62**	**.44**	**.67**	**.67**	**.63**	**.82**⋆	.25	**.75**	**1**†	.40	**.40**	**.75**	**.75**	.23	**1**	**.67**	**.67**	**.60**	**.57**	.50		**1**	.36		**.64**	**.67**	.11	.25	**.67**
UED-ES	.15⋆	.17	**.56**	**.42**	0	.31	**.56**	**.46**	**.50**	**.88**	.11	.33	.29	**.46**	.29	**.67**	**.69**		.31	.36	**.46**	.27	.50	**.54**	.33	.30	**.60**	.43	.29	.50	.38	.13	.14	.17⋆	.23	0	.17	.40	**.83**	**.80**	0	.27		.46	0	.13	
UED-FR	.50	**.56**	**.75**	**.67**	0	.18	**.56**	**.50**	.14	.44	.20	.17	.42	**.53**	.22	**.64**	**.58**	.44	**.57**	.30	.50	.38	.17	.42	**.50**	.22	**.57**	.31	.38	**.44**	**.42**	.23	.40	.33	**.46**	**.39**	.33		**.75**	**1**	.10⋆	.11	.46			.31	.38
UED-HU	**1**	**.75**	**.67**	**.88**	**.67**	.50	**.75**	**1**⋆	**.80**	**1**⋆	**.38**	**.60**	**.83**	**.75**	**1**	**1**⋆	**.86**⋆	**.50**	**.86**	**.75**	**1**†	**.80**	**1**⋆	**1**⋆	**.80**	**.80**	**.83**	**1**	**1**	**1**⋆	**1**	**.80**	**.82**	**.67**	**1**	**1**	**.83**	.50	.33	**.67**	**.86**⋆	**.67**	**1**⋆	**1**		**.83**	.40
UMD-CZ	.40	**.50**	**.64**	**.50**	.50	**.67**	.29	**.50**	.50	**.80**⋆	.25	**.40**	.27	**.71**	**.60**	.44	**.79**†		**.79**†	**.70**	**.80**	**.63**	**.67**	**.50**	**.63**	**.44**	**.63**	**.60**	**.43**	**.75**	**.64**	**.91**†	**.56**	.11	**.54**	**.67**	.44	.25	**.67**	**1**	.27	**.50**	**.88**	**.69**	.17		**.62**
UPC-ES	.40	.42	.44	.21	.15⋆		**.44**	.38	.25	.38	**.54**	.43	.30	.50	.23	.40	**.55**		.33	.33	.22	.47	.13	**.50**	**.50**	.38	.25	**.43**	.14	.33	**.63**	**.56**	.44	.36	**.53**	.25	**.39**	.50	.50	**.50**	.10⋆	.33	**.63**	**.50**	.40	.31	
> OTHERS	.29	.54	.51	.41	.15	.35	.51	.52	.43	.51	.25	.34	.39	.55	.33	.57	.57	.3	.5	.44	.52	.49	.48	.56	.50	.35	.46	.57	.39	.50	.49	.46	.49	.32	.51	.45	.4	.22	.56	**.58**	.19	.28	.53	.48	.11	.3	.52
≥ OTHERS	.41	.7	.66	.55	.28	.44	.67	.64	.57	.65	.38	.47	.54	.7	.45	.71	.70	.39	.63	.55	.63	.61	.57	.7	.64	.49	.60	.71	.51	.62	.62	.57	.62	.45	.68	.58	.55	.33	.70	**.72**	.29	.40	.65	.62	.19	.43	.63

Table 53: Sentence-level ranking for the All-English News Task.