Findings of the 2014 Workshop on Statistical Machine Translation

Findings of the 2014 Workshop on Statistical Machine Translation

Ondřej Bojar Charles University in Prague
Christian Buck University of Edinburgh
Christian Federmann Microsoft Research
Barry Haddow University of Edinburgh
Philipp Koehn JHU / Edinburgh
Johannes Leveling Dublin City University
Christof Monz University of Amsterdam
Pavel Pecina Charles University in Prague
Matt Post Johns Hopkins University
Herve Saint-Amand University of Edinburgh
Radu Soricut Google
Lucia Specia University of Sheffield
Aleš Tamchyna Charles University in Prague

Abstract

This paper presents the results of the WMT14 shared tasks, which included a standard news translation task, a separate medical translation task, a task for run-time estimation of machine translation quality, and a metrics task. This year, 143 machine translation systems from 23 institutions were submitted to the ten translation directions in the standard translation task. An additional 6 anonymized systems were included, and were then evaluated both automatically and manually. The quality estimation task had four subtasks, with a total of 10 teams, submitting 57 entries.

1 Introduction

We present the results of the shared tasks of the Workshop on Statistical Machine Translation (WMT) held at ACL 2014. This workshop builds on eight previous WMT workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013).

This year we conducted four official tasks: a translation task, a quality estimation task, a metrics task[1] and a medical translation task. In the translation task (§2), participants were asked to translate a shared test set, optionally restricting themselves to the provided training data. We held ten translation tasks this year, between English and each of Czech, French, German, Hindi, and Russian. The Hindi translation tasks were new this year, providing a lesser resourced data condition on a challenging language pair. The system outputs for each task were evaluated both automatically and manually.

[1]The metrics task is reported in a separate paper (Macháček and Bojar, 2014).

The human evaluation (§3) involves asking human judges to rank sentences output by anonymized systems. We obtained large numbers of rankings from researchers who contributed evaluations proportional to the number of tasks they entered. Last year, we dramatically increased the number of judgments, achieving much more meaningful rankings. This year, we developed a new ranking method that allows us to achieve the same with fewer judgments.

The quality estimation task (§4) this year included sentence- and word-level subtasks: sentence-level prediction of 1-3 likert scores, sentence-level prediction of percentage of word edits necessary to fix a sentence, sentence-level prediction of post-editing time, and word-level prediction of scores at different levels of granularity (correct/incorrect, accuracy/fluency errors, and specific types of errors). Datasets were released with English-Spanish, English-German, Spanish-English and German-English news translations produced by 2-3 machine translation systems and, for some subtasks, a human translation.

The medical translation task (§5) was introduced this year. Unlike the "standard" translation task, the test sets come from the very specialized domain of medical texts. The aim of this task was not only domain adaptation but also the utilization of translation systems in a larger scenario, namely cross-lingual information retrieval (IR). Extrinsic evaluation in an IR setting was a part of this task (on the other hand, manual evaluation of translation quality was not carried out).

The primary objectives of WMT are to evaluate the state of the art in machine translation, to disseminate common test sets and public training data with published performance numbers, and to refine evaluation and estimation methodologies for machine translation. As before, all of the data,

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 12–58,
Baltimore, Maryland USA, June 26–27, 2014.

translations, and collected human judgments are publicly available.[2] We hope these datasets serve as a valuable resource for research into statistical machine translation and automatic evaluation or prediction of translation quality.

2 Overview of the Translation Task

The recurring task of the workshop examines translation between English and other languages. As in the previous years, the other languages include German, French, Czech and Russian.

We dropped Spanish and added Hindi this year. From a linguistic point of view, Spanish poses similar problems as French, making its prior inclusion less valuable. Hindi is not only interesting since it is a more distant language than the European languages we include, but also because we have much less training data, thus forcing researchers to deal with low resource conditions, but also providing them with a language pair that does not suffer from the computational complexities of having to deal with massive amounts of training data.

We created a test set for each language pair by translating newspaper articles and provided training data.

2.1 Test data

The test data for this year's task was selected from news stories from online sources, as before. However, we changed our method to create the test sets.

In previous years, we took equal amounts of source sentences from all six languages involved (around 500 sentences each), and translated them into all other languages. While this produced a multi-parallel test corpus that could be also used for language pairs (such as Czech-Russian) that we did not include in the evaluation, it did suffer from artifacts from the larger distance between source and target sentences. Most test sentences involved the translation a source sentence that was translated from a their language into a target sentence (which was compared against a translation from that third language as well). Questions have been raised, if the evaluation of, say, French-English translation is best served when testing on sentences that have been originally written in, say, Czech. For discussions about *translationese* please for instance refer to Koppel and Ordan (2011).

This year, we took about 1500 English sentences and translated them into the other 5 languages, and then additional 1500 sentences from each of the other languages and translated them into English. This gave us test sets of about 3000 sentences for our English-X language pairs, which have been either written originally written in English and translated into X, or vice versa.

The composition of the test documents is shown in Table 1. The stories were translated by the professional translation agency Capita, funded by the EU Framework Programme 7 project MosesCore, and by Yandex, a Russian search engine company.[3] All of the translations were done directly, and not via an intermediate language.

2.2 Training data

As in past years we provided parallel corpora to train translation models, monolingual corpora to train language models, and development sets to tune system parameters. Some training corpora were identical from last year (Europarl[4], United Nations, French-English 10^9 corpus, CzEng, Common Crawl, Russian-English Wikipedia Headlines provided by CMU), some were updated (Russian-English parallel data provided by Yandex, News Commentary, monolingual data), and a new corpus was added (Hindi-English corpus, Bojar et al. (2014)), Hindi-English Wikipedia Headline corpus).

Some statistics about the training materials are given in Figure 1.

2.3 Submitted systems

We received 143 submissions from 23 institutions. The participating institutions and their entry names are listed in Table 2; each system did not necessarily appear in all translation tasks. We also included four commercial off-the-shelf MT systems and four online statistical MT systems, which we anonymized.

For presentation of the results, systems are treated as either *constrained* or *unconstrained*, depending on whether their models were trained only on the provided data. Since we do not know how they were built, these online and commercial systems are treated as unconstrained during the automatic and human evaluations.

[2] http://statmt.org/wmt14/results.html

[3] http://www.yandex.com/

[4] As of Fall 2011, the proceedings of the European Parliament are no longer translated into all official languages.

Europarl Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English	
Sentences	2,007,723		1,920,209		646,605	
Words	60,125,563	55,642,101	50,486,398	53,008,851	14,946,399	17,376,433
Distinct words	140,915	118,404	381,583	115,966	172,461	63,039

News Commentary Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	183,251		201,288		146,549		165,602	
Words	5,688,656	4,659,619	5,105,101	5,046,157	3,288,645	3,590,287	4,153,847	4,339,974
Distinct words	72,863	62,673	150,760	65,520	139,477	55,547	151,101	60,801

Common Crawl Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	3,244,152		2,399,123		161,838		878,386	
Words	91,328,790	81,096,306	54,575,405	58,870,638	3,529,783	3,927,378	21,018,793	21,535,122
Distinct words	889,291	859,017	1,640,835	823,480	210,170	128,212	764,203	432,062

United Nations Parallel Corpus

	French ↔ English	
Sentences	12,886,831	
Words	411,916,781	360,341,450
Distinct words	565,553	666,077

Hindi-English Parallel Corpus

	Hindi ↔ English	
Sentences	287,202	
Words	6,002,418	3,953,851
Distinct words	121,236	105,330

10^9 Word Parallel Corpus

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

Yandex 1M Parallel Corpus

	Russian ↔ English	
Sentences	1,000,000	
Words	24,121,459	26,107,293
Distinct words	701,809	387,646

CzEng Parallel Corpus

	Czech ↔ English	
Sentences	14,833,358	
Words	200,658,857	228,040,794
Distinct words	1,389,803	920,824

Wiki Headlines Parallel Corpus

	Russian ↔ English		Hindi ↔ English	
Sentences	514,859		32,863	
Words	1,191,474	1,230,644	141,042	70,075
Distinct words	282,989	251,328	25,678	26,989

Europarl Language Model Data

	English	French	German	Czech
Sentence	2,218,201	2,190,579	2,176,537	668,595
Words	59,848,044	63,439,791	53,534,167	14,946,399
Distinct words	123,059	145,496	394,781	172,461

News Language Model Data

	English	French	German	Czech	Russian	Hindi
Sentence	90,209,983	30,451,749	89,634,193	36,426,900	32,245,651	1,275,921
Words	2,109,603,244	748,852,739	1,606,506,785	602,950,410	575,423,682	36,297,394
Distinct words	4,089,792	1,906,470	10,248,707	3,101,846	2,860,837	258,759

News Test Set

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English		Hindi ↔ English	
Sentences	3003		3003		3003		3003		2507	
Words	81,194	71,147	63,078	67,624	60,240	68,866	62,107	69,329	86,974	55,822
Distinct words	11,715	10,610	13,930	10,458	16,774	9,893	17,009	9,938	8,292	9,217

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

Language	Sources (Number of Documents)
Czech	aktuálně.cz (2), blesk.cz (3), blisty.cz (1), deník.cz (9), e15.cz (1), iDNES.cz (17), ihned.cz (14), lidovky.cz (8), mediafax.cz (2), metro.cz (1), Novinky.cz (5), pravo.novinky.cz (6), reflex.cz (2), tyden.cz (1), zdn.cz (1).
French	BBC French Africa (1), Canoe (9), Croix (4), Cyber Presse (12), Dernieres Nouvelles (1), dhnet.be (5), Equipe (1), Euronews (6), Journal Metro.com (1), La Libre.be (2), La Meuse.be (2), Le Devoir (3), Le Figaro (8), Le Monde (3), Les Echos (15), Lexpress.fr (3), Liberation (1), L'independant (2), Metro France (1), Nice-Matin (6), Le Nouvel Observateur (3), Radio Canada (6), Reuters (7).
English	ABC News (5), BBC (5), CBS News (5), CNN (5), Daily Mail (5), Financial Times (5), Fox News (2), Globe and Mail (1), Independent (1), Los Angeles Times (1), New Yorker (1), News.com Australia (16), Reuters (3), Scotsman (2), smh.com.au (2), stv.tv (1), Telegraph (6), UPI (2).
German	Abendzeitung Nürnberg (1), all-in.de (2), Augsburger Allgemeine (1), AZ Online (1), Börsenzeitung (1), come-on.de (1), Der Westen (2), DZ Online (1), Reutlinger General-Anzeiger (1), Generalanzeiger Bonn (1), Giessener Anzeiger (1), Goslarsche Zeitung (1), Hersfelder Zeitung (1), Jüdische Allgemeine (1), Kreisanzeiger (2), Kreiszeitung (2), Krone (1), Lampertheimer Zeitung (2), Lausitzer Rundschau (1), Mittelbayerische (1), Morgenpost (1), nachrichten.at (1), Neue Presse (1), OP Online (1), Potsdamer Neueste Nachrichten (1), Passauer Neue Presse (1), Recklinghäuser Zeitung (1), Rhein Zeitung (1), salzburg.com (1), Schwarzwälder Bote (29), Segeberger Zeitung (1), Soester Anzeiger (1), Südkurier (17), svz.de (1), Tagesspiegel (1), Usinger Anzeiger (3), Volksblatt.li (1), Westfälischen Anzeiger (3), Wiener Zeitung (1), Wiesbadener Kurier (1), Westdeutsche Zeitung (1), Wilhelmshavener Zeitung (1), Yahoo Deutschland (1).
Hindi	Bhaskar (24), Jagran (61), Navbharat Times / India Times (4), ndtv (2).
Russian	168.ru (1), aif (3), altapress.ru (2), argumenti.ru (2), BBC Russian (3), belta.by (2), communa.ru (1), dp.ru (1), eg-online.ru (1), Euronews (2), fakty.ua (2), gazeta.ru (1), inotv.rt.com (1), interfax (1), Izvestiya (1), Kommersant (7), kp (2), lenta.ru (4), lgng (1), litrossia.ru (1), mirnov.ru (5), mk (8), mn.ru (2), newizv (2), nov-pravda.ru (1), novayagazeta (1), nr2.ru (8), pnp.ru (1), rbc.ru (3), ria.ru (4), rosbalt.ru (1), sovsport.ru (6), Sport Express (10), trud.ru (4), tumentoday.ru (1), vesti.ru (10), zr.ru (1).

Table 1: Composition of the test set. 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.

3 Human Evaluation

As with past workshops, we contend that automatic measures of machine translation quality are an imperfect substitute for human assessments. We therefore conduct a manual evaluation of the system outputs and define its results to be the principal ranking of the workshop. In this section, we describe how we collected this data and compute the results, and then present the official results of the ranking.

This year's evaluation was conducted a bit differently. The main differences are:

- In contrast to the past two years, we collected judgments entirely from researchers participating in the shared tasks and trusted friends of the community. Last year, about two thirds of the data were solicited from random volunteers on the Amazon Mechanical Turk. For some language pairs, the Turkers data had much lower inter-annotator agreement compared to the researchers.
- As a result, we collected about seventy-five percent less data, but were able to obtain good confidence intervals on the clusters with the use of new approaches to ranking.
- We compared three different ranking methodologies, selecting the one with the highest accuracy on held-out data.

We also maintain many of our customs from prior years, including the presentation of the results in terms of a *partial ordering* (clustering) of the systems. Systems in the same cluster could not be meaningfully distinguished and should be considered ties.

3.1 Data collection

The system ranking is produced from a large set of pairwise annotations between system pairs. These pairwise annotations are collected in an evaluation campaign that enlists participants in the shared task to contribute one hundred "Human Intelligence Tasks" (HITs) per system submitted. Each HIT consists of three *ranking tasks*. In a ranking task, an annotator is presented with a source segment, a human reference translation, and the outputs of five anonymized systems, randomly selected from the set of participating systems, and randomly ordered.

To run the evaluation, we use Appraise[5] (Federmann, 2012), an open-source tool built on Python's Django framework. At the top of each HIT, the following instructions are provided:

> *You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).*

[5] `https://github.com/cfedermann/Appraise`

ID	Institution
AFRL, AFRL-PE	Air Force Research Lab (Schwartz et al., 2014)
CIMS	University of Stuttgart / University of Munich (Cap et al., 2014)
CMU	Carnegie Mellon University (Matthews et al., 2014)
CU-*	Charles University, Prague (Tamchyna et al., 2014)
DCU-FDA	Dublin City University (Bicici et al., 2014)
DCU-ICTCAS	Dublin City University (Li et al., 2014b)
DCU-LINGO24	Dublin City University / Lingo24 (wu et al., 2014)
EU-BRIDGE	EU-BRIDGE Project (Freitag et al., 2014)
KIT	Karlsruhe Institute of Technology (Herrmann et al., 2014)
IIT-BOMBAY	IIT Bombay (Dungarwal et al., 2014)
IIIT-HYDERABAD	IIIT Hyderabad
IMS-TTT	University of Stuttgart / University of Munich (Quernheim and Cap, 2014)
IPN-UPV-*	IPN-UPV (Costa-jussà et al., 2014)
KAZNU	Amandyk Kartbayev, FBK
LIMSI-KIT	LIMSI / Karlsruhe Instutute of Technology (Do et al., 2014)
MANAWI-*	Universität des Saarlandes (Tan and Pal, 2014)
MATRAN	Abu-MaTran Project: Promsit / DCU / UA (Rubino et al., 2014)
PROMT-RULE, PROMT-HYBRID	PROMT
RWTH	RWTH Aachen (Peitz et al., 2014)
STANFORD	Stanford University (Neidert et al., 2014; Green et al., 2014)
UA-*	University of Alicante (Sánchez-Cartagena et al., 2014)
UEDIN-PHRASE, UEDIN-UNCNSTR	University of Edinburgh (Durrani et al., 2014b)
UEDIN-SYNTAX	University of Edinburgh (Williams et al., 2014)
UU, UU-DOCENT	Uppsala University (Hardmeier et al., 2014)
YANDEX	Yandex School of Data Analysis (Borisov and Galinskaya, 2014)
COMMERCIAL-[1,2]	Two commercial machine translation systems
ONLINE-[A,B,C,G]	Four online statistical machine translation systems
RBMT-[1,4]	Two rule-based statistical machine translation systems

Table 2: Participants in the shared translation task. Not all teams participated in all language pairs. The translations from the commercial and online systems were not submitted by their respective companies but were obtained by us, and are therefore anonymized in a fashion consistent with previous years of the workshop.

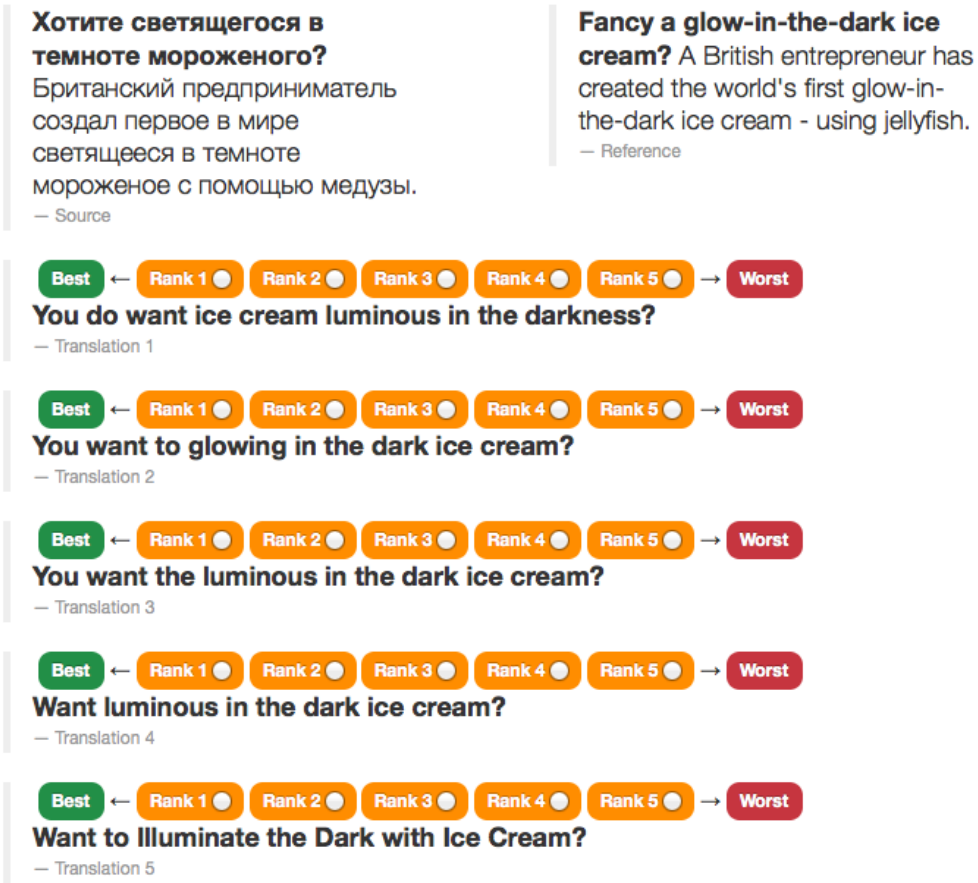

Figure 2: Screenshot of the Appraise interface used in the human evaluation campaign. The annotator is presented with a source segment, a reference translation, and the outputs of five systems (anonymized and randomly ordered), and is asked to rank these according to their translation quality, with ties allowed.

A screenshot of the ranking interface is shown in Figure 2. Annotators are asked to rank the systems from 1 (best) to 5 (worst), with ties permitted. Note that a *lower* rank is better. The rankings provided by a ranking task are then reduced to a set of ten *pairwise rankings* produced by considering all $\binom{5}{2}$ combinations of systems in the ranking task. For example, consider the following annotation provided among systems A, B, F, H, and J:

	1	2	3	4	5
F				•	
A				•	
B		•			
J					•
H			•		

This is reduced to the following set of pairwise judgments:

$$A > B, A = F, A > H, A < J$$
$$B < F, B < H, B < J$$
$$F > H, F < J$$
$$H < J$$

Here, $A > B$ should be read is "A is ranked higher than (worse than) B". Note that by this procedure, the absolute value of ranks and the magnitude of their differences are discarded.

For WMT13, nearly a million pairwise annotations were collected from both researchers and paid workers on Amazon's Mechanical Turk, in a roughly 1:2 ratio. This year, we collected data from researchers only, an ability that was enabled by the use of a new technique for producing the partial ranking for each task (§3.3.3). Table 3 contains more detail.

3.2 Annotator agreement

Each year we calculate annotator agreement scores for the human evaluation as a measure of the reliability of the rankings. We measured pairwise agreement among annotators using Cohen's kappa coefficient (κ) (Cohen, 1960). If $P(A)$ be the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would

LANGUAGE PAIR	Systems	Rankings	Average
Czech–English	5	21,130	4,226.0
English–Czech	10	55,900	5,590.0
German–English	13	25,260	1,943.0
English–German	18	54,660	3,036.6
French–English	8	26,090	3,261.2
English–French	13	33,350	2,565.3
Russian–English	13	34,460	2,650.7
English–Russian	9	28,960	3,217.7
Hindi–English	9	20,900	2,322.2
English–Hindi	12	28,120	2,343.3
TOTAL WMT 14	110	328,830	2,989.3
WMT13	148	942,840	6,370.5
WMT12	103	101,969	999.6
WMT11	133	63,045	474.0

Table 3: Amount of data collected in the WMT14 manual evaluation. The final three rows report summary information from the previous two workshops.

agree by chance, then Cohen's kappa is:

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

Note that κ is basically a normalized version of $P(A)$, one which takes into account how meaningful it is for annotators to agree with each other by incorporating $P(E)$. The values for κ range from 0 to 1, with zero indicating no agreement and 1 perfect agreement.

We calculate $P(A)$ by examining all pairs of systems which had been judged by two or more judges, and calculating the proportion of time that they agreed that $A < B$, $A = B$, or $A > B$. In other words, $P(A)$ is the empirical, observed rate at which annotators agree, in the context of pairwise comparisons.

As for $P(E)$, it captures the probability that two annotators would agree randomly. Therefore:

$$P(E) = P(_{A<B})^2 + P(_{A=B})^2 + P(_{A>B})^2$$

Note that each of the three probabilities in $P(E)$'s definition are squared to reflect the fact that we are considering the chance that *two* annotators would agree by chance. Each of these probabilities is computed empirically, by observing how often annotators actually rank two systems as being tied.

Table 4 gives κ values for inter-annotator agreement for WMT11–WMT14 while Table 5 details intra-annotator agreement scores, including the division of researchers (WMT13$_r$) and MTurk (WMT13$_m$) data. The exact interpretation of the kappa coefficient is difficult, but according to Landis and Koch (1977), 0–0.2 is slight, 0.2–0.4 is fair, 0.4–0.6 is moderate, 0.6–0.8 is substantial, and 0.8–1.0 is almost perfect. The agreement rates are more or less in line with prior years: worse for some tasks, better for others, and on average, the best since WMT11 (where agreement scores were likely inflated due to inclusion of reference translations in the comparisons).

3.3 Models of System Rankings

The collected pairwise rankings are used to produce a ranking of the systems. Machine translation evaluation has always been a subject of contention, and no exception to this rule exists for the WMT manual evaluation. While the precise metric has varied over the years, it has always shared a common idea of computing the average number of times each system was judged better than other systems, and ranking from highest to lowest. For example, in WMT11 Callison-Burch et al. (2011), the metric computed the percentage of the time each system was ranked better than or equal to other systems, and included comparisons to human references. In WMT12 Callison-Burch et al. (2012), comparisons to references were dropped. In WMT13, rankings were produced over 1,000 bootstrap-resampled sets of the training data. A *rank range* was collected for each system across these folds; the average value was used to order the systems, and a 95% confidence interval across these ranks was used to organize the systems into equivalence classes containing systems with over-

LANGUAGE PAIR	WMT11	WMT12	WMT13	WMT13$_r$	WMT13$_m$	WMT14
Czech–English	0.400	0.311	0.244	0.342	0.279	0.305
English–Czech	0.460	0.359	0.168	0.408	0.075	0.360
German–English	0.324	0.385	0.299	0.443	0.324	0.368
English–German	0.378	0.356	0.267	0.457	0.239	0.427
French–English	0.402	0.272	0.275	0.405	0.321	0.357
English–French	0.406	0.296	0.231	0.434	0.237	0.302
Hindi–English	—	—	—	—	—	0.400
English–Hindi	—	—	—	—	—	0.413
Russian–English	—	—	0.278	0.315	0.324	0.324
English–Russian	—	—	0.243	0.416	0.207	0.418
MEAN	0.395	0.330	0.260			0.367

Table 4: κ scores measuring inter-annotator agreement. See Table 5 for corresponding intra-annotator agreement scores.

LANGUAGE PAIR	WMT11	WMT12	WMT13	WMT13$_r$	WMT13$_m$	WMT14
Czech–English	0.597	0.454	0.479	0.483	0.478	0.382
English–Czech	0.601	0.390	0.290	0.547	0.242	0.448
German–English	0.576	0.392	0.535	0.643	0.515	0.344
English–German	0.528	0.433	0.498	0.649	0.452	0.576
French–English	0.673	0.360	0.578	0.585	0.565	0.629
English–French	0.524	0.414	0.495	0.630	0.486	0.507
Hindi–English	—	—	—	—	—	0.605
English–Hindi	—	—	—	—	—	0.535
Russian–English	—	—	0.450	0.363	0.477	0.629
English–Russian	—	—	0.513	0.582	0.500	0.570
MEAN	0.583	0.407	0.479			0.522

Table 5: κ scores measuring intra-annotator agreement, i.e., self-consistency of judges, across for the past few years of the human evaluation.

lapping ranges.

This year, we introduce two new changes. First, we pit the WMT13 method against two new approaches: that of Hopkins and May (2013, §3.3.2), and another based on TrueSkill (Sakaguchi et al., 2014, §3.3.3). Second, we compare these two methods against WMT13's "Expected Wins" approach, and then select among them by determining which of them has the highest accuracy in terms of predicting annotations on a held-out set of pairwise judgments.

3.3.1 Method 1: Expected Wins (EW)

Introduced for WMT13, the EXPECTED WINS has an intuitive score demonstrated to be accurate in ranking systems according to an underlying model of "relative ability" (Koehn, 2012a). The idea is to gauge the probability that a system S_i will be ranked better than another system randomly chosen from a pool of opponents $\{S_j : j \neq i\}$. If we define the function $\text{win}(A, B)$ as the number of times system A is ranked better than system B, then we can define this as follows:

$$\text{score}_{EW}(S_i) = \frac{1}{|\{S_j\}|} \sum_{j,j \neq i} \frac{\text{win}(S_i, S_j)}{\text{win}(S_i, S_j) + \text{win}(S_j, S_i)}$$

Note that this score ignores ties.

3.3.2 Method 2: Hopkins and May (HM)

Hopkins and May (2013) introduced a graphical model formulation of the task, which makes the notion of underlying system ability even more explicit. Each system S_J in the pool $\{S_j\}$ is represented by an associated relative ability μ_j and a variance σ_a^2 (fixed across all systems) which serve as the parameters of a Gaussian distribution. Samples from this distribution represent the quality of sentence translations, with higher quality samples having higher values. Pairwise annotations (S_1, S_2, π) are generated according to the following process:

1. Select two systems S_1 and S_2 from the pool of systems $\{S_j\}$

2. Draw two "translations", adding random Gaussian noise with variance σ^2_{obs} to simulate the subjectivity of the task and the differences among annotators:

$$q_1 \sim \mathcal{N}(\mu_{S_1}, \sigma^2_a) + \mathcal{N}(0, \sigma^2_{obs})$$
$$q_2 \sim \mathcal{N}(\mu_{S_2}, \sigma^2_a) + \mathcal{N}(0, \sigma^2_{obs})$$

3. Let d be a nonzero real number that defines a fixed decision radius. Produce a rating π according to:

$$\pi = \begin{cases} < & q_1 - q_2 > d \\ > & q_2 - q_1 > d \\ = & \text{otherwise} \end{cases}$$

Hopkins and May use Gibbs sampling to infer the set of system means from an annotated dataset. Details of this inference procedure can be found in Sakaguchi et al. (2014). The score used to produce the rankings is simply the system mean associated with each system:

$$\text{score}_{HM}(S_i) = \mu_{S_i}$$

3.3.3 Method 3: TrueSkill (TS)

TrueSkill is an adaptive, online system that employs a similar model of relative ability Herbrich et al. (2006). It was initially developed for Xbox Live's online player community, where it is used to model player ability, assign levels, and select competitive matches. Each player S_j is modeled by two parameters: TrueSkill's current estimate of each system's relative ability, μ_{S_j}, and a per-system measure of TrueSkill's uncertainty of those estimates, $\sigma^2_{S_j}$. When the outcome of a match is observed, TrueSkill uses the relative status of the two systems to update these estimates. If a translation from a system with a high mean is judged better than a system with a greatly lower mean, the result is not surprising, and the update size for the corresponding system means will be small. On the other hand, when an upset occurs in a competition, the means will receive larger updates. Sakaguchi et al. (2014) provide an adaptation of this approach to the WMT manual evaluation, and showed that it performed well on WMT13 data.

Similar to the Hopkins and May model, TrueSkill scores systems by their inferred means:

$$\text{score}_{TS}(S_i) = \mu_{S_i}$$

This score is then used to sort the systems and produce the ranking.

3.4 Method Selection

We have three methods which, provided with the collected data, produce different rankings of the systems. Which of them is correct? More immediately, which one of them should we publish as the official ranking for the WMT14 manual evaluation? As discussed, the method used to compute the ranking has been tweaked a bit each year over the past few years in response to criticisms (e.g., Lopez (2012); Bojar et al. (2011)). While the changes were reasonable (and later corroborated), Hopkins and May (2013) pointed out that this task of model selection should be driven by empirical evaluation on held-out data, and suggested perplexity as the metric of choice.

We choose instead a more direct gold-standard evaluation metric: the accuracy of the rankings produced by each method in predicting pairwise judgments. We use each method to produce a partial ordering of the systems, grouping them into equivalence classes. This partial ordering unambiguously assigns a prediction π_P between any pair of systems (S_i, S_j). By comparing the predicted relationship π_P to the actual annotation for each pairwise judgment in the test data (by token), we can compute an accuracy score for each model.

We predict accuracy in this manner using 100-fold cross-validation. For each task, we split the data into a fixed set of 100 randomly-selected folds. Each fold serves as a test set, with the remaining ninety-nine folds available as training data for each method. Note that the total ordering over systems provided by the score$_*$ functions defined do not predict ties. In order to do enable the models to predict ties, we produce equivalence classes using the following procedure:

- Assign S_1 to a cluster
- For each system S_i, assign it to the current cluster if $\text{score}(S_{i-1}) - \text{score}(S_i) \leq r$; otherwise, assign it to a new cluster

The value of r (the *decision radius* for ties) is tuned using accuracy on the entire training data using grid search over the values $r \in 0, 0.01, 0.02, \ldots, .25$ (26 values in total). This value is tuned separately for each method on each fold. Table 6 contains an example partial ordering.

System	Score	Rank
B	0.60	1
D	0.44	2
E	0.39	2
A	0.25	2
F	-0.09	3
C	-0.22	3

Table 6: The partial ordering computed with the provided scores when $r = 0.15$.

Task	EW	HM	TS	Oracle
Czech–English	40.4	**41.1**	**41.1**	41.2
English–Czech	45.3	45.6	**45.9**	46.8
French–English	49.0	**49.4**	49.3	50.3
English–French	44.6	44.4	**44.7**	46.0
German–English	43.5	**43.7**	**43.7**	45.2
English–German	47.3	**47.4**	47.2	48.2
Hindi–English	**62.5**	62.2	**62.5**	62.6
English–Hindi	53.3	**53.7**	53.5	55.7
Russian–English	47.6	**47.7**	**47.7**	50.6
English–Russian	**46.5**	46.1	46.4	48.2
MEAN	48.0	48.1	**48.2**	49.2

Table 7: Accuracies for each method across 100 folds, for each translation task. The oracle uses the most frequent outcome between each pair of systems, and therefore might not constitute a feasible ranking.

After training, each model has defined a partial ordering over systems.[6] This is then used to compute accuracy on all the pairwise judgments in the test fold. This process yields 100 accuracies for each method; the average accuracy across all the folds can then be used to compute the best method.

Table 7 contains accuracy results for the three methods on the WMT14 tasks. On average, there is a small improvement in accuracy moving from Expected Wins to the H&M model, and then again to the TrueSkill model; however, there is no pattern to the best model for each class. The Oracle column is computed by selecting the most probable outcome ($\pi \in \{<, =, >\}$) for each system pair, and provides an upper bound on accuracy when predicting outcomes using only system-level information. Furthermore, this method of oracle computation might not represent a feasible ranking or clustering,[7].

The TrueSkill approach was best overall, so we used it to produce the official rankings for all language pairs.

3.5 Rank Ranges and Clusters

Above we saw how to produce system scores for each method, which provides a total ordering of the systems. But we would also like to know if the obtained system ranking is statistically significant. Given the large number of systems that participate, and the similarity of the underlying systems resulting from the common training data condition and (often) toolsets, there will be some systems that will be very close in quality. These systems should be grouped together in equivalence classes.

To establish the reliability of the obtained system ranking, we use bootstrap resampling. We sample from the set of pairwise rankings an equal sized set of pairwise rankings (allowing for multiple drawings of the same pairwise ranking), compute a TrueSkill model score for each system based on this sample, and then rank the systems from $1..|\{S_j\}|$. By repeating this procedure 1,000 times, we can determine a range of ranks, into which system falls at least 95% of the time (i.e., at least 950 times) — corresponding to a p-level of $p \leq 0.05$. Furthermore, given the rank ranges for each system, we can cluster systems with overlapping rank ranges.[8]

Table 8 reports all system scores, rank ranges, and clusters for all language pairs and all systems. The official interpretation of these results is that systems in the same cluster are considered tied. Given the large number of judgments that we collected, it was possible to group on average about two systems in a cluster, even though the systems in the middle are typically in larger clusters.

3.6 Cluster analysis

The official ranking results for English-German produced clusters compute at the 90% confidence level due to the presence of a very large cluster (of nine systems). While there is always the possibility that this cluster reflects a true ambiguity, it is more likely due to the fact that we didn't have enough data: English–German had the most sys-

[6] It is a total ordering when $r = 0$, or when all the system scores are outside the decision radius.

[7] For example, if there were a cycle of "better than" judgments among a set of systems.

[8] Formally, given ranges defined by start(S_i) and end(S_i), we seek the largest set of clusters $\{C_c\}$ that satisfies:

$$\forall S \; \exists C : S \in C$$
$$S \in C_a, S \in C_b \rightarrow C_a = C_b$$
$$C_a \neq C_b \rightarrow \forall S_i \in C_a, S_j \in C_b :$$
$$\text{start}(S_i) > \text{end}(S_j) \text{ or } \text{start}(S_j) > \text{end}(S_i)$$

Czech–English

#	score	range	system
1	0.591	1	ONLINE-B
2	0.290	2	UEDIN-PHRASE
3	-0.171	3-4	UEDIN-SYNTAX
	-0.243	3-4	ONLINE-A
4	-0.468	5	CU-MOSES

English–Czech

#	score	range	system
1	0.371	1-3	CU-DEPFIX
	0.356	1-3	UEDIN-UNCNSTR
	0.333	1-4	CU-BOJAR
	0.287	3-4	CU-FUNKY
2	0.169	5-6	ONLINE-B
	0.113	5-6	UEDIN-PHRASE
3	0.030	7	ONLINE-A
4	-0.175	8	CU-TECTO
5	-0.534	9	COMMERCIAL1
6	-0.950	10	COMMERCIAL2

Russian–English

#	score	range	system
1	0.583	1	AFRL-PE
2	0.299	2	ONLINE-B
3	0.190	3-5	ONLINE-A
	0.178	3-5	PROMT-HYBRID
	0.123	4-7	PROMT-RULE
	0.104	5-8	UEDIN-PHRASE
	0.069	5-8	YANDEX
	0.066	5-8	ONLINE-G
4	-0.017	9	AFRL
5	-0.159	10	UEDIN-SYNTAX
6	-0.306	11	KAZNU
7	-0.487	12	RBMT1
8	-0.642	13	RBMT4

English–Russian

#	score	range	system
1	0.575	1-2	PROMT-RULE
	0.547	1-2	ONLINE-B
2	0.426	3	PROMT-HYBRID
3	0.305	4-5	UEDIN-UNCNSTR
	0.231	4-5	ONLINE-G
4	0.089	6-7	ONLINE-A
	0.031	6-7	UEDIN-PHRASE
5	-0.920	8	RBMT4
6	-1.284	9	RBMT1

German–English

#	score	range	system
1	0.451	1	ONLINE-B
2	0.267	2-3	UEDIN-SYNTAX
	0.258	2-3	ONLINE-A
3	0.147	4-6	LIMSI-KIT
	0.146	4-6	UEDIN-PHRASE
	0.138	4-6	EU-BRIDGE
4	0.026	7-8	KIT
	-0.049	7-8	RWTH
5	-0.125	9-11	DCU-ICTCAS
	-0.157	9-11	CMU
	-0.192	9-11	RBMT4
6	-0.306	12	RBMT1
7	-0.604	13	ONLINE-C

French–English

#	score	range	system
1	0.608	1	UEDIN-PHRASE
2	0.479	2-4	KIT
	0.475	2-4	ONLINE-B
	0.428	2-4	STANFORD
3	0.331	5	ONLINE-A
4	-0.389	6	RBMT1
5	-0.648	7	RBMT4
6	-1.284	8	ONLINE-C

English–French

#	score	range	system
1	0.327	1	ONLINE-B
2	0.232	2-4	UEDIN-PHRASE
	0.194	2-5	KIT
	0.185	2-5	MATRAN
	0.142	4-6	MATRAN-RULES
	0.120	4-6	ONLINE-A
3	0.003	7-9	UU-DOCENT
	-0.019	7-10	PROMT-HYBRID
	-0.033	7-10	UA
	-0.069	8-10	PROMT-RULE
4	-0.215	11	RBMT1
5	-0.328	12	RBMT4
6	-0.540	13	ONLINE-C

English–German

#	score	range	system
1	0.264	1-2	UEDIN-SYNTAX
	0.242	1-2	ONLINE-B
2	0.167	3-6	ONLINE-A
	0.156	3-6	PROMT-HYBRID
	0.155	3-6	PROMT-RULE
	0.155	3-6	UEDIN-STANFORD
3	0.094	7	EU-BRIDGE
4	0.033	8-10	RBMT4
	0.031	8-10	UEDIN-PHRASE
	0.012	8-10	RBMT1
5	-0.032	11-12	KIT
	-0.069	11-13	STANFORD-UNC
	-0.100	12-14	CIMS
	-0.126	13-15	STANFORD
	-0.158	14-16	UU
	-0.191	15-16	ONLINE-C
6	-0.307	17-18	IMS-TTT
	-0.325	17-18	UU-DOCENT

Hindi–English

#	score	range	system
1	1.326	1	ONLINE-B
2	0.559	2-3	ONLINE-A
	0.476	2-4	UEDIN-SYNTAX
	0.434	3-4	CMU
3	0.323	5	UEDIN-PHRASE
4	-0.198	6-7	AFRL
	-0.280	6-7	IIT-BOMBAY
5	-0.549	8	DCU-LINGO24
6	-2.092	9	IIIT-HYDERABAD

English–Hindi

#	score	range	system
1	1.008	1	ONLINE-B
2	0.915	2	ONLINE-A
3	0.214	3	UEDIN-UNCNSTR
4	0.120	4-5	UEDIN-PHRASE
	0.054	4-5	CU-MOSES
5	-0.111	6-7	IIT-BOMBAY
	-0.142	6-7	IPN-UPV-CNTXT
6	-0.233	8-9	DCU-LINGO24
	-0.261	8-9	IPN-UPV-NODEV
7	-0.449	10-11	MANAWI-H1
	-0.494	10-11	MANAWI
8	-0.622	12	MANAWI-RMOOV

Table 8: Official results for the WMT14 translation task. Systems are ordered by their inferred system means. Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq .05$, except for English–German, where $p \leq 0.1$. This method is also used to determine the range of ranks into which system falls. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task.

tems (18, compared to 13 for the next languages), yet only an average amount of per-system data. Here, we look at this language pair in more detail, in order to justify this decision, and to shed light on the differences between the ranking methods.

Table 9 presents the 95% confidence-level clusterings for English–German computed with each of the three methods, along with lines that show the reorderings of the systems between them. Reorderings of this type have been used to argue against the reliability of the official WMT ranking (Lopez, 2012; Hopkins and May, 2013). This table shows that these reorderings are captured entirely by the clustering approach we used. This relative *consensus* of these independently-computed and somewhat different models suggests that the published ranking is approaching the true ambiguity underlying systems within the same cluster.

Looking across all language pairs, we find that the total ordering predicted by EW and TS is exactly the same for eight of the ten language pair tasks, and is constrained to reorderings within the official cluster for the other two (German–English — just one adjacent swap — and English–German, depicted in Table 9).

3.7 Conclusions

The official ranking method employed by WMT over the past few years has changed a few times as a result of error analysis and introspection. Until this year, these results were largely based on the intuitions of the community and organizers about deficiencies in the models. In addition to their intuitive appeal, many of these changes (such as the decision to throw out comparisons against references) have been empirically validated Hopkins and May (2013). The actual effect of the refinements in the ranking metric has been minor perturbations in the permutation of systems. The clustering method of Koehn (2012b), in which the official rankings are presented as a partial (instead of total) ordering, alleviated many of the problems observed by Lopez (2012), and also capture all the variance across the new systems introduced this year. In addition, presenting systems as clusters appeals to intuition. As such, we disagree with claims that there is a problem with irreproducibility of the results of the workshop evaluation task, and especially disagree that there is anything approaching a "crisis of confidence" (Hopkins and May, 2013). These claims seem to us to be overstated.

Conducting proper model selection by comparison on held-out data, however, is a welcome suggestion, and our inclusion of this process supports improved confidence in the ranking results. That said, it is notable that the different methods compute very similar orderings. This avoids hallucinating distinctions among systems that are not really there, and captures the intuition that some systems are basically equivalent. The chief benefit of the TrueSkill model is not in outputting a better complete ranking of the systems, but lies in its reduced variance, which allow us to cluster the systems with less data. There is also the unexplored avenue of using TrueSkill to drive the data collection, steering the annotations of judges towards evenly matched systems during the collection phase, potentially allowing confident results to be presented while collecting even less data.

There is, of course, more work to be done. We have produced this year statistically significant clusters with a third of the data required last year, which is an improvement. Models of relative ability are a natural fit for the manual evaluation, and the introduction of an online Bayesian approach to data collection present further opportunities to reduce the amount of data needed. These methods also provide a framework for extending the models in a variety of potentially useful ways, including modeling annotator bias, incorporating sentence metadata (such as length, difficulty, or subtopic), and adding features of the sentence pairs.

4 Quality Estimation Task

Machine translation quality estimation is the task of predicting a quality score for a machine translated text without access to reference translations. The most common approach is to treat the problem as a supervised machine learning task, using standard regression or classification algorithms. The third edition of the WMT shared task on quality estimation builds on the previous editions of the task (Callison-Burch et al., 2012; Bojar et al., 2013), with tasks including both sentence-level and word-level estimation, with new training and test datasets.

The goals of this year's shared task were:

- To investigate the effectiveness of different quality labels.
- To explore word-level quality prediction at

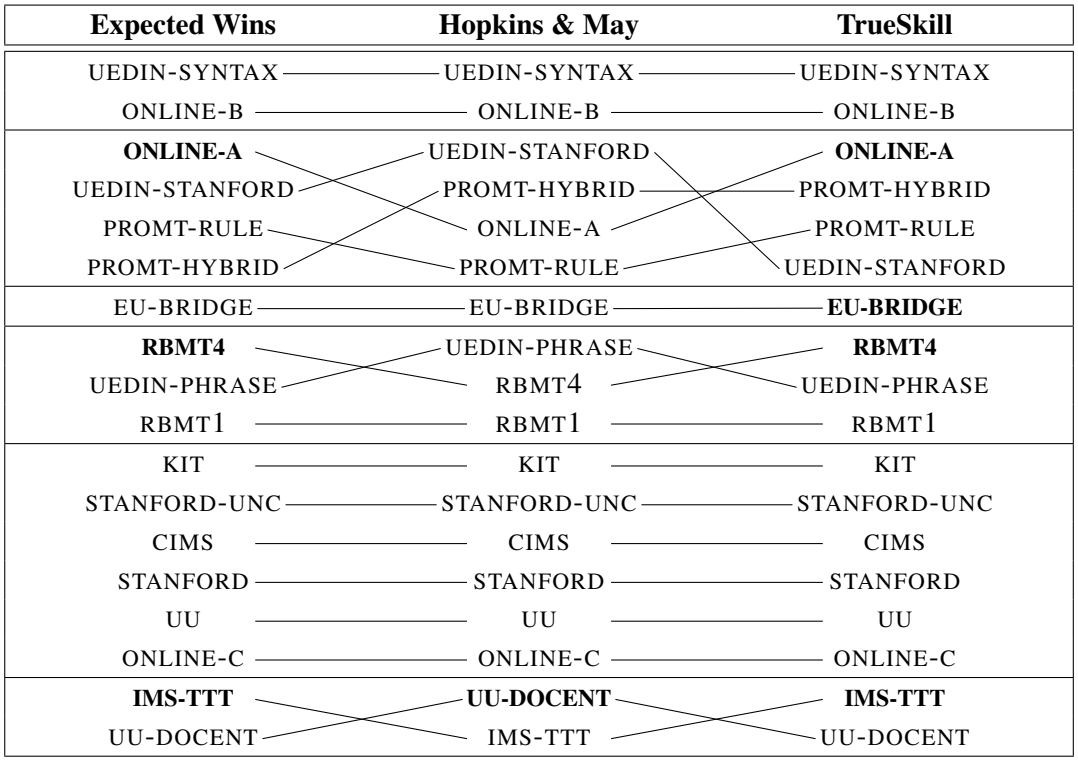

Table 9: A comparison of the rankings produced by Expected Wins, Hopkins & May, and TrueSkill for English–German (the task with the most systems and the largest cluster). The lines extending all the way across mark the official English–German clustering (computed from TrueSkill with 90% confidence intervals), while **bold** entries mark the start of new clusters within each method or column (computed at the 95% confidence level). The TrueSkill clusterings contain all the system reorderings across the other two ranking methods.

different levels of granularity.

- To study the effects of training and test datasets with mixed domains, language pairs and MT systems.
- To examine the effectiveness of quality prediction methods on human translations.

Four tasks were proposed: Tasks 1.1, 1.2, 1.3 are defined at the sentence-level (Sections 4.1), while Task 2, at the word-level (Section 4.2). Each task provides one or more datasets with up to four language pairs each: English-Spanish, English-German, German-English, Spanish-English, and up to four alternative translations generated by: a statistical MT system (SMT), a rule-based MT system (RBMT), a hybrid MT system, and a human. These datasets were annotated with different labels for quality by professional translators as part of the QTLaunchPad[9] project. External resources (e.g. parallel corpora) were provided to participants. Any additional resources, including additional quality estimation training data, could be used by participants (no distinction between *open* and *close* tracks is made). Participants were also provided with a software package to extract quality estimation features and perform model learning, with a suggested list of *baseline* features and learning method for sentence-level prediction. Participants, described in Section 4.3, could submit up to two systems for each task.

Data used for building specific MT systems or internal system information (such as n-best lists) were not made available this year as multiple MT systems were used to produced the datasets, including rule-based systems. In addition, part of the translations were produced by humans. Information on the sources of translations was not provided either. Therefore, as a general rule, participants were only allowed to use black-box features.

4.1 Sentence-level Quality Estimation

For the sentence-level tasks, two variants of the results could be submitted for each task and language pair:

- **Scoring**: An absolute quality score for each sentence translation according to the type of

[9] http://www.qt21.eu/launchpad/

prediction, to be interpreted as an error metric: lower scores mean better translations.

- **Ranking**: A ranking of sentence translations for all source test sentences from best to worst. For this variant, it does not matter how the ranking is produced (from HTER predictions, likert predictions, or even without machine learning).

Evaluation was performed against the true label and/or HTER ranking using the same metrics as in previous years:

- **Scoring**: Mean Average Error (MAE) (primary metric), Root Mean Squared Error (RMSE).
- **Ranking**: DeltaAvg (primary metric) (Bojar et al., 2013) and Spearman's rank correlation.

For all sentence-level these tasks, the same 17 features as in WMT12-13 were used to build baseline systems. The SVM regression algorithm within QUEST (Specia et al., 2013)[10] was applied for that with RBF kernel and grid search for parameter optimisation.

Task 1.1 Predicting post-editing effort

Data in this task is labelled with discrete and absolute scores for perceived post-editing effort, where:

- **1** = Perfect translation, no post-editing needed at all.
- **2** = Near miss translation: translation contains maximum of 2-3 errors, and possibly additional errors that can be easily fixed (capitalisation, punctuation, etc.).
- **3** = Very low quality translation, cannot be easily fixed.

The datasets were annotated in a "triage" phase aimed at selecting translations of type "2" (near miss) that could be annotated for errors at the word-level using the MQM metric (see Task 2, below) for a more fine-grained and systematic translation quality analysis. Word-level errors in translations of type "3" are too difficult if not impossible to annotate and classify, particularly as they often contain inter-related errors in contiguous or overlapping word spans.

[10] http://www.quest.dcs.shef.ac.uk/

For the *training* of prediction models, we provide a new dataset consisting of source sentences and their human translations, as well as two-three versions of machine translations (by an SMT system, an RBMT system and, for English-Spanish/German only, a hybrid system), all in the news domain, extracted from tests sets of various WMT years and MT systems that participated in the translation shared task:

# Source sentences	# Target sentences
954 English	3,816 Spanish
350 English	1,400 German
350 German	1,050 English
350 Spanish	1,050 English

As *test* data, for each language pair and MT system (or human translation) we provide a new set of translations produced by the same MT systems (and humans) as those used for the training data:

# Source sentences	# Target sentences
150 English	600 Spanish
150 English	600 German
150 German	450 English
150 Spanish	450 English

The distribution of true scores in both training and test sets for each language pair is given in Figures 3.

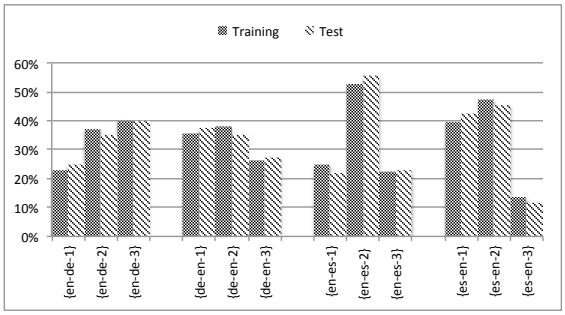

Figure 3: Distribution of true 1-3 scores by langauge pair.

Additionally, we provide some out of domain test data. These translations were annotated in the same way as above, each dataset by one Language Service Provider (LSP), i.e, one professional translator, with two LPSs producing data independently for English-Spanish. They were generated using the LSPs' own source data (a different domain from news), and own MT system (different from the three used for the official datasets). The results on these datasets were not considered

for the official ranking of the participating systems:

# Source sentences	# Target sentences
971 English	971 Spanish
297 English	297 German
388 Spanish	388 English

Task 1.2 Predicting percentage of edits

In this task we use HTER (Snover et al., 2006) as quality score. This score is to be interpreted as the minimum edit distance between the machine translation and its manually post-edited version, and its range is [0, 1] (0 when no edit needs to be made, and 1 when all words need to be edited). We used TERp (default settings: tokenised, case insensitive, etc., but capped to 1)[11] to compute the HTER scores.

For practical reasons, the data is a subset of Task 1.1's dataset: only translations produced by the SMT system English-Spanish. As *training data*, we provide 896 English-Spanish translation suggestions and their post-editions. As *test* data, we provide a new set of 208 English-Spanish translations produced by the same SMT system. Each of the training and test translations was post-edited by a professional translator using the CASMACAT[12] web-based tool, which also collects post-editing time on a sentence-basis.

Task 1.3 Predicting post-editing time

For this task systems are required to produce, for each translation, a real valued estimate of the time (in milliseconds) it takes a translator to post-edit the translation. The training and test sets are a subset of that uses in Task 1.2 (subject to filtering of outliers). The difference is that the labels are now the number of milliseconds that were necessary to post-edit each translation.

As *training data*, we provide 650 English-Spanish translation suggestions and their post-editions. As *test* data, we provide a new set of 208 English-Spanish translations (same test data as for Task 1.2).

4.2 Word-level Quality Estimation

The data for this task is based on a subset of the datasets used for Task 1.1, for all language pairs, human and machine translations: those translations labelled "2" (near misses), plus additional data provided by industry (either on the news domain or on other domains, such as technical documentation, produced using their own MT systems, and also pre-labelled as "2"). All segments were annotated with word-level labels by professional translators using the core categories in MQM (Multidimensional Quality Metrics)[13] as error typology (see Figure 4). Each word or sequence of words was annotated with a single error. For (supposedly rare) cases where a decision between multiple fine-grained error types could not be made, annotators were requested to choose a coarser error category in the hierarchy.

Participants are asked to produce a label for each token that indicates quality at different levels of granularity:

- **Binary classification**: an OK / bad label, where bad indicates the need for editing the token.
- **Level 1 classification**: an OK / accuracy / fluency label, specifying coarser level categories of errors for each token, or "OK" for tokens with no error.
- **Multi-class classification**: one of the labels specifying the error type for the token (terminology, mistranslation, missing word, etc.) in Figure 4, or "OK" for tokens with no error.

As *training* data, we provide tokenised translation output for all language pairs, human and machine translations, with tokens annotated with all issue types listed above, or "OK". The annotation was performed manually by professional translators as part of the QTLaunchPad project. For the coarser variants, fine-grained errors are generalised to Accuracy or Fluency, or "bad" for the binary variant. The amount of available training data varies by language pair:

# Source sentences	# Target sentences
1,957 English	1,957 Spanish
715 English	715 German
350 German	350 English
900 Spanish	900 English

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

[12] http://casmacat.eu/

[13] http://www.qt21.eu/launchpad/content/training

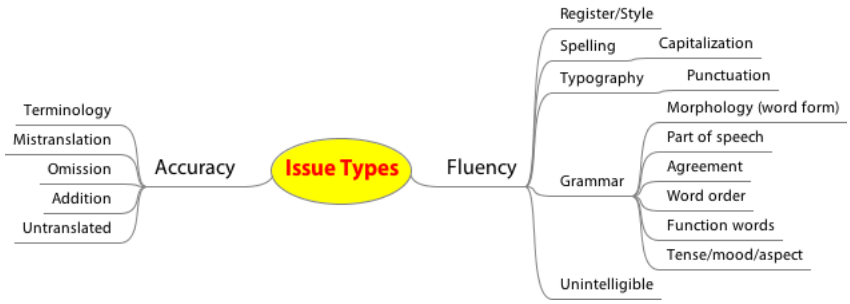

Figure 4: MQM metric as error typology.

As *test* data, we provide additional data points for all language pairs, human and machine translations:

# Source sentences	# Target sentences
382 English	382 Spanish
150 English	150 German
100 German	100 English
150 Spanish	150 English

In contrast to Tasks 1.1–1.3, no baseline feature set is provided to the participants.

Similar to last year (Bojar et al., 2013), the word-level task is primarily evaluated by macro-averaged F-measure (in %). Because the class distribution is skewed – in the test data about 78% of the tokens are marked as "OK" – we compute precision, recall, and F_1 for each class individually, weighting F_1 scores by the frequency of the class in the test data. This avoids giving undue importance to less frequent classes. Consider the following confusion matrix for Level 1 annotation, i.e. the three classes *(O)K*, *(F)luency*, and *(A)ccuracy*:

		reference		
		O	F	A
predicted	O	4172	1482	193
	F	1819	1333	214
	A	198	133	69

For each of the three classes we assume a binary setting (one-vs-all) and derive true-positive (tp), false-positive (fp), and false-negative (fn) counts from the rows and columns of the confusion matrix as follows:

$$
\begin{aligned}
tp_O &= 4172 \\
fp_O &= 1482 + 193 = 1675 \\
fn_O &= 1819 + 198 = 2017 \\
tp_F &= 1333 \\
fp_F &= 1819 + 214 = 2033 \\
fn_F &= 1482 + 133 = 1615 \\
tp_A &= 69 \\
fp_A &= 198 + 133 = 331 \\
fn_A &= 193 + 214 = 407
\end{aligned}
$$

We continue to compute F_1 scores for each class $c \in \{O, F, A\}$:

$$
\begin{aligned}
\text{precision}_c &= tp_c/(tp_c + fp_c) \\
\text{recall}_c &= tp_c/(tp_c + fn_c) \\
F_{1,c} &= \frac{2 \cdot \text{precision}_c \cdot \text{recall}_c}{\text{precision}_c + \text{recall}_c}
\end{aligned}
$$

yielding:

$$
\begin{aligned}
\text{precision}_O &= 4172/(4172 + 1675) = 0.7135 \\
\text{recall}_O &= 4172/(4172 + 2017) = 0.6741 \\
F_{1,O} &= \frac{2 \cdot 0.7135 \cdot 0.6741}{0.7135 + 0.6741} = 0.6932 \\
&\dots \\
F_{1,F} &= 0.4222 \\
F_{1,A} &= 0.1575
\end{aligned}
$$

Finally, we compute the average of $F_{1,c}$ scores weighted by the occurrence count $N(c)$ of c:

$$
\text{weighted } F_{1,\text{ALL}} = \frac{1}{\sum_c N(c)} \sum_c N_c \cdot F_{1,c}
$$

$$
\text{weighted } F_{1,\text{ERR}} = \frac{1}{\sum_{c:c \neq O} N(c)} \sum_{c:c \neq O} N_c \cdot F_{1,c}
$$

which for the above example gives:

$$\text{weighted } F_{1,\text{ALL}} = \frac{1}{6189 + 2948 + 476} \cdot (6189 \cdot 0.6932 + 2948 \cdot 0.4222 + 476 \cdot 0.1575) = 0.5836$$

$$\text{weighted } F_{1,\text{ERR}} = \frac{1}{2948 + 476} \cdot (2948 \cdot 0.4222 + 476 \cdot 0.1575) = 0.3854$$

We choose $F_{1,ERR}$ as our primary evaluation measure because it most closely mimics the common application of F_1 scores in binary classification: one is interested in the performance in detecting a *positive class*, which in this case would be erroneous words. This does, however, ignore the number of correctly classified words of the *OK* class, which is why we also report $F_{1,ALL}$. In addition, we follow Powers (2011) and report Matthews Correlation Coefficient (MCC), averaged in the same way as F_1, as our secondary metric. Finally, for contrast we also report Accuracy (ACC).

4.3 Participants

Table 10 lists all participating teams. Each team was allowed up to two submissions for each task and language pair. In the descriptions below, participation in specific tasks is denoted by a task identifier: T1.1, T1.2, T1.3, and T2.

Sentence-level baseline system (T1.1, T1.2, T1.3): QUEST is used to extract 17 system-independent features from source and translation sentences and parallel corpora (same features as in the WMT12 shared task):

- number of tokens in the source and target sentences.
- average source token length.
- average number of occurrences of the target word within the target sentence.
- number of punctuation marks in source and target sentences.
- language model (LM) probability of source and target sentences based on models for the WMT News Commentary corpus.
- average number of translations per source word in the sentence as given by IBM Model 1 extracted from the WMT News Commentary parallel corpus, and thresholded so that $P(t|s) > 0.2$, or so that $P(t|s) > 0.01$ weighted by the inverse frequency of each word in the source side of the parallel corpus.
- percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source language extracted from the WMT News Commentary corpus.
- percentage of unigrams in the source sentence seen in the source side of the WMT News Commentary corpus.

These features are used to train a Support Vector Machine (SVM) regression algorithm using a radial basis function kernel within the SCIKIT-LEARN toolkit. The γ, ϵ and C parameters were optimised via grid search with 5-fold cross validation on the training set. We note that although the system is referred to as "baseline", it is in fact a strong system. It has proved robust across a range of language pairs, MT systems, and text domains for predicting various forms of post-editing effort (Callison-Burch et al., 2012; Bojar et al., 2013).

DCU (T1.1): DCU-MIXED and DCU-SVR use a selection of features available in QUEST, such as punctuation statistics, LM perplexity, n-gram frequency quartile statistics and coarse-grained POS frequency ratios, and four additional feature types: combined POS and stop word LM features, source-side pseudo-reference features, inverse glass-box features for translating the translation and error grammar parsing features. For machine learning, the QUEST framework is expanded to combine logistic regression and support vector regression and to handle cross- validation and randomisation in a way that training items with the same source side are kept together. External resources are monolingual corpora taken from the WMT 2014 translation task for LMs, the MT system used for the inverse glass-box features (Li et al., 2014b) and, for error grammar parsing, the Penn-Treebank and an error grammar derived from it (Foster, 2007).

ID	Participating team
DCU	Dublin City University Team 1, Ireland (Hokamp et al., 2014)
DFKI	German Research Centre for Artificial Intelligence, Germany (Avramidis, 2014)
FBK-UPV-UEDIN	Fondazione Bruno Kessler, Italy, UPV Universitat Politècnica de València, Spain & University of Edinburgh, UK (Camargo de Souza et al., 2014)
LIG	Laboratoire d'Informatique Grenoble, France (Luong et al., 2014)
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, France (Wisniewski et al., 2014)
MULTILIZER	Multilizer, Finland
RTM-DCU	Dublin City University Team 2, Ireland (Bicici and Way, 2014)
SHEF-lite	University of Sheffield Team 1, UK (Beck et al., 2014)
USHEFF	University of Sheffield Team 2, UK (Scarton and Specia, 2014)
YANDEX	Yandex, Russia

Table 10: Participants in the WMT14 Quality Estimation shared task.

DFKI (T1.2): DFKI/SVR builds upon the baseline system (above) by adding non-redundant data from the WMT13 task for predicting the same label (HTER) and additional features such as (a) rule-based language corrections (language tool) (b), PCFG parsing statistics and counts of tree labels, (c) position statistics of parsing labels, (d) position statistics of trigrams with low probability. DFKI/SVRxdata uses a similar setting, with the addition of more training data from non-minimally post-edited translation outputs (references), filtered based on a threshold on the edit distance between the MT output and the freely-translated reference.

FBK-UPV-UEDIN (T1.2, T1.3, T2): The submissions for the word-level task (T2) use features extracted from word posterior probabilities and confusion network descriptors computed over the 100k-best hypothesis translations generated by a phrase-based SMT system. They also use features from word lexicons, and POS tags of each word for source and translation sentences. The predictions of the Binary model are used as a feature for the Level 1 and Multi-class settings. Both conditional random fields (CRF) and bidirectional long short-term memory recurrent neural networks (BLSTM-RNNs) are used for the Binary setting, and BLSTM-RNNs only for the Level 1 and Multi-class settings.

The sentence-level QE submissions (T1.2 and T1.3) are trained on black-box features extracted using QUEST in addition to features based on word alignments, word posterior probabilities and diversity scores (Souza et al., 2013). These features are computed over 100k-best hypothesis translations also used for task 2. In addition, a set of ratios computed from the word-level predictions of the model trained on the binary setting of task 2 is used. A total of 221 features and the extremely randomised trees (Geurts et al., 2006) learning algorithm are used to train regression models.

LIG (T2): Conditional Random Fields classifiers are trained with features used in LIG's WMT13 systems (Luong et al., 2013): target and source words, alignment information, source and target alignment context, LM scores, target and source POS tags, lexical categorisations (stopword, punctuation, proper name, numerical), constituent label, depth in the constituent tree, target polysemy count, pseudo reference. These are combined with novel features: word occurrence in multiple translation systems and POS tag-based LM scores (longest target/source n-gram length and backoff score for POS tag). These features require external NLP tools and resources such as: TreeTagger, GIZA++, Bekerley parser, Link Grammar parser, WordNet and BabelNet, Google Translate (pseudo-reference). For the binary task, the optimal classification threshold is tuned based on a development set split from the original training set. Feature selection is employed over the all features (for the binary

task only), with the Sequential Backward Selection algorithm. The best performing feature set is then also used for the Level 1 and Multi-class variants.

LIMSI (T2): The submission relies on a random forest classifier and considers only 16 dense and continuous features. To prevent sparsity issues, lexicalised information such as the word or the previous word identities is not included. The features considered are mostly classic MT features and can be categorised into two classes: *association features*, which describe the quality of the association between the source sentence and each target word, and *fluency features*, which describe the 'quality' of the translation hypotheses. The latter rely on different language models (either on POS or on words) and the former on IBM Model 1 translation probabilities and on pseudo- references, i.e. translation produced by an independent MT system. Random forests are known to perform well in tasks like this one, in which only a few dense and continuous features are available, possibly because of their ability to take into account complex interactions between features and to automatically partition the continuous feature values into a discrete set of intervals that achieves the best classification performance. Since they predict the class probabilities, it is possible to directly optimize the F_1 score during training by finding, with a grid search method, the decision threshold that achieved the best F_1 score on the training set.

MULTILIZER (T1.2, T1.3): The 80 black-box features from QUEST are used in addition to new features based on using other MT engines for forward and backward translations. In forward translations, the idea is that different MT engines make different mistakes. Therefore, when several forward translations are similar to each other, these translations are more likely to be correct. This is confirmed by the Pearson correlation of similarities between the forward translations against the true scores (above 0.5). A backward translation is very error-prone and therefore it has to be used in combination with forward translations. A single back-translation similar to original source segment does not bring much information. Instead, when several MT engines give back-translations similar to this source segment, one can conclude that the translation is reliable. Those translations where similarities both in forward translation and backward translation are high are intuitively more likely to be good. A simple feature selection method that omits all features with Pearson correlation against the true scores below 0.2 is used. The systems submitted are obtained using linear regression models.

RTM-DCU (T1.1, T1.2, T1.3, T2): RTM-DCU systems are based on referential translation machines (RTM) (Biçici, 2013) and parallel feature decay algorithms (ParFDA5) (Biçici et al., 2014), which allow language and MT system-independent predictions. For each task, individual RTM models are developed using the parallel corpora and the language model corpora distributed by the WMT14 translation task and the language model corpora provided by LDC for English and Spanish. RTMs use 337 to 437 sentence-level features for coverage and diversity, IBM1 and sentence translation performance, retrieval closeness and minimum Bayes retrieval risk, distributional similarity and entropy, IBM2 alignment, character n-grams, sentence readability, and parse output tree structures. The features use ngrams defined over text or common cover link (CCL) (Seginer, 2007) structures as the basic units of information over which similarity calculations are performed. Learning models include ridge regression (RR), support vector machines (SVR), and regression trees (TREE), which are applied after partial least squares (PLS) or feature selection (FS). For word-level prediction, generalised linear models (GLM) (Collins, 2002) and GLM with dynamic learning (GLMd) (Biçici, 2013) are used with word-level features including CCL links, word length, location, prefix, suffix, form, context, and alignment, totalling up to a couple of million features.

SHEF-lite (T1.1, T1.2, T1.3): These submissions use the framework of Multi-task Gaussian Processes, where multiple datasets are

combined in a multi-task setting similar to the one used by Cohn and Specia (2013). For T1.1, data for all language pairs is put together, and each language is considered a task. For T1.2 and T1.3, additional datasets from previous shared task years are used, each encoded as a different task. For all tasks, the QUEST framework is used to extract a set of 80 black-box features (a superset of the 17 baseline features). To cope with the large size of the datasets, the SHEF-lite-sparse submission uses Sparse Gaussian Processes, which provide sensible sparse approximations using only a subset of instances (inducing inputs) to speed up training and prediction. For this "sparse" submission, feature selection is performed following the approach of Shah et al. (2013) by ranking features according to their learned length-scales and selecting the top 40 features.

USHEFF (T1.1, T1.2, T1.3): USHEFF submissions exploit the use of consensus among MT systems by comparing the MT system output to several alternative translations generated by other MT systems (pseudo-references). The comparison is done using standard evaluation metrics (BLEU, TER, METEOR, ROUGE for all tasks, and two metrics based on syntactic similarities from shallow and dependency parser information for T1.2 and T1.3). Figures extracted from such metrics are used as features to complement prediction models trained on the 17 baseline features. Different from the standard use of pseudo-reference features, these features do not assume that the alternative MT systems are better than the system of interest. A more realistic scenario is considered where the quality of the pseudo-references is not known. For T1, no external systems in addition to those provided for the shared task are used: for a given translation, all alternative translations for the same source segment (two or three, depending on the language pair) are used as pseudo-references. For T1.2 and T1.3, for each source sentence, all alternative translations produced by MT systems on the same data (WMT12/13) are used as pseudo-references. The hypothesis is that by using translations from several MT systems one can find consensual information and this can smooth out the effect of "coincidences" in the similarities between systems' translations. SVM regression with radial basis function kernel and hyper-parameters optimised via grid search is used to build the models.

YANDEX (T1.1): Both submissions are based on the the 80 black-box features, plus an LM score from a larger language model, a pseudo-reference, and several additional features based on POS tags and syntactic parsers. The first attempt uses an extract of the top 5 features selected with a greedy search from the set of all features. SVM regression is used as machine learning algorithm. The second attempt uses the same features processed with Yandex' implementation of the gradient tree boosting (MatrixNet).

4.4 Results

In what follows we give the official results for all tasks followed by a discussion that highlights the main findings for each of the tasks.

Task 1.1 Predicting post-editing effort

Table 11 summarises the results for the ranking variant of Task 1.1. They are sorted from best to worst using the DeltaAvg metric scores as primary key and the Spearman's rank correlation scores as secondary key.

The winning submissions for the ranking variant of Task 1.1 are as follows: for English-Spanish it is RTM-DCU/RTM-TREE, with a DeltaAvg score of 0.26; for Spanish-English it is USHEFF, with a DeltaAvg score of 0.23; for English-German it is again RTM-DCU/RTM-TREE, with a DeltaAvg score of 0.39; and for German-English it is RTM-DCU/RTM-RR, with a DeltaAvg score of 0.38. These winning submissions are better than the baseline system by a large margin, which indicates that current best performance in MT quality estimation has reached levels that are clearly beyond what the baseline system can produce. As for the other systems, according to DeltaAvg, compared to the previous year results a smaller percentage of systems is able to beat the baseline. This might be a consequence of the use of the metric for the prediction of only three discrete labels.

The results for the scoring task are presented in Table 12, sorted from best to worst using the MAE

System ID	DeltaAvg	Spearman Corr
English-Spanish		
• RTM-DCU/RTM-PLS-TREE	0.26	0.38
• RTM-DCU/RTM-TREE	0.26	0.41
• YANDEX/SHAD_BOOSTEDTREES2	0.23	0.35
USHEFF	0.21	0.33
SHEFF-lite	0.21	0.33
YANDEX/SHAD_SVR1	0.18	0.29
SHEFF-lite-sparse	0.17	0.27
Baseline SVM	0.14	0.22
Spanish-English		
• USHEFF	0.23	0.30
• RTM-DCU/RTM-PLS-RR	0.20	0.35
• RTM-DCU/RTM-FS-RR	0.19	0.36
Baseline SVM	0.12	0.21
SHEFF-lite-sparse	0.12	0.17
SHEFF-lite	0.11	0.15
English-German		
• RTM-DCU/RTM-TREE	0.39	0.54
RTM-DCU/RTM-PLS-TREE	0.33	0.42
USHEFF	0.26	0.41
SHEFF-lite	0.26	0.36
Baseline SVM	0.23	0.34
SHEFF-lite-sparse	0.23	0.33
German-English		
• RTM-DCU/RTM-RR	0.38	0.51
• RTM-DCU/RTM-PLS-RR	0.35	0.45
USHEFF	0.28	0.30
SHEFF-lite	0.24	0.27
Baseline SVM	0.21	0.25
SHEFF-lite-sparse	0.14	0.17

Table 11: Official results for the ranking variant of the WMT14 Quality Evaluation Task 1.1. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to bootstrap resampling (1M times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	MAE	RMSE
English-Spanish		
• RTM-DCU/RTM-PLS-TREE	0.49	0.61
• SHEFF-lite	0.49	0.63
• USHEFF	0.49	0.63
• SHEFF-lite/sparse	0.49	0.69
• RTM-DCU/RTM-TREE	0.49	0.61
Baseline SVM	0.52	0.66
YANDEX/SHAD_BOOSTEDTREES2	0.56	0.68
YANDEX/SHAD_SVR1	0.64	0.81
DCU-Chris/SVR	0.66	0.88
DCU-Chris/MIXED	0.94	1.14
Spanish-English		
• RTM-DCU/RTM-FS-RR	0.53	0.64
• SHEFF-lite/sparse	0.54	0.69
• RTM-DCU/RTM-PLS-RR	0.55	0.71
USHEFF	0.57	0.67
Baseline SVM	0.57	0.68
SHEFF-lite	0.62	0.77
DCU-Chris/MIXED	0.65	0.91
English-German		
• RTM-DCU/RTM-TREE	0.58	0.68
RTM-DCU/RTM-PLS-TREE	0.60	0.71
SHEFF-lite	0.63	0.74
USHEFF	0.64	0.75
SHEFF-lite/sparse	0.64	0.75
Baseline SVM	0.64	0.76
DCU-Chris/MIXED	0.69	0.98
German-English		
• RTM-DCU/RTM-RR	0.55	0.67
• RTM-DCU/RTM-PLS-RR	0.57	0.74
USHEFF	0.63	0.76
SHEFF-lite	0.65	0.77
Baseline SVM	0.65	0.78

Table 12: Official results for the scoring variant of the WMT14 Quality Evaluation Task 1.1. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to bootstrap resampling (1M times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

metric scores as primary key and the RMSE metric scores as secondary key.

The winning submissions for the scoring variant of Task 1.1 are as follows: for English-Spanish it is RTM-DCU/RTM-TREE with a MAE of 0.49; for Spanish-English it is RTM-DCU/RTM-FS-RR with a MAE of 0.53; for English-German it is again RTM-DCU/RTM-TREE, with a MAE of 0.58; and for German-English it is RTM-DCU/RTM-RR with a MAE of 0.55. These submissions are again much better than the baseline system, which under the scoring variant seems to perform at a middle-of-the-pack level or lower compared to the overall pool of submissions. Overall, more systems are able to outperform the baseline according to the scoring metric.

The top system for most language pairs are essentially based on the same core techniques (RTM-DCU) according to both the DeltaAvg and MAE metrics. The ranking of other systems, however, can be substantially different according to the two metrics.

Task 1.2 Predicting percentage of edits

Table 13 summarises the results for the ranking variant of Task 1.2. For readability purposes we have used a multiplication-factor of 100 in the scoring script, which makes the HTER numbers (both predicted and gold) to be in the [0, 100] range. They are sorted from best to worst using the DeltaAvg metric scores as primary key and the Spearman's rank correlation scores as secondary key.

The winning submission for the ranking variant of Task 1.2 is RTM-DCU/RTM-SVR, with a DeltaAvg score of 9.31. There is a large margin between this score and the baseline score of DeltaAvg 5.08, which indicates again that current best performance has reached levels that are much beyond what this baseline system can produce. The vast majority of the submissions perform better than the baseline (the only exception is the submission from SHEFF-lite, for which the authors report a major issue with the learning algorithm).

The results for the scoring variant are presented in Table 14, sorted from best to worst by using the MAE metric scores as primary key and the RMSE metric scores as secondary key.

The winning submission for the scoring variant of Task 1.2 is FBK-UPV-UEDIN/WP with a MAE of 12.89, while the baseline system has a MAE of 15.23. Most of the submissions perform better than the baseline.

Task 1.3 Predicting post-editing time

Table 15 summarises the results for the ranking variant of Task 1.3. For readability purposes, we have used a multiplication-factor of 0.001 in the scoring script, which makes the time (both predicted and gold) to be measured in seconds. They are sorted from best to worst using the DeltaAvg metric scores as primary key and the Spearman's rank correlation scores as secondary key.

The winning submission for the ranking variant of Task 1.3 is RTM-DCU/RTM-RR, with a DeltaAvg score of 17.02 (when predicting seconds). The interesting aspect of these results is that the DeltaAvg numbers have a direct real-world interpretation, in terms of time spent (or saved, depending on one's view-point) for post-editing machine-produced translations. A more elaborate discussion on this point can be found in Section 4.5.

The winning submission for the scoring variant of Task 1.3 is RTM-DCU/RTM-SVR, with a MAE of 16.77. Note that all of the submissions perform significantly better than the baseline, which has a MAE of 21.49, and that the majority is not significantly worse than the top scoring submission.

Task 2 Predicting word-level edits

The results for Task 2 are summarised in Tables 17–19. The results are ordered by F_1 score for the Error (BAD) class. For comparison, two trivial baselines are included, one that marks every word as correct and that marks every word with the most common error class found in the training data. Both baselines are clearly useless for any application, but help put the results in perspective. Most teams submitted systems for a single language pair: English-Spanish; only a single team produced predictions for all four pairs.

Table 17 gives the results of the binary (OK vs. BAD) classification variant of Task 2. The winning submissions for this variant are as follows: for English-Spanish it is FBK-UPV-UEDIN/RNN with a weighted F_1 of 48.73; for Spanish-English it is RTM-DCU/RTM-GLMd with a weighted F_1 of 29.14; for English-German it is RTM-DCU/RTM-GLM with a weighted F_1 of 45.30; and for German-English it is again RTM-DCU/RTM-GLM with a weighted F_1 of 26.13.

Remarkably, for three out of four language pairs, the systems fail to beat our trivial baseline of

System ID	DeltaAvg	Spearman Corr
English-Spanish		
• RTM-DCU/RTM-SVR	9.31	0.53
• RTM-DCU/RTM-TREE	8.57	0.48
• USHEFF	7.93	0.45
SHEFF-lite/sparse	7.69	0.43
Baseline	5.08	0.31
SHEFF-lite	0.72	0.09

Table 13: Official results for the ranking variant of the WMT14 Quality Evaluation Task 1.2. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to bootstrap resampling (100k times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	MAE	RMSE
English-Spanish		
• FBK-UPV-UEDIN/WP	12.89	16.74
• RTM-DCU/RTM-SVR	13.40	16.69
• USHEFF	13.61	17.84
RTM-DCU/RTM-TREE	14.03	17.48
DFKI/SVR	14.32	17.74
FBK-UPV-UEDIN/NOWP	14.38	18.10
SHEFF-lite/sparse	15.04	18.38
MULTILIZER	15.04	20.86
Baseline	15.23	19.48
DFKI/SVRxdata	16.01	19.52
SHEFF-lite	18.15	23.41

Table 14: Official results for the scoring variant of the WMT14 Quality Evaluation Task 1.2. The winning submissions are indicated by a •. They are statistically indistinguishable from the top submission according to bootstrap resampling (1M times) with 95% confidence intervals. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	DeltaAvg	Spearman Corr
English-Spanish		
• RTM-DCU/RTM-RR	17.02	0.68
• RTM-DCU/RTM-SVR	16.60	0.67
SHEFF-lite/sparse	16.33	0.63
SHEFF-lite	16.08	0.64
USHEFF	14.98	0.59
Baseline	14.71	0.57

Table 15: Official results for the ranking variant of the WMT14 Quality Evaluation Task 1.3. The winning submissions are indicated by a •. They are statistically indistinguishable from the top submission according to bootstrap resampling (1M times) with a 95% confidence interval. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	MAE	RMSE
English-Spanish		
• RTM-DCU/RTM-SVR	16.77	26.17
• MULTILIZER/MLZ2	17.07	25.83
• SHEFF-lite	17.13	27.33
• MULTILIZER/MLZ1	17.31	25.51
• SHEFF-lite/sparse	17.42	27.35
• FBK-UPV-UEDIN/WP	17.48	25.31
RTM-DCU/RTM-RR	17.50	25.97
FBK-UPV-UEDIN/NOWP	18.69	26.58
USHEFF	21.48	34.28
Baseline	21.49	34.28

Table 16: Official results for the scoring variant of the WMT14 Quality Evaluation Task 1.3. The winning submissions are indicated by a •. They are statistically indistinguishable from the top submission according to bootstrap resampling (1M times) with a 95% confidence interval. The systems in the gray area are not different from the baseline system at a statistically significant level according to the same test.

System ID	weighted F_1 All	F_1 Bad ↑	MCC	ACC
English-Spanish				
Baseline (always OK)	50.43	0.00	0.00	64.38
Baseline (always Bad)	18.71	52.53	0.00	35.62
• FBK-UPV-UEDIN/RNN	62.00	48.73	18.23	61.62
LIMSI/RF	60.55	47.32	15.44	60.09
LIG/FS	63.55	44.47	19.41	64.67
LIG/BL ALL	63.77	44.11	19.91	65.12
FBK-UPV-UEDIN/RNN+tandem+crf	62.17	42.63	16.32	63.26
RTM-DCU/RTM-GLM	60.68	35.08	13.45	63.74
RTM-DCU/RTM-GLMd	60.24	32.89	12.98	63.97
Spanish-English				
Baseline (always OK)	74.41	0.00	0.00	82.37
Baseline (always Bad)	5.28	29.98	0.00	17.63
• RTM-DCU/RTM-GLMd	79.54	29.14	25.47	82.98
RTM-DCU/RTM-GLM	79.42	26.91	25.93	83.43
English-German				
Baseline (always OK)	59.39	0.00	0.00	71.33
Baseline (always Bad)	12.78	44.57	0.00	28.67
• RTM-DCU/RTM-GLM	71.51	45.30	28.61	72.97
RTM-DCU/RTM-GLMd	68.73	36.91	21.32	71.41
German-English				
Baseline (always OK)	67.82	0.00	0.00	77.60
Baseline (always Bad)	8.20	36.60	0.00	22.40
• RTM-DCU/RTM-GLM	72.41	26.13	16.08	76.14
RTM-DCU/RTM-GLMd	71.42	22.97	12.63	75.46

Table 17: Official results for the binary part of the WMT14 Quality Evaluation Task 2. The winning submissions are indicated by a •. All values are given as percentages.

marking all the words as wrong. This may either indicate that the predictions themselves are of low quality or the chosen evaluation approach is misleading. On the other hand F_1 scores are a common measure of binary classification performance and no averaging is performed here.

Table 18 gives the results of the Level 1 classification (OK, Fluency, Accuracy) variant of Task 2. Here the second baseline is to always predict Fluency errors, as this is the most common error category in the training data. The winning submissions of this variant are as follows: for English-Spanish it is FBK-UPV-UEDIN/RNN+tandem+crf with a weighted F_1 of 23.94 and for Spanish-English, English-German, and German-English it is RTM-DCU/RTM-GLMd with weighted F_1 scores of 23.94, 21.94, and 8.57 respectively.

As before, all systems fail to outperform the single-class baseline for the Spanish-English language pair according to our primary metric. However, for Spanish-English and English-German both submissions are able to beat the baseline by large margin. We also observe that the absolute numbers vary greatly between language pairs.

Table 19 gives the results of the Multi-class classification variant of Task 2. Again, the second baseline is to always predict the most common error category in the training data, which varies depending on language pair and produces and increasingly weak baseline as the number of classes rises.

The winning submissions of this variant are as follows: for English-Spanish, Spanish-English, and English-German it is RTM-DCU/RTM-GLM with weighted F_1 scores of 26.84, 8.75, and 15.02 respectively and and for German-English it is RTM-DCU/RTM-GLMd with a weighted F_1 of 3.08. Not only do these systems perform above our baselines for all but the German-English language pair, they also outperform all other submissions for English-Spanish. Remarkably, RTM-DCU/RTM-GLM wins English-Spanish for all of the proposed metrics by a sizeable margin.

4.5 Discussion

In what follows, we discuss the main accomplishments of this year's shared task starting from the goals we had previously identified for it.

Investigating the effectiveness of different quality labels

For the sentence-level tasks, the results of this year's shared task allow us to investigate the effectiveness of predicting translation quality using three very different quality labels: perceived post-editing effort on a scale of [1-3] (Task 1.1); HTER scores (Task 1.2); and the time that a translator takes to post-edit the translation (Task 1.3). One of the ways one can compare the effectiveness across all these different labels is to look at how well the models can produce predictions that correlate with the gold label that we have at our disposal. A measure of correlation that does not depend on the value of the labels is Spearman's ranking correlation. From this perspective, the label that seems the most effective appears to be post-editing time (Task 1.3), with the best system (RTM-DCU/RTM-RR) producing a Spearman's ρ of 0.68 (English-Spanish translations, see Table 15). In comparison, when perceived post-editing effort labels are used (Task 1.1), the best systems achieve a Spearman's ρ of 0.38 and 0.30 for English-Spanish and Spanish-English translations, respectively, and ρ of 0.54 and 0.51 for English-German and German-English, respectively (Table 11); for HTER scores (Task 1.2) the best systems achieve a Spearman's ρ of 0.53 for English-Spanish translations (Table 13).

This comparison across tasks seems to indicate that, among the three labels we have proposed, post-editing time seems to be the most *learnable*, in the sense that automatic predictions can vest match the gold labels (in this case, with respect to the rankings they induce). A possible reason for this is that post-editing time correlates with the length of the source sentence whereas HTER is a normalised measure.

Compared to the results regarding time prediction in the Quality Evaluation shared task from 2013 (Bojar et al., 2013), we note that this time all submissions were able to beat the baseline system (compared to only 1/3 of the submissions in 2013). In addition, better handling of the data acquisition reduced the number of outliers in this year's dataset allowing for numbers that are more reliably *interpretable*. As an example of its interpretability, consider the following: the winning submission for the ranking variant of Task 1.3 is RTM-DCU/RTM-RR, with a a Spearman's ρ of 0.68 and a DeltaAvg score of 17.02 (when predict-

System ID	weighted F_1 All	weighted F_1 Errors ↑	weighted MCC All	weighted MCC Errors	ACC
English-Spanish					
Baseline (always OK)	50.43	0.00	0.00	0.00	64.38
Baseline (always fluency)	14.39	40.41	0.00	0.00	30.67
• FBK-UPV-UEDIN/RNN+tandem+crf	58.36	38.54	16.63	13.89	57.98
FBK-UPV-UEDIN/RNN	60.32	37.25	18.22	15.51	61.75
LIG/BL ALL	58.97	31.79	14.95	11.48	61.13
LIG/FS	58.95	31.78	14.92	11.46	61.10
RTM-DCU/RTM-GLMd	58.23	26.62	12.60	12.76	62.94
RTM-DCU/RTM-GLM	56.47	29.91	8.11	7.96	58.56
Spanish-English					
Baseline (always OK)	74.41	0.00	0.00	0.00	82.37
Baseline (always fluency)	2.67	15.13	0.00	0.00	12.24
• RTM-DCU/RTM-GLMd	78.89	23.94	25.41	25.45	83.17
RTM-DCU/RTM-GLM	78.78	21.96	26.31	26.99	83.69
English-German					
Baseline (always OK)	59.39	0.00	0.00	0.00	71.33
Baseline (always fluency)	3.83	13.35	0.00	0.00	14.82
• RTM-DCU/RTM-GLMd	64.58	21.94	17.69	15.92	69.26
RTM-DCU/RTM-GLM	64.43	21.10	16.99	14.93	69.34
German-English					
Baseline (always OK)	67.82	0.00	0.00	0.00	77.60
Baseline (always fluency)	3.34	14.92	0.00	0.00	13.79
• RTM-DCU/RTM-GLMd	69.17	8.57	10.61	5.76	75.91
RTM-DCU/RTM-GLM	69.09	8.26	9.95	5.76	75.97

Table 18: Official results for the Level 1 classification part of the WMT14 Quality Evaluation Task 2. The winning submissions are indicated by a •. All values are given as percentages.

System ID	weighted F_1 All	weighted F_1 Errors ↑	weighted MCC All	weighted MCC Errors	ACC
English-Spanish					
Baseline (always OK)	50.43	0.00	0.00	0.00	64.38
Baseline (always unintelligible)	7.93	22.26	0.00	0.00	21.99
• RTM-DCU/RTM-GLM	60.52	26.84	23.77	21.45	66.83
FBK-UPV-UEDIN/RNN+tandem+crf	52.96	23.07	15.17	10.74	52.13
LIG/BL ALL	56.66	20.50	18.56	13.39	60.39
LIG/FS	56.66	20.50	18.56	13.39	60.39
FBK-UPV-UEDIN/RNN	52.84	17.09	7.66	4.24	57.18
RTM-DCU/RTM-GLMd	51.87	3.22	10.16	4.04	64.42
Spanish-English					
Baseline (always OK)	74.41	0.00	0.00	0.00	82.37
Baseline (always word order)	0.34	1.96	0.00	0.00	4.24
• RTM-DCU/RTM-GLM	76.34	8.75	19.82	13.43	83.27
RTM-DCU/RTM-GLMd	76.21	8.19	19.35	15.32	83.17
English-German					
Baseline (always OK)	59.39	0.00	0.00	0.00	71.33
Baseline (always mistranslation)	2.48	8.66	0.00	0.00	11.78
• RTM-DCU/RTM-GLM	63.57	15.02	17.57	15.08	70.82
RTM-DCU/RTM-GLMd	63.33	12.48	18.70	13.20	71.45
German-English					
Baseline (always OK)	67.82	0.00	0.00	0.00	77.60
Baseline (always word order)	1.56	6.96	0.00	0.00	9.23
• RTM-DCU/RTM-GLMd	67.62	3.08	7.19	1.48	74.73
RTM-DCU/RTM-GLM	67.86	2.36	7.55	1.79	75.75

Table 19: Official results for the Multi-class classification part of the WMT14 Quality Evaluation Task 2. The winning submissions are indicated by a •. All values are given as percentages.

ing seconds). This number has a direct real-world interpretation: using the order proposed by this system, a human translator would spend, on average, about 17 seconds less on a sentence taken from the top of the ranking compared to a sentence picked randomly from the set.[14] To put this number into perspective, for this dataset the average time to complete a sentence post-editing is 39 seconds. As such, one has an immediate interpretation for the usefulness of using such a ranking: translating around 100 sentences taken from the top of the rankings would take around 36min (at about 22 seconds/sentence), while translating the same number of sentences extracted randomly from the same dataset would take around 1h5min (at about 39 seconds/sentence). It is in this sense that we consider post-editing time an interpretable label.

Another desirable property of label predictions is *usefulness*; this property, however, it highly task-dependent and therefore cannot be judged in the absence of a specific task. For instance, an interpretable label like post-editing time may not be that useful in a task the requires one to place the machine translations into "ready to publish" and "not ready to publish" bins. For such an application, labels such as the ones used by Task 1.1 are clearly more useful, and also very much interpretable within the scope of the task. Our attempt at presenting the Quality Prediction task with a variety of prediction labels illustrates a good range of properties for the proposed labels and enables one to draw certain conclusions depending on the needs of the specific task at hand.

For the word-level tasks, different quality labels equate with using different levels of granularity for the predictions, which we discuss next.

Exploring word-level quality prediction at different levels of granularity

Previous work on word-level predictions, e.g. (Bojar et al., 2013) has focused on prediction of automatically derived labels, generally due to practical considerations as the manual annotation is labour intensive. While easily applicable, automatic annotations, using for example TER alignment between the machine translation and reference (or post-edition), face the same problems as automatic MT evaluation metrics as they fail to account for different word choices and lack the ability to reliably distinguish meaning preserving reorderings from those that change the semantics of the output. Furthermore, previous automatic annotation for word-level quality estimation has focused on binary labels: correct / incorrect, or at most, the main edit operations that can be captured by alignment metrics like TER: correct, insertion, deletion, substitution.

In this year's task we were able to provide manual fine-grained annotations at the word-level produced by humans irrespective of references or post-editions. Error categories range from frequent ones, such as *unintelligible*, *mistranslation*, and *terminology*, to rare ones such as *additions* or *omissions*. For example, only 10 out of more than 3,400 errors in the English-Spanish test set fall into the latter categories, while over 2,000 words are marked as *unintelligible*. By hierarchically grouping errors into coarser categories we aimed to find a compromise between data sparsity and the expressiveness of the labels. What marks a good compromise depends on the use case, which we do not specify here, and the quality of the finer grained predictions: if a system is able to predict even rare errors these may be grouped later if necessary.

Overall, word-level error prediction seems to remain a challenging task as evidenced by the fact that many submissions were unable to beat a trivial baseline. We hypothesise that this is at least partially due to a mismatch in loss-functions used in training and testing. We know from the system descriptions that some systems were tuned to optimise squared error or accuracy, while evaluation was performed using weighted F_1 scores. On the other hand, even a comparison of just accuracy shows that systems struggle to obtain a lower error rates than the "all-OK" baseline.

Such performance problems are consistent over the three levels of granularity, contrary to the intuition that binary classification would be easier. A notable exception is the RTM-DCU/RTM-GLM system, which is able to beat both the baseline and all other systems on the Multi-Class variant of the English-Spanish task – cf. Table 19 – with regard to all metrics. For this and most other submissions we observe that labels are not consistent for different granularities, i.e. at token marked with a specific error in the multi-class variant may still

[14] Note that the 17.02 seconds figure is a difference in real-time, not predicted time; what is considered in this variant of Task 1.3 is only the predicted ranking of data points, not the absolute values of the predictions.

carry an "OK" label in binary annotation. Thus, additional coarse grained annotations may be derived by automatic means. For example, mapping the multi-class predictions of the above system to coarser categories improves the $F_{1,ERR}$ score in Table 17 from 35.08 to 37.02 but does not change the rank with respect to the other entries.

The fact that coarse grained predictions seem not to be derived from the fine-grained ones leads us to believe that most participants treated the different granularities as independent classification tasks. The FBK-UPV-UEDIN team transfers information in the opposite direction by using their binary predictions as features for Level-1 and multi-class.

Given the current quality of word-level prediction it remains unclear if these systems can already be employed in a practical setting, e.g. to focus the attention of post-editors.

Studying the effects of training and test datasets with mixed domains, language pairs and MT systems

This year's shared task made available datasets for more than one language pair with the same or different types of annotation, 2-3 multiple MT systems (plus a human translation) per language pair, and out-of-domain test data (Tasks 1.1 and 2). Instances for each language pair were kept in separate datasets and thus the "language pair" variable can be analysed independently. However, for a given language pair, datasets mix translation systems (and humans) in Task 1.1, and also text domains in Task 2.

Directly comparing the performance across language pairs is not possible, given that their datasets have different numbers of instances (produced by 3 or 4 systems) and/or different true score distributions (see Figure 3). For a relative comparison (although not all systems submitted results for all language pairs, which is especially true in Task 2), we observe in Task 1.1 that for all language pairs generally at least half of the systems did better than the baseline. To our surprise, only one submission combined data for multiple languages together for Task 1.1: SHEF-lite, treating each language pair data as a different task in a multi-task learning setting. However, only for the 'sparse' variant of the submission significant gains were reported over modelling each task independently (with the tasks still sharing the same data kernel and the same hyperparameters).

The interpretation of the results for Task 2 is very dependent on the evaluation metric used, but generally speaking a large variation in performance was found between different languages, with English-Spanish performing the best, possibly given the much larger number of training instances. Data for Task 2 also presented varied true score distributions (as shown by the performance of the baseline (e.g. always "OK") in Tables 17-19.

One of the main goals with Task 1.1 (and Task 2 to some extent) was to test the robustness of models in a blind setting where multiple MT systems (and human translations) are put together and their identifiers are now known. All submissions for these tasks were therefore translation system agnostic, with no submission attempting to perform meta-identification of the origins of the translations. For Task 1.1, data from multiple MT systems was explicitly used by USHEFF though the idea of consensus translations. Translations from all but the system of interest for the same source segment were used as pseudo-references. The submission significantly outperformed the baseline for all language pairs and did particularly well for Spanish-English and English-Spanish.

An in depth analysis of Task 1.1's datasets on the difference in prediction performance between models built and applied for individual translation systems and models built and tested for all translations pooled together is presented in (Shah and Specia, 2014). Not surprisingly, the former models perform significantly better, with MAE scores ranging between 0.35 and 0.5 for different language pairs and MT systems, and significantly lower scores for models trained and tested on human translations only (MAE scores between 0.2 and 0.35 for different language pairs), against MAE scores ranging between 0.5 and 0.65 for models with pooled data.

For Tasks 1.2 and 1.3, two submissions included English-Spanish data which had been produced by yet different MT systems (SHEF-lite and DFKI). While using these additional instances seemed attractive given the small number of instances available for these tasks, it is not clear what their contribution was. For example, with a reduced set of instances (only 400) from the combined sets, SHEF-lite/sparse performed significantly better than its variant SHEF-lite.

Finally, with respect to out-of-domain (different

text domain and MT system) test data, for Task 1.1, none of the papers submitted included experiments. (Shah and Specia, 2014) applied the models trained on pooled datasets (as explained above) for each language pair to the out-of-domain test sets. The results were surprisingly positive, with average MAE score of 0.5, compared to the 0.5-0.65 range for in-domain data (see above). Further analysis is necessary to understand the reasons for that.

In Task 2, the official training and test sets already include out-of-domain data because of the very small amount of in-domain data available, and thus is is hard to isolate the effect of this data on the results.

Examining the effectiveness of quality prediction methods on human translations

Datasets for Tasks 1.1 and 2 contain human translations, in addition to the automatic translations from various MT systems. Predicting human translation quality is an area that has been largely unexplored. Previous work has looked into distinguishing human from machine translations (e.g. (Gamon et al., 2005)), but this problem setting is somehow artificial, and moreover arguably harder to solve nowadays given the higher general quality of current MT systems (Shah and Specia, 2014). Although human translations are obviously of higher quality in general, many segments are translated by MT systems with the same or similar levels of quality as human translation. This is particularly true for Task 2, since data had been previously categorised and only "near misses" were selected for the word-level annotation, i.e., human and machine translations that were both nearly perfect in this case.

While no distinction was made between human and machine translations in our tasks, we believe the mix of these two types of translations has had a negative impact in prediction performance. Intuitively, one can expect errors in human translation to be more subtle, and hence more difficult to capture via standard quality estimation features. For example, an incorrect lexical choice (due to, e.g., ambiguity) which still fits the context and does not make the translation ungrammatical is unlikely to be captured. We hoped that participants would design features for this particular type of translation, but although linguistically motivated features have been exploited, they did not seem appropriate for human translations.

It is interesting to mention the indirect use of human translations by USHEFF for Tasks 1.1-1.3: given a translation for a source segment, all other translations for the same segment were used as pseudo-references. Apart from when this translation was actually the human translation, the human translation was effectively used as a reference. While this reference was mixed with 2-3 other pseudo-references (other machine translations) for the feature computations, these features led to significant gains in performance over the baseline features Scarton and Specia (2014).

We believe that more investigation is needed for human translation quality prediction. Tasks dedicated to this type of data at both sentence- and word-level in the next editions of this shared task would be a possible starting point. The acquisition of such data is however much more costly, as it is arguably hard to find examples of low quality human translation, unless specific settings, such as translation learner corpora, are considered.

5 Medical Translation Task

The Medical Translation Task addresses the problem of domain-specific and genre-specific machine translation. The task is split into two subtasks: **summary translation**, focused on translation of sentences from summaries of medical articles, and **query translation**, focused on translation of queries entered by users into medical information search engines.

In general, texts of specific domains and genres are characterized by the occurrence of special vocabulary and syntactic constructions which are rare or even absent in traditional (general-domain) training data and therefore difficult for MT. Specific training data (containing such vocabulary and syntactic constructions) is usually scarce or not available at all. Medicine, however, is an example of a domain for which in-domain training data (both parallel and monolingual) is publicly available in amounts which allow to train a complete SMT system or to adapt an existing one.

5.1 Task Description

In the Medical Translation Task, we provided links to various medical-domain training resources and asked participants to use the data to train or adapt their systems to translate unseen test sets for both subtasks between English and Czech (CS), German (DE), and French (FR), in both directions.

The summary translation test data is domain-specific, but otherwise can be considered as ordinary sentences. On the other hand, the query translation test data is also specific for its genre (general style) – it contains short sequences of (more or less) of independent terms rather than complete and grammatical sentences, the usual target of current MT systems.

Similarly to the standard Translation Task, the participants of the Medical Translation Task were allowed to use only the provided resources in the *constrained task* (in addition to data allowed in the constrained standard Translation Task), but could exploit any additional resources in the *unconstrained task*. The submissions were expected with true letter casing and detokenized. The translation quality was measured using automatic evaluation metrics, manual evaluation was not performed.

5.2 Test and Development Data

The test and development data sets for this task were provided by the EU FP7 project Khresmoi.[15] This projects develops a multi-lingual multi-modal search and access system for biomedical information and documents and its MT component allows users to use non-English queries to search in English documents and see summaries of retrieved documents in their preferred language (Czech, German, or French). The statistics of the data sets are presented in Tables 20 and 21.

For the summary translation subtask, 1,000 and 500 sentences were provided for test development purposes, respectively. The sentences were randomly sampled from *automatically* generated summaries (extracts) of English documents (web pages) containing medical information relevant to 50 topics provided for the CLEF 2013 eHealth Task 3.[16] Out-of-domain and ungrammatical sentences were manually removed. The sentences were then translated by medical experts into Czech, German and French, and the translations were reviewed. Each sentence was provided with the corresponding document ID and topic ID. The set also included a description for each of the 50 topics. The data package (Khresmoi Summary Translation Test Data 1.1) is now available from the LINDAT/CLARIN repository[17] and more details can be found in Zdeňka Urešová and Pecina (2014).

For the query translation subtask, the main test set contains 1,000 queries for test and 508 queries for development purposes. The original English queries were extracted at random from real user query logs provided by the Health on the Net foundation[18] (queries by general public) and the Trip database[19] (queries by medical experts). Each query was translated into Czech, German, and French by medical experts and the translations were reviewed. The data package (Khresmoi Query Translation Test Data 1.0) is available from the LINDAT/CLARIN repository.[20]

An additional test set for the query translation subtask was adopted from the CLEF 2013 eHealth Task 3 (Pecina et al., 2014). It contains 50 queries constructed from titles of the test topics (originally in English) translated into Czech, German, and French by medical experts. The participants were asked to translate the queries back to English and the resulting translations were used in an information retrieval (IR) experiment for extrinsic evaluation.

5.3 Training Data

This section reviews the in-domain resources which were allowed for the constrained Medical Translation Task in addition to resources for the constrained standard Translation Task (see Section 2). Most of the corpora are available for direct download, others can be obtained upon registration. The corpora usually employ their own, more or less complex data format. To lower the entry barrier, we provided a set of easy-to-use scripts to convert the data to a plain text format suitable for MT training.

5.3.1 Parallel Training Data

The medical-domain parallel data includes the following corpora (see Table 22 for statistics): The *EMEA* corpus (Tiedemann, 2009) contains documents from the European Medicines Agency, automatically processed and aligned on sentence level. It is available for many language pairs, including those relevant to this task. *UMLS* is a multilingual metathesaurus of health and biomed-

[15] http://khresmoi.eu/

[16] https://sites.google.com/site/shareclefehealth/

[17] http://hdl.handle.net/11858/00-097C-0000-0023-866E-1

[18] http://www.hon.ch/

[19] http://www.tripdatabase.com/

[20] http://hdl.handle.net/11858/00-097C-0000-0022-D9BF-5

	sents	tokens			
	total	Czech	German	French	English
dev	500	9,209	9,924	12,369	10,350
test	1,000	19,191	20,831	26,183	21,423

Table 20: Statistics of summary test data.

	queries			tokens			
	total	general	expert	Czech	German	French	English
dev	508	249	259	1,128	1,041	1,335	1,084
test	1,000	500	500	2,121	1,951	2,490	2,067

Table 21: Statistics of query test data.

L1–L2	Czech–English			DE–EN			FR–EN		
data set	sents	L1 tokens	L2 tokens	sents	L1 tokens	L2 tokens	sents	L1 tokens	L2 tokens
EMEA	1,053	13,872	14,378	1,108	13,946	14,953	1,092	17,605	14,786
UMLS	1,441	4,248	5,579	2,001	6,613	8,153	2,171	8,505	8,524
Wiki	3	5	6	10	19	22	8	19	17
MuchMore				29	688	740			
PatTr				1,848	102,418	106,727	2,201	127,098	108,665
COPPA							664	49,016	39,933

Table 22: Statistics of the in-domain parallel training data allowed for the constrained task (in thousands).

data set	English	Czech	German	French
PatTR	121,592		53,242	54,608
UMLS	7,991	63	24	37
Wiki	26,945	1,784	10,232	8,376
AACT	13,341			
DrugBank	953			
FMA	884			
GENIA	557			
GREC	62			
PIL	662			

Table 23: Sizes of monolingual training data allowed for the constrained tasks (in thousands of tokens).

ical vocabularies and standards (U.S. National Library of Medicine, 2009). The UMLS dataset was constructed by selecting the concepts which have translations in the respective languages. The *Wiki* dataset contains bilingual pairs of titles of Wikipedia articles belonging to the categories identified to be medical-domain within the Khresmoi project. It is available for all three language pairs. The *MuchMore Springer Corpus* is a German–English parallel corpus of medical journals abstracts published by Springer (Buitelaar et al., 2003). *PatTR* is a parallel corpus extracted from the MAREC patent collection (Wäschle and Riezler, 2012). It is available for German–English and French–English. For the medical domain, we only consider text from patents indicated to be from the medicine-related categories (A61, C12N, C12P). *COPPA* (Corpus of Parallel Patent Applications (Pouliquen and Mazenc, 2011) is a French–English parallel corpus extracted from the MAREC patent collection (Wäschle and Riezler, 2012). The medical-domain subset is identified by the same categories as in PatTR.

5.3.2 Monolingual Training Data

The medical-domain monolingual data consists of the following corpora (statistics are presented in Table 23): The monolingual *UMLS* dataset contains concept descriptions in CS, DE, and FR extracted from the UMLS Metathesaurus (see Section 5.3.1). The monolingual *Wiki* dataset consists of articles belonging to the categories identified to be medical-domain within the Khresmoi project. The *PatTR* dataset contains non-parallel data extracted from the medical patents included in the PatTR corpus (see Section 5.3.1). *AACT* is a collection of restructured and reformatted English texts publicly available and downloadable from ClinicalTrials.gov, containing clinical studies conducted around the world. *DrugBank* is a bioinformatics and cheminformatics resource containing drug descriptions (Knox et al., 2011). *GENIA* is a corpus of biomedical literature compiled and annotated within the GENIA project (Kim et al., 2003). *FMA* stands for the Foundational Model of Anatomy Ontology, a knowledge source for biomedical informatics concerned with symbolic representation of the phenotypic structure of the human body (Rosse and Mejino Jr., 2008). *GREC* (Gene Regulation Event Corpus) is a semantically annotated English corpus of abstracts of biomedical papers (Thompson et al., 2009). The *PIL* corpus is a collection of documents giving instructions to patients about their medication (Bouayad-Agha et al., 2000).

5.4 Participants

A total of eight teams participated in the Medical Translation Task by submitting their systems to at least one subtask for one or more translation directions. A list of the participants is given in Table 24; we provide short descriptions of their systems in the following.

CUNI was involved in the organization of the task, and their primary goal was to set up a baseline for both the subtasks and for all translation directions.

ID	Participating team
CUNI	Charles University in Prague (Dušek et al., 2014)
DCU-Q	Dublin City University (Okita et al., 2014)
DCU-S	Dublin City University (Zhang et al., 2014)
LIMSI	Laboratoire dInformatique pour la Mecanique et les Sciences de lIngénieur (Pécheux et al., 2014)
POSTECH	Pohang University of Science and Technology (Li et al., 2014a)
UEDIN	University of Edinburgh (Durrani et al., 2014a)
UM-DA	University of Macau (Wang et al., 2014)
UM-WDA	University of Macau (Lu et al., 2014)

Table 24: Participants in the WMT14 Medical Translation Task.

Their systems are based on the Moses phrase-based toolkit and linear interpolation of in-domain and out-of-domain language models and phrase tables. The constrained/unconstrained systems differ in the training data only. The constrained ones are built using all allowed training data; the unconstrained ones take advantage of additional web-crawled monolingual data used for training of the language models, and additional parallel non-medical data from the PatTr and COPPA patent collections.

DCU-Q submitted a system designed specifically for terminology translation in the query translation task for EN–FR and FR–EN. This system supports six terminology extraction methods and is able to detect rare word pairs including zero-appearance word pairs. It uses monotonic decoding with lattice inputs, avoiding unnecessary hypothesis expansions by the reordering model.

DCU-S submitted a system to the FR–EN summary translation subtask only. The system is similar to DCU's system for patent translation (phrased-based using Moses) but adapted to translate medical summaries and reports.

LIMSI took part in the summary translation subtask for English to French.Their primary submission uses a combination of two translation systems: NCODE, based on bilingual n-gram translation models; and an on-the-fly estimation of the parameters of Moses along with a vector space model to perform domain adaptation. A continuous-space language model is also used in a post-processing step for each system.

POSTECH submitted a phrase-based SMT system and query translation system for the DE–EN language pair in both subtasks. They analysed three types of query formation, generated query translation candidates using term-to-term dictionaries and a phrase-based system, and then scored them using a co-occurrence word frequency measure to select the best candidate.

UEDIN applied the Moses phrase-based system to all language pairs and both subtasks. They used the hierarchical reordering model and the OSM feature, same as in UEDIN's news translation system, and applied compound splitting to German input. They used separate language models built on in-domain and out-of-domain data with linear interpolation. For all language pairs except CS-EN and DE-EN, they selected data for the translation model using modified Moore-Lewis filtering. For DE-EN and CS-EN, they concatenated all the supplied parallel training data.

UM-DA submitted systems for all language pairs in the summary translation subtask based on a combination of different adaptation steps, namely domain-specific pre-processing, language model adaptation, translation model adaptation, numeric adaptation, and hyphenated word adaptation. Data for the domain-adapted language and translation models were selected using various data selection techniques.

UM-WDA submitted systems for all language pairs in the summary translation subtask. Their systems are domain-adapted using web-crawled in-domain resources: bilingual dictionaries and monolingual data. The translation model and language model trained on the crawled data were interpolated with the best-performing language and translation model employed in the UM-DA systems.

5.5 Results

MT quality in the Medical Translation Task is evaluated using automatic evaluation metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), PER (Tillmann et al., 1997), and CDER (Leusch et al., 2006). BLEU scores are reported as percentage and all error rates are reported as one minus the original value, also as percentage, so that all metrics are in the 0-100 range, and higher scores indicate better translations.

The main reason for not conducting human evaluation, as it happens in the standard Trans-

	original	normalized truecased				normalized lowercased			
ID	**BLEU**	**BLEU**	**1-TER**	**1-PER**	**1-CDER**	**BLEU**	**1-TER**	**1-PER**	**1-CDER**
					Czech→English				
CUNI	29.64	29.79±1.07	47.45±1.15	61.64±1.06	52.18±0.98	31.68±1.14	49.84±1.10	64.38±1.06	54.10±0.96
CUNI	22.44	22.57±0.95	41.43±1.16	55.46±1.09	46.42±0.96	32.34±1.12	50.24±1.20	65.07±1.10	54.42±0.96
UEDIN	36.65	36.87±1.23	54.35±1.19	67.16±1.00	**57.61±1.01**	38.02±1.24	**56.14±1.17**	69.24±1.01	**58.96±0.96**
UM-DA	**37.62**	**37.79±1.26**	**54.55±1.20**	**68.29±0.88**	57.28±1.03	**38.81±1.28**	56.04±1.20	**70.06±0.82**	58.45±1.05
CUNI	22.92	23.06±0.97	42.49±1.10	56.10±1.12	47.13±0.95	33.18±1.15	51.48±1.15	66.00±1.03	55.30±0.96
CUNI	22.69	22.84±0.98	42.21±1.14	56.01±1.11	46.79±0.94	32.84±1.13	51.10±1.11	65.79±1.07	54.81±0.96
UM-WDA	37.35	37.53±1.26	54.39±1.19	68.21±0.83	57.16±1.07	38.61±1.27	55.92±1.17	70.02±0.81	58.36±1.07
ONLINE		39.57±1.21	58.24±1.14	70.16±0.78	60.04±1.02	40.62±1.23	59.72±1.11	71.94±0.74	61.26±1.01
					German→English				
CUNI	28.20	28.34±1.12	46.66±1.13	61.53±1.03	50.57±0.93	30.69±1.19	48.91±1.16	64.12±1.04	52.52±0.95
CUNI	28.85	28.99±1.15	47.12±1.15	61.98±1.07	50.72±0.98	31.37±1.21	49.29±1.13	64.53±1.05	52.64±0.98
POSTECH	25.92	25.99±1.06	43.66±1.14	59.62±0.92	47.13±0.90	26.97±1.06	45.13±1.12	61.53±0.89	48.37±0.88
UEDIN	**37.31**	**37.53±1.19**	**55.72±1.14**	**68.82±0.99**	**58.35±0.95**	**38.60±1.25**	**57.18±1.12**	**70.46±0.98**	**59.53±0.94**
UM-DA	35.71	35.81±1.23	53.08±1.16	66.82±0.98	55.91±0.96	36.55±1.27	54.01±1.13	68.05±0.97	56.78±0.95
CUNI	30.58	30.71±1.10	48.68±1.09	63.19±1.08	52.72±0.94	33.14±1.19	50.98±1.06	65.88±1.04	54.74±0.94
CUNI	30.22	30.32±1.12	47.71±1.18	62.20±1.10	52.17±0.91	32.75±1.20	50.00±1.14	64.87±1.06	54.19±0.92
UM-WDA	32.70	32.88±1.19	49.60±1.18	63.74±1.01	53.50±0.96	33.95±1.23	51.05±1.19	65.54±0.98	54.73±0.96
ONLINE		41.18±1.24	59.33±1.09	70.95±0.92	61.92±1.01	42.29±1.23	60.76±1.08	72.51±0.88	63.06±0.96
					French→English				
CUNI	34.42	34.55±1.20	52.24±1.17	64.52±1.03	56.48±0.91	36.52±1.23	54.35±1.12	67.07±1.00	58.34±0.91
CUNI	33.67	33.59±1.16	50.39±1.23	61.75±1.16	56.74±0.97	35.55±1.21	52.55±1.26	64.45±1.13	58.63±0.91
DCU-B	44.85	45.01±1.24	62.57±1.12	74.11±0.78	64.33±0.99	46.12±1.26	64.04±1.06	75.84±0.74	65.55±0.94
UEDIN	46.44	46.68±1.26	**64.12±1.16**	74.47±0.87	**66.40±0.96**	48.01±1.29	**65.70±1.15**	76.30±0.86	**67.76±0.91**
UM-DA	**47.08**	**47.22±1.33**	64.08±1.16	**75.41±0.88**	66.15±0.96	**48.23±1.31**	65.36±1.10	**76.95±0.89**	67.18±0.93
CUNI	34.74	34.89±1.12	52.39±1.16	63.76±1.09	57.29±0.94	36.84±1.17	54.56±1.13	66.43±1.07	59.14±0.90
CUNI	35.04	34.99±1.18	52.11±1.24	63.24±1.09	57.51±0.97	37.04±1.18	54.38±1.17	66.02±1.05	59.55±0.93
UM-WDA	43.84	44.06±1.32	61.14±1.18	73.13±0.87	63.09±1.00	45.17±1.36	62.63±1.15	74.94±0.84	64.37±0.99
ONLINE		46.99±1.35	64.31±1.12	76.07±0.78	66.09±1.00	47.99±1.33	65.65±1.07	77.65±0.75	67.20±0.96
					English→Czech				
CUNI	17.36	17.65±0.96	37.17±1.02	49.13±0.98	40.31±0.95	18.75±0.96	38.32±1.02	50.82±0.91	41.39±0.94
CUNI	16.64	16.89±0.93	36.57±1.05	48.79±0.98	39.46±0.90	17.94±0.96	37.74±1.03	50.50±0.97	40.59±0.91
UEDIN	**23.45**	**23.74±1.00**	**44.20±1.10**	**55.38±0.88**	**46.23±0.99**	**24.20±1.00**	**44.92±1.08**	**56.38±0.90**	**46.78±1.00**
UM-DA	22.61	22.72±0.98	42.73±1.16	54.12±0.93	44.73±1.01	23.12±1.01	43.41±1.14	55.11±0.93	45.32±1.02
CUNI	20.56	20.84±1.01	39.98±1.09	51.98±0.99	42.86±1.00	22.03±1.05	41.19±1.08	53.66±0.97	43.93±1.01
CUNI	19.50	19.72±0.97	38.09±1.10	50.12±1.06	41.50±0.96	20.91±1.02	39.26±1.12	51.79±1.04	42.59±0.96
UM-WDA	22.14	22.33±0.96	42.30±1.11	53.89±0.92	44.48±1.01	22.72±0.97	43.02±1.09	54.89±0.95	45.08±0.99
ONLINE		33.45±1.28	51.64±1.28	61.82±1.10	53.97±1.18	34.02±1.31	52.35±1.22	62.84±1.08	54.52±1.18
					English→German				
CUNI	12.52	12.64±0.77	29.84±0.99	45.38±1.14	34.69±0.81	16.63±0.91	33.63±1.07	50.03±1.24	38.43±0.87
CUNI	12.42	12.53±0.77	29.02±1.05	44.27±1.16	34.62±0.78	16.41±0.91	32.87±1.08	48.99±1.21	38.37±0.86
POSTECH	15.46	15.59±0.91	34.41±1.01	49.00±0.83	37.11±0.90	15.98±0.92	34.98±1.00	49.94±0.81	37.60±0.87
UEDIN	20.88	21.01±1.03	40.03±1.08	**55.54±0.91**	42.95±0.90	21.40±1.03	40.55±1.08	56.33±0.92	43.41±0.90
UM-DA	**20.89**	**21.09±1.07**	**40.76±1.03**	55.45±0.89	**43.02±0.93**	**21.52±1.08**	**41.31±1.01**	**56.38±0.90**	**43.58±0.91**
CUNI	14.29	14.42±0.81	31.82±1.03	47.01±1.13	36.81±0.79	18.87±0.90	35.76±1.11	51.76±1.17	40.65±0.87
CUNI	13.44	13.58±0.75	30.37±1.03	45.80±1.14	35.80±0.76	17.84±0.89	34.41±1.13	50.75±1.18	39.85±0.78
UM-WDA	18.77	18.91±1.00	37.92±1.02	53.59±0.85	40.90±0.86	19.30±1.02	38.42±1.01	54.40±0.85	41.34±0.86
ONLINE		23.92±1.06	44.33±0.97	57.47±0.80	46.35±0.91	24.29±1.07	44.83±0.98	58.20±0.80	46.71±0.92
					English→French				
CUNI	30.30	30.67±1.11	46.59±1.09	59.83±1.04	50.51±0.93	32.06±1.12	48.01±1.09	61.66±1.00	51.83±0.94
CUNI	29.35	29.71±1.10	45.84±1.07	58.81±1.04	50.00±0.96	31.02±1.10	47.24±1.09	60.57±1.02	51.31±0.94
LIMSI	40.14	43.54±1.22	59.70±1.04	69.45±0.86	61.35±0.96	44.04±1.22	60.32±1.03	70.20±0.85	61.90±0.94
LIMSI	38.83	42.21±1.13	58.88±1.01	68.70±0.81	60.59±0.93	42.69±1.12	59.53±0.98	69.50±0.80	61.17±0.91
UEDIN	40.74	**44.24±1.16**	**60.66±1.07**	**70.35±0.82**	**62.28±0.95**	**44.85±1.17**	**61.43±1.05**	**71.27±0.81**	**62.94±0.91**
UM-DA	**41.24**	41.68±1.12	58.72±1.06	69.37±0.78	60.12±0.95	42.16±1.11	59.39±1.05	70.21±0.77	60.71±0.92
CUNI	32.23	32.61±1.09	48.48±1.08	61.13±1.01	52.24±0.93	34.08±1.10	49.93±1.11	62.92±0.99	53.65±0.92
CUNI	32.45	32.84±1.06	48.68±1.06	61.32±0.98	52.35±0.94	34.22±1.07	50.09±1.04	63.04±0.96	53.67±0.91
UM-WDA	40.78	41.16±1.13	58.20±0.99	68.93±0.84	59.64±0.94	41.79±1.12	59.10±0.96	70.01±0.84	60.39±0.91
ONLINE		58.63±1.26	70.70±1.12	78.22±0.81	71.89±0.96	59.27±1.26	71.50±1.10	79.16±0.81	72.63±0.94

Table 25: Official results of translation quality evaluation in the medical **summary translation** subtask.

	original	normalized truecased				normalized lowercased			
ID	**BLEU**	**BLEU**	**1-TER**	**1-PER**	**1-CDER**	**BLEU**	**1-TER**	**1-PER**	**1-CDER**
					Czech→English				
CUNI	10.71	10.57±3.42	15.72±2.77	23.37±3.03	18.68±2.42	**30.13±4.85**	53.38±3.01	62.53±2.84	55.44±2.87
CUNI	9.92	9.78±3.04	16.84±2.84	23.80±3.08	19.85±2.40	28.21±4.56	**54.15±3.04**	**62.56±2.99**	**55.91±2.79**
UEDIN	**24.66**	**24.68±4.52**	**39.88±3.05**	**49.97±3.29**	**41.81±2.80**	28.25±4.94	45.31±3.14	55.66±3.06	46.67±2.77
CUNI	12.00	11.86±3.42	18.49±2.74	24.67±2.85	21.08±2.29	31.91±4.81	**57.61±3.13**	65.02±2.99	**59.24±2.69**
CUNI	10.54	10.39±3.48	18.86±2.48	26.65±2.05	20.53±2.08	**32.39±5.45**	56.79±3.02	**65.52±2.26**	57.96±2.56
ONLINE		28.88±4.96	47.31±3.35	55.19±3.21	49.88±2.89	35.33±5.20	55.80±3.20	64.05±2.97	57.94±2.85
					German→English				
CUNI	10.90	10.74±3.41	18.89±2.39	26.09±2.00	20.29±2.07	**32.15±5.23**	**55.56±2.90**	**63.68±2.34**	**56.45±2.62**
CUNI	10.71	10.55±3.47	18.40±2.35	25.45±2.04	19.84±2.07	32.06±5.19	54.85±2.91	62.87±2.39	55.52±2.61
POSTECH	18.06	17.97±4.38	28.57±3.30	40.38±2.77	31.79±2.80	21.99±4.65	35.76±3.35	47.84±2.82	38.84±2.92
POSTECH	17.99	17.88±4.72	29.79±3.04	41.15±2.48	32.49±2.63	24.41±4.83	41.72±3.19	53.33±2.55	44.06±2.88
UEDIN	**23.33**	**23.39±4.37**	**38.55±3.65**	**48.21±3.43**	**40.75±3.05**	27.17±4.63	43.87±3.52	53.76±3.48	45.72±3.03
CUNI	10.54	10.39±3.48	18.86±2.48	26.65±2.05	20.53±2.08	**32.39±5.45**	**56.79±3.02**	**65.52±2.26**	**57.96±2.56**
CUNI	8.75	8.49±3.60	19.10±2.27	24.98±1.95	19.95±2.02	30.00±5.59	56.07±2.92	62.92±2.32	56.27±2.56
ONLINE		19.97±4.46	37.03±3.26	43.91±3.22	40.95±2.93	33.86±4.87	53.28±3.28	60.86±3.22	56.33±2.98
					French→English				
CUNI	13.90	13.79±3.61	18.49±2.55	28.35±2.81	20.36±2.20	34.97±5.34	59.54±2.94	72.30±2.63	58.86±2.76
CUNI	12.10	11.95±3.41	17.23±2.57	27.12±2.88	19.15±2.28	33.74±5.01	58.95±2.96	71.25±2.76	58.20±2.81
DCU-Q	**30.85**	**31.24±5.08**	**58.88±2.97**	**67.94±2.62**	**59.19±2.62**	**36.88±5.07**	**66.38±2.85**	**75.86±2.37**	**66.29±2.55**
DCU-Q	26.51	26.16±4.40	48.02±3.72	57.34±3.24	53.56±2.79	28.61±4.52	53.65±3.73	63.51±3.21	59.07±2.79
UEDIN	27.20	27.60±3.98	38.54±3.22	48.81±3.26	39.77±2.95	32.23±4.27	43.66±3.20	54.31±3.17	44.53±2.79
CUNI	14.03	14.00±3.30	20.11±2.38	29.00±2.71	21.62±2.22	**38.98±5.08**	62.90±2.87	74.49±2.45	62.12±2.64
CUNI	13.38	13.16±3.52	17.79±2.56	28.84±2.81	19.17±2.23	35.00±5.20	59.52±2.98	73.08±2.57	58.41±2.68
ONLINE		32.96±5.04	53.68±3.21	64.27±2.80	54.40±2.66	38.09±5.52	61.44±3.08	72.59±2.61	61.60±2.78
					English→Czech				
CUNI	8.37	8.00±3.65	17.74±2.23	26.46±1.96	19.48±2.10	19.49±4.60	41.53±2.94	51.34±2.51	42.54±2.74
CUNI	9.04	8.75±3.64	18.25±2.27	26.97±1.92	19.69±2.11	**21.46±5.05**	**42.36±3.09**	**51.99±2.40**	**43.18±2.68**
UEDIN	**12.57**	**12.40±3.61**	**21.15±2.96**	**33.56±2.80**	**22.30±2.67**	14.06±3.80	24.92±2.90	37.85±2.72	25.58±2.70
UEDIN	6.64	6.21±4.73	-2.35±3.06	5.95±3.48	-0.97±3.12	14.35±3.52	14.51±3.19	24.96±3.50	15.11±3.10
CUNI	9.06	8.64±3.82	19.92±2.24	26.97±1.94	20.82±2.06	**22.42±5.24**	**44.89±2.94**	**52.89±2.40**	**45.36±2.78**
CUNI	8.49	8.01±6.05	18.13±2.28	25.19±1.86	19.19±2.01	21.04±4.80	42.66±2.87	50.34±2.47	43.30±2.74
ONLINE		21.09±4.60	48.56±2.82	54.72±2.51	48.30±2.83	24.37±4.80	51.93±2.74	58.10±2.50	51.62±2.80
					English→German				
CUNI	10.17	10.01±3.92	**26.48±3.24**	**36.71±3.37**	**29.26±2.96**	13.02±4.17	**31.96±3.41**	**42.39±3.21**	**34.61±2.95**
CUNI	9.98	9.69±3.94	26.16±3.19	35.50±3.23	28.86±2.94	12.90±4.28	31.75±3.33	41.24±3.21	34.38±3.05
POSTECH	**13.43**	13.01±5.91	26.38±3.09	35.75±3.16	27.86±2.82	**15.05±5.71**	30.45±3.10	39.89±3.14	31.79±3.00
POSTECH	13.41	**13.15±5.21**	22.18±3.09	30.89±3.31	24.17±3.06	14.96±5.15	26.13±3.19	34.92±3.40	27.98±3.12
UEDIN	10.45	10.14±3.86	23.44±3.43	34.55±3.34	25.46±3.17	11.91±4.42	27.91±3.45	39.08±3.42	29.63±3.31
CUNI	8.91	7.72±6.48	**30.05±3.22**	**40.65±2.71**	**31.91±2.88**	13.66±5.37	**35.51±3.28**	**46.12±2.74**	**37.27±3.01**
CUNI	9.14	8.69±6.44	27.66±3.31	37.95±3.45	31.00±2.82	14.03±5.92	33.53±3.45	44.03±3.53	36.73±3.00
ONLINE		20.07±6.06	41.07±3.23	47.41±2.86	41.61±3.02	21.67±6.23	43.78±3.23	50.18±2.95	44.26±3.06
					English→French				
CUNI	13.12	12.92±2.84	21.95±2.41	33.19±2.09	23.70±2.24	**28.42±3.98**	51.43±2.90	63.74±2.35	52.64±2.58
CUNI	12.80	12.65±2.81	19.16±2.61	31.61±2.21	21.91±2.32	27.52±4.05	47.47±3.08	61.43±2.37	49.82±2.72
DCU-Q	**27.69**	**27.84±4.11**	**48.97±3.06**	**60.90±2.55**	**51.84±2.83**	28.98±4.16	**51.73±3.10**	**63.84±2.47**	**54.43±2.76**
UEDIN	20.16	21.76±3.42	31.66±4.23	44.37±4.13	44.29±2.73	23.25±3.49	35.38±4.19	48.52±4.07	47.94±2.75
CUNI	13.78	13.57±3.00	21.92±2.51	33.47±2.03	24.16±2.32	30.07±4.10	51.12±3.08	63.61±2.45	52.96±2.67
CUNI	15.27	15.24±3.12	23.58±2.54	34.39±2.54	25.79±2.32	**31.40±4.15**	**53.60±2.96**	**65.39±2.57**	**55.47±2.69**
ONLINE		28.93±3.66	49.20±3.08	60.85±2.69	51.68±2.78	30.88±3.66	52.25±3.08	64.06±2.62	54.59±2.68

Table 26: Official results of translation quality evaluation in the medical **query translation** subtask.

source lang.	ID	P@5	P@10	NDCG@5	NDCG@10	MAP	Rprec	bpref	rel
Czech→English	**CUNI**	0.3280	0.3340	0.2873	0.2936	0.2217	0.2362	0.3473	1461
German→English	**CUNI**	0.2800	0.3000	0.2467	0.2630	0.2057	0.2077	0.3310	1426
French→English	**CUNI**	0.3280	0.3380	0.2811	0.2882	0.2206	0.2284	0.3504	1481
	DCU-Q	0.3480	0.3460	0.3060	0.3072	0.2252	0.2358	0.3659	1524
	UEDIN	**0.4440**	**0.4300**	**0.3793**	**0.3826**	**0.2843**	**0.2935**	**0.3936**	**1544**
English (monolingual)		0.4600	0.4700	0.4091	0.4205	0.3035	0.3198	0.3858	1638

Table 27: Official results of retrieval evaluation in the **query translation** subtask.

lation Task, was the lack of domain expertise of prospective raters. While in the standard task, the only requirement for the raters was to be a native speaker of the target language, in the Medical Translation Task, a very good knowledge of the domain would be necessary to provide reliable judgements and the raters with such an expertise (medical doctors and native speakers) were not available.

The complete results of the task are presented in Table 25 (for summary translation) and Tables 26 and 27 (for query translation). Participant IDs given in bold indicate primary submissions, IDs in normal font refer to contrastive submissions. The first section for each translation direction (white background) refers to constrained submissions and the second one (light-gray background) to unconstrained submissions. The column denoted as "original" contains BLEU scores as reported by the Matrix submission system obtained on the original submitted translations. Due to punctuation inconsistency in the original reference translations, we decided to perform punctuation normalization before calculating the official scores. The columns denoted as "normalized truecased" contain scores obtained on the submitted translations after punctuation normalization and the columns denoted as "normalized lowercased" contain scores obtained after punctuation normalization and lowercasing. The normalization script is available in the package with summary translation test data. The confidence intervals were obtained by bootstrap resampling with a confidence level of 95%. Figures in bold denote the best constrained system and, if its score is higher, the best unconstrained system for each translation direction and each metric. For comparison, we also present results of a major on-line translation system (denoted as ONLINE).

The results of the extrinsic evaluation of query translation submissions are given in 27. We used the CLEF 2013 eHealth Task 3 test collection containing about 1 million web pages (in English), 50 test queries (originally in English and translated to Czech, German, and French), and their relevance assessments. Some of the participants of the WMT Medical Task (three teams with five submissions in total) submitted translations of the queries (from Czech, German, and French) into English and these translations were used to query the CLEF 2013 eHealth Task 3 test collection using a state-of-the-art system based on a BM25 model, described in Pecina et al. (2014). Originally, we asked for 10 best translations for each query, but only the best one were used for the evaluation. The results are provided in terms of standard IR evaluation measures: precision at a cut-off of 5 and 10 documents (P@5, P@10), normalized discounted cumulative gain (Järvelin and Kekäläinen, 2002) at 5 and 10 documents (NDCG@5, NDCG@10), mean average precision (MAP) (Voorhees and Harman, 2005), precision reached after R documents retrieved, where R indicates the number of the relevant documents for each query in the entire collection (Rprec), binary preference (bpref) (Buckley and Voorhees, 2004), and number or relevant documents retrieved (rel). The cross-lingual results are also compared with the monolingual one (obtained by using the reference (English) translations of the test topics) to see how the system would perform if the queries were translated perfectly.

5.6 Discussion and Conclusion

Both the subtasks turned out to be quite challenging not only because of the specific domain – in summary sentences, we can observe much higher density of terminology than in ordinary sentences; the queries, which are also rich in terminology, do not form sentences at all.

Most submissions were based on systems participating in the standard Translation Task and trained on the provided data or its subsets CUNI provided baseline systems for all language pairs in both subtasks, which turned to be relatively strong for the query translation task, especially in translation to English, but only in terms of scores obtained on normalized and lowercased translations since their truecasing component did not perform well.

In the summary translation subtask, the best overall results were achieved by the UEDIN team which won for DE–EN, EN–CS, and EN–FR, followed by the UM-DA team, which performed on par with UEDIN in all other translation.

The unconstrained submissions in almost all cases did not outperform the results of the constrained submissions. Some improvements were observed in the query translations subtasks by the CUNI's unconstrained system with language models trained on larger in-domain data.

The ONLINE system outperforms all other sub-

missions with only two exceptions – the UM-DA's and UEDIN's systems for the summary translation in the FR–EN direction, though the score differences are within the 95% confidence interval.

In the query translation subtask, DCU-Q built a system designed specifically for terminology translation between French and English and outperformed all other participants in translation into English; however, the confidence intervals in the query translation task are much wider and most of the differences in scores of the automatic metrics are not statistically significant.

The extrinsic evaluation in the cross-lingual information retrieval was conducted for translations into English only. CUNI provided the baselines for all directions, but other submissions were done for FR–EN only. Here, the winner is UEDIN, who outperformed both CUNI and DCU-Q, and their scores are very close to those obtained using the reference English translations.

Acknowledgments

This work was supported in parts by the MosesCore, Casmacat, Khresmoi, Matecat and QTLaunchPad projects funded by the European Commission (7th Framework Programme), and by gifts from Yandex.

We would also like to thank our colleagues Matouš Macháček and Martin Popel for detailed discussions.

References

Avramidis, E. (2014). Efforts on machine learning over human-mediated translation edit rate. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Beck, D., Shah, K., and Specia, L. (2014). Sheffield 2.0: Sparse multi-task gaussian processes for translation quality estimation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Biçici, E. (2013). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Biçici, E., Liu, Q., and Way, A. (2014). Parallel FDA5 for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Bicici, E., Liu, Q., and Way, A. (2014). Parallel fda5 for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bicici, E. and Way, A. (2014). Referential translation machines for predicting translation quality. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–42, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Tamchyna, A., and Zeman, D. (2014). Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference*, Reykjavik, Iceland. ELRA.

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. (2011). A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland. Association for Computational Linguistics.

Borisov, A. and Galinskaya, I. (2014). Yandex school of data analysis russian-english machine translation system for wmt14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bouayad-Agha, N., Scott, D. R., and Power, R. (2000). Integrating content and style in documents: A case study of patient information leaflets. *Information Design Journal*, 9(2–3):161–176.

Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information.

In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, United Kingdom.

Buitelaar, P., Sacaleanu, B., Špela Vintar, Steffen, D., Volk, M., Dejean, H., Gaussier, E., Widdows, D., Weiser, O., and Frederking, R. (2003). Multilingual concept hierarchies for medical information organization and retrieval. Public deliverable, MuchMore project.

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

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

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. F. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT10)*, Uppsala, Sweden.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*, Athens, Greece.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Camargo de Souza, J. G., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014). Fbk-upv-uedin participation in the wmt14 quality estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Cap, F., Weller, M., Ramm, A., and Fraser, A. (2014). Cims – the cis and ims joint submission to wmt 2014 translating from english into german. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurment*, 20(1):37–46.

Cohn, T. and Specia, L. (2013). Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL-2013, pages 32–42, Sofia, Bulgaria.

Collins, M. (2002). Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Costa-jussà, M. R., Gupta, P., Rosso, P., and Banchs, R. E. (2014). English-to-hindi system description for wmt 2014: Deep source-context features for moses. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Do, Q. K., Herrmann, T., Niehues, J., Allauzen, A., Yvon, F., and Waibel, A. (2014). The kit-limsi translation system for wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Dungarwal, P., Chatterjee, R., Mishra, A., Kunchukuttan, A., Shah, R., and Bhattacharyya, P. (2014). The iit bombay hindi-english translation system at wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Durrani, N., Haddow, B., Koehn, P., and Heafield, K. (2014a). Edinburgh's phrase-based machine translation systems for wmt-14. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

Durrani, N., Haddow, B., Koehn, P., and Heafield, K. (2014b). Edinburghs phrase-based machine translation systems for wmt-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Dušek, O., Hajič, J., Hlaváčová, J., Novák, M., Pecina, P., Rosa, R., Tamchyna, A., Urešová, Z., and Zeman, D. (2014). Machine translation of medical texts in the khresmoi project. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

Federmann, C. (2012). Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:25–35.

Foster, J. (2007). Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal on Document Analysis and Recognition*, 10(3-4):129–145.

Freitag, M., Peitz, S., Wuebker, J., Ney, H., Huck, M., Sennrich, R., Durrani, N., Nadejde, M., Williams, P., Koehn, P., Herrmann, T., Cho, E., and Waibel, A. (2014). Eu-bridge mt: Combined machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Gamon, M., Aue, A., and Smets, M. (2005). Sentence-level MT evaluation without reference translations: beyond language modeling. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, Budapest.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Green, S., Cer, D., and Manning, C. (2014). Phrasal: A toolkit for new directions in statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hardmeier, C., Stymne, S., Tiedemann, J., Smith, A., and Nivre, J. (2014). Anaphora models and reordering for phrase-based smt. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Herbrich, R., Minka, T., and Graepel, T. (2006). TrueSkill™: A Bayesian Skill Rating System. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 569–576, Vancouver, British Columbia, Canada. MIT Press.

Herrmann, T., Mediani, M., Cho, E., Ha, T.-L., Niehues, J., Slawik, I., Zhang, Y., and Waibel, A. (2014). The karlsruhe institute of technology translation systems for the wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hokamp, C., Calixto, I., Wagner, J., and Zhang, J. (2014). Target-centric features for translation quality estimation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hopkins, M. and May, J. (2013). Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for Omics research on drugs. *Nucleic acids research*, 39(suppl 1):D1035–D1041.

Koehn, P. (2012a). Simulating human judgment in

machine translation evaluation campaigns. In *International Workshop on Spoken Language Translation (IWSLT)*.

Koehn, P. (2012b). Simulating Human Judgment in Machine Translation Evaluation Campaigns. In *Proceedings of the Ninth International Workshop on Spoken Language Translation*, pages 179–184, Hong Kong, China.

Koehn, P. and Monz, C. (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.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*, pages 1318–1326, Portland, Oregon.

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

Leusch, G., Ueffing, N., and Ney, H. (2006). Cder: Efficient mt evaluation using block movements. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 241–248, Trento, Italy.

Li, J., Kim, S.-J., Na, H., and Lee, J.-H. (2014a). Postech's system description for medical text translation task. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

Li, L., Wu, X., Vaillo, S. C., Xie, J., Way, A., and Liu, Q. (2014b). The dcu-ictcas mt system at wmt 2014 on german-english translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Lopez, A. (2012). Putting Human Assessments of Machine Translation Systems in Order. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Lu, Y., Wang, L., Wong, D. F., Chao, L. S., Wang, Y., and Oliveira, F. (2014). Domain adaptation for medical text translation using web resources. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

Luong, N. Q., Besacier, L., and Lecouteux, B. (2014). Lig system for word level qe task at wmt14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Luong, N. Q., Lecouteux, B., and Besacier, L. (2013). LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 384–389, Sofia, Bulgaria. Association for Computational Linguistics.

Macháček, M. and Bojar, O. (2014). Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Matthews, A., Ammar, W., Bhatia, A., Feely, W., Hanneman, G., Schlinger, E., Swayamdipta, S., Tsvetkov, Y., Lavie, A., and Dyer, C. (2014). The cmu machine translation systems at wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Neidert, J., Schuster, S., Green, S., Heafield, K., and Manning, C. (2014). Stanford universitys submissions to the wmt 2014 translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Okita, T., Vahid, A. H., Way, A., and Liu, Q. (2014). Dcu terminology translation system for medical query subtask at wmt14. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA. Association for Computational Linguistics.

Pécheux, N., Gong, L., Do, Q. K., Marie, B., Ivanishcheva, Y., Allauzen, A., Lavergne, T.,

Niehues, J., Max, A., and Yvon, Y. (2014). LIMSI @ WMT'14 Medical Translation Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA.

Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G., Kelly, L., Leveling, J., Mareček, D., Novák, M., Popel, M., Rosa, R., Tamchyna, A., and Urešová, Z. (2014). Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, (0):–.

Peitz, S., Wuebker, J., Freitag, M., and Ney, H. (2014). The rwth aachen german-english machine translation system for wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Pouliquen, B. and Mazenc, C. (2011). COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 24–30, Xiamen, China. Asia-Pacific Association for Machine Translation.

Powers, D. M. W. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*.

Quernheim, D. and Cap, F. (2014). Large-scale exact decoding: The ims-ttt submission to wmt14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Rosse, C. and Mejino Jr., J. L. V. (2008). The foundational model of anatomy ontology. In Burger, A., Davidson, D., and Baldock, R., editors, *Anatomy Ontologies for Bioinformatics*, volume 6 of *Computational Biology*, pages 59–117. Springer London.

Rubino, R., Toral, A., Sánchez-Cartagena, V. M., Ferrández-Tordera, J., Ortiz Rojas, S., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Way, A. (2014). Abu-matran at wmt 2014 translation task: Two-step data selection and rbmt-style synthetic rules. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland.

Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2014). The ua-prompsit hybrid machine translation system for the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Scarton, C. and Specia, L. (2014). Exploring consensus in machine translation for quality estimation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. (2014). Machine translation and monolingual postediting: The afrl wmt-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Seginer, Y. (2007). *Learning Syntactic Structure*. PhD thesis, University of Amsterdam.

Shah, K., Cohn, T., and Specia, L. (2013). An investigation on the effectiveness of features for translation quality estimation. In *Proceedings of the Machine Translation Summit XIV*, pages 167–174, Nice, France.

Shah, K. and Specia, L. (2014). Quality estimation for translation selection. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, Dubrovnik, Croatia.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

Souza, J. G. C. d., Espl-Gomis, M., Turchi, M., and Negri, M. (2013). Exploiting qualitative information from automatic word alignment for cross-lingual nlp tasks. In *The 51st Annual*

Meeting of the Association for Computational Linguistics - Short Papers (ACL Short Papers 2013).

Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst - A Translation Quality Estimation Framework. In *Proceedings of the 51th Conference of the Association for Computational Linguistics (ACL), Demo Session*, Sofia, Bulgaria.

Tamchyna, A., Popel, M., Rosa, R., and Bojar, O. (2014). Cuni in wmt14: Chimera still awaits bellerophon. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tan, L. and Pal, S. (2014). Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Thompson, P., Iqbal, S., McNaught, J., and Ananiadou, S. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10(1):349.

Tiedemann, J. (2009). News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, Borovets, Bulgaria. John Benjamins.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated DP based search for statistical translation. In Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors, *Proceedings of the Fifth European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece. International Speech Communication Association.

U.S. National Library of Medicine (2009). UMLS reference manual. Metathesaurus. Bethesda, MD, USA.

Voorhees, E. M. and Harman, D. K., editors (2005). *TREC: Experiment and evaluation in information retrieval*, volume 63 of *Digital libraries and electronic publishing series*. MIT press Cambridge, Cambridge, MA, USA.

Wang, L., Lu, Y., Wong, D. F., Chao, L. S., Wang, Y., and Oliveira., F. (2014). Combining domain adaptation approaches for medical text translation. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

Wäschle, K. and Riezler, S. (2012). Analyzing parallelism and domain similarities in the MAREC patent corpus. In Salampasis, M. and Larsen, B., editors, *Multidisciplinary Information Retrieval*, volume 7356 of *Lecture Notes in Computer Science*, pages 12–27. Springer Berlin Heidelberg.

Williams, P., Sennrich, R., Nadejde, M., Huck, M., Hasler, E., and Koehn, P. (2014). Edinburghs syntax-based systems at wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Wisniewski, G., Pécheux, N., Allauzen, A., and Yvon, F. (2014). Limsi submission for wmt'14 qe task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

wu, x., Haque, R., Okita, T., Arora, P., Way, A., and Liu, Q. (2014). Dcu-lingo24 participation in wmt 2014 hindi-english translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA. Association for Computational Linguistics.

Zdeňka Urešová, Ondřej Dušek, J. H. and Pecina, P. (2014). Multilingual test sets for machine translation of search queries for cross-lingual information retrieval in the medical domain. In *To appear in Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Zhang, J., Wu, X., Calixto, I., Vahid, A. H., Zhang, X., Way, A., and Liu, Q. (2014). Experiments in medical translation shared task at wmt 2014. In *Proceedings of the ACL 2014 Ninth Workshop of Statistical Machine Translation*, Baltimore, USA.

A Pairwise System Comparisons by Human Judges

Tables 28–37 show pairwise comparisons between systems for each language pair. The numbers in each of the tables' cells indicate the percentage of times that the system in that column was judged to be better than the system in that row, ignoring ties. Bolding indicates the winner of the two systems.

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

Each table contains final rows showing how likely a system would win when paired against a randomly selected system (the expected win ratio score) and the rank range according the official method used in Table 8. Gray lines separate clusters based on non-overlapping rank ranges.

	ONLINE-B	UEDIN-PHRASE	UEDIN-SYNTAX	ONLINE-A	CU-MOSES
ONLINE-B	–	.47‡	.43‡	.42‡	.39‡
UEDIN-PHRASE	**.53**‡	–	.44‡	.44‡	.41‡
UEDIN-SYNTAX	**.57**‡	**.56**‡	–	.49	.48†
ONLINE-A	**.58**‡	**.56**‡	**.51**	–	.48⋆
CU-MOSES	**.61**‡	**.59**‡	**.52**†	**.52**⋆	–
score	.57	.54	.47	.46	.44
rank	1	2	3-4	3-4	5

Table 28: Head to head comparison, ignoring ties, for Czech-English systems

	CU-DEPFIX	UEDIN-UNCNSTR	CU-BOJAR	CU-FUNKY	ONLINE-B	UEDIN-PHRASE	ONLINE-A	CU-TECTO	COMMERCIAL1	COMMERCIAL2
CU-DEPFIX	–	.50	.42‡	.48	.44‡	.43‡	.41‡	.35‡	.30‡	.24‡
UEDIN-UNCNSTR	.50	–	**.51**	.48	.42‡	.37‡	.42‡	.39‡	.31‡	.26‡
CU-BOJAR	**.58**‡	.49	–	.49	.45‡	.44‡	.40‡	.36‡	.32‡	.24‡
CU-FUNKY	**.52**	**.52**	**.51**	–	.48	.47†	.44‡	.34‡	.33‡	.26‡
ONLINE-B	**.56**‡	**.58**‡	**.55**‡	**.52**	–	.48	.47†	.41‡	.31‡	.26‡
UEDIN-PHRASE	**.57**‡	**.63**‡	**.56**‡	**.53**†	**.52**	–	.48	.44‡	.32‡	.27‡
ONLINE-A	**.59**‡	**.58**‡	**.60**‡	**.56**‡	**.53**†	**.52**	–	.45‡	.37‡	.30‡
CU-TECTO	**.65**‡	**.61**‡	**.64**‡	**.66**‡	**.59**‡	**.56**‡	**.55**‡	–	.42‡	.30‡
COMMERCIAL1	**.70**‡	**.69**‡	**.68**‡	**.67**‡	**.69**‡	**.68**‡	**.63**‡	**.58**‡	–	.40‡
COMMERCIAL2	**.76**‡	**.74**‡	**.76**‡	**.74**‡	**.74**‡	**.73**‡	**.70**‡	**.70**‡	**.60**‡	–
score	.60	.59	.58	.57	.54	.52	.50	.44	.36	.28
rank	1-3	1-3	1-4	3-4	5-6	5-6	7	8	9	10

Table 29: Head to head comparison, ignoring ties, for English-Czech systems

	ONLINE-B	UEDIN-SYNTAX	ONLINE-A	LIMSI-KIT	EU-BRIDGE	UEDIN-PHRASE	KIT	RWTH	DCU-ICTCAS	CMU	RBMT4	RBMT1	ONLINE-C
ONLINE-B	–	.46	.40‡	.41‡	.35‡	.42‡	.38‡	.35‡	.40‡	.31‡	.33‡	.32‡	.22‡
UEDIN-SYNTAX	**.54**	–	**.51**	.47	.47	.45	.45⋆	.39‡	.36‡	.38‡	.35‡	.34‡	.27‡
ONLINE-A	**.60**‡	.49	–	.42†	.44†	**.51**	.41‡	.38‡	.44⋆	.42‡	.38‡	.31‡	.20‡
LIMSI-KIT	**.59**‡	**.53**	**.58**†	–	**.55**	**.53**	.31‡	.45⋆	.39‡	.41‡	.37‡	.35‡	.29‡
EU-BRIDGE	**.65**‡	**.53**	**.56**†	.45	–	.45	.44⋆	.48	.40‡	.37‡	.39‡	.37‡	.30‡
UEDIN-PHRASE	**.58**‡	**.55**	.49	.47	**.55**	–	.48	.39‡	.34‡	.45⋆	.40‡	.40‡	.34‡
KIT	**.62**‡	**.55**⋆	**.59**‡	**.69**‡	**.56**⋆	**.52**	–	.45⋆	.41‡	.45⋆	.47	.40‡	.31‡
RWTH	**.65**‡	**.61**‡	**.62**‡	**.55**⋆	**.52**	**.61**‡	**.55**⋆	–	**.54**	.44†	.44†	.38‡	.37‡
DCU-ICTCAS	**.60**‡	**.64**‡	**.56**⋆	**.61**‡	**.60**‡	**.66**‡	**.59**‡	.46	–	**.51**	.49	.46⋆	.40‡
CMU	**.69**‡	**.62**‡	**.58**‡	**.59**‡	**.63**‡	**.55**⋆	**.55**⋆	**.56**†	.49	–	**.53**	.42‡	.43†
RBMT4	**.67**‡	**.65**‡	**.62**‡	**.63**‡	**.61**‡	**.60**‡	**.53**	**.56**†	**.51**	.47	–	**.51**	.37‡
RBMT1	**.68**‡	**.66**‡	**.69**‡	**.65**‡	**.63**‡	**.60**‡	**.60**‡	**.62**‡	**.54**⋆	**.58**‡	.49	–	.38‡
ONLINE-C	**.78**‡	**.73**‡	**.80**‡	**.71**‡	**.70**‡	**.66**‡	**.69**‡	**.63**‡	**.60**‡	**.57**†	**.63**‡	**.62**‡	–
score	.63	.58	.58	.55	.55	.54	.49	.47	.45	.44	.44	.40	.32
rank	1	2-3	2-3	4-6	4-6	4-6	7-8	7-8	9-11	9-11	9-11	12	13

Table 30: Head to head comparison, ignoring ties, for German-English systems

	UEDIN-SYNTAX	ONLINE-B	ONLINE-A	PROMT-HYBRID	PROMT-RULE	UEDIN-STANFORD	EU-BRIDGE	RBMT4	UEDIN-PHRASE	RBMT1	KIT	STANFORD-UNC	CIMS	STANFORD	UU	ONLINE-C	IMS-TTT	UU-DOCENT
UEDIN-SYNTAX	–	**.55**⋆	.46⋆	.45⋆	.46⋆	.44†	.41‡	.45†	.43‡	.41‡	.38‡	.38‡	.36‡	.33‡	.38‡	.30‡	.30‡	.25‡
ONLINE-B	.45⋆	–	.50	.48	.50	.47	.43†	.46⋆	.41‡	.45†	.39‡	.39‡	.37‡	.32‡	.35‡	.34‡	.30‡	.29‡
ONLINE-A	**.54**⋆	.50	–	.44†	**.52**	.50	.45⋆	.43‡	.43†	.42‡	.39‡	.41‡	.42‡	.42‡	.37‡	.44†	.38‡	.33‡
PROMT-HYBRID	**.55**⋆	**.52**	**.56**†	–	.45⋆	.47	.47	.46⋆	.50	.44†	.42‡	.40‡	.41‡	.38‡	.39‡	.39‡	.33‡	.34‡
PROMT-RULE	**.54**⋆	.50	.48	**.55**⋆	–	**.51**	.47	.47	.45⋆	.38‡	.42‡	.40‡	.43†	.41‡	.43†	.38‡	.35‡	.29‡
UEDIN-STANFORD	**.56**†	**.53**	.50	**.53**	.49	–	.48	.50	.47	.44†	.46	.36‡	.36‡	.36‡	.36‡	.35‡	.30‡	.32‡
EU-BRIDGE	**.59**‡	**.57**†	**.55**⋆	**.53**	**.53**	**.52**	–	.46⋆	.43†	**.52**	.42‡	.42‡	.45⋆	.35‡	.36‡	.41‡	.38‡	.30‡
RBMT4	**.55**†	**.54**⋆	**.57**‡	**.54**⋆	**.53**	.50	**.54**⋆	–	**.53**	.49	.44†	.49	.50	.47	.40‡	.42‡	.38‡	.40‡
UEDIN-PHRASE	**.57**‡	**.59**‡	**.57**†	.50	**.55**⋆	**.53**	**.57**†	.47	–	.50	**.55**⋆	.47	.45⋆	.44†	.43†	.42‡	.37‡	.34‡
RBMT1	**.59**‡	**.55**†	**.58**‡	**.56**†	**.62**‡	**.56**†	.48	**.51**	.50	–	.47	.47	.45†	.47	.43‡	.42‡	.38‡	.41‡
KIT	**.62**‡	**.61**‡	**.61**‡	**.58**‡	**.58**‡	**.54**	**.58**‡	**.56**†	.45⋆	**.53**	–	.47	.49	.46	.43‡	.48	.34‡	.37‡
STANFORD-UNC	**.62**‡	**.61**‡	**.59**‡	**.60**‡	**.60**‡	**.64**‡	**.58**‡	**.51**	**.53**	**.53**	**.53**	–	.48	.47	.45†	.45⋆	.39‡	.41‡
CIMS	**.64**‡	**.63**‡	**.58**‡	**.59**‡	**.57**†	**.64**‡	**.55**⋆	.50	**.55**⋆	**.55**†	**.51**	**.52**	–	**.53**	.42‡	**.52**	.47	.42‡
STANFORD	**.67**‡	**.68**‡	**.58**‡	**.62**‡	**.59**‡	**.64**‡	**.65**‡	**.53**	**.56**†	**.53**	**.54**	**.53**	.47	–	**.53**	.42‡	.39‡	.48
UU	**.62**‡	**.65**‡	**.62**‡	**.61**‡	**.57**†	**.64**‡	**.64**‡	**.60**‡	**.57**†	**.57**‡	**.57**‡	**.55**†	**.58**‡	.47	–	.46⋆	.45†	.38‡
ONLINE-C	**.70**‡	**.66**‡	**.56**†	**.61**‡	**.62**‡	**.65**‡	**.59**‡	**.58**‡	**.58**‡	**.58**‡	**.52**	**.55**⋆	.48	**.58**‡	**.54**⋆	–	.48	.47
IMS-TTT	**.70**‡	**.70**‡	**.62**‡	**.67**‡	**.65**‡	**.70**‡	**.62**‡	**.62**‡	**.63**‡	**.62**‡	**.66**‡	**.61**‡	**.53**	**.61**‡	**.55**†	**.52**	–	.49
UU-DOCENT	**.75**‡	**.71**‡	**.67**‡	**.66**‡	**.71**‡	**.68**‡	**.70**‡	**.60**‡	**.66**‡	**.59**‡	**.63**‡	**.59**‡	**.58**‡	**.52**	**.62**‡	**.53**	**.51**	–
score	.60	.59	.56	.56	.56	.56	.54	.51	.51	.50	.48	.47	.46	.44	.43	.42	.38	.37
rank	1-2	1-2	3-6	3-6	3-6	3-6	7	8-10	8-10	8-10	11-12	11-13	12-14	13-15	14-16	15-16	17-18	17-18

Table 31: Head to head comparison, ignoring ties, for English-German systems

	UEDIN-PHRASE	KIT	ONLINE-B	STANFORD	ONLINE-A	RBMT1	RBMT4	ONLINE-C
UEDIN-PHRASE	–	.48	.48	.45‡	.43‡	.28‡	.28‡	.19‡
KIT	**.52**	–	**.54**†	.48	.44‡	.31‡	.29‡	.21‡
ONLINE-B	**.52**	.46†	–	**.51**	.47	.31‡	.30‡	.24‡
STANFORD	**.55**‡	**.52**	.49	–	.46†	.34‡	.30‡	.23‡
ONLINE-A	**.57**‡	**.56**‡	**.53**	**.54**†	–	.32‡	.29‡	.21‡
RBMT1	**.72**‡	**.69**‡	**.69**‡	**.66**‡	**.68**‡	–	.42‡	.33‡
RBMT4	**.72**‡	**.71**‡	**.70**‡	**.70**‡	**.71**‡	**.58**‡	–	.39‡
ONLINE-C	**.81**‡	**.79**‡	**.76**‡	**.77**‡	**.79**‡	**.67**‡	**.61**‡	–
score	.63	.60	.59	.58	.57	.40	.35	.25
rank	1	2-4	2-4	2-4	5	6	7	8

Table 32: Head to head comparison, ignoring ties, for French-English systems

	ONLINE-B	UEDIN-PHRASE	KIT	MATRAN	MATRAN-RULES	ONLINE-A	UU-DOCENT	PROMT-HYBRID	UA	PROMT-RULE	RBMT1	RBMT4	ONLINE-C
ONLINE-B	–	.46⋆	.48	.46⋆	.50	.41‡	.39‡	.39‡	.37‡	.38‡	.37‡	.35‡	.27‡
UEDIN-PHRASE	**.54**⋆	–	.50	.47	.46	.46⋆	.42‡	.41‡	.46⋆	.42‡	.35‡	.34‡	.33‡
KIT	**.52**	.50	–	**.53**	**.51**	.50	.43‡	.49	.41‡	.42‡	.35‡	.37‡	.29‡
MATRAN	**.54**⋆	**.53**	.47	–	.49	.50	.43‡	.43‡	.38‡	.48	.40‡	.34‡	.32‡
MATRAN-RULES	.50	**.54**	.49	**.51**	–	**.53**	.40‡	.45†	.46⋆	.42‡	.44†	.40‡	.34‡
ONLINE-A	**.59**‡	**.54**⋆	.50	.50	.47	–	.44†	.49	.47	.45⋆	.42‡	.37‡	.34‡
UU-DOCENT	**.61**‡	**.58**‡	**.57**‡	**.57**‡	**.60**‡	**.56**†	–	.43‡	**.52**	.46⋆	.39‡	.44†	.33‡
PROMT-HYBRID	**.61**‡	**.59**‡	**.51**	**.57**‡	**.55**†	**.51**	**.57**‡	–	.50	.41‡	.46⋆	.44†	.35‡
UA	**.63**‡	**.54**⋆	**.59**‡	**.62**‡	**.54**⋆	**.53**	.48	.50	–	.49	.46⋆	.43‡	.34‡
PROMT-RULE	**.62**‡	**.58**‡	**.58**‡	**.52**	**.58**‡	**.55**⋆	**.54**⋆	**.59**‡	**.51**	–	.47	.39‡	.37‡
RBMT1	**.63**‡	**.65**‡	**.65**‡	**.60**‡	**.56**†	**.58**‡	**.61**‡	**.54**⋆	**.54**⋆	**.53**	–	.46⋆	.45†
RBMT4	**.65**‡	**.66**‡	**.63**‡	**.66**‡	**.60**‡	**.63**‡	**.56**†	**.56**†	**.57**‡	**.61**‡	**.54**⋆	–	.45⋆
ONLINE-C	**.73**‡	**.67**‡	**.71**‡	**.67**‡	**.66**‡	**.66**‡	**.67**‡	**.65**‡	**.66**‡	**.63**‡	**.55**†	**.55**⋆	–
score	.59	.57	.55	.55	.54	.53	.49	.49	.48	.47	.43	.40	.34
rank	1	2-4	2-5	2-5	4-6	4-6	7-9	7-10	7-10	8-10	11	12	13

Table 33: Head to head comparison, ignoring ties, for English-French systems

	ONLINE-B	ONLINE-A	UEDIN-SYNTAX	CMU	UEDIN-PHRASE	AFRL	IIT-BOMBAY	DCU-LINGO24	IIIT-HYDERABAD
ONLINE-B	–	.36‡	.33‡	.37‡	.31‡	.21‡	.20‡	.14‡	.00
ONLINE-A	**.64**‡	–	.48	.47⋆	.44‡	.31‡	.30‡	.24‡	.12‡
UEDIN-SYNTAX	**.67**‡	**.52**	–	.47	.46†	.33‡	.29‡	.24‡	.12‡
CMU	**.63**‡	**.53**⋆	**.53**	–	.47	.37‡	.31‡	.26‡	.11‡
UEDIN-PHRASE	**.69**‡	**.56**‡	**.54**†	**.53**	–	.40‡	.33‡	.25‡	.11‡
AFRL	**.79**‡	**.69**‡	**.67**‡	**.63**‡	**.60**‡	–	**.53**	.40‡	.16‡
IIT-BOMBAY	**.80**‡	**.70**‡	**.71**‡	**.69**‡	**.67**‡	.47	–	.44‡	.19‡
DCU-LINGO24	**.86**‡	**.76**‡	**.76**‡	**.74**‡	**.75**‡	**.60**‡	**.56**‡	–	.19‡
IIIT-HYDERABAD	**.94**‡	**.88**‡	**.88**‡	**.89**‡	**.89**‡	**.84**‡	**.81**‡	**.81**‡	–
score	.75	.62	.61	.60	.57	.44	.41	.34	.13
rank	1	2-3	2-4	3-4	5	6-7	6-7	8	9

Table 34: Head to head comparison, ignoring ties, for Hindi-English systems

	ONLINE-B	ONLINE-A	UEDIN-UNCNSTR	UEDIN-PHRASE	CU-MOSES	IIT-BOMBAY	IPN-UPV-CNTXT	DCU-LINGO24	IPN-UPV-NODEV	MANAWI-H1	MANAWI	MANAWI-RMOOV
ONLINE-B	–	.49	.28‡	.29‡	.27‡	.23‡	.22‡	.20‡	.17‡	.12‡	.13‡	.13‡
ONLINE-A	**.51**	–	.31‡	.29‡	.27‡	.25‡	.20‡	.20‡	.21‡	.19‡	.16‡	.15‡
UEDIN-UNCNSTR	**.72**‡	**.69**‡	–	.44†	.49	.39‡	.40‡	.34‡	.39‡	.29‡	.30‡	.27‡
UEDIN-PHRASE	**.71**‡	**.71**‡	**.56**†	–	.48	.45†	.44†	.39‡	.37‡	.31‡	.31‡	.32‡
CU-MOSES	**.73**‡	**.73**‡	**.51**	**.52**	–	.47	.42‡	.40‡	.45⋆	.36‡	.35‡	.33‡
IIT-BOMBAY	**.77**‡	**.75**‡	**.61**‡	**.55**†	**.53**	–	.50	.47	.45†	.41‡	.40‡	.36‡
IPN-UPV-CNTXT	**.78**‡	**.80**‡	**.60**‡	**.56**†	**.58**‡	.50	–	**.51**	.41‡	.40‡	.40‡	.37‡
DCU-LINGO24	**.80**‡	**.80**‡	**.66**‡	**.61**‡	**.60**‡	**.53**	.49	–	**.52**	.41‡	.41‡	.39‡
IPN-UPV-NODEV	**.83**‡	**.79**‡	**.61**‡	**.63**‡	**.55**⋆	**.55**†	**.59**‡	.48	–	.46⋆	.44†	.38‡
MANAWI-H1	**.88**‡	**.81**‡	**.71**‡	**.69**‡	**.64**‡	**.59**‡	**.60**‡	**.59**‡	**.54**⋆	–	.35‡	.34‡
MANAWI	**.87**‡	**.84**‡	**.70**‡	**.69**‡	**.65**‡	**.60**‡	**.60**‡	**.59**‡	**.56**†	**.65**‡	–	.39‡
MANAWI-RMOOV	**.87**‡	**.85**‡	**.73**‡	**.68**‡	**.67**‡	**.64**‡	**.63**‡	**.61**‡	**.62**‡	**.66**‡	**.61**‡	–
score	.77	.75	.57	.54	.52	.47	.46	.43	.42	.38	.35	.31
rank	1	2	3	4-5	4-5	6-7	6-7	8-9	8-9	10-11	10-11	12

Table 35: Head to head comparison, ignoring ties, for English-Hindi systems

	AFRL-PE	ONLINE-B	ONLINE-A	PROMT-HYBRID	PROMT-RULE	UEDIN-PHRASE	YANDEX	ONLINE-G	AFRL	UEDIN-SYNTAX	KAZNU	RBMT1	RBMT4
AFRL-PE	–	.42‡	.40‡	.39‡	.39‡	.41‡	.35‡	.39‡	.28‡	.26‡	.26‡	.29‡	.21‡
ONLINE-B	**.58**‡	–	.42‡	.43‡	.45†	.45†	.42‡	.43‡	.46⋆	.37‡	.33‡	.29‡	.31‡
ONLINE-A	**.60**‡	**.58**‡	–	.50	.45†	**.51**	.47	.45†	.42‡	.40‡	.33‡	.32‡	.30‡
PROMT-HYBRID	**.61**‡	**.57**‡	.50	–	.47	.45⋆	.49	.44†	.43‡	.44†	.39‡	.31‡	.27‡
PROMT-RULE	**.61**‡	**.55**†	**.55**†	**.53**	–	.46⋆	.47	.49	.48	.42‡	.36‡	.34‡	.30‡
UEDIN-PHRASE	**.59**‡	**.55**†	.49	**.55**⋆	**.54**⋆	–	.49	.50	.47	.44†	.32‡	.37‡	.29‡
YANDEX	**.65**‡	**.58**‡	**.53**	**.51**	**.53**	**.51**	–	.48	.50	.43‡	.34‡	.36‡	.34‡
ONLINE-G	**.61**‡	**.57**‡	**.55**†	**.56**†	**.51**	.50	**.52**	–	.48	.43‡	.39‡	.35‡	.30‡
AFRL	**.72**‡	**.54**⋆	**.58**‡	**.57**‡	**.52**	**.53**	.50	**.52**	–	.44†	.41‡	.41‡	.37‡
UEDIN-SYNTAX	**.74**‡	**.63**‡	**.60**‡	**.56**†	**.58**‡	**.56**†	**.57**‡	**.57**‡	**.56**†	–	**.51**	.36‡	.37‡
KAZNU	**.74**‡	**.67**‡	**.67**‡	**.61**‡	**.64**‡	**.68**‡	**.66**‡	**.61**‡	**.59**‡	.49	–	.44†	.38‡
RBMT1	**.71**‡	**.71**‡	**.68**‡	**.69**‡	**.66**‡	**.63**‡	**.64**‡	**.65**‡	**.59**‡	**.64**‡	**.56**†	–	.47
RBMT4	**.79**‡	**.69**‡	**.70**‡	**.73**‡	**.70**‡	**.71**‡	**.66**‡	**.70**‡	**.63**‡	**.63**‡	**.62**‡	**.53**	–
score	.66	.58	.55	.55	.53	.53	.52	.51	.49	.45	.40	.36	.32
rank	1	2	3-5	3-5	4-7	5-8	5-8	5-8	9	10	11	12	13

Table 36: Head to head comparison, ignoring ties, for Russian-English systems

	PROMT-RULE	ONLINE-B	PROMT-HYBRID	UEDIN-UNCNSTR	ONLINE-G	ONLINE-A	UEDIN-PHRASE	RBMT4	RBMT1
PROMT-RULE	–	**.51**	.45†	.43‡	.43‡	.39‡	.38‡	.15‡	.00
ONLINE-B	.49	–	.50	.47⋆	.38‡	.36‡	.38‡	.16‡	.13‡
PROMT-HYBRID	**.55**†	.50	–	.49	.47	.39‡	.40‡	.18‡	.15‡
UEDIN-UNCNSTR	**.57**‡	**.53**⋆	**.51**	–	.50	.44‡	.36‡	.25‡	.18‡
ONLINE-G	**.57**‡	**.62**‡	**.53**	.50	–	.46⋆	.44‡	.23‡	.18‡
ONLINE-A	**.61**‡	**.64**‡	**.61**‡	**.56**‡	**.54**⋆	–	.49	.24‡	.18‡
UEDIN-PHRASE	**.62**‡	**.62**‡	**.60**‡	**.64**‡	**.56**‡	**.51**	–	.30‡	.21‡
RBMT4	**.85**‡	**.84**‡	**.82**‡	**.75**‡	**.77**‡	**.76**‡	**.70**‡	–	.42‡
RBMT1	**.91**‡	**.87**‡	**.85**‡	**.82**‡	**.82**‡	**.82**‡	**.79**‡	**.58**‡	–
score	.64	.64	.61	.58	.55	.51	.49	.26	.19
rank	1-2	1-2	3	4-5	4-5	6-7	6-7	8	9

Table 37: Head to head comparison, ignoring ties, for English-Russian systems