English→Russian MT evaluation campaign

Pavel Braslavski
Kontur Labs / Ural Federal University,Russia
pbras@yandex.ru

Alexander Beloborodov
Ural Federal University
Russia
xander-beloborodov@yandex.ru

Maxim Khalilov
TAUS Labs
The Netherlands
maxim@tauslabs.com

Serge Sharoff
University of Leeds
UK
s.sharoff@leeds.ac.uk

Abstract

This paper presents the settings and the results of the ROMIP 2013 MT shared task for the English→Russian language direction. The quality of generated translations was assessed using automatic metrics and human evaluation. We also discuss ways to reduce human evaluation efforts using pairwise sentence comparisons by human judges to simulate sort operations.

1 Introduction

Machine Translation (MT) between English and Russian was one of the first translation directions tested at the dawn of MT research in the 1950s (Hutchins, 2000). Since then the MT paradigms changed many times, many systems for this language pair appeared (and disappeared), but as far as we know there was no systematic quantitative evaluation of a range of systems, analogous to DARPA'94 (White et al., 1994) and later evaluation campaigns. The Workshop on Statistical MT (WMT) in 2013 has announced a Russian evaluation track for the first time.[1] However, this evaluation is currently ongoing, it should include new methods for building statistical MT (SMT) systems for Russian from the data provided in this track, but it will not cover the performance of existing systems, especially rule-based (RBMT) or hybrid ones.

Evaluation campaigns play an important role in promotion of the progress for MT technologies. Recently, there have been a number of MT shared tasks for combinations of several European, Asian and Semitic languages (Callison-Burch et al., 2011; Callison-Burch et al., 2012; Federico et al., 2012), which we took into account in designing the campaign for the English-Russian direction. The evaluation has been held in the context of ROMIP,[2] which stands for Russian Information Retrieval Evaluation Seminar and is a TREC-like[3] Russian initiative started in 2002.

One of the main challenges in developing MT systems for Russian and for evaluating them is the need to deal with its free word order and complex morphology. Long-distance dependencies are common, and this creates problems for both RBMT and SMT systems (especially for phrase-based ones). Complex morphology also leads to considerable sparseness for word alignment in SMT.

The language direction was chosen to be English→Russian, first because of the availability of native speakers for evaluation, second because the systems taking part in this evaluation are mostly used in translation of English texts for the Russian readers.

2 Corpus preparation

In designing the set of texts for evaluation, we had two issues in mind. First, it is known that the domain and genre can influence MT performance (Langlais, 2002; Babych et al., 2007), so we wanted to control the set of genres. Second, we were aiming at using sources allowing distribution of texts under a Creative Commons licence. In the end two genres were used coming from two sources. The newswire texts were collected from the English Wikinews website.[4] The second genre was represented by 'regulations' (laws, contracts, rules, etc), which were collected from the Web using a genre classification method described in (Sharoff, 2010). The method provided a sufficient accuracy (74%) for the initial selection of texts under the category of 'regulations,' which was followed by a manual check to reject texts clearly outside of this genre category.

[1] http://www.statmt.org/wmt13/

[2] http://romip.ru/en/

[3] http://trec.nist.gov/

[4] http://en.wikinews.org/

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 262–267,
Sofia, Bulgaria, August 4-9 2013.

The initial corpus consists of 8,356 original English texts that make up 148,864 sentences. We chose to retain the entire texts in the corpus rather than individual sentences, since some MT systems may use information beyond isolated sentences. 100,889 sentences originated from Wikinews; 47,975 sentences came from the 'regulations' corpus. The first 1,002 sentences were published in advance to allow potential participants time to adjust their systems to the corpus format. The remaining 147,862 sentences were the corpus for testing translation into Russian. Two examples of texts in the corpus:

90237 *Ambassadors from the United States of America, Australia and Britain have all met with Fijian military officers to seek insurances that there wasn't going to be a coup.*

102835 *If you are given a discount for booking more than one person onto the same date and you later wish to transfer some of the delegates to another event, the fees will be recalculated and you will be asked to pay additional fees due as well as any administrative charge.*

For automatic evaluation we randomly selected 947 'clean' sentences, i.e. those with clear sentence boundaries, no HTML markup remains, etc. (such flaws sometimes occur in corpora collected from the Web). 759 sentences originated from the 'news' part of the corpus, the remaining 188 came from the 'regulations' part. The sentences came from sources without published translations into Russian, so that some of the participating systems do not get unfair advantage by using them for training. These sentences were translated by professional translators. For manual evaluation, we randomly selected 330 sentences out of 947 used for automatic evaluation, specifically, 190 from the 'news' part and 140 from the 'regulations' part.

The organisers also provided participants with access to the following additional resources:

- 1 million sentences from the English-Russian parallel corpus released by Yandex (the same as used in WMT13)[5];
- 119 thousand sentences from the English-Russian parallel corpus from the TAUS Data Repository.[6]

These resources are not related to the test corpus of the evaluation campaign. Their purpose was to make it easier to participate in the shared task for teams without sufficient data for this language pair.

3 Evaluation methodology

The main idea of manual evaluation was (1) to make the assessment as simple as possible for a human judge and (2) to make the results of evaluation unambiguous. We opted for pairwise comparison of MT outputs. This is different from simultaneous *ranking* of several MT outputs, as commonly used in WMT evaluation campaigns. In case of a large number of participating systems each assessor ranks only a subset of MT outputs. However, a fair overall ranking cannot be always derived from such partial rankings (Callison-Burch et al., 2012). The pairwise comparisons we used can be directly converted into unambiguous overall rankings. This task is also much simpler for human judges to complete. On the other hand, pairwise comparisons require a larger number of evaluation decisions, which is feasible only for few participants (and we indeed had relatively few submissions in this campaign). Below we also discuss how to reduce the amount of human efforts for evaluation.

In our case the assessors were asked to make a pairwise comparison of two sentences translated by two different MT systems against a gold standard translation. The question for them was to judge translation adequacy, i.e., which MT output conveys information from the reference translation better. The source English sentence was not presented to the assessors, because we think that we can have more trust in understanding of the source text by a professional translator. The translator also had access to the entire text, while the assessors could only see a single sentence.

For human evaluation we employed the multifunctional TAUS DQF tool[7] in the 'Quick Comparison' mode.

Assessors' judgements resulted in rankings for each sentence in the test set. In case of ties the ranks were averaged, e.g. when the ranks of the systems in positions 2-4 and 7-8 were tied, their ranks became: `1 3 3 3 5 6 7.5 7.5`. To produce the final ranking, the sentence-level ranks were averaged over all sentences.

Pairwise comparisons are time-consuming: n

[5] https://translate.yandex.ru/corpus?lang=en

[6] https://www.tausdata.org

[7] https://tauslabs.com/dynamic-quality/dqf-tools-mt

Metric	OS1	OS2	OS3	OS4	P1	P2	P3	P4	P5	P6	P7
Automatic metrics ALL (947 sentences)											
BLEU	0.150	0.141	0.133	0.124	**0.157**	0.112	0.105	0.073	0.094	0.071	0.073
NIST	**5.12**	4.94	4.80	4.67	5.00	4.46	4.11	2.38	4.16	3.362	3.38
Meteor	**0.258**	0.240	0.231	0.240	0.251	0.207	0.169	0.133	0.178	0.136	0.149
TER	**0.755**	0.766	0.764	0.758	0.758	0.796	0.901	0.931	0.826	0.934	0.830
GTM	**0.351**	0.338	0.332	0.336	0.349	0.303	0.246	0.207	0.275	0.208	0.230
Automatic metrics NEWS (759 sentences)											
BLEU	0.137	0.131	0.123	0.114	**0.153**	0.103	0.096	0.070	0.083	0.066	0.067
NIST	**4.86**	4.72	4.55	4.35	4.79	4.26	3.83	2.47	3.90	3.20	3.19
Meteor	0.241	0.224	0.214	0.222	**0.242**	0.192	0.156	0.127	0.161	0.126	0.136
TER	0.772	0.776	0.784	0.777	**0.768**	0.809	0.908	0.936	0.844	0.938	0.839
GTM	0.335	0.324	0.317	0.320	**0.339**	0.290	0.233	0.201	0.257	0.199	0.217

Table 1: Automatic evaluation results

cases require $\frac{n(n-1)}{2}$ pairwise decisions. In this study we also simulated a 'human-assisted' insertion sort algorithm and its variant with binary search. The idea is to run a standard sort algorithm and ask a human judge each time a comparison operation is required. This assumes that human perception of quality is transitive: if we know that $A < B$ and $B < C$, we can spare evaluation of A and C. This approach also implies that sentence pairs to judge are generated and presented to assessors on the fly; each decision contributes to selection of the pairs to be judged in the next step. If the systems are pre-sorted in a reasonable way (e.g. by an MT metric, under assumption that automatic pre-ranking is closer to the 'ideal' ranking than a random one), then we can potentially save even more pairwise comparison operations. Presorting makes ranking somewhat biased in favour of the order established by an MT metric. For example, if it favours one system against another, while in human judgement they are equal, the final ranking will preserve the initial order. Insertion sort of n sentences requires $n - 1$ comparisons in the best case of already sorted data and $\frac{n(n-1)}{2}$ in the worst case (reversely ordered data). Insertion sort with binary search requires $\sim n \log n$ comparisons regardless of the initial order. For this study we ran exhaustive pairwise evaluation and used its results to simulate human-assisted sorting.

In addition to human evaluation, we also ran system-level automatic evaluations using BLEU (Papineni et al., 2001), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2009), and GTM (Turian et al., 2003). We also wanted to estimate the correlations of these metrics with human judgements for the English→Russian pair on the corpus level and on the level of individual sentences.

4 Results

We received results from five teams, two teams submitted two runs each, which totals seven participants' runs (referred to as P1..P7 in the paper). The participants represent SMT, RBMT, and hybrid approaches. They included established groups from academia and industry, as well as new research teams. The evaluation runs also included the translations of the 947 test sentences produced by four free online systems in their default modes (referred to as OS1..OS4). For 11 runs automatic evaluation measures were calculated; eight runs underwent manual evaluation (four online systems plus four participants' runs; no manual evaluation was done by agreement with the participants for the runs P3, P6, and P7 to reduce the workload).

ID	Name and information
OS1	Phrase-based SMT
OS2	Phrase-based SMT
OS3	Hybrid (RBMT+statistical PE)
OS4	Dependency-based SMT
P1	Compreno, Hybrid, ABBYY Corp
P2	Pharaon, Moses, Yandex&TAUS data
P3,4	Balagur, Moses, Yandex&news data
P5	ETAP-3, RBMT, (Boguslavsky, 1995)
P6,7	Pereved, Moses, Internet data

OS3 is a hybrid system based on RBMT with SMT post-editing (PE). P1 is a hybrid system with analysis and generation driven by statistical evaluation of hypotheses.

All (330 sentences)							
OS3 (highest)	P1	OS1	OS2	OS4	P5	P2	P4 (lowest)
3.159	3.350	3.530	3.961	4.082	5.447	5.998	6.473
News (190 sentences)							
OS3 (highest)	P1	OS1	OS2	OS4	P5	P2	P4 (lowest)
2.947	3.450	3.482	4.084	4.242	5.474	5,968	6,353
Regulations (140 sentences)							
P1 (highest)	OS3	OS1	OS2	OS4	P5	P2	P4 (lowest)
3.214	3.446	3.596	3.793	3.864	5.411	6.039	6.636
Simulated dynamic ranking (insertion sort)							
P1 (highest)	OS1	OS3	OS2	OS4	P5	P4	P2 (lowest)
3.318	3.327	3.588	4.221	4.300	5.227	5.900	6.118
Simulated dynamic ranking (binary insertion sort)							
OS1 (highest)	P1	OS3	OS2	OS4	P5	P2	P4 (lowest)
2.924	3.045	3.303	3.812	4.267	5.833	5.903	6.882

Table 2: Human evaluation results

Table 1 gives the automatic scores for each of participating runs and four online systems. OS1 usually has the highest overall score (except BLEU), it also has the highest scores for 'regulations' (more formal texts), P1 scores are better for the news documents.

14 assessors were recruited for evaluation (participating team members and volunteers); the total volume of evaluation is 10,920 pairwise sentence comparisons. Table 2 presents the rankings of the participating systems using averaged ranks from the human evaluation. There is no statistically significant difference (using Welch's t-test at $p \leq 0.05$) in the overall ranks within the following groups: (OS1, OS3, P1) $<$ (OS2, OS4) $<$ P5 $<$ (P2, P4). OS3 (mostly RBMT) belongs to the troika of leaders in human evaluation contrary to the results of its automatic scores (Table 1). Similarly, P5 is consistently ranked higher than P2 by the assessors, while the automatic scores suggest the opposite. This observation confirms the well-known fact that the automatic scores underestimate RBMT systems, e.g., (Béchar et al., 2012).

To investigate applicability of the automatic measures to the English-Russian language direction, we computed Spearman's ρ correlation between the ranks given by the evaluators and by the respective measures. Because of the amount of variation for each measure on the sentence level, robust estimates, such as the median and the trimmed mean, are more informative than the mean, since they discard the outliers (Huber, 1996). The results are listed in Table 3. All measures exhibit reasonable correlation on the corpus level (330 sentences), but the sentence-level results are less impressive. While TER and GTM are known to provide better correlation with post-editing efforts for English (O'Brien, 2011), free word order and greater data sparseness on the sentence level makes TER much less reliable for Russian. METEOR (with its built-in Russian lemmatisation) and GTM offer the best correlation with human judgements.

The lower part of Table 2 also reports the results of simulated dynamic ranking (using the NIST rankings as the initial order for the sort operation). It resulted in a slightly different final ranking of the systems since we did not account for ties and 'averaged ranks'. However, the ranking is practically the same up to the statistically significant rank differences in reference ranking (see above). The advantage is that it requires a significantly lower number of pairwise comparisons. Insertion sort yielded 5,131 comparisons (15.5 per sentence; 56% of exhaustive comparisons for 330 sentences and 8 systems); binary insertion sort yielded 4,327 comparisons (13.1 per sentence; 47% of exhaustive comparisons).

Out of the original set of 330 sentences for human evaluation, 60 sentences were evaluated by two annotators (which resulted in 60*28=1680 pairwise comparisons), so we were able to calculate the standard Kohen's κ and Krippendorff's α scores (Artstein and Poesio, 2008). The results of inter-annotator agreement are: percentage agreement 0.56, $\kappa = 0.34, \alpha = 0.48$, which is simi-

Metric	Sentence level			Corpus level
	Median	Mean	Trimmed	
BLEU	0.357	0.298	0.348	0.833
NIST	0.357	0.291	0.347	0.810
Meteor	0.429	0.348	0.393	0.714
TER	0.214	0.186	0.204	0.619
GTM	0.429	0.340	0.392	0.714

Table 3: Correlation to human judgements

lar to sentence ranking reported in other evaluation campaigns (Callison-Burch et al., 2012; Callison-Burch et al., 2011). It was interesting to see the agreement results distinguishing the top three systems against the rest, i.e. by ignoring the assessments for the pairs within each group, $\alpha = 0.53$, which indicates that the judges agree on the difference in quality between the top three systems and the rest. On the other hand, the agreement results within the top three systems are low: $\kappa = 0.23, \alpha = 0.33$, which is again in line with the results for similar evaluations between closely performing systems (Callison-Burch et al., 2011).

5 Conclusions and future plans

This was the first attempt at making proper quantitative and qualitative evaluation of the English→Russian MT systems. In the future editions, we will be aiming at developing a new test corpus with a wider genre palette. We will probably complement the campaign with Russian→English translation direction. We hope to attract more participants, including international ones and plan to prepare a 'light version' for students and young researchers. We will also address the problem of tailoring automatic evaluation measures to Russian — accounting for complex morphology and free word order. To this end we will re-use human evaluation data gathered within the 2013 campaign. While the campaign was based exclusively on data in one language direction, the correlation results for automatic MT quality measures should be applicable to other languages with free word order and complex morphology.

We have made the corpus comprising the source sentences, their human translations, translations by participating MT systems and the human evaluation data publicly available.[8]

[8] http://romip.ru/mteval/

Acknowledgements

We would like to thank the translators, assessors, as well as Anna Tsygankova, Maxim Gubin, and Marina Nekrestyanova for project coordination and organisational help. Research on corpus preparation methods was supported by EU FP7 funding, contract No 251534 (HyghTra). Our special gratitude goes to Yandex and ABBYY who partially covered the expenses incurred on corpus translation. We're also grateful to the anonymous reviewers for their useful comments.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bogdan Babych, Anthony Hartley, Serge Sharoff, and Olga Mudraya. 2007. Assisting translators in indirect lexical transfer. In *Proc. of 45th ACL*, pages 739–746, Prague.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June.

Hanna Béchar, Raphaël Rubino, Yifan He, Yanjun Ma, and Josef van Genabith. 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In *Proceedings of COLING'12*, Mumbai.

Igor Boguslavsky. 1995. A bi-directional Russian-to-English machine translation system (ETAP-3). In *Proceedings of the Machine Translation Summit V*, Luxembourg.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 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, June.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology*, pages 138–145, San Diego, CA.

Marcelo Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stuker. 2012. Overview of the IWSLT 2012 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 12–34, Hong Kong, December.

Peter J. Huber. 1996. *Robust Statistical Procedures*. Society for Industrial and Applied Mathematics.

John Hutchins, editor. 2000. *Early years in machine translation: Memoirs and biographies of pioneers*. John Benjamins, Amsterdam, Philadelphia. `http://www.hutchinsweb.me.uk/EarlyYears-2000-TOC.htm`.

Philippe Langlais. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *Proceedings of Second international workshop on computational terminology (COMPUTERM 2002)*, pages 1–7, Taipei, Taiwan. `http://acl.ldc.upenn.edu/W/W02/W02-1405.pdf`.

Sharon O'Brien. 2011. Towards predicting post-editing productivity. *Machine translation*, 25(3):197–215.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Thomas J. Watson Research Center.

Serge Sharoff. 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 149–166. Springer, Berlin/New York.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, March.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of Machine Translation Summit IX*, New Orleans, LA, USA, September.

John S. White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. In *Proceedings of AMTA'94*, pages 193–205.