Applying HMEANT to English-Russian Translations

Alexander Chuchunkov **Alexander Tarelkin** **Irina Galinskaya**
{madfriend,newtover,galinskaya}@yandex-team.ru

Yandex LLC
Leo Tolstoy st. 16, Moscow, Russia

Abstract

In this paper we report the results of first experiments with HMEANT (a semi-automatic evaluation metric that assesses translation utility by matching semantic role fillers) on the Russian language. We developed a web-based annotation interface and with its help evaluated practicability of this metric in the MT research and development process. We studied reliability, language independence, labor cost and discriminatory power of HMEANT by evaluating English-Russian translation of several MT systems. Role labeling and alignment were done by two groups of annotators - with linguistic background and without it. Experimental results were not univocal and changed from very high inter-annotator agreement in role labeling to much lower values at role alignment stage, good correlation of HMEANT with human ranking at the system level significantly decreased at the sentence level. Analysis of experimental results and annotators' feedback suggests that HMEANT annotation guidelines need some adaptation for Russian.

1 Introduction

Measuring translation quality is one of the most important tasks in MT, its history began long ago but most of the currently used approaches and metrics have been developed during the last two decades. BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and METEOR (Banerjee and Lavie, 2005)metric require reference translation to compare it with MT output in fully automatic mode, which resulted in a dramatical speed-up for MT research and development. These metrics correlate with manual MT evaluation and provide reliable evaluation for many languages and for different types of MT systems.

However, the major problem of popular MT evaluation metrics is that they aim to capture lexical similarity of MT output and reference translation (fluency), but fail to evaluate the semantics of translation according to the semantics of reference (adequacy) (Lo and Wu, 2011a). An alternative approach that is worth mentioning is the one proposed by Snover et al. (2006), known as HTER, which measures the quality of machine translation in terms of post-editing. This method was proved to correlate well with human adequacy judgments, though it was not designed for a task of gisting. Moreover, HTER is not widely used in machine translation evaluation because of its high labor intensity.

A family of metrics called MEANT was proposed in 2011 (Lo and Wu, 2011a), which approaches MT evaluation differently: it measures how much of an event structure of reference does machine translation preserve, utilizing shallow semantic parsing (MEANT metric) or human annotation (HMEANT) as a gold standard.

We applied HMEANT to a new language — Russian — and evaluated the usefulness of metric. The practicability for the Russian language was studied with respect to the following criteria provided by Birch et al. (2013):

Reliability – measured as inter-annotator agreement for individual stages of evaluation task.

Discriminatory Power – the correlation of rankings of four MT systems (by manual evaluation, BLEU and HMEANT) measured on a sentence and test set levels.

Language Independence – we collected the problems with the original method and guidelines and compared these problems to those reported by Bojar and Wu (2012) and Birch et al. (2013).

Efficiency – we studied the labor cost of annotation task, i. e. average time required to evaluate

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 43–50,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

translations with HMEANT. Besides, we tested the statement that semantic role labeling (SRL) does not require experienced annotators (in our case, with linguistic background).

Although the problems of HMEANT were outlined before (by Bojar and Wu (2012) and Birch et al. (2013)) and several improvements were proposed, we decided to step back and conduct experiments with HMEANT in its original form. No changes to the metric, except for the annotation interface enhancements, were made.

This paper has the following structure. Section 2 reports the previous experiments with HMEANT; section 3 summarizes the methods behind HMEANT; section 4 – the settings for our own experiments; sections 5 and 6 are dedicated to results and discussion.

2 Related Work

Since the beginning of the machine translation era the idea of semantics-driven approach for translation wandered around in the MT researchers community (Weaver, 1955). Recent works by Lo and Wu (2011a) claim that this approach is still perspective. These works state that in order for machine translation to be useful, it should convey the shallow semantic structure of the reference translation.

2.1 MEANT for Chinese-English Translations

The original paper on MEANT (Lo and Wu, 2011a) proposes the semi-automatic metric, which evaluates machine translations utilizing annotated event structure of a sentence both in reference and machine translation. The basic assumption behind the metric can be stated as follows: translation shall be considered "good" if it preserves shallow semantic (predicate-argument) structure of reference. This structure is described in the paper on shallow semantic parsing (Pradhan et al., 2004): basically, we approach the evaluation by asking simple questions about events in the sentence: "*Who did what to whom, when, where, why and how?*". These structures are annotated and aligned between two translations. The authors of MEANT reported results of several experiments, which utilized both human annotation and semantic role labeling (as a gold standard) and automatic shallow semantic parsing. Experiments show that HMEANT correlates with human adequacy judgments (for three MT systems) at the value of 0.43 (Kendall tau, sentence level), which is very close to the correlation of HTER (BLEU has only 0.20). Also inter-annotator agreement was reported for two stages of annotation: role identification (selecting the word span) and role classification (labeling the word span with role). For the former, IAA ranged from 0.72 to 0.93 (which can be interpreted as a good agreement) and for the latter, from 0.69 to 0.88 (still quite good, but should be put in doubt). IAA for the alignment stage was not reported.

2.2 HMEANT for Czech-English Translations

MEANT and HMEANT metrics were adopted for an experiment on evaluation of Czech-English and English-Czech translations by Bojar and Wu (2012). These experiments were based on a human-evaluated set of 40 translations from WMT12[1], which were submitted by 13 systems; each system was evaluated by exactly one annotator, plus an extra annotator for reference translations. This setting implied that inter-annotator agreement could not be examined. HMEANT correlation with human assessments was reported as 0.28, which is significantly lower than the value obtained by Lo and Wu (2011a).

2.3 HMEANT for German-English Translations

Birch et al. (2013) examined HMEANT thoroughly with respect to four criteria, which address the usefulness of a task-based metric: reliability, efficiency, discriminatory power and language independence. The authors conducted an experiment to evaluate three MT systems: rule-based, phrase-based and syntax-based on a set of 214 sentences (142 German and 72 English). IAA was broken down into the different stages of annotation and alignment. The experimental results showed that whilst the IAA for HMEANT is satisfying at the first stages of the annotation, the compounding effect of disagreement at each stage (up to the alignment stage) greatly reduced the effective overall IAA — to 0.44 on role alignment for German, and, only slightly better, 0.59 for English. HMEANT successfully distinguished three types of systems, however, this result could not be considered reliable as IAA is not very high (and rank

[1] http://statmt.org/wmt12

correlation was not reported). The efficiency of HMEANT was stated as reasonably good; however, it was not compared to the labor cost of (for example) HTER. Finally, the language independence of the metric was implied by the fact that original guidelines can be applied both to English and German translations.

3 Methods

3.1 Evaluation with HMEANT

The underlying annotation cycle of HMEANT consists of two stages: semantic role labeling (SRL) and alignment. During the SRL stage, each annotator is asked to mark all the frames (a predicate and associated roles) in reference translation and hypothesis translation. To annotate a frame, one has to mark the frame head – predicate (which is a verb, but not a modal verb) and its arguments, role fillers, which are linked to that predicate. These role fillers are given a role from the inventory of 11 roles (Lo and Wu, 2011a). The role inventory is presented in Table 1, where each role corresponds to a specific question about the whole frame.

Who?	**What?**	**Whom?**
Agent	Patient	Benefactive
When?	**Where?**	**Why?**
Temporal	Locative	Purpose
How?		
Manner, Degree, Negation, Modal, Other		

Table 1. The role inventory.

On the second stage, the annotators are asked to align the elements of frames from reference and hypothesis translations. The annotators link both actions and roles, and these alignments can be matched as "Correct" or "Partially Correct" depending on how well the meaning was preserved. We have used the original minimalistic guidelines for the SRL and alignment provided by Lo and Wu (2011a) in English with a small set of Russian examples.

3.2 Calculating HMEANT

After the annotation, HMEANT score of the hypothesis translation can be calculated as the F-score from the counts of matches of predicates and their role fillers (Lo and Wu, 2011a). Predicates (and roles) without matches are not accounted, but they result in the lower value overall. We have used the uniform model of HMEANT, which is defined as follows.

$\#F_i$ – number of correct role fillers for predicate i in machine translation;

$\#F_i(partial)$ – number of partially correct role fillers for predicate i in MT;

$\#MT_i$, $\#REF_i$ – total number of role fillers in MT or reference for predicate i;

N_{mt}, N_{ref} – total number of predicates in MT or reference;

w – weight of the partial match (0.5 in the uniform model).

$$P = \sum_{matched\ i} \frac{\#F_i}{\#MT_i} \quad R = \sum_{matched\ i} \frac{\#F_i}{\#REF_i}$$

$$P_{part} = \sum_{matched\ i} \frac{\#F_i(partial)}{\#MT_i}$$

$$R_{part} = \sum_{matched\ i} \frac{\#F_i(partial)}{\#REF_i}$$

$$P_{total} = \frac{P + w * P_{part}}{N_{mt}} \quad R_{total} = \frac{R + w * R_{part}}{N_{ref}}$$

$$HMEANT = \frac{2 * P_{total} * R_{total}}{P_{total} + R_{total}}$$

3.3 Inter-Annotator Agreement

Like Lo and Wu (2011a) and Birch et al. (2013) we studied inter-annotator agreement (IAA). It is defined as an F1-measure, for which we consider one of the annotators as a gold standard:

$$IAA = \frac{2 * P * R}{P + R}$$

Where precision (P) is the number of labels (roles, predicates or alignments) that match between annotators divided by the total number of labels by annotator 1; recall (R) is the number of matching labels divided by the total number of labels by annotator 2. Following Birch et al. (2013), we consider only exact word span matches. Also we have adopted the individual stages of the annotation procedure that are described in (Birch et al. 2013): *role identification* (selecting the word span), *role classification* (marking the word span with a role), *action identification* (marking the word span as a predicate), *role alignment* (linking roles between translations) and *action alignment* (linking frame heads). Calculating IAA for each stage separately helped to isolate the disagreements and to see, which stages resulted in a low agreement value overall. To look at the most common role disagreements we also created the pairwise agreement matrix, every cell (i, j) of which is the number of times the role i was confused with the role j by any pair of annotators.

3.4 Kendall's Tau Rank Correlation With Human Judgments

For the set of translations used in our experiments, we had a number of relative human judgments (the set was taken from WMT13[2]). We used the rank aggregation method described in (Callison-Burch et al., 2012) to build up one ranking from these judgments. This method is called *Expected Win Score (EWS)* and for MT system S_i from the set $\{S_j\}$ it is defined the following way:

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

Where $win(S_i, S_j)$ is the number of times system i was given a rank higher than system j. This method of aggregation was used to obtain the comparisons of systems, which outputs were never presented together to assessors during the evaluation procedure at WMT13.

After we had obtained the ranking of systems by human judgments, we compared this ranking to the ranking by HMEANT values of machine translations. To do that, we used Kendall's tau (Kendall, 1938) rank correlation coefficient and reported the results as Lo and Wu (2011a) and Bojar (Bojar and Wu, 2012).

4 Experimental Setup

4.1 Test Set

For our experiments we used the set of translations from WMT13. We tested HMEANT on a set of four best MT systems (Bojar et al., 2013) for the English-Russian language pair (Table 2).

From the set of direct English-Russian translations (500 sentences) we picked those which allowed to build a ranking for the four systems (94 sentences); then out of these we randomly picked 50 and split them into 6 tasks of 25 so that each of the 50 sentences was present in exactly three tasks. Each task consisted of 25 reference translations and 100 hypothesis translations.

System	EWS (WMT)
PROMT	0.4949
Online-G	0.475
Online-B	0.3898
CMU-Primary	0.3612

Table 2. The top four MT systems for the en-ru translation task at WMT13. The scores were calculated for the subset of translations which we used in experiments.

4.2 Annotation Interface

As far as we know there is no publically available interface for HMEANT annotation. Thus, first of all, having the prototype (Lo and Wu, 2011b) and taking into account comments and suggestions of Bojar and Wu (2012) (e.g., ability to go back within the phases of annotation), we created a web-based interface for role labeling and alignment. This interface allows to annotate a set of references with one machine translation at a time (Figure 1) and to align actions and roles. We also provided a timer which allowed to measure the time required to label the predicates and roles.

4.3 Annotators

We asked to participate two groups of annotators: 6 researchers with linguistic background (linguists) and 4 developers without it. Every annotator did exactly one task; each of the 50 sentences was annotated by three linguists and at least two developers.

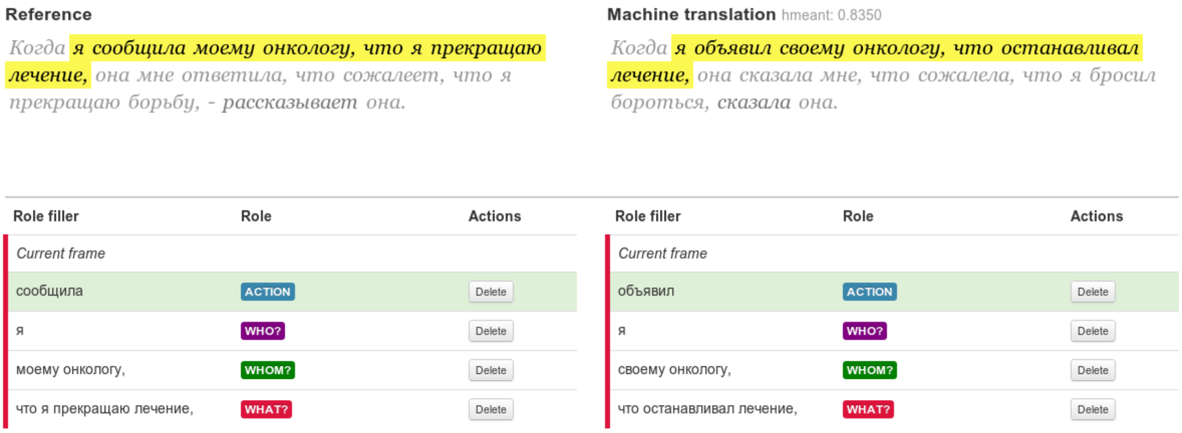
5 Results

As a result of the experiment, 638 frames were annotated in reference translations (overall) and 2 016 frames in machine translations. More detailed annotation statistics are presented in Table 3. A closer look indicates that the ratio of aligned frames and roles in references was larger than in any of machine translations.

5.1 Manual Ranking

After the test set was annotated, we compared manual ranking and ranking by HMEANT; on the system level, these rankings were similar; however, on the sentence level, there was no correlation between rankings at all. Thus we decided to take a closer look at the manual assessments. For the selected 4 systems most of the pairwise com-

[2] http://statmt.org/wmt13



Reference

*Когда **я сообщила моему онкологу, что я прекращаю лечение,** она мне ответила, что сожалеет, что я прекращаю борьбу, - **рассказывает** она.*

Role filler	Role	Actions
Current frame		
сообщила	ACTION	Delete
я	WHO?	Delete
моему онкологу,	WHOM?	Delete
что я прекращаю лечение,	WHAT?	Delete

Machine translation hmeant: 0.8350

*Когда **я объявил своему онкологу, что останавливал лечение,** она сказала мне, что сожалела, что я бросил бороться, **сказала** она.*

Role filler	Role	Actions
Current frame		
объявил	ACTION	Delete
я	WHO?	Delete
своему онкологу,	WHOM?	Delete
что останавливал лечение,	WHAT?	Delete

Figure 1. The screenshot of SRL interface. The tables under the sentences contain the information about frames (the active frame has a red border and is highlighted in the sentence, inactive frames (not shown) are semi-transparent).

Source	# Frames	# Roles	Aligned frames, %	Aligned roles, %
Reference	638	1 671	86.21 %	74.15 %
PROMT	609	1 511	79.97 %	67.57 %
Online-G	499	1 318	77.96 %	66.46 %
Online-B	469	1 257	78.04 %	68.42 %
CMU-Primary	439	1 169	75.17 %	66.30 %

Table 3. Annotation statistics.

System	Manual	HMEANT	BLEU
PROMT	0.532	0.443	0.126
Online-G	0.395	0.390	0.146
Online-B	0.306	0.374	0.147
CMU-Primary	0.267	0.292	0.136

Table 4. EWS over manual assessments, EWS over HMEANT and BLEU scores for MT systems.

parisons were obtained in a transitive way, i. e. using comparisons with other systems. Furthermore, we encountered a number of useless rankings, where all the outputs were given the same rank. After all, for many sentences the ranking of systems was based on a few pairwise comparisons provided by one or two annotators. These rankings seemed to be not very reliable, thus we decided to rank four machine translations for each of the 50 sentences manually to make sure that the ranking has a strong ground. We asked 6 linguists to do that task. The average pairwise rank correlation (between assessors) reached 0.77, making the overall ranking reliable; we aggregated 6 rankings for each sentence using EWS.

5.2 Correlation with Manual Assessments

To look at HMEANT on a system level, we compared rankings produced during manual assessment and HMEANT annotation tasks. Those rankings were then aggregated with EWS (Table 4).

It should be noticed that HMEANT allowed to rank systems correctly. This fact indicates that HMEANT has a good discriminatory power on the level of systems, which is a decent argument for the usage of this metric. Also it is worth to note that ranking by HMEANT matched the ranking by the number of frames and roles (Table 3).

On a sentence level, we studied the rank correlation of ranking by manual assessments and by HMEANT values for each of the annotators. The manual ranking was aggregated by EWS from the manual evaluation task (see Section 5.1). Results are reported in Table 5.

We see that resulting correlation values are significantly lower than those reported by Lo and Wu (2011a) – our rank correlation values did not reach 0.43 on average across all the annotators (and even 0.28 as reported by Bojar and Wu (2012)).

Annotator	τ
Linguist 1	0.0973
Linguist 2	**0.3845**
Linguist 3	0.1157
Linguist 4	-0.0302
Linguist 5	**0.1547**
Linguist 6	0.1468
Developer 1	**0.1794**
Developer 2	**0.2411**
Developer 3	0.1279
Developer 4	0.1726

Table 5. The rank correlation coefficients for HMEANT and human judgments. Reliable results (with p-value >0.05) are in bold.

5.3 Inter-Annotator Agreement

Following Lo and Wu (2011a) and Birch et al. (2013) we report the IAA for the individual stages of annotation and alignment. These results are shown in Table 6.

Stage	**Linguists**		**Developers**	
	Max	**Avg**	**Max**	**Avg**
REF, id	0.959	0.803	0.778	0.582
MT, id	0.956	0.795	0.667	0.501
REF, class	0.862	0.715	0.574	0.466
MT, class	0.881	0.721	0.525	0.434
REF, actions	0.979	0.821	0.917	0.650
MT, actions	0.971	0.839	0.700	0.577
Actions – align	0.908	0.737	0.429	0.332
Roles – align	0.709	0.523	0.378	0.266

Table 6. The inter-annotator agreement for the individual stages of annotation and alignment procedures. Id, class, align stand for identification, classification and alignment respectively.

The results are not very different from those reported in the papers mentioned above, except for even lower agreement for developers. The fact that the results could be reproduced on a new language seems very promising, however, the lack of training for the annotators without linguistic background resulted in lower inter-annotator agreement.

Also we studied the most common role disagreements for each pair of annotators (either linguists or developers). As it can be deduced from the IAA values, the agreement on all roles is lower for linguists, however, both groups of annotators share the roles on which the agreement is best of all: *Predicate, Agent, Locative, Negation, Temporal*. Most common disagreements are presented in Table 7.

Role A	**Role B**	**%, L**	**%, D**
Whom	What	18.0	15.2
Whom	Who	13.7	23.1
Why	None	17.0	22.3
How (manner)	What	10.5	-
How (manner)	How (degree)	-	19.0
How (modal)	Action	18.1	16.3

Table 7. Most common role disagreements. Last columns (L for linguists, D for developers) stand for the ratio of times Role A was confused with Role B across all the label types (roles, predicate, none).

These disagreements can be explained by the fact that some annotators looked "deeper" in the sentence semantics, whereas other annotators only tried to capture the shallow structure as fast as possible. This fact explains, for example, disagreement on the *Whom* role – for some sentences, e. g. "*могли бы убедить политических лидеров*" ("*could persuade the political leaders*") it requires some time to correctly mark *политических лидеров* (*political leaders*) as an answer to *Whom*, not *What*. The disagreement on the *Purpose* (a lot of times it was annotated only by one expert) is explained by the fact that there were no clear instructions on how to mark clauses. As for the *Action* and *Modal*, this disagreement is based on the requirement that *Action* should consist of one word only; this requirement raised questions about complex verbs, e.g. "*закончил делать*" ("*stopped doing*"). It is ambiguous how to annotate these verbs: some annotators decided to mark it as *Modal*+*Action*, some – as *Action*+*What*. Probably, the correct way to mark it should be just as *Action*.

5.4 Efficiency

Additionnaly, we conducted an efficiency experiment in the group of linguists. We measured the average time required to annotate a predicate (in reference or machine translation) and a role. Results are presented in Table 8.

Annotator	REF		MT	
	Role	Action	Role	Action
Linguist 1	14	26	11	36
Linguist 2	10	12	8	12
Linguist 3	13	14	8	23
Linguist 4	16	15	9	15
Linguist 5	13	20	11	24
Linguist 6	17	35	9	32

Table 8. Average times (in seconds) required to annotate actions and roles.

These results look very promising; using the numbers in Table 3, we get the average time required to annotate a sentence: 1.5 – 2 minutes for a reference (and even up to 4 minutes for slower linguists) and 1.5 – 2.5 minutes for a machine translation. Also for a group of "slower" linguists (1, 5, 6) inter-annotator agreement was lower (-0.05 on average) than between "faster" linguists (2, 3, 4) for all stages of annotation and alignment. Average time to annotate an action is similar for the reference and MT outputs, but it takes more time to annotate roles in references than in machine translations.

6 Discussion

6.1 Problems with HMEANT

As we can see, HMEANT is an acceptably reliable and efficient metric. However, we have met some obstacles and problems with original instructions during the experiments with Russian translations. We believe that these obstacles are the main causes of low inter-annotator agreement at the last stages of annotation procedure and low correlation of rankings.

Frame head (predicate) is required. This requirement does not allow frames without predicate at all, e.g. "*Он мой друг*" ("*He is my friend*") – the Russian translation of "*is*" (present tense) is a null verb.

One-word predicates. There are cases where complex verbs (e.g., which consist of two verbs) can be correctly translated as a one-word verb. For example, "*остановился*" ("*stopped*") is correctly rephrased as "*перестал делать*" ("*ceased doing*").

Roles only of one type can be aligned. Sometimes one role can be correctly rephrased as another role, but roles of different type can not be aligned. For example, "*Он уехал из города*" ("*He went away from the town*") means the same as "*Он покинул город*" ("*He left the town*"). The former has a structure of *Who* + *Action* + *Where*, the latter – *Who* + *Action* + *What*.

Should we annotate as much as possible? It is not clear from the guideline whether we should annotate almost everything that looks like a frame or can be interpreted as a role. There are some prepositional phrases which can not be easily classified as one role or another. Example: "*Нам не стоит об этом волноваться*" ("*We should not worry about this*") – it is not clarified how to deal with "*об этом*" ("*about this*") prepositional phrase.

7 Conclusion

In this paper we describe a preliminary series of experiments with HMEANT, a new metric for semantic role labeling. In order to conduct these experiments we developed a special web-based annotation interface with a timing feature. A team of 6 linguists and 4 developers annotated Russian MT output of 4 systems. The test set of 50 English sentences along with reference translations was taken from the WMT13 data. We measured IAA for each stage of annotation process, compared HMEANT ranking with manual assessment and calculated the correlation between HMEANT and manual evaluation. We also measured annotation time and collected a feedback from annotators, which helped us to locate the problems and better understand the SRL process. Analysis of the preliminary experimental results of Russian MT output annotation led us to the following conclusions about HMEANT as a metric.

Language Independence. For a relatively small set of Russian sentences, we encountered problems with the guidelines, but they were not specific to the Russian language. This can be naively interpreted as language independence of the metric.

Reliability. Inter-annotator agreement is high for the first stages of SRL, but we noted that it decreases on the last stages because of the compound effect of disagreements on previous stages.

Efficiency. HMEANT proved to be really effective in terms of time required to annotate references and MT outputs and can be used in production environment, though the statement that HMEANT annotation task does not require qualified annotators was not confirmed.

Discriminatory Power. On the system level, HMEANT allowed to correctly rank MT systems (according to the results of manual assessment task). On the sentence level, correlation with human rankings is low.

To sum up, first experience with HMEANT was considered to be successful and allowed us to make a positive decision about applicability of the new metric to the evaluation of English-Russian machine translations. We have to say that HMEANT guidelines, annotation procedures and the inventory of roles work in general, however, low inter-annotator agreement at the last stages of annotation task and low correlation with human judgments on the sentence level suggest us to make respective adaptations and conduct new series of experiments.

Acknowledgements

We would like to thank our annotators for their efforts and constructive feedback. We also wish to express our great appreciation to Alexey Baytin and Maria Shmatova for valuable ideas and advice.

References

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.

Alexandra Birch, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. 2013. The Feasibility of HMEANT as a Human MT Evaluation Metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, page 52–61, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondrej Bojar and Dekai Wu. 2012. Towards a Predicate-Argument Evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, page 30–38, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, page 1–44, Sofia, Bulgaria, August. 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*, page 10–51, Montréal, Canada, June. Association for Computational Linguistics.

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 Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, pages 81–93.

Chi-kiu Lo and Dekai Wu. 2011a. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 220–229, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2011b. A radically simple, effective annotation and alignment methodology for semantic frame based smt and mt evaluation. *LIHMT 2011*, page 58.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sameer S Pradhan, Wayne Ward, Kadri Hacioglu, James H Martin, and Daniel Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *HLT-NAACL*, pages 233–240.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Warren Weaver. 1955. Translation. *Machine translation of languages*, 14:15–23.