Computer Science • 18(2) 2017 http://dx.doi.org/10.7494/csci.2017.18.2.129

Krzysztof Wołk
Danijel Koržinek

COMPARISON AND ADAPTATION OF AUTOMATIC EVALUATION METRICS FOR QUALITY ASSESSMENT OF RE-SPEAKING

Abstract *Re-speaking is a mechanism for obtaining high-quality subtitles for use in live broadcasts and other public events. Because it relies on humans to perform the actual re-speaking, the task of estimating the quality of the results is non-trivial. Most organizations rely on human effort to perform the actual quality assessment, but purely automatic methods have been developed for other similar problems (like Machine Translation). This paper will try to compare several of these methods: BLEU, EBLEU, NIST, METEOR, METEOR-PL, TER, and RIBES. These will then be matched to the human-derived NER metric, commonly used in re-speaking. The purpose of this paper is to assess whether the above automatic metrics normally used for MT system evaluation can be used in lieu of the manual NER metric to evaluate re-speaking transcripts.*

Keywords speech, re-speaking, machine translation, evaluation

Citation Computer Science 18(2) 2017: 129–144

1. Introduction

Over the last several years, one of the main driving forces in speech technology has come from the efforts of various groups and organizations trying to tackle the concerns of the disabled; specifically, deaf and hard-of-hearing people. Most notably, a long-term effort by such organizations has lead to a plan by the European Commission to enable "Subtitling of 100% of programs in public TV all over the EU by 2020, with simple technical standards and consumer-friendly rules" [6]. This ambitious task would not be possible to achieve without the aid of speech technology. While there has recently been a considerable improvement in the quality of Automatic Speech Recognition (ASR) technology, many of the tasks present in real-life are simply beyond complete automation. On the other hand, there are tasks that are also impossible to achieve by humans without the aid of ASR. For example, movie subtitles are usually done by human transcribers and can take an entire day (or up to a week) to complete one film. Live subtitling, however, can sustain only a few-seconds delay between the time an event is recorded and the time it appears on the viewer's screen. This is where re-speaking comes into play.

The idea of re-speaking is to use ASR to create live and fully annotated subtitles; but rather than risking misrecognition of the speech happening in the live recording, a specially trained individual (the so-called re-speaker) repeats the speech from the recorded event in a quiet and controlled environment. This approach has many advantages that guarantee excellent results: a controlled environment, the ability to adapt the speaker, the ability of the speaker to adapt and learn to better use the software, solving problems like double-speak, the cocktail-party effect, and non-grammatical speech by paraphrasing. The standard of quality required by many jurisdictions demands fewer than 2% of errors [19]. From the point of view of a re-speaker, this problem is very similar to that of an interpreter; only instead of translating from one language to another, re-speaking is usually done within one language only. There are many aspects of re-speaking worthy of their own insight [5], but this paper will deal only with the issue of quality assessment.

Similarly to Machine Translation (MT), assessment of the accuracy of re-speaking is not a trivial task because there are many possible ways to paraphrase an utterance, just like there are many ways to translate any given sentence. Measuring the accuracy of such data has to take semantic meaning into account rather than blindly performing a simple word-to-word comparison.

One option is to use humans to perform this evaluation, as in the NER model described later in this paper. This has been recognized as very expensive and time-consuming [9]. As a result, human effort cannot keep up with the growing and continual need for evaluation. This has led to the recognition that the development of automated evaluation techniques is critical [9, 18].

Unfortunately, most of the automatic evaluation metrics were developed for other purposes than re-speaking (mostly for machine translation) and are not suited for languages that differ semantically and structurally from English (like Polish). The

Polish language has complex declension, 7 cases, 15 gender forms, and complicated grammatical construction. This leads to larger vocabulary and greater complexity in data requirements for such tasks. Unlike Polish, English does not have declensions. In addition, word order (esp. the Subject-Verb-Object (SVO) pattern) is absolutely crucial to determining the meaning of an English sentence [22]. While these automatic metrics have already been thoroughly studied, we feel that there is still much to be learned, especially in different languages and for different tasks (like re-speaking).

This paper will compare some of these automated and human-assisted metrics while also considering issues that are related specifically to Polish. To our knowledge, these metrics used for evaluating MT systems were never utilized in the context of re-speaking. A small introduction to all of the metrics is presented in the beginning of the paper, followed by an experiment using actual re-speaking data.

2. Machine translation metrics

This chapter will describe many of the state-of-the-art metrics for machine translation. Most of them were tested using software that is available online. The Meteor metric was also adapted to Polish for the purpose of this paper. This is described in detail in section 2.5. The EBLEU metric, described in section 2.7, is an extension of the popular BLEU metric; it was created by the first author of this paper as outlined in the cited references.

2.1. BLEU metric

BLEU was developed based on a premise similar to that used for speech recognition, described in [16] as: "The closer a machine translation is to a professional human translation, the better it is." So, the BLEU metric is designed to measure how close SMT output is to that of human reference translations. It is important to note that translations (SMT or human) may differ significantly in word usage, word order, and phrase length [16]. To address these complexities, BLEU attempts to match variable-length phrases between SMT output and reference translations. Weighted match averages are used to determine the translation score [1]. A number of variations of the BLEU metric exist; however, the basic metric requires the calculation of a brevity penalty (P_B), which is calculated as follows:

$$P_B = \begin{cases} 1 & c > r \\ e^{(1-r/c)} & c <= r \end{cases} \tag{1}$$

where r is the length of the reference corpus, and the candidate (reference) translation length is given by c. [1] The basic BLEU metric is then determined as shown in [1]:

$$\text{BLEU} = P_B \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

where w_n are positive weights summing to one, and n-gram precision p_n is calculated using n-grams with a maximum length of N.

There are several other important features of BLEU. First, word and phrase position within the text are not evaluated by this metric. To prevent SMT systems from artificially inflating their scores by overusing words known with high confidence, each candidate word is constrained by the word count of the corresponding reference translation. A geometric mean of individual sentence scores is then calculated for the entire corpus (with consideration of the brevity penalty) [1].

2.2. NIST metric

The NIST metric was designed to improve BLEU by rewarding the translation of infrequently used words. This was intended to further prevent the inflation of SMT evaluation scores by focusing on common words and high-confidence translations. As a result, the NIST metric uses heavier weights for rarer words. The final NIST score is calculated using the arithmetic mean of the n-gram matches between SMT and the reference translations. In addition, a smaller brevity penalty is used for smaller variations in phrase lengths. The reliability and quality of the NIST metric has been shown to be superior to the BLEU metric in many cases [4].

2.3. Translation Edit Rate (TER)

The Translation Edit Rate (TER) was designed to provide a very intuitive SMT evaluation metric, requiring less data than other techniques while avoiding the labor intensity of human evaluation. It calculates the number of edits required to make a machine translation match exactly to the closest reference translation in fluency and semantics [2, 12]. Calculation of the TER metric is defined in [12]:

$$\mathrm{TER} = \frac{E}{R} \tag{3}$$

where E represents the minimum number of edits required for an exact match, and the average length of the reference text is given by R. Edits may include the deletion of words, word insertion, word substitutions, and changes in word or phrase order [12].

2.4. METEOR metric

The Metric for the Evaluation of Translation with Explicit Ordering (METEOR) is intended to take several factors that are indirect in BLEU into account more directly. Recall (the proportion of matched n-grams to total reference n-grams) is used directly in this metric. In addition, METEOR explicitly measures higher-order n-grams, considers word-to-word matches, and applies arithmetic averaging for a final score. The best matches against multiple reference translations can also be used [2].

The METEOR method uses a sophisticated and incremental word alignment method that starts by considering exact word-to-word matches, word stem matches,

and synonym matches. Alternative word-order similarities are then evaluated based on these matches.

The calculation of precision is similar in the METEOR and NIST metrics. Recall is calculated at the word level. To combine the precision and recall scores, METEOR uses a harmonic mean. METEOR rewards longer n-gram matches [2].

The METEOR metric is calculated as shown in [2]:

$$\text{METEOR} = F\text{mean} * (1 - Penalty) = \left(\frac{10 \cdot P \cdot R}{P + R}\right)(1 - P_M) \tag{4}$$

where the unigram recall and precision are given by R and P, respectively. Brevity penalty P_M is determined by:

$$P_M = 0.5 \cdot \left(\frac{C}{M_U}\right) \tag{5}$$

where M_U is the number of matching unigrams, and C is the minimum number of phrases required to match unigrams in the SMT output to those found in the reference translations.

An important factor in METEOR is the liberal use of weights for the numerous aspects of the system. These weights need to be tuned for specific tasks to match human judgment precisely. The tool comes with a couple of pre-tuned parameter sets for some common task and language pairs, but these obviously include neither re-speaking nor Polish.

2.5. METEOR-PL

While often used in a language-independent manner, the greatest advantage of METEOR is its ability to model the features of a specific language, like the aforementioned synonyms, stems, and paraphrasing. These features are enabled through the use of standard language tools, easily obtainable for many languages. In order to adapt METEOR to Polish, several steps needed to be made.

The synonym matcher uses a special script for extracting relevant information from the Princeton WordNet [15] project. In Polish, there is an equivalent project developed at the Wroclaw University of Technology [14], and it works exactly the same as the Princeton original.

The standard METEOR stemmer is implemented using the Snowball [17] tool, but this only supports a limited set of languages. Other languages can be implemented by hand-crafting rules using a special finite grammar. For Polish, Morfologik [21] (a well-known morphological analyzer and stemmer developed at the IPI PAN) was used instead. This meant that the METEOR source code needed to be slightly modified, specifically to support multiple stems per word.

One final modification of the system with respect to Polish was the creation of a list of function words. This list isn't very long (this is true for other languages as

well) and may have to be adjusted for specific uses. In its current state, it contains punctuation, some common abbreviations, and common conjunctions.

The last feature of METEOR is its ability to model paraphrasing. This uses a system that is trained on a parallel corpus. At the moment of writing this paper, the amount of re-speaking data was so small that no such corpus could be readily produced. This is something that could significantly improve the measure in the future, however.

2.6. RIBES

The focus of the RIBES metric is word order. It uses rank correlation coefficients based on word order to compare SMT and reference translations. The primary rank correlation coefficients used are Spearman's ρ, which measures the distance of differences in rank, and Kendall's τ, which measures the direction of differences in rank [10].

These rank measures can be normalized to ensure positive values [10]:

- normalized Spearman's ρ (NSR) $= (\rho + 1)/2$,
- normalized Kendall's τ (NKT) $= (\tau + 1)/2$.

These measures can be combined with precision P and modified to avoid overestimating the correlation of only corresponding words in the SMT and reference translations:

$$\text{NSR}P\alpha \text{ and } \text{NKT}P\alpha \tag{6}$$

where α is a parameter in the range $0 < \alpha < 1$.

2.7. EBLEU

We now discuss the enhancements to the BLEU metric that we introduced in [23]. The goal was to make this metric more reliable when it comes to morphologically rich languages like Polish. In particular, our enhanced metric rewards synonyms and rare-word translations while modifying the calculation of cumulative scores.

2.7.1. Consideration of synonyms

In our enhanced metric, we would like to reward matches of synonyms, since the correct meaning is still conveyed.

Consider this test phrase: "this is an exam"; and this reference phrase: "this is a quiz".

The BLEU score is calculated as follows:

$$BLEU = (1 + 1 + 1 + 0)/4 = 3/4 = 0.75 \tag{7}$$

BLEU does not count the word "exam" as a match, because it does not find it in the reference phrase. However, this word is not a bad choice. In our method, we want to score the synonym "exam" higher than zero and lower than the exact word "quiz".

During the BLEU evaluation, we check each word for an exact match. If the word is a synonym and not an exact match, we do not give a full score to that word. The score for a synonym will be the default BLEU score for an original word multiplied by a constant (synonym-score).

For example, if this constant equals 0.90, the new score with synonyms is:

$$(1 + 1 + 1 + 0.9)/4 = 3.9/4 = 0.975 \tag{8}$$

With this algorithm, we have synonym scores for all n-grams, because we have "a quiz" in 2-gram and "is a quiz" in 3-gram in both the test and reference phrases. The score values and other details were determined empirically; information about this can be found in [23].

2.7.2. Consideration of rare words

Our algorithm gives extra points to rare word matches. First, it obtains the rare words found in the reference corpus. If we sort all distinct words of the reference with their repetition order (descending), the last words in this list are rare words. The algorithm takes a specific percentage of the whole sorted list as rare words (rare-word-percent).

When the default BLEU algorithm tries to score a word, the score is multiplied by a constant (rare-word-score) if this word is on the rare word list. This action applies to all n-grams. So, if we have a rare word in a 2-gram, the algorithm increases the score for this 2-gram. For example, if the word "Roman" is rare, the "Roman Empire" 2-gram gets an increased score. The algorithm is careful that the score for each sentence falls within a range of 0.0 to 1.0.

2.7.3. Determination of cumulative score

The cumulative score of our algorithm combines the default BLEU scores using logarithms and exponentials as follows: where B_i is the default BLEU score and C_i is

Initialize $s = 0$;
for *each* i^{th}*-gram:* **do**
  $s = s + \log(B_i)$;
  $C_i = \exp(s/i)$;
end

the cumulative score. In addition, we know that:

$$\exp(\log(a) + \log(b)) = a \cdot b \text{ and } \exp(\log(a)/b) = a^{(\frac{1}{b})} \tag{9}$$

3. NER subtitle accuracy model

The NER model [19] is a simple extension of the word-accuracy metric adapted specifically for measuring the accuracy of subtitles. It is one of two measures that is of particular importance for media providers (television companies, movie distributors, etc); the other one being the reduction rate. Generally, the aim of good subtitles is to

reduce the length of the written text as much as possible (in order to preserve space on the screen and make it easier to read) while maintaining almost-perfect accuracy (usually above 98%).

Since we are dealing with paraphrasing, it is very difficult to perform accurate measurements by only comparing the text. The NER model gets around this problem by counting errors using a simple formula (which inspired its name):

$$\text{NER accuracy} = \frac{N - E - R}{N} \cdot 100\% \tag{10}$$

where N is the number of analyzed tokens (usually also includes punctuation), E is the number of errors performed by the re-speaker, and R is the number of errors performed by the ASR system (on the re-speaker's speech). Additionally, the errors in E are weighted: 0.25 for minor errors, 0.5 for normal, and 1 for serious ones. There are user-friendly tools available for making these assessments, so obviously there may be a certain level of human bias involved in these estimates. Nevertheless, all of the decisions are thoroughly checked and explainable using this method, which makes it one of the most-popular techniques for the subtitle quality assessment used by many regulatory bodies worldwide.

It is also worth noting that the primary goal of NER was to create a measure that would correlate with user perception of subtitle quality. Many studies in various languages confirm this [5] by finding a high correlation with user experience assessment derived from questionnaires. Another important aspect is the consistency and reproducibility of the measure, which has also been confirmed in separate studies [19].

4. Dataset and experiment setup

The data used in the experiments described in this paper was collected during a study performed in the Institute of Applied Linguistics at the University of Warsaw [5]. This ongoing study aims to determine the relevant features of good re-speakers and their techniques. One of the obvious measures is naturally the quality of re-speaking discussed in this paper.

The subjects were studied in several hour-long sessions, where they had to perform various tasks and psychological exams. The final test was to do an actual re-speaking of pre-recorded material. This was simultaneously recorded and recognized by a commercial, off-the-shelf ASR suite. The software was moderately adapted to each re-speaker during a several-hour session that was held a few weeks before the test.

The selection of the participants was performed in such a way as to fairly represent both the professional and less-experienced re-speakers. At a minimum, all of the participants took part in a short, couple of hour training spanning 2–3 days in order to learn the basics and most important rules of this profession (compared to the several month training taken by professionals). The main goal of this selection is related to the project mentioned above; more details can be learned from there. This

paper, however, makes no distinction between the specific re-speaking sessions and treats them all equally.

The complete set of materials included four different 5-minute segments in the speaker's native language and one in English (where the task was also a translation). The recordings were additionally transcribed by a human to convey exactly what the person said. The final dataset contains three sets of transcriptions:

- the transcription of the original recorded material,
- the transcription of the re-speaker transcribed by a human,
- the output of the ASR recognizing the re-speaker.

5. Result comparison

In our experiments, we used 57 transcripts prepared using the protocol above. Each transcript was evaluated with all of the metrics described in this paper as well as manually using the NER metric. Table 1 presents an evaluation between human-made text transcription and the original texts. Table 2 presents an evaluation between the ASR system and original texts.

Table 1

Evaluation of human-made text transcription and original text. RED stands for reduction rate and MTR for METEOR. For brevity, only the first ten and last ten transcripts are shown.

BLEU	NIST	TER	MTR	MTR-PL	EBLEU	RIBES	NER	RED.
56.20	6.90	29.10	79.62	67.07	62.32	85.59	92.38	10.15
56.58	6.42	30.96	78.38	67.44	58.82	86.13	94.86	17.77
71.56	7.86	18.27	88.28	76.19	79.58	92.48	94.71	12.01
76.64	8.27	13.03	90.34	79.29	87.38	93.07	93.1	3.72
34.95	5.32	44.50	61.86	47.74	37.06	71.60	91.41	17.03
61.73	7.53	20.47	83.10	72.55	69.11	92.43	92.33	4.89
61.74	6.93	28.26	78.26	69.78	63.99	78.32	95.3	10.29
33.52	4.28	46.02	63.06	47.61	36.75	77.55	93.95	26.81
68.97	7.46	22.50	83.15	76.56	71.83	88.78	94.73	4.05
70.02	7.80	18.78	86.12	78.71	75.16	88.15	95.23	6.41
...	...	...	...	...	...	...	...	...
40.73	4.93	41.96	68.61	54.34	42.96	83.09	89.97	21.15
29.03	2.65	51.27	61.24	47.21	39.53	67.21	86.07	31.98
68.75	7.78	18.78	86.24	77.18	72.40	90.65	94.95	5.41
75.24	7.97	16.07	88.88	81.51	81.27	90.88	95.75	7.45
78.71	8.24	11.51	91.33	83.99	86.08	94.77	98.69	4.23
37.60	4.31	44.84	66.32	51.59	42.22	78.73	89.37	−6.6
73.20	8.07	14.38	88.78	79.55	77.39	93.93	93.73	2.2
67.43	7.67	20.30	85.64	75.90	70.28	87.57	94.91	12.07
70.06	7.90	18.44	87.29	76.67	78.93	90.79	93.49	7.11
71.88	7.83	17.77	88.24	78.07	77.51	89.21	97.74	8.46
80.05	8.3	11.00	91.81	85.72	83.94	95.03	96.12	2.88

Table 2
Evaluation between ASR and original text. This was only performed for the first 20 re-speakers.

BLEU	NIST	TER	MTR	MTR-PL	EBLEU	RIBES	NER	RED.
41.89	6.05	44.33	66.10	54.05	44.77	78.94	92.38	10.15
48.94	5.94	37.39	71.14	60.24	49.79	81.29	94.86	17.77
57.38	7.11	27.24	78.41	67.08	62.87	89.42	94.71	12.01
59.15	7.07	27.24	77.21	67.94	65.31	87.71	93.1	3.72
26.08	4.57	55.33	52.33	39.14	26.89	69.22	91.41	17.03
44.17	6.32	36.38	69.16	60.15	47.97	86.04	92.33	4.89
51.79	6.39	34.86	71.42	65.47	52.31	79.19	95.3	10.29
22.03	3.17	61.93	45.27	33.90	22.14	59.93	93.95	26.81
52.35	6.09	39.93	68.02	63.07	53.87	78.53	94.73	4.05
54.44	6.65	33.50	73.16	65.42	57.28	82.11	95.23	6.41
65.95	7.57	19.63	84.68	76.30	72.76	92.45	97.01	3.89
59.12	7.26	24.53	81.63	69.57	61.59	89.13	95.83	8.63
17.08	2.96	68.19	42.14	30.55	21.97	59.69	85.76	28.09
49.78	6.53	32.32	72.88	64.10	51.98	86.56	93.98	8.63
46.01	6.3	34.69	71.10	61.70	46.11	87.96	95.79	11.68
35.50	5.03	44.33	65.64	50.58	36.53	79.16	93.61	22.77
34.42	4.51	56.01	52.80	41.15	34.17	63.30	94.09	14.33
58.58	6.95	28.93	77.96	69.47	59.22	85.73	94.78	9.31
49.06	6.50	31.64	72.94	63.35	47.49	85.64	94.82	6.09
19.86	2.58	65.48	46.48	31.29	21.38	60.96	85.26	32.83

Backward linear regression is used to find the most reliable combination of metrics that could be used to replace NER [20]. This regression has been selected for the following reasons:

- the data is linear (correlation analysis present linear relation),
- the data is ratio level data; thus, good for linear regression,
- this regression analysis provides more than one regression result; thus, the best variables can be found,
- reliable variables are extracted by excluding irrelevant variables from the first to the last stages (give the most-reliable variables).

Table 3 represents the regression summary for estimating NER using various metrics [11]. The significance (p-value) is estimated with an alpha 0.05 to investigate the most-significant variables. The Adjusted R^2 [7, 24] provides an idea of how much the variables explain the NER variances. The table also contains several section, representing subsequent iterations of the model with the least significant variable removed at each step. The removed variables are shown in bold at each step. The metrics in the last model are all significant, as their p-values are less than alpha, therefore no more iterations need to be executed.

The R^2 value of the last model accounts for the 76.1% of NER variance. All of the models have statistically acceptable R^2 values.

Table 3

Regression result summary for predicting NER using six metrics. The boldface metrics are least-significant and are removed in subsequent models. The italicized metrics are significant predictors of NER.

Model		Unstandardized Coeff.		Std. Coeff.	t-score	Signif.	Adj. R^2
		Beta	Std. Error	Beta			
1st model	(Constant)	100.068	19.933		5.020	.000	0.759
	BLEU	.176	.122	1.060	1.439	.157	
	NIST	1.162	.530	.738	2.192	.033	
	TER	−.072	.186	−.353	−.386	**.701**	
	METEOR	−.205	.168	−.797	−1.216	.230	
	EBLEU	−.218	.078	−1.293	−2.806	.007	
	RIBES	−.067	.093	−.219	−.723	.473	
	METEOR_pl	.194	.160	.954	1.216	.230	
2nd model	(Constant)	92.622	4.895		18.920	.000	0.763
	BLEU	.189	.117	1.136	1.614	.113	
	NIST	1.220	.504	.775	2.420	*.019*	
	METEOR	−.190	.163	−.741	−1.170	.248	
	EBLEU	−.215	.077	−1.270	−2.803	*.007*	
	RIBES	−.051	.082	−.167	−.620	**.538**	
	METEOR_pl	.217	.147	1.067	1.479	.145	
3rd model	(Constant)	91.541	4.547		20.133	.000	0.766
	BLEU	.184	.116	1.107	1.586	.119	
	NIST	1.068	.438	.678	2.440	*.018*	
	METEOR	−.237	.143	−.923	−1.655	.104	
	EBLEU	−.193	.068	−1.145	−2.841	*.006*	
	METEOR_pl	.222	.146	1.093	1.527	**.133**	
4th model	(Constant)	90.471	4.550		19.885	.000	0.760
	BLEU	.285	.096	1.716	2.955	*.005*	
	NIST	1.083	.443	.687	2.444	*.018*	
	METEOR	−.099	.112	−.385	−.878	**.384**	
	EBLEU	−.204	.069	−1.210	−2.982	*.004*	
5th model	(Constant)	86.556	.913		94.814	.000	0.761
	BLEU	.254	.090	1.531	2.835	*.006*	
	NIST	.924	.404	.587	2.289	*.026*	
	EBLEU	−.221	.066	−1.310	−3.370	*.001*	

6. Conclusion

The goal of this paper was to evaluate automatic metrics that are normally used for assessing the quality of Machine Translation systems and apply them for assessing the performance of re-speaking professionals. The ground-truth for this endeavor was the human-derived NER metric used commonly in the re-speaking industry. Note that this paper does not attempt to establish which metric is directly best at evaluating re-speaking, but simply which metric is the best estimator of NER (which is widely confirmed to correlate with human judgement regarding its quality of subtitles).

From the regression, it has been found that BLEU is the most-significant predictor of NER. After BLEU, NIST is significant; and finally, EBLEU is also a significant metric that can predict NER better. Thus, these can be alternatives to the NER metric. The regression equation that can compute the value of NER based on these three statistical metrics is:

$$\mathrm{NER} = 86.55 + 0.254 \cdot \mathrm{BLEU} + 0.924 \cdot \mathrm{NIST} - 0.221 \cdot \mathrm{EBLEU} \tag{11}$$

Moving forward, regression plots from the regression above for the significant metrics are presented in Figures 1, 2, and 3. These plots show the relationships between the dependent metrics with each significant metric. The closer the dots are to the regression line in the plot, the better the R^2 value and, thus, the relationship.

It is worth noting that the METEOR metric should be fine-tuned to work properly. This would, however, require more data, and more research into this is planned in the future. The results presented in this paper are derived from a very small data sample and may not be very representative of the task in general. The process of acquiring more data is still ongoing, so these experiments are going to be repeated at some point when more data becomes available.

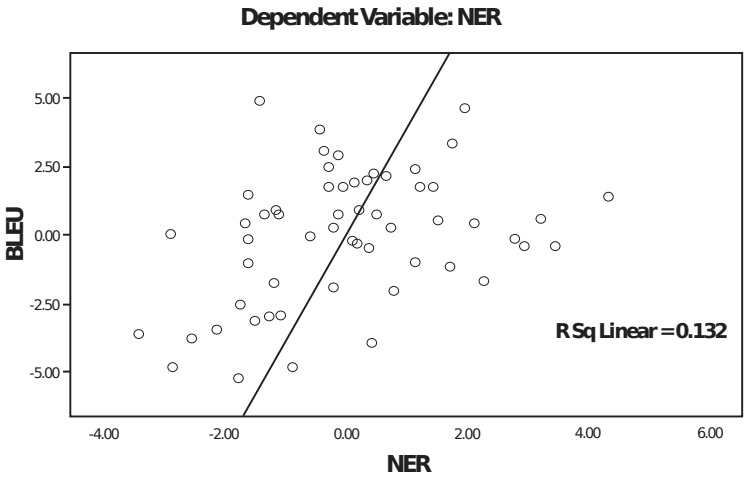

Figure 1. Partial regression plot for NER and BLEU.

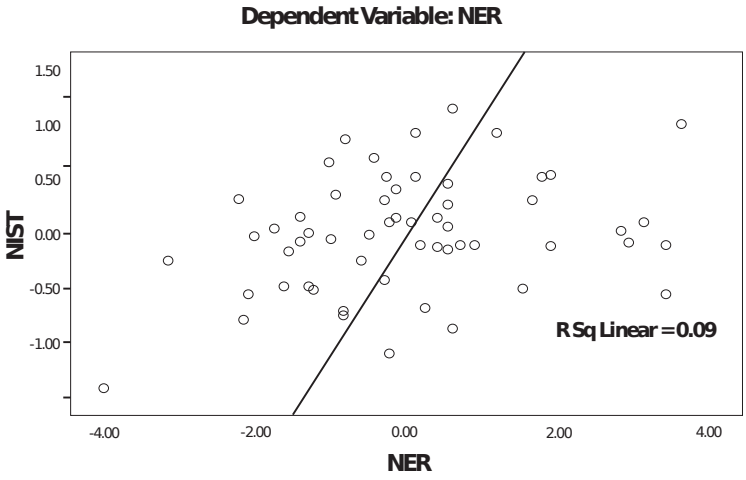

Figure 2. Partial regression plot for NER and NIST.

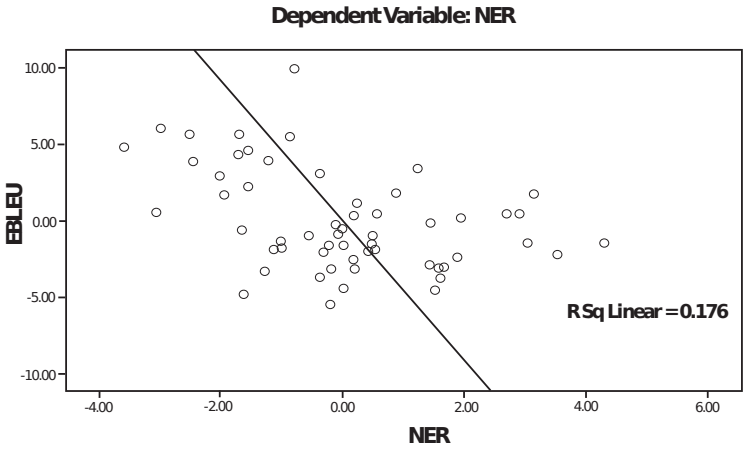

Figure 3. Partial regression plot for NER and EBLEU.

There are several additional topics that will be attempted in the future: the use of other metrics, including inter-lingual, like MEANT [13] or LEPOR [8], as well as intra-lingual, like CollGram [3] or language model perplexity and other model-based techniques. Furthermore, we would also like to develop methods for assessing an inter-lingual form of re-speaking (i.e., re-speaking combined with translation). The methods would be similar but may require some form of machine translation to allow for full automation.

Acknowledgements

We would like to thank Dr. Agnieszka Szarkowska and Łukasz Dutka for sharing the data that allowed us to perform these experiments. The data analyzed in the paper was collected as part of a study supported by research grant no. 2013/11/B/HS2/02762 from the National Science Center Poland for the years of 2014-2017. Some of the research used in this paper was funded by the CLARIN-PL project. This project was partially supported by the infrastructure bought within the project "Heterogenous Computation Cloud" funded by the Regional Operational Program of Mazovia Voivodeship.

References

[1] Axelrod A.: *Factored language model for statistical machine translation*, Master of Science by Research, Institute for Communicating and Collaborative System, Division of Informatics, University of Edinburgh, 2006.

[2] Banerjee S., Lavie A.: 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*, vol. 29, pp. 65–72, 2005.

[3] Bestgen Y., Granger S.: Quantifying the development of phraseological competence in L2 English writing: An automated approach, *Journal of Second Language Writing*, vol. 26, pp. 28–41, 2014.

[4] Doddington G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *Proceedings of the second international conference on Human Language Technology Research*, pp. 138–145. Morgan Kaufmann Publishers, 2002.

[5] Dutka Ł., Szarkowska A., Chmiel A., Lijewska A., Krejtz K., Marasek K., Brocki Ł.: Are interpreters better respeakers? An exploratory study on respeaking competences, *Respeaking, live subtitling and accessibility*, Rome, 12 June 2015.

[6] European Federation of Hard of Hearing People: State of subtitling access in EU. 2011 Report. `http://ec.europa.eu/internal_market/consultations/2011/audiovisual/non-registered-organisations/european-federation-of-hard-of-hearing-people-efhoh-_en.pdf`, 2011. [Online; accessed 30 Jan. 2016].

[7] Frost J.: Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables, *The Minitab Blog*, 2013. `http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables`.

[8] Han A.L.F., Wong D.F., Chao L.S.: LEPOR: A robust evaluation metric for machine translation with augmented factors. In: *Proceedings of COLING 2012: Posters*, pp. 441–450, 2012.

[9] Hovy E.: Toward finely differentiated evaluation metrics for machine translation. In: *Proceedings of the EAGLES Workshop on Standards and Evaluation*, Pisa, Italy, 1999.

[10] Isozaki H., Hirao T., Duh K., Sudoh K., Tsukada H.: Automatic evaluation of translation quality for distant language pairs. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 944–952. Association for Computational Linguistics, 2010.

[11] Kim J.O., Mueller C.W.: Standardized and unstandardized coefficients in causal analysis. An expository note, *Sociological Methods & Research*, vol. 4(4), pp. 423–438, 1976.

[12] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., et al.: Moses: Open source toolkit for statistical machine translation. In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177–180, Association for Computational Linguistics, 2007.

[13] Lo C.-k., Wu D.: MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 220–229, Association for Computational Linguistics, 2011.

[14] Maziarz M., Piasecki M., Szpakowicz S.: Approaching plWordNet 2.0. In: *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan, pp. 50–62, 2012.

[15] Miller G.A.: WordNet: a lexical database for English, *Communications of the ACM*, vol. 38(11), pp. 39–41, 1995.

[16] Papineni K., Roukos S., Ward T., Zhu W.J.: BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[17] Porter M.F.: Snowball: A language for stemming algorithms, 2001. `http://snowball.tartarus.org/texts/introduction.html`.

[18] Reeder F.: Additional mt-eval references. In: *International Standards for Language Engineering, Evaluation Working Group*, 2001.

[19] Romero-Fresco P., Martínez J.: Accuracy rate in live subtitling. The NER model. In: *Audiovisual Translation in a Global Context*, pp. 28–50, 2011.

[20] Seber G.A., Lee A.J.: *Linear regression analysis*, John Wiley & Sons, 2012.

[21] Woliński M., Miłkowski M., Ogrodniczuk M., Przepiórkowski A., Szałkiewicz Ł.: PoliMorf: a (not so) new open morphological dictionary for Polish. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), 23–25 May, Istanbul, Turkey, pp. 860–864.

[22] Wołk K., Marasek K.: Polish-English Speech Statistical Machine Translation Systems for the IWSLT 2013. In: *Proceedings of the 10th International Workshop on Spoken Language Translation*, Heidelberg, Germany, pp. 113–119, 2013.

[23] Wołk K., Marasek K.: Enhanced Bilingual Evaluation Understudy, *Lecture Notes on Information Theory*, vol. 2(2), 2014.

[24] Zimmerman D.W.: Teachers corner: A note on interpretation of the paired-samples t test, *Journal of Educational and Behavioral Statistics*, vol. 22(3), pp. 349–360, 1997.

Affiliations

Krzysztof Wołk
Polish-Japanese Academy of Information Technology, Faculty of Information Technology, Department of Multimedia, Warsaw, Poland, kwolkw@pja.edu.pl

Danijel Koržinek
Polish-Japanese Academy of Information Technology, Faculty of Information Technology, Department of Multimedia, Warsaw, Poland, danijel@pja.edu.pl

Received: 31.01.2016
Revised: 16.11.2016
Accepted: 16.11.2016