EUROPEAN
ASSOCIATION
FOR MACHINE
TRANSLATION

EAMT2018

Proceedings of the 21st Annual Conference of the European Association for Machine Translation

28–30 May 2018
Universitat d'Alacant
Alacant, Spain

Edited by
Juan Antonio Pérez-Ortiz
Felipe Sánchez-Martínez
Miquel Esplà-Gomis
Maja Popović
Celia Rico
André Martins
Joachim Van den Bogaert
Mikel L. Forcada

Organised by

Universitat d'Alacant
Universidad de Alicante

transducens
research group

The papers published in this proceedings are —unless indicated otherwise— covered by the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 International (CC-BY-ND 3.0). You may copy, distribute, and transmit the work, provided that you attribute it (authorship, proceedings, publisher) in the manner specified by the author(s) or licensor(s), and that you do not use it for commercial purposes. The full text of the licence may be found at https://creativecommons.org/licenses/by-nc-nd/3.0/deed.en.

© 2018 The authors

ISBN: 978-84-09-01901-4

Machine Translation Evaluation beyond the Sentence Level

Jindřich Libovický[*] Faculty of Mathematic and Physics
Charles University
libovicky@ufal.mff.cuni.cz

Thomas Brovelli (Meyer) **Bruno Cartoni**
Google
{thomasmeyer,
brunocartoni}@google.com

Abstract

Automatic machine translation evaluation was crucial for the rapid development of machine translation systems over the last two decades. So far, most attention has been paid to the evaluation metrics that work with text on the sentence level and so did the translation systems. Across-sentence translation quality depends on discourse phenomena that may not manifest at all when staying within sentence boundaries (e.g. coreference, discourse connectives, verb tense sequence etc.). To tackle this, we propose several document-level MT evaluation metrics: generalizations of sentence-level metrics, language-(pair)-independent versions of lexical cohesion scores and coreference and morphology preservation in the target texts. We measure their agreement with human judgment on a newly created dataset of pairwise paragraph comparisons for four language pairs.

1 Introduction

Automatic machine translation (MT) evaluation is a crucial technique that accompanied the development of machine translation systems over the last two decades. It allows replacing accurate, but prohibitively slow manual evaluation by a fast and replicable automatic evaluation routine approximating human judgment. So far, the most attention has been paid to the evaluation metrics that work with text on the sentence level and most of the MT systems work at sentence level as well.

The recent advances in neural machine translation (Wu et al., 2016) demonstrated that the state-of-the-art systems, are not too far from the human-level quality on the sentence level. Translating paragraphs or even entire documents is thus becoming a new challenge for MT systems. While this progress is underway, one also needs to assess the translation quality at the paragraph level.

Quality of coherent text translation depends on discourse phenomena that cannot be resolved within sentence boundaries. For instance, the correct sequence of events in the text or the correct placement of gendered pronouns needs to be retained in the target language text to provide a correct translation. Recent experiments with incorporating a broader context into neural machine translation (Wang et al., 2017; Jean et al., 2017) brought only a modest improvement. As these approaches were evaluated using only sentence-level metrics, some important properties of the models might have been missed.

Another important motivation for developing paragraph- or document-level metrics is the growing popularity of reinforcement learning in neural MT, optimizing the model directly towards a given metric (Ranzato et al., 2015; Shen et al., 2016; Gu et al., 2017). If we want to take advantage of this setup at the paragraph level, more elaborated metrics are necessary.

In this paper, we propose several paragraph-level MT evaluation metrics. We evaluate how these metrics agree with human judgment while deciding which translation is better when only a single paragraph of text is used for the comparison on four different language pairs. Because of the lack of annotated data, we create our own

© 2018 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

[*] Work done during an internship at Google.

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 179–188
Alacant, Spain, May 2018.

dataset consisting of the system outputs submitted to the shared translation tasks of the Workshop on Machine Translation (WMT) between 2014 and 2016 (Bojar et al., 2014; Bojar et al., 2015; Bojar et al., 2016). The dataset with anonymized paragraph translation ratings will be published with the final version of this paper.

The remainder of the paper is organized as follows: Section 2 summarizes the previous work, Section 3 introduces the paragraph-level level metrics, Section 4 describes the evaluation dataset. In Section 5, we describe the experiments we conducted to estimate agreement of the proposed metrics with human judgment.

2 Previous Work

There have been a few attempts so far to measure translation quality beyond the sentence level. With most of the MT frameworks still translating sentence by sentence, there was no urgent need to measure quality at higher levels. The fact that the standard MT scoring methods such as BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), seem to correlate well with human judgment further supported and established that practice.

With the advent of high-quality sentence-level machine translation (Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017), one of the next challenges is to translate entire paragraphs and documents consistently, i.e. in a lexically coherent and pragmatically appropriate manner. Argumentative structure of text, consistency of lexical choice, and the right 'tone' for its pragmatic intent are the next problems to focus on.

Simple n-gram matching (as with BLEU) and/or allowing for certain word order and synonym variants (as with METEOR (Lavie and Agarwal, 2007)), will likely not be able to capture the aforementioned linguistic phenomena that are crucial for the coherence of the entire text. More aggravatingly, both BLEU and METEOR heavily rely on comparison against one (or sometimes up to 4) human reference translations. These are however not usually available for an entire document. The BLUE score is technically a corpus-level metric because it computes the brevity penalty over the whole corpus. Nevertheless, it does not make use of cross-sentence information in a particularly useful way.

Carpuat (2009) empirically showed that enforcing the one-sense-per-discourse hypothesis by repeating the same words in an MT output can potentially improve the MT quality. Wong and Kit (2012) proposed measuring the semantic similarity of previously seen words in a text in order to capture lexical cohesion of documents in the target language. Lexical cohesion relates to word choice, that Wong and Kit measure by tracking collocation and reiteration (of word stems), additionally allowing for synonyms, near-synonyms and superordinates (for collocation). We take on this approach as well and provide a *language-independent* variant in Section 3.1.

Soricut and Echihabi (2010) on the other hand, viewed the document-level MT evaluation as a ranking problem. They built an MT system that relies on regression models to find BLEU-like numbers for good translations at the document-level which are then ranked higher than others. Similarly to what we will find below, Soricut and Echihabi have shown that an averaged BLEU score over a document is a useful indicator of actual good translation quality and can be used as a feature to find pseudo-reference translations (coming from a secondary MT system) that in turn can be used to estimate the quality of the former MT system.

Similarly, Scarton and Specia (2014) are concerned with quality estimation at the document level, especially when no human reference translations are available. They use a mix of pseudo-reference scores, as Soricut and Echihabi (2010), together with the lexical cohesion features by Wong and Kit (2012). They take the word form repetitions to make the metric language-independent, while we rely on word embeddings that account for richer encoding of synonyms, antonyms etc. than just pure repeated mentions. The main discursive features Scarton and Specia use are LSA scores. They rely on Spearman rank correlation of the word vector of a current sentence compared to all sentences of the document. Whereas both Soricut and Echihabi's and Scarton and Specia's papers need human reference translations or at least pseudo-references for training their regression models, our metrics below can be deployed fully automatically and rely mostly on a monolingual (but automatic) word aligner and freely available, automatic syntactic and semantic parsers.

Hardmeier and Federico (2010) and Miculicich Werlen and Popescu-Belis (2017) use F-score based metrics for pronoun translation evaluation. In Sections 3.2 we take a similar approach from computing coreference preservation.

Besides the approaches presented above, there have also been a few attempts to measure translation quality for certain discourse phenomena in isolation. Meyer et al. (2015) have developed a metric to measure improvements on MT for discourse connectives, whereas for example Gojun and Fraser (2012) and Loaiciga et al. (2014) specifically looked at measuring translation quality for verb tense. Although these approaches have presented interesting results, they can unfortunately not point to the overall translation quality of an entire paragraph.

3 Implemented Paragraph-Level Metrics

We implement two sets of metrics. The first ones operate on the paragraph level and are mostly generalizations of existing MT evaluation metrics (see Section 3.1).

The second set of metrics relies on monolingual word alignment between the reference paragraph and the translation hypothesis (see Section 3.2). Word alignment allows us to measure linguistically motivated statistics about the translation. Nevertheless, alignment errors can pose the danger of bringing additional noise to the evaluation. Moreover, word alignment is only an approximation of what we would really need for thorough document level statistics which would be phrase-level alignment.

In order to find linguistic features (especially entities, coreference and morphology) for the metrics described in the following section, we have been analyzing the respective texts with the Google Cloud Natural Language API[1].

3.1 Metrics without the Monolingual Alignment

Paragraph-Level BLEU. We implemented a simple extension of the standard sentence-level BLEU score (Papineni et al., 2002). Unlike the standard BLEU score, we compute the n-gram statistic throughout the whole paragraph.

The BLEU score is a product of modified n-gram precision and a brevity penalty. The modified n-gram precision approximates the lexical adequacy of the translation and its local fluency. Note that the longer the text is, the less reliable the short n-gram precision becomes because the most frequent words from a language are more likely to get covered by chance. The brevity penalty prevents overrating of longer texts as the probability of accidental covering of the reference text by the hypotheses' n-grams grows with the text length.

Lexical Cohesion Score. One of the features we attribute to a good translation is its stylistic consistency which also includes lexical cohesion. Especially in non-fiction text, we expect the same terms to be used for the same concepts as well as their belonging to the same language register.

Wong and Kit (2012) tried to capture these phenomena in a lexical cohesion score for MT evaluation. The original metric is an average ratio of semantically similar content words observed previously in the text. We propose a language independent extension of the metric.

Formally, we define the score in the following way:

$$\frac{1}{|C|-1}\sum_{i=2}^{|C|}\mathbf{1}\left[\exists c_j : j < i \;\&\; c_i \text{ is related to } c_j\right] \tag{1}$$

where $C = (c_1, \ldots, c_{|C|})$ is a sequence of content words in the text. Semantic similarity was originally defined by a graph distance threshold in WordNet (Fellbaum, 1998) which does not have sufficiently high coverage for languages other than English.

In order to overcome this drawback we reformulate the score:

$$\frac{1}{|C|-1}\sum_{i=2}^{|C|}\max_{j=1..i-1}\operatorname{sim}(c_i, c_j). \tag{2}$$

As function sim, we use cosine similarity of pretrained word embeddings (Mikolov et al., 2013) instead of the binary indication of semantic similarity based on WordNet.

3.2 Metrics Requiring Monolingual Alignment

For monolingual alignment, we re-implemented the state-of-the-art rule-based monolingual aligner (Sultan et al., 2014). In order to make the aligner language-independent, we transferred the rules for finding equivalent dependency structures from Stanford-style dependencies to Universal

[1] Publicly available at: `https://cloud.google.com/natural-language/docs/`

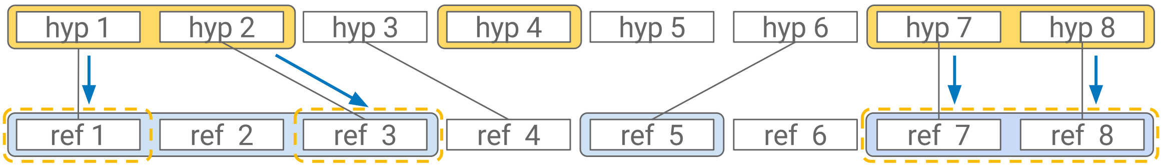


Figure 1: Examples of coreference chain projection via monolingual alignment.

	Prec.	Recall	F_1	Acc.
METEOR	89.7	71.0	79.3	10.8
Our aligner	88.2	65.7	75.3	10.1

Table 1: Comparison of the METEOR aligner and our aligner on the Edinburgh++ dataset.

Dependencies (Marneffe et al., 2014). Unlike the original aligner, our aligner does not require explicitly aligned sentences and is agnostic to the sentence boundaries as it is treating the paragraphs as dependency forests.

The alignment algorithm is a pipeline of rule-based steps. In the first step, it aligns identical word sequences and named entities. In the second step, the dependency surroundings of already aligned content words are aligned if their dependency labels belong to manually designed categories. Then, linear surroundings of the content words are aligned if the words in the surroundings are similar enough. For that purpose, we use the lexical paraphrases table of PPDB 2.0 paraphrase database (Pavlick et al., 2015) and a word embedding distance. We repeat the procedure for non-content words, with the only difference that we use semantic similarity in the dependency context alignment as well.

The aligner has similar results to the METEOR aligner (Lavie and Agarwal, 2007) when comparing against the Edinburgh++ dataset (Cohn et al., 2008) (see Table 1). It does not use longer phrases from the paraphrase database which would increase the aligning complexity prohibitively in case of long texts.

Paragraph-Level METEOR. We extended the METEOR score to operate on the paragraph level in a straightforward manner. As the standard METEOR, it is a product of a disfluency score d and an adequacy score a. The disfluency score is computed as

$$d = \frac{1}{2}\left(\frac{\text{\# alignment steps}}{\text{\# unigrams matched}}\right)^3 \qquad (3)$$

and captures how much the hypothesis paragraph would need to be torn apart in order to be aligned with the reference.

Lexical adequacy is computed as a weighted harmonic mean of precision and recall:

$$a = \frac{10 \cdot P \cdot R}{R + 9P} \qquad (4)$$

where P and R stands for precision and recall of the hypothesis words computed over the monolingual alignment.

We evaluate two methods of computing precision and recall. In the standard way, which we refer to as *Hard METEOR*, we assign a unit weight to all alignment links. As an alternative, we introduce *Soft METEOR* where we weight the alignment links by word similarity estimated from word embeddings distance and weight the precision and recall accordingly.

Morphology Preservation. Similarly to the METEOR score, where we compute the lexical adequacy of all words in the text, we can measure preservation of morphological categories that can provide information about phenomena that are crossing sentence boundaries.

As in METEOR, we measure the F-score of the morphological categories being the same. The F-score takes into account also the false negatives and false positives. Alternatively, we calculate the accuracy of only those word pairs that have been aligned together. Computing the accuracy instead of the F-measure is more appropriate in cases where morphological categories are not well-covered by the monolingual alignment, e.g. pronouns.

We measure the preservation of pronoun number and gender, which should capture the extent of coreference chains throughout the text. Additionally, verb gender, tense and number will also capture how the sequence of described events is preserved between the translation hypothesis and the reference.

Computing morphology preservation is not only limited by the quality of the monolingual align-

English source: *The fertile ground and the rainforest climate of Isla del Rey are ideal for growing marijuana plants. Three days ago the authorities in Panama tore out the 4,500 plants and burnt them.*

German reference	System A	System B
Der fruchtbare Boden und das regenwaldtypische Klima der Isla del Rey sind für das Gedeihen der Marihuana-Pflanzen bestens geeignet. Seit drei Tagen reissen die Behörden Panamas die 4500 Pflanzen aus und verbrennen sie.	*Der fruchtbare Boden und das Regenwaldklima von Isla del Rey sind ideal, um Marihuanaanlagen zu wachsen. Vor drei Tagen rissen die Behörden in Panama die 4.500 Anlagen heraus und verbrannten sie.*	*Der fruchtbare Boden und das Regenwaldklima von Isla del Rey sind ideal für wachsende Marihuana-Pflanzen. Vor drei Tagen haben die Behörden in Panama die 4.500 Pflanzen ausgebrannt und verbrannt.*
Sentence-level BLEU	.229	.233
Sentence-level METEOR	.392	.376
Sentence-level TER	.545	.545
Paragraph BLEU score	.203	.229
Coreference: BLANC	.791	.890
Coreference: Non-link F_1 score	.966	.980
Hypotheses lexical cohesion	.594	.632
Meteor – hard	.447	.530
Meteor – soft	.434	.522
Pronoun gender accuracy	.941	.900
Pronoun number accuracy	1.000	.950
Verb gender F_1 score	.667	.500
Verb number F_1 score	.667	.250
Verb tense F_1 score	.222	.250

Table 2: Score values for the implemented document-level metrics. This illustrates proof-of-concept and good correlation with sentence-level metrics.

ment. It can also generate false positives, e.g. in cases where grammatical gender is not preserved because of different but still correct lexical choice. We believe that averaged over a longer dataset, this type of metrics can still bring interesting linguistic insight.

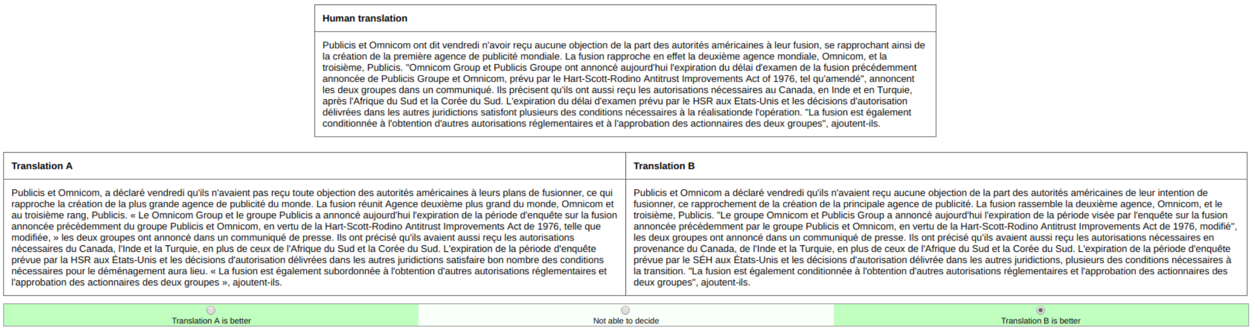
Coreference Preservation. Coreference chains can easily get broken during machine translation, especially when the translation is done on the sentence level. Except for indirect measurements of the coreference preservation via morphological categories of pronouns and verbs, we also explicitly compute coreference preservation via projection of the reference coreference chains to the translation hypothesis.

We apply entity and coreference resolution on the translation hypothesis (by detecting all nominal elements such as noun phrases, proper names and pronouns, as well as their coreference links). We project these mentions of entities in the hypothesis text to the reference translation using the alignment links as illustrated in Figure 1. No restrictions are imposed on this projection, so that the projected mentions do not even have to be continuous chunks of text. This also gives a mention matching that can be used during metric computation.

Once the projection is done, we treat the coreference chains in the reference text as the ground truth and measure the quality of the projected chains (i.e. treat them as the response).

There are two main approaches to the evaluation of coreference resolution. We can either measure how well the resolver spotted the words in the entity mentions or how well it preserved the coreference links. We therefore implemented two coreference metrics: the B^3 average F_1 score for treating the problem as retrieval of mentions, and the BLANC score (Luo et al., 2014), which is an average of the F_1 score of the coreference links and the F_1 measure of the complements of the coreference links (complement of the complete graph).

Table 2 shows the values of the document-level translation evaluation metrics in a real example from the WMT 2016 test set. When judged by a human, the hypothesis from system A is slightly



Human translation

Publicis et Omnicom ont dit vendredi n'avoir reçu aucune objection de la part des autorités américaines à leur fusion, se rapprochant ainsi de la création de la première agence de publicité mondiale. La fusion rapproche en effet la deuxième agence mondiale, Omnicom, et la troisième, Publicis. "Omnicom Group et Publicis Groupe ont annoncé aujourd'hui l'expiration du délai d'examen de la fusion précédemment annoncée de Publicis Groupe et Omnicom, prévu par le Hart-Scott-Rodino Antitrust Improvements Act of 1976, tel qu'amendé", annoncent les deux groupes dans un communiqué. Ils précisent qu'ils ont aussi reçu les autorisations nécessaires au Canada, en Inde et en Turquie, après l'Afrique du Sud et la Corée du Sud. L'expiration du délai d'examen prévu par le HSR aux Etats-Unis et les décisions d'autorisation délivrées dans les autres juridictions satisfont plusieurs des conditions nécessaires à la réalisationde l'opération. "La fusion est également conditionnée à l'obtention d'autres autorisations réglementaires et à l'approbation des actionnaires des deux groupes", ajoutent-ils.

Translation A	Translation B
Publicis et Omnicom, a déclaré vendredi qu'ils n'avaient pas reçu toute objection des autorités américaines à leurs plans de fusionner, ce qui rapproche la création de la plus grande agence de publicité du monde. La fusion réunit Agence deuxième plus grand du monde, Omnicom et au troisième rang, Publicis. « Le Omnicom Group et le groupe Publicis a annoncé aujourd'hui l'expiration de la période d'enquête sur la fusion annoncée précédemment du groupe Publicis et Omnicom, en vertu de la Hart-Scott-Rodino Antitrust Improvements Act de 1976, telle que modifiée, » les deux groupes ont annoncé dans un communiqué de presse. Ils ont précisé qu'ils avaient aussi reçu les autorisations nécessaires du Canada, l'Inde et la Turquie, en plus de ceux de l'Afrique du Sud et la Corée du Sud. L'expiration de la période d'enquête prévue par la HSR aux États-Unis et les décisions d'autorisation délivrées dans les autres juridictions satisfaire bon nombre des conditions nécessaires pour le déménagement aura lieu. « La fusion est également subordonnée à l'obtention d'autres autorisations réglementaires et l'approbation des actionnaires des deux groupes », ajoutent-ils.	Publicis et Omnicom a déclaré vendredi qu'ils n'avaient reçu aucune objection de la part des autorités américaines de leur intention de fusionner, ce rapprochement de la création de la principale agence de publicité. La fusion rassemble la deuxième agence, Omnicom, et le troisième, Publicis. "Le groupe Omnicom et Publicis Group a annoncé aujourd'hui l'expiration de la période visée par l'enquête sur la fusion annoncée précédemment par le groupe Publicis et Omnicom, en vertu de la Hart-Scott-Rodino Antitrust Improvements Act de 1976, modifié", les deux groupes ont annoncé dans un communiqué de presse. Ils ont précisé qu'ils avaient aussi reçu les autorisations nécessaires en provenance du Canada, de l'Inde et la Turquie, en plus de ceux de l'Afrique du Sud et la Corée du Sud. L'expiration de la période d'enquête prévue par le SÉH aux États-Unis et les décisions d'autorisation délivrée dans les autres juridictions, plusieurs des conditions nécessaires à la transition. "La fusion est également conditionnée à l'obtention d'autres autorisations réglementaires et l'approbation des actionnaires des deux groupes", ajoutent-ils.

○ Translation A is better ○ Not able to decide ● Translation B is better

Figure 2: Example evaluation task for human annotators

better, but has a lower sentence-level BLEU score than system B. Our document-level metrics can hint at the better quality of A with e.g. the lexical cohesion score as well as the pronoun and verb morphology scores.

4 Dataset

Unlike sentence-level MT evaluation which can benefit from evaluation campaigns like the WMT tasks of annual metrics evaluation (Bojar et al., 2017), there is no dataset consisting of human judgments on machine translation quality beyond the sentence level. Even the metrics that were discussed in Section 2 were only evaluated against human judgments collected at the sentence level.

In order to evaluate our metrics reliably, we created a new dataset consisting of pairwise paragraph comparisons of machine translation outputs that have been rated by several human annotators per pair. The paragraphs are extracted from the freely accessible test sets provided for the WMT workshops (years 2014 to 2016). Our rated data sets will be made available publicly with the final version of this paper.

4.1 Pilot Annotation

In order to determine a reasonable length for paragraphs to be evaluated by human raters, we conducted a pilot experiment where we sampled 30 paragraphs from the WMT datasets for the English to German, German to English, English to French and French to English translation directions. The length of these paragraphs has arbitrarily been set to approximately 180 words each. At this stage, the target side translations have been sampled randomly from system outputs submitted to the WMT shared news tasks of the years 2014 to 2016. The annotators were provided with a simple user interface that showed them the human reference translation, a system output A to the left and a system output B to the right. The annotators task was to select either *system A is better*, *undecided* or *system B is better* compared to the reference translation (Figure 2). In the pilot round, the evaluators were trained linguists and native speakers of the target languages. The annotators were afterwards informally interviewed.

We learned from the feedback of annotators that the sampled paragraph length of 180 words is enough to capture phenomena in translation that cross sentence boundaries. Metric-wise, our paragraph-level extensions of BLEU and METEOR are reasonable choices, especially for English to French and French to English translation and align well with the human judgment (which is not to be expected to be perfect either when rating over several sentences). Lexical cohesion difference and linked-based coreference scores also confirm that the more lexically coherent a paragraph is, the higher it is rated by humans, independently of the reference translation. The annotators relative agreement was over 70 % ($\kappa = 0.4$) and only a minority of paragraph pairs remained undecided.

4.2 Large-Scale Annotation

The annotation of a bigger evaluation dataset was done for four language pairs: English to Czech, English to German, English to French and English to Russian. The paragraphs were randomly sampled from the same set of WMT system submissions as in the pilot round[2]. In addition to the MT systems submitted to WMT, we also translated the sampled paragraphs with Google's neural MT (Wu et al., 2016).

Unlike the pilot round, which was conducted

[2] If you would like to use the dataset, please use the following form: https://goo.gl/forms/zvpOddi9FelFkJxJ2.

language pair		agr.	κ	BLEU	ΔBLEU
en → cs	all	.68	.53	12.3	6.1
	good	.42	.12	22.9	5.8
en → de	all	.55	.33	17.0	6.1
	good	.37	.06	26.0	5.5
en → fr	all	.58	.37	25.0	8.6
	good	.40	.05	36.5	6.5
en → ru	all	.58	.37	20.9	8.9
	good	.42	.13	30.5	6.7

Table 3: Statistics on the collected dataset: annotator agreement (agr.) as a proportion of cases when all three annotators agreed, Cohen's κ, average BLEU score and average BLEU score difference (ΔBLEU). Labels 'good' and 'all' refer the quality of the translation the paragraphs were sampled from. The former contains pairs of paragraphs only from outputs of systems that achieved a total sentence-level BLEU score of over 30 points on the selected paragraphs. The latter contains samples irrespective of BLEU scores (also see Section 4.2).

by trained linguists, the only requirement for this larger crowd-sourced annotation was that the raters must be native speakers of the target language and must understand English. Every paragraph pair was evaluated independently by three raters and the majority vote was used as final rating decision.

To be able to better evaluate how the document-level metrics behave under different circumstances, we created two test sets for each of the language pairs. In the first test set, the paragraphs are sampled randomly from the WMT submissions which are often of different quality. The second, more challenging test set, contains pairs of paragraphs only from outputs of systems that achieved a total sentence-level BLEU score of over 30 points on the selected paragraphs. Both variants contain 400 paragraph pairs for all the four language pairs. The statistics of the dataset are tabulated in Table 3. One notable fact is that the annotator agreement (proportion of cases when all three annotators agreed) is relatively low and even decreases when using a higher quality system.

5 Experiments

We evaluated the metrics proposed in Section 3 on the collected datasets on English to German and English to French translation directions. For every metric, we computed the proportion of cases when the paragraph annotated as the better one has also been assigned the higher score, i.e. which of the two system outputs provides a better entire paragraph translation when comparing to the reference. All the paragraphs were also evaluated with the standard sentence-level metrics (BLEU, METEOR, TER)[3]. The detailed results are presented in Table 4.

If we interpret the annotator agreement as probability that all three annotators agree, we can factorize this probability into two steps: first that two agreed (and thus did the majority vote) and that the third annotator agreed with them. Therefore, we can estimate the probability of the third annotator agreeing with the majority vote as a square root of the annotator agreement. These are presented in the first line of Table 4.

The main finding of the analysis is that the agreement of both the traditional sentence-level metrics and the proposed metrics with the human judgment is relatively low in pairwise comparison. In fact, only a small majority of the pairwise comparisons is done correctly. This particular finding contradicts the training techniques based on the REINFORCE algorithm (Williams, 1992) where the update rule explicitly contains the pairwise comparison. Moreover, it is not clear whether there is a room for improvement given that for good translation systems, the performance of the automatic metrics is on par with the estimated human agreements.

The other interesting result is that it is possible to estimate which translation is better almost equally well when focusing only on a particular phenomenon (coreference, lexical cohesion, morphology) as with metrics that should capture the translation quality holistically (METEOR, BLEU).

The metrics based on morphological analysis achieved better performance on paragraph pairs consisting of good translations. It might be so because the morphological analysis is more likely to fail in case of malformed translation outputs where the monolingual alignment is more difficult, because the hypothesis is different from the reference.

A similar trend can also be observed for coreference preservation. The BLANC score used for coreference evaluation is an average of F_1 scores of estimating correctly the coreference links and

[3] We used the metrics as implemented in MultEval, `https://github.com/jhclark/multeval`

Metric	en → de good	en → de all	en → fr good	en → fr all
Estimated human agreement	.610	.743	.629	.762
Sentence-level BLEU	.615	.594	.643	.629
Sentence-level METEOR	.612	.594	.640	.629
Sentence-level TER	.567	.559	.610	.594
Paragraph BLEU score	.610	.572	.658	.629
Coreference: BLANC	.577	.428	.533	.542
Coreference: Non-link F_1 score	.584	.425	.538	.548
Hypotheses lexical cohesion	.542	.489	.635	.499
Meteor – hard	.587	.562	.640	.598
Meteor – soft	.584	.562	.643	.601
Pronoun gender accuracy	.484	.438	.495	.505
Pronoun number accuracy	.524	.348	.443	.433
Verb gender F_1 score	.529	.198	.510	.492
Verb number F_1 score	.537	.214	.510	.464
Verb tense F_1 score	.537	.208	.508	.495

Table 4: Average agreement of the proposed metrics with the majority vote on human judgment on pairwise paragraph comparison. Columns denotes as 'all' contain randomly sampled system pairs, columns denoted as 'good' contain only pairs where both compared paragraphs achieved a BLEU score of at least 30.

its complement-non-link relations. Often, a better aggrement was achieved with the score computed only over the non-link relations which are much denser than the coreference links. We hypothesize this makes the score more robust to alignment errors.

6 Conclusions

The presented study focused on two main new contributions.

First, we implemented an entire package of automatic paragraph-level MT quality metrics that are language-(pair)-independent and track MT quality at different levels throughout entire paragraphs. Our extensions of the METEOR and lexical cohesion scores thereby showed promising results for most adequately and consistently measuring paragraph-level MT quality. We also experimented with more linguistically motivated scores, such as coreference preservation that could be interesting for future experiments, once the alignment of pronouns and referential expressions is more reliable.

Second, we prepared a dataset of human judgments on pairwise comparisons of MT quality at the paragraph level which can be used for new metrics evaluation. The dataset consists of system translations from English to Czech, French, German and Russian submitted to WMT in recent years. For each language pair, 400 pairwise comparisons of randomly selected paragraphs and another 400 pairs of more similar, high-quality translation pairs have been rated humanly for paragraph translation quality.

Future work will try to improve the monolingual alignment. Better performance of parsers and coreference resolvers would indirectly also help the presented metrics. Integration of pseudo-references (where no human reference translations are available) and training an ensemble of all the metrics in our package can also be a promising direction.

References

[Bojar et al.2014] Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

[Bojar et al.2015] Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

[Bojar et al.2016] Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

[Bojar et al.2017] Bojar, Ondřej, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Carpuat2009] Carpuat, Marine. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27. Association for Computational Linguistics.

[Cohn et al.2008] Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4):597–614, dec.

[Fellbaum1998] Fellbaum, Christiane. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

[Gehring et al.2017] Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In Precup, Doina and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 06–11 Aug. PMLR.

[Gojun and Fraser2012] Gojun, Anita and Alexander Fraser. 2012. Determining the placement of german verbs in english-to-german smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 726–735, Avignon, France. Association for Computational Linguistics.

[Gu et al.2017] Gu, Jiatao, Kyunghyun Cho, and Victor O.K. Li. 2017. Trainable greedy decoding for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1968, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Hardmeier and Federico2010] Hardmeier, Christian and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation, Paris.*

[Jean et al.2017] Jean, S., S. Lauly, O. Firat, and K. Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135, apr.

[Lavie and Agarwal2007] Lavie, Alon and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Loaiciga et al.2014] Loaiciga, Sharid, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-french verb phrase alignment in europarl for tense translation modeling. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

[Luo et al.2014] Luo, Xiaoqiang, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of blanc to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, Baltimore, Maryland, June. Association for Computational Linguistics.

[Marneffe et al.2014] Marneffe, Marie-Catherine De, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: a cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[Meyer et al.2015] Meyer, T., N. Hajlaoui, and A. Popescu-Belis. 2015. Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197.

[Miculicich Werlen and Popescu-Belis2017] Miculicich Werlen, Lesly and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Mikolov et al.2013] Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

[Papineni et al.2002] Papineni, Kishore, 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*, pages 311–318. Association for Computational Linguistics.

[Pavlick et al.2015] Pavlick, Ellie, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July. Association for Computational Linguistics.

[Ranzato et al.2015] Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732.

[Scarton and Specia2014] Scarton, Carolina and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 101–108. Association for Computational Linguistics.

[Shen et al.2016] Shen, Shiqi, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany, August. Association for Computational Linguistics.

[Soricut and Echihabi2010] Soricut, Radu and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621. Association for Computational Linguistics.

[Sultan et al.2014] Sultan, Md, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

[Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

[Wang et al.2017] Wang, Longyue, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *CoRR*, abs/1704.04347.

[Williams1992] Williams, Ronald J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

[Wong and Kit2012] Wong, Billy T. M. and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July. Association for Computational Linguistics.

[Wu et al.2016] Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.