Machine Translation for Multilingual Summary Content Evaluation

Josef Steinberger and Marco Turchi
Joint Research Centre,
European Commission,
Via E. Fermi 2749,
21027 Ispra (VA), Italy
[name].[surname]@jrc.ec.europa.eu

Abstract

The multilingual summarization pilot task at TAC'11 opened a lot of problems we are facing when we try to evaluate summary quality in different languages. The additional language dimension greatly increases annotation costs. For the TAC pilot task English articles were first translated to other 6 languages, model summaries were written and submitted system summaries were evaluated. We start with the discussion whether ROUGE can produce system rankings similar to those received from manual summary scoring by measuring their correlation. We study then three ways of projecting summaries to a different language: projection through sentence alignment in the case of parallel corpora, simple summary translation and summarizing machine translated articles. Building such summaries gives opportunity to run additional experiments and reinforce the evaluation. Later, we investigate whether an evaluation based on machine translated models can perform close to an evaluation based on original models.

1 Introduction

Evaluation of automatically produced summaries in different languages is a challenging problem for the summarization community, because human efforts are multiplied to create model summaries for each language. Unavailability of parallel corpora suitable for news summarization adds even another annotation load because documents need to be translated to other languages. At the last TAC'11 campaign, six research groups spent a lot of work on creating evaluation resources in seven languages (Giannakopoulos et al., 2012). Thus compared to the monolingual evaluation, which requires writing model summaries and evaluating outputs of each system by hand, in the multilingual setting we need to obtain translations of all documents into the target language, write model summaries and evaluate the peer summaries for all the languages.

In the last fifteen years, research on Machine Translation (MT) has made great strides allowing human beings to understand documents written in various languages. Nowadays, on-line services such as *Google Translate* and *Bing Translator*[1] can translate text into more than 50 languages showing that MT is not a pipe-dream.

In this paper we investigate how machine translation can be plugged in to evaluate quality of summarization systems, which would reduce annotation efforts. We also discuss projecting summaries to different languages with the aim to reinforce the evaluation procedure (e.g. obtaining additional peers for comparison in different language or studying their language-independence).

This paper is structured as follows: after discussing the related work in section 2, we give a short overview of the TAC'11 multilingual pilot task (section 3). We compare average model and system manual scores and we also study ROUGE correlation to the manual scores. We run our experiments on a subset of languages of the TAC multilingual task corpus (English, French and Czech). Section 4 introduces our translation system. We mention its

[1] http://translate.google.com/ and http://www.microsofttranslator.com/

Proceedings of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pages 19–27,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

translation quality for language pairs used later in this study. Then we move on to the problem of projecting summaries to different languages in section 5. We discuss three approaches: projecting summary through sentence alignment in a parallel corpus, translating a summary, and summarizing translated source texts. Then, we try to answer the question whether using translated models produces similar system rankings as when using original models (section 6), accompanied by a discussion of discriminative power difference and cross-language model comparison.

2 Related work

Attempts of using machine translation in different natural language processing tasks have not been popular due to poor quality of translated texts, but recent advance in Machine Translation has motivated such attempts. In Information Retrieval, Savoy and Dolamic (2009) proposed a comparison between Web searches using monolingual and translated queries. On average, the results show a limited drop in performance, around 15% when translated queries are used.

In cross-language document summarization, Wan et al. (2010) and Boudin et al. (2010) combined the MT quality score with the informativeness score of each sentence to automatically produce summary in a target language. In Wan et al. (2010), each sentence of the source document is ranked according to both scores, the summary is extracted and then the selected sentences translated to the target language. Differently, in Boudin et al. (2010), sentences are first translated, then ranked and selected. Both approaches enhance the readability of the generated summaries without degrading their content.

Automatic evaluation of summaries has been widely investigated in the past. In the task of cross-lingual summarization evaluation Saggion et al. (2002) proposed different metrics to assess the content quality of a summary. Evaluation of summaries without the use of models has been introduced by Saggion et al. (2010). They showed that substituting models by full document in the computation of the Jensen-Shannon divergence measure can produce reliable rankings. Yeloglu et al. (2011) concluded that the pyramid method partially reflects the manual inspection of the summaries and ROUGE can only be used when there is a manually created summary. A method, and related resources, which allows saving precious annotation time and that makes the evaluation results across languages directly comparable was introduced by Turchi et al. (2010). This approach relies on parallel data and it is based on the manual selection of the most important sentences in a cluster of documents from a sentence-aligned parallel corpus, and by projecting the sentence selection to various target languages.

Our work addresses the same problem of reducing annotation time and generating models, but from a different prospective. Instead of using parallel data and annotation projection or full documents, we investigate the use of machine translation at different level of summary evaluation. While the aproach of Turchi et al. (2010) is focussed on sentence selection evaluation our strategy can also evaluate generative summaries, because it works on summary level.

3 TAC'11 Multilingual Pilot

The Multilingual task of TAC'11 (Giannakopoulos et al., 2012) aimed to evaluate the application of (partially or fully) language-independent summarization algorithms on a variety of languages. The task was to generate a representative summary (250 words) of a set of 10 related news articles.

The task included 7 languages (English, Czech, French, Hebrew, Hindi, Greek and Arabic). Annotation of each language sub-corpus was performed by a different group. English articles were manually translated to the target languages, 3 model summaries were written for each topic.

8 groups (systems) participated in the task, however, not all systems produced summaries for all languages. In addition there were 2 baselines: Centroid Baseline – the start of the centroid article and GA Topline – summary based on genetic algorithm using model summary information, which should serve as an upper bound.

Human annotators scored each summary, both models and peers, on the 5-to-1 scale (5 = the best, 1 = the worst) – human grades. The score corresponds to the overall responsiveness of the main TAC task – equal weight of content and readability. [2]

[2] In this article we focus on raw human grades. The task

	English	French	Czech	**average**	English	French	Czech	**average**
	Manual grades				Manual grades			
average model	4.06	4.03	4.73	**4.27**	4.06	4.03	4.73	**4.27**
average peer	2.73	2.18	2.56	**2.50**	2.73	2.18	2.56	**2.50**
	ROUGE-2				ROUGE-SU4			
average model	.194	.222	.206	**.207**	.235	.255	.237	**.242**
average peer	.139	.167	.182	**.163**	.183	.207	.211	**.200**
correlation to manual grading – peers and models not stemmed								
peers only	.574	.427	.444	**.482**	.487	.362	.519	**.456**
(p-value)	(< .1)							
models & peers	.735	.702	.484	**.640**	.729	.703	.549	**.660**
(p-value)	(< .01)	(< .02)			(< .02)	(< .02)		
correlation to manual grading – peers and models stemmed								
Peers only	.573	.445	.500	**.506**	.484	.336	.563	**.461**
(p-value)	(< .1)							
models & peers	.744	.711	.520	**.658**	.723	.700	.636	**.686**
(p-value)	(< .01)	(< .01)			(< .02)	(< .02)	(< .1)	

Table 1: Average ROUGE-2 and ROUGE-SU4 scores for models and peers, and their correlation to the manual evaluation (grades). We report levels of significance (p) for two-tailed test. Cells with missing p-values denote non-significant correlations ($p > .1$).

3.1 Manual Evaluation

When we look at the manually assigned grades we see that there is a clear gap between human and automatic summaries (see the first two rows in table 1). While the average grade for models were always over 4, peers were graded lower by 33% for English and by 54% for French and Czech. However, there were 5 systems for English and 1 system for French which were not significantly worse than at least one model.

3.2 ROUGE

The first question is: can an automatic metric rank the systems similarly as manual evaluation? This would be very useful when we test different configurations of our systems, in which case manual scoring is almost impossible. Another question is: can the metric distinguish well the gap between models and peers? ROUGE is widely used because of its simplicity and its high correlation with manually assigned content quality scores on overall system rankings, although per-case correlation is lower.

We investigated how the two most common ROUGE scores (ROUGE-2 and ROUGE-SU4) correlate with human grades. Although using n-grams with n greater than 1 gives limited possibility to reflect readability in the scores when compared to reference summaries, ROUGE is considered mainly as a content evaluation metric. Thus we cannot expect a perfect correlation because half of the grade assigned by humans reflects readability issues. ROUGE could not also evaluate properly the baselines. The centroid baseline contains a continuous text (the start of an article) and it thus gets higher grades by humans because of its good readability, but from the ROUGE point of view the baseline is weak. On the other hand, the topline used information from models and it is naturally more similar to them when evaluated by ROUGE. Its low readability ranked it lower in the case of human evaluation. Because of these problems we include in the correlation figures only the submitted systems, neither the baseline nor the topline.

Table 1 compares average model and peer ROUGE scores for the three analyzed languages. It adds two correlations[3] to human grades: for *models+systems* and for *systems only*. The first case should answer the question whether the automatic metric can distinguish between human and automatic summaries. The second settings could show

overview paper (Giannakopoulos et al., 2012) discusses, in addition, scaling down the grades of shorter summaries to avoid assigning better grades to shorter summaries.

[3] We used the classical Pearson correlation.

whether the automatic metric accurately evaluates the quality of automatic summaries. To ensure a fair comparison of models and non-models, each model summary is evaluated against two other models, and each non-model summary is evaluated three times, each time against a different couple of models, and these three scores are averaged out (the jackknifing procedure).[4] The difference of the model and system ROUGE scores is significant, although it is not that distinctive as in the case of human grades. The distinction results in higher correlations when we include models than in the more difficult *systems only* case. This is shown by both correlation figures and their confidence. The only significant correlation for the *systems only* case was for English and ROUGE-2. Other correlations did not cross the 90% confidence level. If we run ROUGE for morphologically rich languages (e.g. Czech), stemming plays more important role than in the case of English. In the case of French, which stands in between, we found positive effect of stemming only for ROUGE-2. ROUGE-2 vs. ROUGE-SU4: for English and French we see better correlation with ROUGE-2 but the free word ordering in Czech makes ROUGE-SU4 correlate better.

4 In-house Translator

Our translation service (Turchi et al., 2012) is based on the most popular class of Statistical Machine Translation systems (SMT): the Phrase-Based model (Koehn et al., 2003). It is an extension of the noisy channel model introduced by Brown et al. (1993), and uses phrases rather than words. A source sentence f is segmented into a sequence of I phrases $f^I = \{f_1, f_2, \ldots f_I\}$ and the same is done for the target sentence e, where the notion of phrase is not related to any grammatical assumption; a phrase is an n-gram. The best translation e_{best} of f is obtained by:

$$e_{best} = arg \max_e p(e|f) = arg \max_e p(f|e) p_{LM}(e)$$

$$= arg \max_e \prod_{i=1}^{I} \phi(f_i|e_i)^{\lambda_\phi} d(a_i - b_{i-1})^{\lambda_d} \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \ldots e_{i-1})^{\lambda_{LM}}$$

where $\phi(f_i|e_i)$ is the probability of translating a phrase e_i into a phrase f_i. $d(a_i - b_{i-1})$ is the distance-based reordering model that drives the system to penalize significant word reordering during translation, while allowing some flexibility. In the reordering model, a_i denotes the start position of the source phrase that is translated into the ith target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)$th target phrase. $p_{LM}(e_i|e_1 \ldots e_{i-1})$ is the language model probability that is based on the Markov's chain assumption. It assigns a higher probability to fluent/grammatical sentences. λ_ϕ, λ_{LM} and λ_d are used to give a different weight to each element. For more details see (Koehn et al., 2003). In this work we use the open-source toolkit Moses (Koehn et al., 2007).

Furthermore, our system takes advantage of a large in-house database of multi-lingual named and geographical entities. Each entity is identified in the source language and its translation is suggested to the SMT system. This solution avoids the wrong translation of those words which are part of a named entity and also common words in the source language, (e.g. "Bruno Le Maire" which can be wrongly translated to "Bruno Mayor"), and enlarges the source language coverage.

We built four models covering the following language pairs: En-Fr, En-Cz, Fr-En and Cz-En. To train them we use the freely available corpora: Europarl (Koehn, 2005), JRC-Acquis (Steinberger et al., 2006), CzEng0.9 (Bojar and Žabokrtský, 2009), Opus (Tiedemann, 2009), DGT-TM[5] and News Corpus (Callison-Burch et al., 2010), which results in more than 4 million sentence pairs for each model. Our system was tested on the News test set (Callison-Burch et al., 2010) released by the organizers of the 2010 Workshop on Statistical Machine Translation. Performance was evaluated using the Bleu score (Papineni et al., 2002): En-Fr 0.23, En-Cz 0.14, Fr-En 0.26 and Cz-En 0.22. The Czech

[4] In our experiments we used the same ROUGE settings as at TAC. The summaries were truncated to 250 words. For English we used the Porter stemmer included in the ROUGE package, for Czech the aggressive version from `http://members.unine.ch/jacques.savoy/clef/index.html` and for French `http://jcs.mobile-utopia.com/jcs/19941\_FrenchStemmer.java`.

[5] `http://langtech.jrc.it/DGT-TM.html`

language is clearly more challenging than French for the SMT system, this is due to the rich morphology and the partial free word order. These aspects are more evident when we translate to Czech, for which we have poor results.

5 Mapping Peers to Other Languages

When we want to generate a summary of a set of articles in a different language we have different possibilities. The first case is when we have articles in the target language and we run our summarizer on them. This was done in the Multilingual TAC task. If we have parallel corpora we can take advantage of projecting a sentence-extractive summary from one language to another (see Section 5.1).

If we do not have the target language articles we can apply machine translation to get them and run the summarizer on them (see Section 5.3). If we miss a crucial resource for running the summarizer for the target language we can simply translate the summaries (see Section 5.2).

In the case of the TAC Multilingual scenario these projections can also give us summaries for all languages from the systems which were applied only on some languages.

5.1 Aligned Summaries

Having a sentence-aligned (parallel) corpus gives access to additional experiments. Because the current trend is still on the side of pure sentence extraction we can investigate whether the systems select the same sentences across the languages. While creating the TAC corpus each research group translated the English articles into their language, thus the resulting corpus was close to be parallel. However, sentences are not always aligned one-to-one because a translator may decide, for stylistic or other reasons, to split a sentence into two or to combine two sentences into one. Translations and original texts are never perfect, so that it is also possible that the translator accidentally omits or adds some information, or even a whole sentence. For these reasons, aligners such as Vanilla[6], which implements the Gale and Church algorithm (Gale and Church, 1994), typically also allow two-to-one, one-to-two, zero-to-one and one-to-zero sentence alignments. Alignments other than one-to-one thus present a challenge for the method of aligning two text, in particular one-to-two and two-to-one alignments. We used Vanilla to align Czech and English article sentences, but because of high error rate we corrected the alignment by hand.

The English summaries were then aligned to Czech (and the opposite direction as well) according to the following approach. Sentences in a source language system summary were split. For each sentence we found the most similar sentence in the source language articles based on 3-gram overlap. The alignment information was used to select sentences for the target language summary. Some simplification rules were applied: if the most similar sentence found in the source articles was aligned with more sentences in the target language articles, all the projected sentences were selected (one-to-two alignment); if the sentence to be projected covered only a part of sentences aligned with one target language sentence, the target language sentence was selected (two-to-one alignment).

The 4th row in table 2 shows average peer ROUGE scores of aligned summaries.[7] When comparing the scores to the peers in original language (3rd row) we notice that the average peer score is slightly better in the case of English (cz→en projection) and significantly worse for Czech (en→cz projection) indicating that Czech summaries were more similar to English models than English summaries to Czech models.

Having the alignment we can study the overlap of the same sentences selected by a summarizer in different languages. The peer average for the en-cz language pair was 31%, meaning that only a bit less than one third of sentences was selected both to English and Czech summaries by the same system. The percentage differed a lot from a summarizer to another one, from 13% to 57%. This number can be seen as an indicator of summarizer's language independence.

However, the system rankings of aligned summaries did not correlate well with human grades. There are many inaccuracies in the alignment summary creation process. At first, finding the sentence

[6] `http://nl.ijs.si/telri/Vanilla/`

[7] Models are usually not sentence-extractive and thus aligning them would not make much sense.

	ROUGE-2					ROUGE-SU4				
	fr→en	cz→en	en→fr	en→cz	**avg.**	fr→en	cz→en	en→fr	en→cz	**avg.**
	average ROUGE scores									
orig. model	.194	.194	.222	.206	**.207**	.235	.235	.255	.237	**.242**
transl. model	.128	.162	.187	.123	**.150**	.184	.217	.190	.160	**.188**
orig. peer	.139	.139	.167	.182	**.163**	.183	.183	.207	.211	**.200**
aligned peer		.148		.146	**.147**		.175		.140	**.180**
transl. peer	.100	.119	.128	.102	**.112**	.155	.174	.179	.140	**.162**
	correlation to source language manual grading for translated summaries									
peers only	.411	.483	.746	.456	**.524**	.233	.577	.754	.571	**.534**
(p-value)			(< .05)					(< .05)		
models & peers	.622	.717	.835	.586	**.690**	.581	.777	.839	.620	**.704**
(p-value)	(< .05)	(< .05)	(< .01)	(< .1)		(< .05)	(< .02)	(< .01)	(< .05)	
	correlation to target language manual grading for translated summaries									
peers only	.685	.708	.555	.163	**.528**	.516	.754	.529	.267	**.517**
(p-value)	(< .1)									

Table 2: ROUGE results of translated summaries, evaluated against target language models (e.g., cz→en against English models).

in the source data that was probably extracted is strongly dependent on the sentence splitting each summarizer used. At second, alignment relations different from one-to-one results in selecting content with different length compared to the original summary. And since ROUGE measures recall, and truncates the summaries, it introduces another inaccuracy. There were also relations one-to-zero (sentences not translated to the target language). In that case no content was added to the target summary.

5.2 Translated Summaries

The simplest way to obtain a summary in a different language is to apply machine translation software on summaries. Here we investigate (table 2) whether machine translation errors affect the system order by correlation to human grades again. In this case we have two reference human grade sets: one for the source language (from which we translate) and one for the target language (to which we translate). Since there were different models for each language we can include models only in computing the correlation against source language manual grading.

At first, we can see that ROUGE scores are affected by the translation errors. Average model ROUGE-2 score went down by 28% and average peer ROUGE-2 by 31%. ROUGE-SU4 seems to be more robust to deal with the translation errors: models went down by 21%, peers by 19%. The gap between models and peers is still distinguishable, system ranking correlation to human grades holds similar levels although less statistically significant correlations can be seen. Clearly, quality of the translator affects these results because our worst translator (en→cz) produced the worst summaries. Correlation to the source language manual grades indicates how the ranking of the summarizers is affected (changed) by translation errors. For example it compares ranking for English based on manual grades with ranking computed on the same summaries translated from English to French. The second scenario (correlation to target language scores) shows how similar is the ranking of summarizers based on translated summaries with the target language ranking based on original summaries. If we omit translation inaccuracies, low correlation in the latter case indicates qualitatively different output of participating peers (e.g. en and cz summaries).

5.3 Summarizing Translated Articles

To complete the figure we tested the configuration in which we first translate the full articles to the target language and then apply a summarizer. As we have at disposal an implementation of system 3 from the TAC multilingual task we used it on 4 translated document sets (en→cz, cz→en, fr→en, en→fr). This system was the best according to human grades in all three discussed languages.

method	ROUGE-2	ROUGE-SU4
en	.177	.209
cz → en alignment	.200	.235
cz → en translation	.142	.194
en from (cz → en source translation)	**.132**	**.181**
fr → en translation	.120	.172
en from (fr → en source translation)	**.129**	**.185**
fr	.214	.241
en → fr translation	.167	.212
fr from (en → fr source translation)	**.156**	**.202**
cz	.204	.225
en → cz alignment	.176	.196
en → cz translation	.115	.150
cz from (en → cz source translation)	**.138**	**.178**

Table 3: ROUGE results of different variants of summaries produced by system 3. The first line shows the ROUGE scores of the original English summaries submitted by system 3. The second line gives average scores of the cz→en aligned summaries (see Section 5.1), in the 3rd and 5th lines there are figures of cz→en and fr→en translated summaries, and 4th and 6th lines show scores when the summarizer was applied on translated source texts (cz→en and fr→en). Similarly, lines further down show performance for French and Czech.

The system is based on the latent semantic analysis framework originally proposed by Gong and Liu (2002) and later improved by J. Steinberger and Ježek (2004). It first builds a term-by-sentence matrix from the source articles, then applies Singular Value Decomposition (SVD) and finally uses the resulting matrices to identify and extract the most salient sentences. SVD finds the latent (orthogonal) dimensions, which in simple terms correspond to the different topics discussed in the source (for details see (Steinberger et al., 2011)).

Table 3 shows all results of summaries generated by the summarizer. The first part compares English summaries. We see that when projecting the summary through alignment from Czech, see Section 5.1, a better summary was obtained. When using translation the summaries are always significantly worse compared to original (TAC) summaries, with the lowest performing en→cz translation. It is interesting that in the case of this low-performing translator it was significantly better to translate the source articles and to use the summarizer afterwards. The advantage of this configuration is that the core of the summarizer (LSA) treats all terms the same way, thus even English terms that were not translated work well for sentence selection. On the other hand, when translating the summary ROUGE will not match the English terms in Czech models.

6 Using Translated Models

With growing number of languages the annotation effort rises (manual creation of model summaries). Now we investigate whether we can produce models in one pivot language (e.g., English) and translate them automatically to all other languages. The fact that in the TAC corpus we have manual summaries for each language gives us opportunity to reinforce the evaluation by translating all model summaries to a common language and thus obtaining a larger number of models. This way we can also evaluate similarity among models coming from different languages and it lowers the annotators' subjectivity.

6.1 Evaluation Against Translated Models

Table 4 shows ROUGE figures when peers were evaluated against translated models. We discuss also the case when English peer summaries (and models as well) are evaluated against both French and Czech models translated to English. We can see again lower ROUGE scores caused by translation errors, however, there is more or less the same gap between peers and models and the correlation holds similar levels as when using the original target language models. Exceptions are using English models translated to French and Czech models translated to English in combination with the *systems only* correlation. If we used both French and Czech mod-

	ROUGE-2						ROUGE-SU4					
peers from	en			fr	cz	**avg.**	en			fr	cz	**avg.**
models tr. from	fr	cz	fr / cz	en	en		fr	cz	fr / cz	en	en	
average model	.144	.167	.155	.165	.144	**.155**	.207	.221	.206	.215	.190	**.208**
average peer	.110	.111	.104	.135	.125	**.117**	.170	.162	.153	.186	.172	**.169**
	correlation to target language manual grading											
peers only	.639	.238	.424	.267	.541	**.422**	.525	.136	.339	.100	.624	**.345**
(p-value)	$< .1$											
models & peers	.818	.717	.782	.614	.520	**.690**	.785	.692	.759	.559	.651	**.793**
(p-value)	$< .01$	$< .02$	$< .01$	$< .05$			$< .01$	$< .02$	$< .01$	$< .1$	$< .1$	

Table 4: ROUGE results of using translated model summaries, which evaluate both peer and model summaries in the particular language.

els translated to English, higher correlation of English peers with translated French models was averaged out by lower correlation with Czech models. And because the TAC Multilingual task contained 7 languages the experiment can be extended to using translated models from 6 languages. However, our results rather indicate that using the best translator is better choice.

Given the small scale of the experiment we cannot draw strong conclusions on discriminative power[8] when using translated models. However, our experiments indicate that by using translated summaries we are partly loosing discriminative power (i.e. ROUGE finds fewer significant differences between systems).

6.2 Comparing Models Across Languages

By translating both Czech and French models to English we could compare all models against each other. For each topic we had 9 models: 3 original English models, 3 translated from French and 3 from Czech. In this case we reached slightly better correlations for the *models+systems* case: ROUGE-2: .790, ROUGE-SU4: .762. It was mainly because of the fact that this time also *models only* rankings from ROUGE correlated with human grades (ROUGE-2: .475, ROUGE-SU4: .445). When we used only English models, the models ranking did not correlate at all (≈ -0.1). Basically, one English model was less similar to the other two, but it did not mean that it was worse which was shown by adding models from other languages. If we do not have enough reference summaries this could be a way to lower subjectivity in the evaluation process.

7 Conclusion

In this paper we discuss the synergy between machine translation and multilingual summarization evaluation. We show how MT can be used to obtain both peer and model evaluation data.

Summarization evaluation mostly aims to achieve two main goals a) to identify the absolute performance of each system and b) to rank all the systems according to their performances. Our results show that the use of translated summaries or models does not alter much the overall system ranking. It maintains a fair correlation with the source language ranking although without statistical significance in most of the *systems only* cases given the limited data set. A drop in ROUGE score is evident, and it strongly depends on the translation performance. The use of aligned summaries, which limits the drop, requires high quality parallel data and alignments, which are not always available and have a significant cost to be created.

The study leaves many opened questions: What is the required translation quality which would let us substitute target language models? Are translation errors averaged out when using translated models from more languages? Can we add a new language to the TAC multilingual corpus just by using MT having in mind lower quality (→ lower scores) and being able to quantify the drop? Experimenting with a larger evaluation set could try to find the answers.

[8] Discriminative power measures how successful the automatic measure is in finding the same significant differences between systems as manual evaluation.

References

O. Bojar and Z. Žabokrtský. 2009. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

F. Boudin, S. Huet, J.M. Torres-Moreno, and J.M. Torres-Moreno. 2010. A graph-based approach to cross-language multi-document summarization. *Research journal on Computer science and computer engineering with applications (Polibits)*, 43:113–118.

P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O.F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53. Association for Computational Linguistics.

W.A. Gale and K.W. Church. 1994. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19.

G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma. 2012. Tac 2011 multiling pilot overview. In *Proceedings of TAC'11*. NIST.

Y. Gong and X. Liu. 2002. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of ACM SIGIR*, New Orleans, US.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT summit*, volume 5.

K. Papineni, S. Roukos, T. Ward, and W.J. 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.

H. Saggion, D. Radev, S. Teufel, W. Lam, and S.M. Strassel. 2002. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In *Proceedings of LREC 2002*, pages 747–754.

H. Saggion, J.M. Torres-Moreno, I. Cunha, and E. SanJuan. 2010. Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1059–1067. Association for Computational Linguistics.

J. Savoy and L. Dolamic. 2009. How effective is google's translation service in search? *Communications of the ACM*, 52(10):139–143.

J. Steinberger and K. Ježek. 2004. Text summarization and singular value decomposition. In *Proceedings of the 3rd ADVIS conference*, Izmir, Turkey.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *Arxiv preprint cs/0609058*.

J. Steinberger, M. Kabadjov, R. Steinberger, H. Tanev, M. Turchi, and V. Zavarella. 2011. Jrcs participation at tac 2011: Guided and multilingual summarization tasks. In *Proceedings of the Text Analysis Conference (TAC)*.

J. Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of the Recent Advances in Natural Language Processing Conference*, volume 5, pages 237–248. John Benjamins Amsterdam.

M. Turchi, J. Steinberger, M. Kabadjov, and R. Steinberger. 2010. Using parallel corpora for multilingual (multi-document) summarisation evaluation. In *Proceedings of the Multilingual and Multimodal Information Access Evaluation Conference*, pages 52–63. Springer.

M. Turchi, M. Atkinson, A. Wilcox, B. Crawley, S. Bucci, R. Steinberger, and E. Van der Goot. 2012. Onts:optima news translation system. In *Proceedings of EACL 2012*, page 25.

X. Wan, H. Li, and J. Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926. Association for Computational Linguistics.

O. Yeloglu, E. Milios, and N. Zincir-Heywood. 2011. Multi-document summarization of scientific corpora. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 252–258. ACM.