SUMT: A Framework of Summarization and MT

Houda Bouamor **Behrang Mohit** **Kemal Oflazer**
Carnegie Mellon University
Doha, Qatar
hbouamor@qatar.cmu.edu, behrang@cmu.edu, ko@cs.cmu.edu

Abstract

We present a novel system combination of machine translation and text summarization which provides high quality summary translations superior to the baseline translation of the entire document. We first use supervised learning and build a classifier that predicts if the translation of a sentence has high or low translation quality. This is a reference-free estimation of MT quality which helps us to distinguish the subset of sentences which have better translation quality. We pair this classifier with a state-of-the-art summarization system to build an MT-aware summarization system. To evaluate summarization quality, we build a test set by summarizing a bilingual corpus. We evaluate the performance of our system with respect to both MT and summarization quality and, demonstrate that we can balance between improving MT quality and maintaining a decent summarization quality.

1 Introduction

Machine Translation (MT) has been championed as an effective technology for knowledge transfer from English to languages with less digital content. An example of such efforts is the automatic translation of English Wikipedia to languages with smaller collections. However, MT quality is still far from ideal for many of the languages and text genres. While translating a document, there are many poorly translated sentences which can provide incorrect context and confuse the reader. Moreover, some of these sentences are not as *informative* and could be summarized to make a more cohesive document. Thus, for tasks in which complete translation is not mandatory, MT can be effective if the system can provide a more informative subset of the content with higher translation quality.

In this work, we demonstrate a framework of MT and text summarization which replaces the baseline translation with a proper summary that has higher translation quality than the full translation. For this, we combine a state of the art English summarization system and a novel framework for prediction of MT quality without references.

Our research contributions are:

(a) We extend a classification framework for reference-free prediction of translation quality at the sentence-level.

(b) We incorporate MT knowledge into a summarization system which results in high quality translation summaries.

(c) For evaluation purposes, we conduct a bilingual manual summarization of a parallel corpus.[1]

Our English-Arabic system reads in an English document along with its baseline Arabic translation and outputs, as a summary, a subset of the Arabic sentences based on their informativeness and also their translation quality. We demonstrate the utility of our system by evaluating it with respect to both its MT and the summarization quality. For summarization, we conduct both reference-based and reference-free evaluations and observe a performance in the range of the state of the art system. Moreover, the translation quality of the summaries shows an important improvement against the baseline translation of the entire documents.

This MT-aware summarization can be applied to translation of texts such as Wikipedia articles.

[1]The bilingually summarized corpora could be found at: http://nlp.qatar.cmu.edu/resources/SuMT

International Joint Conference on Natural Language Processing, pages 270–278,
Nagoya, Japan, 14-18 October 2013.

For such domain-rich articles, there is a large variation of translation quality across different sections. An intelligent reduction of the translation tasks results in improved final outcome. Finally, the framework is mostly language independent and can be customized for different target languages and domains.

2 Related work

Our approach draws on insights from problems related to text summarization and also automatic MT evaluation. Earlier works on Arabic summarization in campaigns and competitions such as DUC (Litkowski, 2004) or Multi-Ling (Giannakopoulos et al., 2011) were focused on abstractive summarization which involves the generation of new sentences from the original document. The fluency of such generated summaries might not be perfect. However, having a noisy source language text for an MT system can degrade the translation quality dramatically. Thus, extractive summarization like our framework is more suitable for MT summarization. In retrospect our annotated Arabic-English summaries is a unique bilingual resource as most other Arabic-English summarization corpora (e.g. DUC) are abstractive summaries.

There has been a body of recent work on the reference-free prediction of translation quality both as confidence estimation metrics and also direct prediction of human judgment scores (Bojar et al., 2013; Specia, 2012) or the range of the BLEU score (Soricut and Echihabi, 2010; Mohit and Hwa, 2007). These works mostly use supervised learning frameworks with a rich set of source and target language features. Our binary classification of MT quality is closer to the classification system of Mohit and Hwa (2007) to estimate translation difficulty of phrases. However, there are several modifications such as the method of labeling, the focus on sentence level prediction and finally the use of a different metric for both the labeling and final evaluation (which reduces the metric bias). For learning features, we cumulatively explore and optimize most of the reported features, and add document-level features to model the original document properties for each sentence.

Another line of research constrained by the lack of access to reference translations is confidence estimation for MT which is simply system's judgments of its own performance. The confidence measure is a score for N-grams (substrings of the hypothesis) which are generated by an MT system. Confidence estimation is performed at the word level (Blatz et al., 2003) or phrase level (Zens and Ney, 2006). The measure is based on feature values extracted from the underlying SMT system and also its training data. There are many overlaps between the features used in confidence estimation and the MT quality prediction. However, the two frameworks use different learning methods. Confidence estimation systems usually do not have gold standard data and are mostly a linear interpolation of a large group of scores. In contrast, MT quality predictors such as our framework usually use supervised learning and rely on gold standard data.

Text summarization has been successfully paired with different NLP applications such as MT in cross-language summarization. Wan et al. (2010) and Boudin et al. (2011) proposed cross-language summarization frameworks in which for each sentence, in a source language text, an MT quality and informativeness scores are combined to produce summary in a target language (Chinese and French, respectively). In the latter, sentences are first translated, ranked and then summaries are generated. Differently, in Wan et al. (2010), each sentence of the source document is ranked based on an *a posteriori* combination of both scores. The selected summarized sentences are then translated to the target language using Google Translate. In contrast, we go a step further and design a hybrid approach in which we incorporate our MT quality classifier into the state-of-the-art summarization system. Moreover, we use SMT beyond a black-box and actually incorporate its knowledge in prediction of translation quality along with other set of features such as document-related and Arabic morphological information. Finally we demonstrate that our approach outperforms Wan et al. (2010) by conducting automatic evaluation of MT and summarization systems.

3 An overview of the approach

Given a source language document and its translation, our aim is to find a high quality summary of the translation with a quality superior to translating the entire document. Figure 1 illustrates an overview of our framework composed of the following major components: (a) a standard

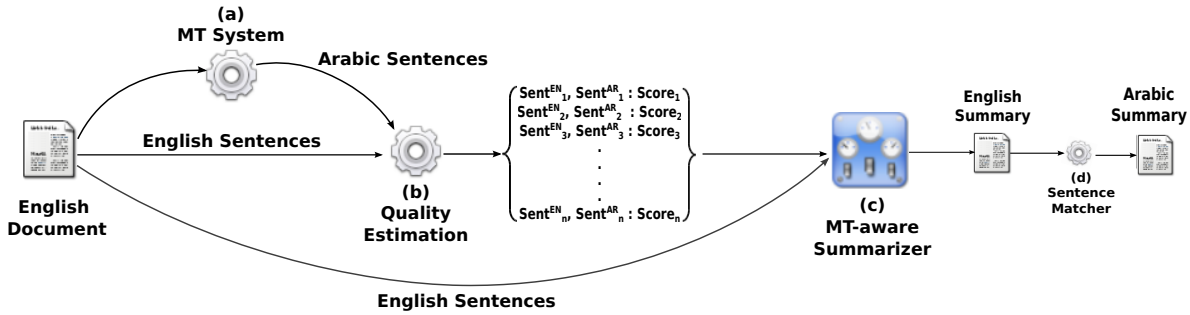

Figure 1: An overview of our MT-aware summarization system

SMT system; (b) our reference-free MT quality estimation system; (c) our MT-aware summarization system; and (d) the English-Arabic sentence matcher. Our system provides the translation summary through the following steps:

1. We translate an input English document into Arabic using the SMT system.

2. The quality estimation system (b) predicts if a translated sentence has high or low translation quality and assigns a quality score to each sentence.

3. We summarize the English document using our MT-aware summarization system (c), which incorporates the translation quality score (output of (b)) in its sentence selection process.

4. We produce the final Arabic translation summary by matching the English summarized sentences with the corresponding Arabic translations (d).

5. We automatically evaluate the quality of our MT-aware summarization system using MT and summarization metrics.

Our contributions are mainly related to the second and third components which will be discussed in Sections 4 and 5.

4 Reference-free quality estimation of MT

Our system needs to estimate the translation quality without access to the Arabic reference translations. The reference-free MT evaluation has been investigated extensively in the past decade. A valuable gold-standard resource for many of these studies are human judgment scores which have been developed in evaluation programs like NIST, and workshops such as WMT (Koehn and Monz, 2006). Since such human judgments do not exist for English to Arabic translations, we adapt the framework of Mohit and Hwa (2007) for predicting the translation quality. This framework uses only reference translations and the automatic MT evaluation scores to create labeled data for training a classifier. The binary classifier reads in a source language sentence, with its automatically obtained translation and predicts if the target sentence has *high* or *low* translation quality. We describe details of this framework in the following section.

4.1 Labeling gold-standard data

In order to train the binary classifier, we need gold standard data with English source sentences labeled as having high or low translation quality when translated into Arabic. For this labeling, we estimate translation quality by the Translation Edit Rate TER metric (Snover et al., 2006).[2] We deliberately use two different metrics for gold standard labeling (TER) and the final MT evaluation (using BLEU (Papineni et al., 2002)) to reduce the bias that a metric can introduce to the framework. In this task, we use a parallel corpus that is composed of a set of documents. We automatically translate each document and label its sentences based on the following procedure:

(a) Measure the TER score of the document against its reference translation.

(b) For each sentence within the document, measure its TER score: If this score is higher than the document score, it has *low* translation quality. Otherwise it has *high* translation quality.

[2]This automatic labeling framework exempts us from the manual labeling of translation quality like Wan et al. (2010).

This provides a simple estimate of the translation quality for a source language sentence relative to the document that it belongs to. This document-level relevance is a deliberate choice to build a classifier that ranks the translation quality of a sentence with respect to other sentences in the document (similar to a summarization system). We also note that the quality labeling is obviously non-absolute and relative to the specific SMT engine used in this work.

4.2 MT quality classifier

We use a Support Vector Machine (SVM) classifier and exploit a rich set of features to represent a source language sentence and its translation. We use the default configuration with a linear kernel function. In order to estimate a score for the translation quality, we use a normalized form of the classifier's score for each sentence. The score is the distance from the separating surface and is proper estimate of the intensity of the class label.

4.3 Learning features

We use a suite of features that have been extensively used in works related to translation quality estimation. We adapt the feature extraction procedure from the Quest framework (Specia et al., 2013) to our English-Arabic translation setup, and extract the following groups of features:

General features: For each sentence we use different features modeling its length in terms of words, the ratio of source-target length, source-target punctuation marks, numerical characters, and source-target content words.

Language model scores: The likelihood of a target language (Arabic) sentence can be a good indicator of its grammaticality. In our experiments, we used the SRILM toolkit (Stolcke, 2002) to build 5-gram language model using the LDC Arabic Gigaword corpus. We then, apply this model to obtain log-likelihood and perplexity scores for each sentence.

MT-based scores: We extract a set of features from the generated MT output. These include the absolute number and the ratios of out of vocabulary terms and the ratio of Arabic detokenization that is performed on the Arabic MT output.

Morphosyntactic features: We use features to model the difference of sequences of POS tags for a pair of source-target sentences. These features measure the POS preservation in the translation process (e.g. measuring if the proper nouns in the source sentence are kept and also translated as proper nouns in Arabic). We compute the absolute difference between the number of different POS tags. The source and target sentences are tagged respectively using the TreeTagger (Schmid, 1994) and AMIRA (Diab, 2009) toolkits. We also, indicate the percentage of nouns, verbs, proper nouns in the source and target sentences.

Document-level features: We extend Mohit and Hwa (2007) framework by incorporating a set of document-level features (in addition to the sentence-level ones) which scales the sentence's classification relative to its document. In a linear model, these document-level features rescale and shift the feature space relative to the given document which helps us to classify the sentence with respect to the document. These features consist of the average of the sentence-level features described above.

5 MT-aware Summarization

We pair summarization and MT (SUMT) by including information about the MT quality into the summarization system. Our MT-aware summarizer focuses on the linguistic and translation quality of a given sentence, as well as its position, length, and the content in its sentence ranking procedure. The main goal of this system is to obtain an informative summary of a source document with an improved translation quality that could replace the complete, yet less fluent translation of the document.

We explore various configurations and find the sweet spot of the translation and summarization qualities in the system illustrated in Figure 1. This includes converting the MEAD summarizer into an MT-aware summarization framework by including information from the classifier into the sentence ranking procedure.

5.1 The MEAD Summarization system

In our experiments, the summary for each document is generated using MEAD (Radev et al., 2004), a state-of-the-art single- and multi-document summarization system. MEAD has been widely used both as a platform for developing summarization systems and as a baseline system for testing novel summarizers. It is a

centroid-based extractive summarizer which selects the most important sentences from a sequence of sentences based on a linear combination of three parameters: the sentence length, the centroid score and the position score (Radev et al., 2001). MEAD also employs a cosine reranker to eliminate redundant sentences. We create summaries at 50% length (a fixed ratio for all documents) using MEAD's default configurations.

5.2 SUMT system

Our MT-aware summarizer (SUMT) represents an approach of adapting the basic sentence scoring/ranking approach of MEAD. We extend the default MEAD sentence ranking procedure by incorporating information about the translation quality of the sentence. This score is provided by our SVM-based classifier. The selected sentences generally correspond to those having high translation quality (estimated by TER).

Typically, the ranking score of a sentence is defined by a linear combination of the weighted sentence position, centroid and length scores. We used the default weights defined for each feature in the default version of MEAD. The additional quality feature weight is optimized automatically towards the improvement of BLEU, using a held-out development set of documents. Finally, sentences in each document are ranked based on the final obtained score. In this work, we take a hard 50% summarization ratio which is applied to MEAD, SUMT and our gold standard summaries.

6 Experimental Setup

In this section, we explain details of the data and the general setting for different components of our system.

6.1 Translation and Summarization Corpora

For our experiments, we use the standard English-Arabic NIST test corpora which are commonly used MT evaluations.[3] We use the documents provided in NIST 2008 and 2009 for the training and development, and those in the NIST 2005 for testing. Each collection contains an Arabic and four English reference translations. Since we work on English to Arabic translation, we only use the first translation as the reference.

[3]All of the different MT corpora can be accessed from Linguistic Data Consortium (LDC).

6.2 Annotation of gold-standard summaries

The automatic summarization should be able to reduce the complexity of documents length wise, while keeping the essential information from the original documents like important events, person names, location, organizations and dates. In order to evaluate the quality of the summaries, we conducted a bilingual summarization of our test corpus (the NIST 2005). This parallel corpus is composed of 100 parallel documents containing each in average 10 sentences. We asked two native speakers (one per language) to summarize each side of the corpora independent of each other and independent of the MT output. We set a hard 50% ratio for annotators to choose approximately half of the sentences per document. Annotators followed a brief guideline to completely understand the entire document and examine and select summary sentences based on the following criteria: (a) Being informative with respect to the main story and the topic (b) minimizing the redundancy of information (c) preserving key information such as the named entities and dates. We obtain as inter-judge agreement a value of $\kappa = 0.61$ corresponding to a moderate agreement according to the literature.[4]

6.3 MT Setup

The baseline MT system is the open-source MOSES phrase-based decoder trained on a standard English-Arabic parallel corpus. This 18 million word parallel corpus consists of the non-UN parts of the NIST corpus distributed by the LDC. We perform the standard preprocessing and tokenization on the English side using simple punctuation-based rules. We also use the MADA+TOKAN morphological analyzer (Habash et al., 2009) to preprocess and tokenize the Arabic side of the corpus. The corpus is word-aligned using the standard setting of GIZA++ and the grow-diagonal-final heuristic of MOSES. We use the 5-gram language model with modified Kneser-Ney smoothing. The language model for our system is trained using the LDC Arabic Gigaword corpus. A set of 500 sentences is used to tune the decoder parameters using the MERT (Och, 2003). After decoding, we use the El Kholy and Habash (2010) Arabic deto-

[4]This Cohen's kappa value is obtained using the MEAD evaluation tool designed to assess the agreement between two summaries.

kenization framework to prepare the Arabic output for evaluation.

6.4 MT-quality classifier

We use the models described in Section 4 to build a Support Vector Machine (SVM) binary classifier using the LIBSVM package (Chang and Lin, 2011). To train our classifier we use a total of 2670 sentence pairs extracted from 259 documents of NIST 2008 and 2009 data sets. The sentences are labeled following our TER-based procedure. The automatic labeling procedure (section 4.1) enforces a rough 50-50 high and low quality translations. Thus, we obtaine 1370 negative examples and 1363 positive ones. For all tests, we use a set of 100 documents from the NIST 2005 test set, containing 1056 sentences.

7 Evaluation and results

We experimented with different configurations of the MT and the summarization system with the goal of achieving a balanced performance in both dimensions. We reached the sweet spot of performance in both dimensions in our MT-aware summarization system in which we achieved major (over 4 points BLEU score) improvements while maintaining an acceptable summarization quality. In the following we discuss the performance of the MT and summarization systems.

7.1 MT evaluation

Table 1 presents MT quality for the baseline system and different summarization frameworks measured by BLEU, TER and METEOR (Lavie and Agarwal, 2007) scores.[5]

The remaining MT experiments are conducted on summarized documents. These include summaries provided by: (a) a length-based baseline system that simply chooses the subset of sentences with the shortest length (**Length**); (b) the state of the art MEAD summarizer (**MEAD**); (c) our MT quality estimation classifier (**Classifier**); (d) a linear interpolation of informativeness and MT quality scores in the spirit of Wan et al. (2010) (**Interpol**)[6]; (e) our MT-aware summarizer (**SuMT**); and (f) an oracle classifier which chooses the subset of sentences with the highest translation quality (**Oracle**). This oracle provides an upper bound estimate of room that we have to improve translation quality of the summaries.

	BLEU	**TER**	**METEOR**
Baseline	*27.52*	*58.00*	*28.51*
Length	26.33	58.13	27.81
MEAD	28.42	55.00	28.82
Classifier	31.36	52.00	29.22
Interpol	28.45	55.00	29.05
SuMT	**32.12**	**51.00**	**30.48**
Oracle	34.75	47.00	32.42

Table 1: A comparison of MT quality for full and summarized documents.

We are aware that the comparison of the MT baseline system with these summarization systems is not a completely fair comparison as the test sets are not comparable. However, with a ballpark comparison of the baseline (for full documents) with the summarized documents, we demonstrate the average range of improvement in translation quality. Moreover, we compare different summarization systems with each other to reach the best combination of MT and summarization quality.

We set a 50% summarization ratio in all experiments and also in creation of the gold-standard to create similar comparable conditions. For example, for evaluating our quality estimation classifier as a summarizer, we filter out the bottom 50% of the sentences (based on their classification scores) for each document and evaluate the translation quality of the top 50% Arabic translation sentences.

The MT results for the MEAD summarizer indicate that summarization of MT does not necessarily improve MT quality. In contrast, the comparison between the baseline, the oracle summarizer and SuMT system demonstrates a major improvement in MT quality that is competitive with the oracle summarizer (an improvement of almost +5 BLEU scores). The results given in Table 1 show also that our system produce better MT quality sentences than Interpol (+4.67 BLEU points). This could be explained by the higher weight assigned to the informativeness score in the linear interpolation. In the following sections we demonstrate that we maintain a decent summarization quality while we achieve these MT improvements.

[5] Our English to Arabic baseline system shows a performance in the ballpark of the reported score for the state of the art systems (e.g. El Kholy and Habash (2010)).

[6] The overall score of a sentence is defined as follows: $score = (1 - \lambda) * InfoScore + \lambda * TransScore$ where $\lambda = 0.3$ and $TransScore$ and $InfoScore$ denote the MT quality score and the informativeness score of a sentence.

	English					*Arabic*				
	Length	**MEAD**	**Classifier**	**Interpol**	**SuMT**	**Length**	**MEAD**	**Classifier**	**Interpol**	**SuMT**
ROUGE-1	54.21	**75.93**	67.41	73.72	72.51	36.01	45.66	44.94	45.33	**46.43**
ROUGE-2	38.15	**67.77**	56.72	66.01	62.83	15.19	22.83	22.23	22.46	**23.28**
ROUGE-SU4	38.99	**67.96**	57.03	54.14	63.17	15.81	23.56	23.09	20.33	**24.07**
ROUGE-L	51.77	**74.92**	65.92	72.79	71.17	33.74	43.20	42.33	42.81	**43.84**

Table 2: ROUGE F-Scores for different summarization systems providing 50% length for English and Arabic summaries for each document.

7.2 Model-based summarization evaluation

We evaluate the quality of our summarization systems for both English and Arabic. We first focus on English summaries generated using different summarization configurations, and then evaluate the quality of Arabic summaries obtained by matching the English summarized sentences with the corresponding Arabic translations. It is not surprising that summarizing a noisy Arabic MT output would not produce high quality Arabic summaries. Instead, we use the parallel corpus to project the summarization from the source language (English) to the corresponding Arabic translations.

For evaluating our summarization systems, we use ROUGE (Lin, 2004), a metric based on n-gram similarity scores between a model summary generated by human and an automatically generated peer summary. We use the ROUGE-1, ROUGE-2, ROUGE-SU4 and ROUGE-L F-scores with the two human summaries described in Section 6 as models.[7] We use the same parameters and options in ROUGE as in the DUC 2007 summarization evaluation task.[8] Table 2 presents the ROUGE F-scores obtained on our test datasets for the different summarization systems for both languages.

Similar to section 7.1, we experiment with five summarizers: **Length**, **MEAD**, **Classifier**, **Interpol**, **SuMT**. As expected, the MEAD summarizer shows the best summarization performance. Also, the length-based baseline system generates poor quality summaries (about 22 score ROUGE-1 reduction from MEAD). This is not surprising since the baseline only uses the length of the sentence regardless its content. Furthermore, the performance of the classifier-based summarizer is lower than the MEAD, because it does not use the summarization feature and only relies on an estimated translation quality to select the sentences.

Reviewing different values of the ROUGE metric in the left side Table 2, we observe that SuMT and Interpol summaries maintain a decent quality, comparable to the state of the art MEAD. For example, they give promising results in terms of ROUGE-L (71.17% and 72.79%, respectively), which consistently indicates that the sentences produced are closer to the reference summary in linguistic surface structure than those of the classifier (65.92). In addition to the quality of the English summaries, we are more interested in assessing the quality of the Arabic summaries. This comes back to our main goal of producing a fluent Arabic summary with good translation quality. We evaluate the Arabic summaries by measuring different ROUGE metrics against our model summaries. The results in the right side of Table 2 show that our MT-aware summarization framework achieves the best results in different ROUGE configurations and outperforms the state-of-the-art summarizer (+1 point ROUGE-1). In other words, our Arabic translated summaries generated using SuMT, are the most fluent and have the most similar structure compared with the Arabic model summaries.

7.3 Model-free summarization evaluation

In addition to the reference-based summarization evaluation described above, we conducted model-free experiments evaluating the summary quality for both languages. Recently, Louis and Nenkova (2013) proposed SIMetrix, a framework that does not require gold standard summaries for measuring the summarization quality. The framework is based on the idea that higher similarity with the source document would be indicative of high quality summary. SIMetrix is a suite of model-free similarity metrics for comparing a generated sum-

[7] A study conducted by Lin and Hovy (2003) shows that automatic evaluation using unigram and bigram co-occurrences between summary pairs have the highest correlation with human evaluations and have high recall and precision in significance test with manual evaluation results.

[8] `http://duc.nist.gov/duc2007/tasks.html`.

mary with the source document for which it was produced. That includes cosine similarity, distributional similarity and also use of topic signature words. SIMetrix is shown to produce summary scores that correlate accurately with human assessments.[9]

We used SIMetrix to evaluate the quality of the summaries generated by different systems. We report in Table 3, **%TopicTokens** referring to the percentage of tokens in the summary that are topic words of the input document; the Kullback Leibler divergence (**KL**); and the Jensen Shannon divergence (**JS**) between vocabulary distributions of the input and summary texts, which was found to produce the best predictions of summary quality. Since KL divergence is not symmetric, we measure it both ways Input-Summary ($\mathbf{KL}_{IS}$) and Summary-Input ($\mathbf{KL}_{SI}$). Based on these metrics, a good summary is expected to have low divergence between probability distributions of words in the input and summary, and high similarity with the input.

Table 3 illustrates these similarity results for both English and Arabic summaries. The results are consistent with those found in the model-based evaluation. For Arabic, our MT-aware system achieves the best results in terms of different divergence (0.14 JS against 0.17 for MEAD) and topic related scores (73.37% of tokens in the SUMT Arabic summaries are topic words in the input document against 71.78% in MEAD summaries). It is important to note that lower divergence scores indicate higher quality summaries.

	Length	MEAD	Classifier	Interpol	SuMT
			English		
%TopicTokens	63.21	63.70	63.28	63.50	63.33
$\mathbf{KL}_{IS}$	0.37	0.18	0.33	0.19	0.25
$\mathbf{KL}_{SI}$	0.14	0.02	0.12	0.07	0.09
JS	0.04	0.01	0.03	0.02	0.02
			Arabic		
%TopicTokens	71.51	71.61	73.18	72.44	**73.37**
$\mathbf{KL}_{IS}$	1.30	1.24	1.28	1.20	**1.19**
$\mathbf{KL}_{SI}$	1.11	1.07	1.03	0.96	**0.94**
JS	0.17	0.15	0.16	0.15	**0.14**

Table 3: Distribution similarity scores for each system summaries evaluated against the input document for English and Arabic.

[9]SIMetrix is available at: `http://www.seas.upenn.edu/~lannie/IEval2.html`.

8 Conclusion and Future work

We presented our approach in pairing automatic text summarization with machine translation to generate a higher quality content. We demonstrated an English to Arabic MT aware summarization framework with high summarization quality and greatly improved translation quality.

We plan to extend our current system in the following directions: (a) We will examine alternative learning frameworks and features to improve our prediction of the translation quality. (b) We will explore different methods to incorporate and optimize the MT quality information with the summarization system. (c) We will explore alternative text domains such as Wikipedia in which there is a larger variation of translation quality in different parts of the document. Considering the poor translation quality of many language pairs, text summarization can provide effective support for MT in various end-user applications. We believe there are many avenues to explore in this direction of research.

9 Acknowledgements

We thank Nizar Habash and anonymous reviewers for their valuable comments and suggestions. We thank Mollie Kauffer and Wajdi Zaghouani for preparing the Arabic and English summaries. This publication was made possible by grants YSREP-1-018-1-004 and NPRP-09-1140-1-177 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence Estimation for Machine Translation.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the ACL-WMT-2013*.

Florian Boudin, Stéphane Huet, and Juan-Manuel Torres-Moreno. 2011. A Graph-based Approach to Cross-language Multi-document Summarization. *Polibits*, (43):113–118.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines.

ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27.

Mona Diab. 2009. Second generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS Tagging, and Base Phrase Chunking. In *Proceedings of MEDAR*, Cairo, Egypt.

Ahmed El Kholy and Nizar Habash. 2010. Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of LREC*.

George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011. TAC 2011 Multiling Pilot Overview. In *Proceedings of the TAC 2011*.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, Pos Tagging, Stemming and Lemmatization. In *Proceedings of MEDAR*.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the WMT*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of NAACL*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL-04 Text Summarization Workshop (Text Summarization Branches Out)*.

Kenneth C Litkowski. 2004. Summarization Experiments in DUC 2004. In *Proceedings of the HLT-NAACL Workshop on Automatic Summarization, DUC-2004*, pages 6–7.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):1–34.

Behrang Mohit and Rebecca Hwa. 2007. Localization of Difficult-to-Translate Phrases. In *Proceedings of ACL WMT-07*.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

Dragomir Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multidocument summarization using MEAD. In *Proceedings of DUC*.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, et al. 2004. MEAD-a platform for multidocument multilingual text summarization. In *Proceedings of LREC*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of ACL*.

Lucia Specia, Kashif Shah, Jose Guilherme Camargo de Souza, and Trevor Cohn. 2013. QUEST-A Translation Quality Estimation Framework. In *Proceedings of the ACL, demo session*, Sofia, Bulgaria.

Lucia Specia. 2012. Estimating Machine Translation Quality. In *MT Marathon*.

Andreas Stolcke. 2002. SRILM-an Extensible Language Modeling Toolkit. In *Proceedings of ICLSP*.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language Document Summarization Based on Machine Translation Quality Prediction. In *Proceedings of ACL*, Uppsala, Sweden.

Richard Zens and Hermann Ney. 2006. N-Gram Posterior Probabilities for Statistical Machine Translation. In *Proceedings of WMT*.