Cross-Language Document Summarization Based on Machine Translation Quality Prediction

Xiaojun Wan, Huiying Li and Jianguo Xiao
Institute of Compute Science and Technology, Peking University, Beijing 100871, China
Key Laboratory of Computational Linguistics (Peking University), MOE, China
{wanxiaojun,lihuiying,xiaojianguo}@icst.pku.edu.cn

Abstract

Cross-language document summarization is a task of producing a summary in one language for a document set in a different language. Existing methods simply use machine translation for document translation or summary translation. However, current machine translation services are far from satisfactory, which results in that the quality of the cross-language summary is usually very poor, both in readability and content. In this paper, we propose to consider the translation quality of each sentence in the English-to-Chinese cross-language summarization process. First, the translation quality of each English sentence in the document set is predicted with the SVM regression method, and then the quality score of each sentence is incorporated into the summarization process. Finally, the English sentences with high translation quality and high informativeness are selected and translated to form the Chinese summary. Experimental results demonstrate the effectiveness and usefulness of the proposed approach.

1 Introduction

Given a document or document set in one source language, cross-language document summarization aims to produce a summary in a different target language. In this study, we focus on English-to-Chinese document summarization for the purpose of helping Chinese readers to quickly understand the major content of an English document or document set. This task is very important in the field of multilingual information access.

Till now, most previous work focuses on monolingual document summarization, but cross-language document summarization has received little attention in the past years. A straightforward way for cross-language document summarization is to translate the summary from the source language to the target language by using machine translation services. However, though machine translation techniques have been advanced a lot, the machine translation quality is far from satisfactory, and in many cases, the translated texts are hard to understand. Therefore, the translated summary is likely to be hard to understand by readers, i.e., the summary quality is likely to be very poor. For example, the translated Chinese sentence for an ordinary English sentence ("It is also Mr Baker who is making the most of presidential powers to dispense largesse.") by using Google Translate is "同时，也是贝克是谁提出了对总统权力免除最慷慨。". The translated sentence is hard to understand because it contains incorrect translations and it is very disfluent. If such sentences are selected into the summary, the quality of the summary would be very poor.

In order to address the above problem, we propose to consider the translation quality of the English sentences in the summarization process. In particular, the translation quality of each English sentence is predicted by using the SVM regression method, and then the predicted MT quality score of each sentence is incorporated into the sentence evaluation process, and finally both informative and easy-to-translate sentences are selected and translated to form the Chinese summary.

An empirical evaluation is conducted to evaluate the performance of machine translation quality prediction, and a user study is performed to evaluate the cross-language summary quality. The results demonstrate the effectiveness of the proposed approach.

The rest of this paper is organized as follows: Section 2 introduces related work. The system is overviewed in Section 3. In Sections 4 and 5, we present the detailed algorithms and evaluation

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 917–926,
Uppsala, Sweden, 11-16 July 2010.

results of machine translation quality prediction and cross-language summarization, respectively. We discuss in Section 6 and conclude this paper in Section 7.

2 Related Work

2.1 Machine Translation Quality Prediction

Machine translation evaluation aims to assess the correctness and quality of the translation. Usually, the human reference translation is provided, and various methods and metrics have been developed for comparing the system-translated text and the human reference text. For example, the BLEU metric, the NIST metric and their relatives are all based on the idea that the more shared substrings the system-translated text has with the human reference translation, the better the translation is. Blatz et al. (2003) investigate training sentence-level confidence measures using a variety of fuzzy match scores. Albrecht and Hwa (2007) rely on regression algorithms and reference-based features to measure the quality of sentences.

Transition evaluation without using reference translations has also been investigated. Quirk (2004) presents a supervised method for training a sentence level confidence measure on translation output using a human-annotated corpus. Features derived from the source sentence and the target sentence (e.g. sentence length, perplexity, etc.) and features about the translation process are leveraged. Gamon et al. (2005) investigate the possibility of evaluating MT quality and fluency at the sentence level in the absence of reference translations, and they can improve on the correlation between language model perplexity scores and human judgment by combing these perplexity scores with class probabilities from a machine-learned classifier. Specia et al. (2009) use the ICM theory to identify the threshold to map a continuous predicted score into "good" or "bad" categories. Chae and Nenkova (2009) use surface syntactic features to assess the fluency of machine translation results.

In this study, we further predict the translation quality of an English sentence before the machine translation process, i.e., we do not leverage reference translation and the target sentence.

2.2 Document Summarization

Document summarization methods can be generally categorized into extraction-based methods and abstraction-based methods. In this paper, we focus on extraction-based methods. Extraction-based summarization methods usually assign each sentence a saliency score and then rank the sentences in a document or document set.

For single document summarization, the sentence score is usually computed by empirical combination of a number of statistical and linguistic feature values, such as term frequency, sentence position, cue words, stigma words, topic signature (Luhn 1969; Lin and Hovy, 2000). The summary sentences can also be selected by using machine learning methods (Kupiec et al., 1995; Amini and Gallinari, 2002) or graph-based methods (ErKan and Radev, 2004; Mihalcea and Tarau, 2004). Other methods include mutual reinforcement principle (Zha 2002; Wan et al., 2007).

For multi-document summarization, the centroid-based method (Radev et al., 2004) is a typical method, and it scores sentences based on cluster centroids, position and TFIDF features. NeATS (Lin and Hovy, 2002) makes use of new features such as topic signature to select important sentences. Machine Learning based approaches have also been proposed for combining various sentence features (Wong et al., 2008). The influences of input difficulty on summarization performance have been investigated in (Nenkova and Louis, 2008). Graph-based methods have also been used to rank sentences in a document set. For example, Mihalcea and Tarau (2005) extend the TextRank algorithm to compute sentence importance in a document set. Cluster-level information has been incorporated in the graph model to better evaluate sentences (Wan and Yang, 2008). Topic-focused or query biased multi-document summarization has also been investigated (Wan et al., 2006). Wan et al. (2010) propose the EUSUM system for extracting easy-to-understand English summaries for non-native readers.

Several pilot studies have been performed for the cross-language summarization task by simply using document translation or summary translation. Leuski et al. (2003) use machine translation for English headline generation for Hindi documents. Lim et al. (2004) propose to generate a Japanese summary without using a Japanese summarization system, by first translating Japanese documents into Korean documents, and then extracting summary sentences by using Korean summarizer, and finally mapping Korean summary sentences to Japanese summary sentences. Chalendar et al. (2005) focuses on semantic analysis and sentence generation techniques for cross-language summarization. Orasan

and Chiorean (2008) propose to produce summaries with the MMR method from Romanian news articles and then automatically translate the summaries into English. Cross language query based summarization has been investigated in (Pingali et al., 2007), where the query and the documents are in different languages. Other related work includes multilingual summarization (Lin et al., 2005), which aims to create summaries from multiple sources in multiple languages. Siddharthan and McKeown (2005) use the information redundancy in multilingual input to correct errors in machine translation and thus improve the quality of multilingual summaries.

3 The Proposed Approach

Previous methods for cross-language summarization usually consist of two steps: one step for summarization and one step for translation. Different order of the two steps can lead to the following two basic English-to-Chinese summarization methods:

Late Translation (LateTrans): Firstly, an English summary is produced for the English document set by using existing summarization methods. Then, the English summary is automatically translated into the corresponding Chinese summary by using machine translation services.

Early Translation (EarlyTrans): Firstly, the English documents are translated into Chinese documents by using machine translation services. Then, a Chinese summary is produced for the translated Chinese documents.

Generally speaking, the LateTrans method has a few advantages over the EarlyTrans method:

1) The LateTrans method is much more efficient than the EarlyTrans method, because only a very few summary sentences are required to be translated in the LateTrans method, whereas all the sentences in the documents are required to be translated in the EarlyTrans method.

2) The LateTrans method is deemed to be more effective than the EarlyTrans method, because the translation errors of the sentences have great influences on the summary sentence extraction in the EarlyTrans method.

Thus in this study, we adopt the LateTrans method as our baseline method. We also adopt the late translation strategy for our proposed approach.

In the baseline method, a translated Chinese sentence is selected into the summary because the original English sentence is informative. However, an informative and fluent English sentence is likely to be translated into an uninformative and disfluent Chinese sentence, and therefore, this sentence cannot be selected into the summary.

In order to address the above problem of existing methods, our proposed approach takes into account a novel factor of each sentence for cross-language summary extraction. Each English sentence is associated with a score indicating its translation quality. An English sentence with high translation quality score is more likely to be selected into the original English summary, and such English summary can be translated into a better Chinese summary. Figure 1 gives the architecture of our proposed approach.

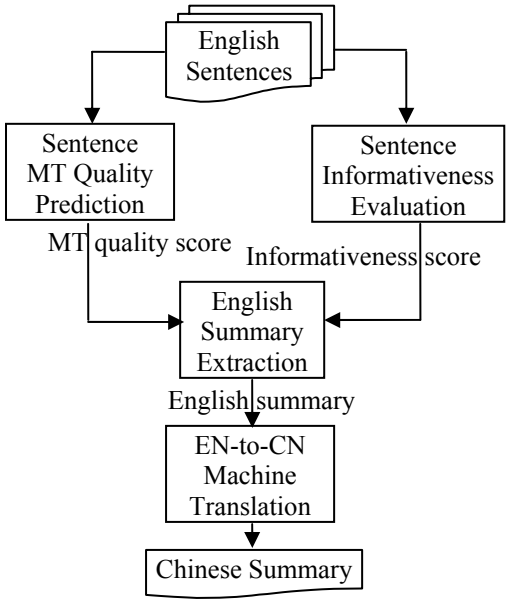

Figure 1: Architecture of our proposed approach

Seen from the figure, our proposed approach consists of four main steps: 1) The machine translation quality score of each English sentence is predicted by using regression methods; 2) The informativeness score of each English sentence is computed by using existing methods; 3) The English summary is produced by making use of both the machine translation quality score and the informativeness score; 4) The extracted English summary is translated into Chinese summary by using machine translation services.

In this study, we adopt *Google Translate*[1] for English-to-Chinese translation. *Google Translate* is one of the state-of-the-art commercial machine translation systems used today. It applies statistical learning techniques to build a translation

[1] http://translate.google.com/translate_t

model based on both monolingual text in the target language and aligned text consisting of examples of human translations between the languages.

The first step and the evaluation results will be described in Section 4, and the other steps and the evaluation results will be described together in Section 5.

4 Machine Translation Quality Prediction

4.1 Methodology

In this study, machine translation (MT) quality reflects both the translation accuracy and the fluency of the translated sentence. An English sentence with high MT quality score is likely to be translated into an accurate and fluent Chinese sentence, which can be easily read and understand by Chinese readers. The MT quality prediction is a task of mapping an English sentence to a numerical value corresponding to a quality level. The larger the value is, the more accurately and fluently the sentence can be translated into Chinese sentence.

As introduced in Section 2.1, several related work has used regression and classification methods for MT quality prediction without reference translations. In our approach, the MT quality of each sentence in the documents is also predicted without reference translations. The difference between our task and previous work is that previous work can make use of both features in source sentence and features in target sentence, while our task only leverages features in source sentence, because in the late translation strategy, the English sentences in the documents have not been translated yet at this step.

In this study, we adopt the ε-support vector regression (ε-SVR) method (Vapnik 1995) for the sentence-level MT quality prediction task. The SVR algorithm is firmly grounded in the framework of statistical learning theory (VC theory). The goal of a regression algorithm is to fit a flat function to the given training data points.

Formally, given a set of training data points $D=\{(x_i,y_i)|\ i=1,2,\ldots,n\} \subset R^d \times R$, where x_i is input feature vector and y_i is associated score, the goal is to fit a function f which approximates the relation inherited between the data set points. The standard form is:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C\sum_{i=1}^{n} \xi_i + C\sum_{i=1}^{n} \xi_i^*$$

Subject to

$$w^T f(x_i) + b - y_i \le \varepsilon + \xi_i$$
$$y_i - w^T f(x_i) - b \le \varepsilon + \xi_i^*$$
$$\varepsilon, \xi_i, \xi_i^* \ge 0,\ i = 1,\ldots,n.$$

The constant $C>0$ is a parameter for determining the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated.

In the experiments, we use the LIBSVM tool (Chang and Lin, 2001) with the RBF kernel for the task, and we use the parameter selection tool of 10-fold cross validation via grid search to find the best parameters on the training set with respect to mean squared error (MSE), and then use the best parameters to train on the whole training set.

We use the following two groups of features for each sentence: the first group includes several basic features, and the second group includes several parse based features[2]. They are all derived based on the source English sentence.

The basic features are as follows:

1) **Sentence length**: It refers to the number of words in the sentence.
2) **Sub-sentence number**: It refers to the number of sub-sentences in the sentence. We simply use the punctuation marks as indicators of sub-sentences.
3) **Average sub-sentence length**: It refers to the average number of words in the sub-sentences within the sentence.
4) **Percentage of nouns and adjectives**: It refers to the percentage of noun words or adjective words in the in the sentence.
5) **Number of question words**: It refers to the number of question words (who, whom, whose, when, where, which, how, why, what) in the sentence.

We use the Stanford Lexicalized Parser (Klein and Manning, 2002) with the provided English PCFG model to parse a sentence into a parse tree. The output tree is a context-free phrase structure grammar representation of the sentence. The parse features are then selected as follows:

1) **Depth of the parse tree**: It refers to the depth of the generated parse tree.
2) **Number of SBARs in the parse tree**: SBAR is defined as a clause introduced by a (possibly empty) subordinating conjunction. It is an indictor of sentence complexity.

[2] Other features, including n-gram frequency, perplexity features, etc., are not useful in our study. MT features are not used because *Google Translate* is used as a black box.

3) **Number of NPs in the parse tree**: It refers to the number of noun phrases in the parse tree.
4) **Number of VPs in the parse tree**: It refers to the number of verb phrases in the parse tree.

All the above feature values are scaled by using the provided svm-scale program.

At this step, each English sentence s_i can be associated with a MT quality score *TransScore*(s_i) predicted by the ε-SVR method. The score is finally normalized by dividing by the maximum score.

4.2 Evaluation

4.2.1 Evaluation Setup

In the experiments, we first constructed the gold-standard dataset in the following way:

DUC2001 provided 309 English news articles for document summarization tasks, and the articles were grouped into 30 document sets. The news articles were selected from TREC-9. We chose five document sets (d04, d05, d06, d08, d11) with 54 news articles out of the DUC2001 document sets. The documents were then split into sentences and we used 1736 sentences for evaluation. All the sentences were automatically translated into Chinese sentences by using the *Google Translate* service.

Two Chinese college students were employed for data annotation. They read the original English sentence and the translated Chinese sentence, and then manually labeled the overall translation quality score for each sentence, separately. The translation quality is an overall measure for both the translation accuracy and the readability of the translated sentence. The score ranges between 1 and 5, and 1 means "very bad", and 5 means "very good", and 3 means "normal". The correlation between the two sets of labeled scores is 0.646. The final translation quality score was the average of the scores provided by the two annotators.

After annotation, we randomly separated the labeled sentence set into a training set of 1428 sentences and a test set of 308 sentences. We then used the LIBSVM tool for training and testing.

Two metrics were used for evaluating the prediction results. The two metrics are as follows:

Mean Square Error (MSE): This metric is a measure of how correct each of the prediction values is on average, penalizing more severe errors more heavily. Given the set of prediction scores for the test sentences: $\hat{Y} = \{\hat{y}_i \mid i = 1,...n\}$, and the manually assigned scores for the sentences: $Y = \{y_i \mid i = 1,...n\}$, the MSE of the prediction result is defined as

$$MSE(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Pearson's Correlation Coefficient (ρ): This metric is a measure of whether the trends of prediction values matched the trends for human-labeled data. The coefficient between Y and $\hat{Y}$ is defined as

$$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{n s_y s_{\hat{y}}}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the sample means of Y and $\hat{Y}$, s_y and $s_{\hat{y}}$ are the sample standard deviations of Y and $\hat{Y}$.

4.2.2 Evaluation Results

Table 1 shows the prediction results. We can see that the overall results are promising. And the correlation is moderately high. The results are acceptable because we only make use of the features derived from the source sentence. The results guarantee that the use of MT quality scores in the summarization process is feasible.

We can also see that both the basic features and the parse features are beneficial to the overall prediction results.

Feature Set	MSE	ρ
Basic features	0.709	0.399
Parse features	0.702	0.395
All features	**0.683**	**0.433**

Table 1: Prediction results

5 Cross-Language Document Summarization

5.1 Methodology

In this section, we first compute the informativeness score for each sentence. The score reflect how the sentence expresses the major topic in the documents. Various existing methods can be used for computing the score. In this study, we adopt the centroid-based method.

The centroid-based method is the algorithm used in the MEAD system. The method uses a heuristic and simple way to sum the sentence scores computed based on different features. The score for each sentence is a linear combination of

the weights computed based on the following three features:

Centroid-based Weight. The sentences close to the centroid of the document set are usually more important than the sentences farther away. And the centroid weight $C(s_i)$ of a sentence s_i is calculated as the cosine similarity between the sentence text and the concatenated text for the whole document set D. The weight is then normalized by dividing the maximal weight.

Sentence Position. The leading several sentences of a document are usually important. So we calculate for each sentence a weight to reflect its position priority as $P(s_i)=1-(i-1)/n$, where i is the sequence of the sentence s_i and n is the total number of sentences in the document. Obviously, i ranges from 1 to n.

First Sentence Similarity. Because the first sentence of a document is very important, a sentence similar to the first sentence is also important. Thus we use the cosine similarity value between a sentence and the corresponding first sentence in the same document as the weight $F(s_i)$ for sentence s_i.

After all the above weights are calculated for each sentence, we sum all the weights and get the overall score for the sentence as follows:

$$InfoScore(s_i) = \alpha \cdot C(s_i) + \beta \cdot P(s_i) + \gamma \cdot F(s_i)$$

where α, β and γ are parameters reflecting the importance of different features. We empirically set $\alpha=\beta=\gamma=1$.

After the informativeness scores for all sentences are computed, the score of each sentence is normalized by dividing by the maximum score.

After we obtain the MT quality score and the informativeness score of each sentence in the document set, we linearly combine the two scores to get the overall score of each sentence.

Formally, let $TransScore(s_i) \in [0,1]$ and $InfoScore(s_i) \in [0,1]$ denote the MT quality score and the informativeness score of sentence s_i, the overall score of the sentence is:

$$OverallScore(s_i) = (1-\lambda) \times InfoScore(s_i) + \lambda \times TransScore(s_i)$$

where $\lambda \in [0,1]$ is a parameter controlling the influences of the two factors. If λ is set to 0, the summary is extracted without considering the MT quality factor. In the experiments, we empirically set the parameter to 0.3 in order to balance the two factors of content informativeness and translation quality.

For multi-document summarization, some sentences are highly overlapping with each other, and thus we apply the same greedy algorithm in (Wan et al., 2006) to penalize the sentences highly overlapping with other highly scored sentences, and finally the informative, novel, and easy-to-translate sentences are chosen into the English summary.

Finally, the sentences in the English summary are translated into the corresponding Chinese sentences by using *Google Translate*, and the Chinese summary is formed.

5.2 Evaluation

5.2.1 Evaluation Setup

In this experiment, we used the document sets provided by DUC2001 for evaluation. As mentioned in Section 4.2.1, DUC2001 provided 30 English document sets for generic multi-document summarization. The average document number per document set was 10. The sentences in each article have been separated and the sentence information has been stored into files. Generic reference English summaries were provided by NIST annotators for evaluation. In our study, we aimed to produce Chinese summaries for the English document sets. The summary length was limited to five sentences, i.e. each summary consisted of five sentences.

The DUC2001 dataset was divided into the following two datasets:

Ideal Dataset: We have manually labeled the MT quality scores for the sentences in five document sets (d04-d11), and we directly used the manually labeled scores in the summarization process. The ideal dataset contained these five document sets.

Real Dataset: The MT quality scores for the sentences in the remaining 25 document sets were automatically predicted by using the learned SVM regression model. And we used the automatically predicted scores in the summarization process. The real dataset contained these 25 document sets.

We performed two evaluation procedures: one based on the ideal dataset to validate the feasibility of the proposed approach, and the other based on the real dataset to demonstrate the effectiveness of the proposed approach in real applications.

To date, various methods and metrics have been developed for English summary evaluation by comparing system summary with reference summary, such as the pyramid method (Nenkova et al., 2007) and the ROUGE metrics (Lin and Hovy, 2003). However, such methods or metrics cannot be directly used for evaluating Chinese summary without reference Chinese summary.

Instead, we developed an evaluation protocol as follows:

The evaluation was based on human scoring. Four Chinese college students participated in the evaluation as subjects. We have developed a friendly tool for helping the subjects to evaluate each Chinese summary from the following three aspects:

Content: This aspect indicates how much a summary reflects the major content of the document set. After reading a summary, each user can select a score between 1 and 5 for the summary. 1 means "very uninformative" and 5 means "very informative".

Readability: This aspect indicates the readability level of the whole summary. After reading a summary, each user can select a score between 1 and 5 for the summary. 1 means "hard to read", and 5 means "easy to read".

Overall: This aspect indicates the overall quality of a summary. After reading a summary, each user can select a score between 1 and 5 for the summary. 1 means "very bad", and 5 means "very good".

We performed the evaluation procedures on the ideal dataset and the read dataset, separately. During each evaluation procedure, we compared our proposed approach (λ=0.3) with the baseline approach without considering the MT quality factor (λ=0). And the two summaries produced by the two systems for the same document set were presented in the same interface, and then the four subjects assigned scores to each summary after they read and compared the two summaries. And the assigned scores were finally averaged across the documents sets and across the subjects.

5.2.2 Evaluation Results

Table 2 shows the evaluation results on the ideal dataset with 5 document sets. We can see that based on the manually labeled MT quality scores, the Chinese summaries produced by our proposed approach are significantly better than that produced by the baseline approach over all three aspects. All subjects agree that our proposed approach can produce more informative and easy-to-read Chinese summaries than the baseline approach.

Table 3 shows the evaluation results on the real dataset with 25 document sets. We can see that based on the automatically predicted MT quality scores, the Chinese summaries produced by our proposed approach are significantly better than that produced by the baseline approach over the readability aspect and the overall aspect. Almost all subjects agree that our proposed approach can produce more easy-to-read and high-quality Chinese summaries than the baseline approach.

Comparing the evaluation results in the two tables, we can find that the performance difference between the two approaches on the ideal dataset is bigger than that on the real dataset, especially on the content aspect. The results demonstrate that the more accurate the MT quality scores are, the more significant the performance improvement is.

Overall, the proposed approach is effective to produce good-quality Chinese summaries for English document sets.

	Baseline Approach			**Proposed Approach**		
	content	**readability**	**overall**	**content**	**readability**	**overall**
Subject1	3.2	2.6	2.8	3.4	3.0	3.4
Subject2	3.0	3.2	3.2	3.4	3.6	3.4
Subject3	3.4	2.8	3.2	3.6	3.8	3.8
Subject4	3.2	3.0	3.2	3.8	3.8	3.8
Average	3.2	2.9	3.1	3.55*	3.55*	3.6*

Table 2: Evaluation results on the ideal dataset (5 document sets)

	Baseline Approach			**Proposed Approach**		
	content	**readability**	**overall**	**content**	**readability**	**overall**
Subject1	2.64	2.56	2.60	2.80	3.24	2.96
Subject2	3.60	2.76	3.36	3.52	3.28	3.64
Subject3	3.52	3.72	3.44	3.56	3.80	3.48
Subject4	3.16	2.96	3.12	3.16	3.44	3.52
Average	3.23	3.00	3.13	3.26	3.44*	3.40*

Table 3: Evaluation results on the real dataset (25 document sets)

(* indicates the difference between the average score of the proposed approach and that of the baseline approach is statistically significant by using t-test.)

5.2.3 *Example Analysis*

In this section, we give two running examples to better show the effectiveness of our proposed approach. The Chinese sentences and the original English sentences in the summary are presented together. The normalized MT quality score for each sentence is also given at the end of the Chinese sentence.

Document set 1: D04 from the ideal dataset

Summary by baseline approach:

s1: 预计美国的保险公司支付，估计在佛罗里达州的73亿美元（37亿英镑），作为安德鲁飓风的结果-迄今为止最昂贵的灾难曾经面临产业。(0.56)
(US INSURERS expect to pay out an estimated Dollars 7.3bn (Pounds 3.7bn) in Florida as a result of Hurricane Andrew - by far the costliest disaster the industry has ever faced.)
s2: 有越来越多的迹象表明安德鲁飓风，不受欢迎的，因为它的佛罗里达和路易斯安那州的受灾居民，最后可能不伤害到连任的布什总统竞选。(0.67)
(THERE are growing signs that Hurricane Andrew, unwelcome as it was for the devastated inhabitants of Florida and Louisiana, may in the end do no harm to the re-election campaign of President George Bush.)
s3: 一般事故发生后，英国著名保险公司昨日表示，保险索赔的安德鲁飓风所引发的成本也高达4000万美元'。 (0.44)
(GENERAL ACCIDENT said yesterday that insurance claims arising from Hurricane Andrew could 'cost it as much as Dollars 40m'.)
s4: 在巴哈马，政府发言人麦库里说，4人死亡已离岛东部群岛报告。 (0.56)
(In the Bahamas, government spokesman Mr Jimmy Curry said four deaths had been reported on outlying eastern islands.)
s5: 新奥尔良的和1.6万人，是特别脆弱，因为该市位于海平面以下，有密西西比河通过其中心的运行和一个大型湖泊立即向北方。(0.44)
(New Orleans, with a population of 1.6m, is particularly vulnerable because the city lies below sea level, has the Mississippi River running through its centre and a large lake immediately to the north.)

Summary by proposed approach:

s1: 预计美国的保险公司支付，估计在佛罗里达州的73亿美元（37亿英镑），作为安德鲁飓风的结果-迄今为止最昂贵的灾难曾经面临产业。(0.56)
(US INSURERS expect to pay out an estimated Dollars 7.3bn (Pounds 3.7bn) in Florida as a result of Hurricane Andrew - by far the costliest disaster the industry has ever faced.)
s2: 有越来越多的迹象表明安德鲁飓风，不受欢迎的，因为它的佛罗里达和路易斯安那州的受灾居民，最后可能不伤害到连任的布什总统竞选。(0.67)
(THERE are growing signs that Hurricane Andrew, unwelcome as it was for the devastated inhabitants of Florida and Louisiana, may in the end do no harm to the re-election campaign of President George Bush.)
s3: 在巴哈马，政府发言人麦库里说，4人死亡已离岛东部群岛报告。(0.56)
(In the Bahamas, government spokesman Mr Jimmy Curry said four deaths had been reported on outlying eastern islands.)
s4: 在首当其冲的损失可能会集中在美国的保险公司，业内分析人士昨天说。 (0.89)
(The brunt of the losses are likely to be concentrated among US insurers, industry analysts said yesterday.)
s5: 在北迈阿密，损害是最小的。(1.0)
(In north Miami, damage is minimal.)

Document set 2: D54 from the real dataset

Summary by baseline approach:

s1: 两个加州11月6日投票的主张，除其他限制外，全州成员及州议员的条件。(0.57)
(Two propositions on California's Nov. 6 ballot would, among other things, limit the terms of statewide officeholders and state legislators.)
s2: 原因之一是任期限制将开放到现在的政治职务任职排除了许多人的职业生涯。(0.36)
(One reason is that term limits would open up politics to many people now excluded from office by career incumbents.)
s3: 建议限制国会议员及州议员都很受欢迎，越来越多的条件是，根据专家和投票。(0.20)
(Proposals to limit the terms of members of Congress and of state legislators are popular and getting more so, according to the pundits and the polls.)
s4: 国家法规的酒吧首先从运行时间为国会候选人已举行了加入的资格规定了宪法规定，并已失效。(0.24)
(State statutes that bar first-time candidates from running for Congress have been held to add to the qualifications set forth in the Constitution and have been invalidated.)
s5: 另一个论点是，公民的同时，不断进入新的华盛顿国会将面临流动更好的结果，比政府的任期较长的代表提供的。(0.20)
(Another argument is that a citizen Congress with its continuing flow of fresh faces into Washington would result in better government than that provided by representatives with lengthy tenure.)

Summary by proposed approach:

s1: 两个加州 11 月 6 日投票的主张，除其他限制外，全州成员及州议员的条件。(0.57)
(Two propositions on California's Nov. 6 ballot would, among other things, limit the terms of statewide officeholders and state legislators.)
s2: 原因之一是任期限制将开放到现在的政治职务任职排除了许多人的职业生涯。(0.36)
(One reason is that term limits would open up politics to many people now excluded from office by career incumbents.)
s3: 另一个论点是，公民的同时，不断进入新的华盛顿国会将面临流动更好的结果，比政府的任期较长的代表提供的。(0.20)
(Another argument is that a citizen Congress with its continuing flow of fresh faces into Washington would result in better government than that provided by representatives with lengthy tenure.)
s4: 有两个国会任期限制，经济学家，至少公共选择那些劝说，要充分理解充分的理由。(0.39)
(There are two solid reasons for congressional term limitation that economists, at least those of the public-choice persuasion, should fully appreciate.)
s5: 与国会的问题的根源是，除非有重大丑闻，几乎是不可能战胜现任。(0.47)
(The root of the problems with Congress is that, barring major scandal, it is almost impossible to defeat an incumbent.)

6 Discussion

In this study, we adopt the late translation strategy for cross-document summarization. As mentioned earlier, the late translation strategy has some advantages over the early translation strategy. However, in the early translation strategy, we can use the features derived from both the source English sentence and the target Chinese sentence to improve the MT quality prediction results.

Overall, the framework of our proposed approach can be easily adapted for cross-document summarization with the early translation strategy.

And an empirical comparison between the two strategies is left as our future work.

Though this study focuses on English-to-Chinese document summarization, cross-language summarization tasks for other languages can also be solved by using our proposed approach.

7 Conclusion and Future Work

In this study we propose a novel approach to address the cross-language document summarization task. Our proposed approach predicts the MT quality score of each English sentence and then incorporates the score into the summarization process. The user study results verify the effectiveness of the approach.

In future work, we will manually translate English reference summaries into Chinese reference summaries, and then adopt the ROUGE metrics to perform automatic evaluation of the extracted Chinese summaries by comparing them with the Chinese reference summaries. Moreover, we will further improve the sentence's MT quality by using sentence compression or sentence reduction techniques.

Acknowledgments

This work was supported by NSFC (60873155), Beijing Nova Program (2008B03), NCET (NCET-08-0006), RFDP (20070001059) and National High-tech R&D Program (2008AA01Z421). We thank the students for participating in the user study. We also thank the anonymous reviewers for their useful comments.

References

J. Albrecht and R. Hwa. 2007. A re-examination of machine learning approaches for sentence-level mt evaluation. In *Proceedings of ACL2007*.

M. R. Amini, P. Gallinari. 2002. The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In *Proceedings of SIGIR2002*.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for statistical machine translation. *Johns Hopkins Summer Workshop Final Report*.

J. Chae and A. Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of EACL2009*.

G. de Chalendar, R. Besançon, O. Ferret, G. Grefenstette, and O. Mesnard. 2005. Crosslingual summarization with thematic extraction, syntactic sentence simplification, and bilingual generation. In *Workshop on Crossing Barriers in Text Summarization Research, 5th International Conference on Recent Advances in Natural Language Processing (RANLP2005)*.

C.-C. Chang and C.-J. Lin. 2001. LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

G. ErKan, D. R. Radev. LexPageRank. 2004. Prestige in Multi-Document Text Summarization. In *Proceedings of EMNLP2004*.

M. Gamon, A. Aue, and M. Smets. 2005. Sentence-level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT2005*.

D. Klein and C. D. Manning. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Proceedings of NIPS2002*.

J. Kupiec, J. Pedersen, F. Chen. 1995. A.Trainable Document Summarizer. In *Proceedings of SIGIR1995*.

A. Leuski, C.-Y. Lin, L. Zhou, U. Germann, F. J. Och, E. Hovy. 2003. Cross-lingual C*ST*RD: English access to Hindi information. *ACM Transactions on Asian Language Information Processing*, 2(3): 245-269.

J.-M. Lim, I.-S. Kang, J.-H. Lee. 2004. Multi-document summarization using cross-language texts. In *Proceedings of NTCIR-4*.

C. Y. Lin, E. Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 17th Conference on Computational Linguistics*.

C..-Y. Lin and E.. H. Hovy. 2002. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In *Proceedings of ACL-02*.

C.-Y. Lin and E.H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of HLT-NAACL -03*.

C.-Y. Lin, L. Zhou, and E. Hovy. 2005. Multilingual summarization evaluation 2005: automatic evaluation report. In *Proceedings of MSE (ACL-2005 Workshop)*.

H. P. Luhn. 1969. The Automatic Creation of literature Abstracts. *IBM Journal of Research and Development*, 2(2).

R. Mihalcea, P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of EMNLP2004*.

R. Mihalcea and P. Tarau. 2005. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP-05*.

A. Nenkova and A. Louis. 2008. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. In *Proceedings of ACL-08:HLT*.

A. Nenkova, R. Passonneau, and K. McKeown. 2007. The Pyramid method: incorporating human content selection variation in summarization evaluation.

ACM Transactions on Speech and Language Processing (TSLP), 4(2).

C. Orasan, and O. A. Chiorean. 2008. Evaluation of a Crosslingual Romanian-English Multi-document Summariser. In *Proceedings of 6th Language Resources and Evaluation Conference (LREC2008)*.

P. Pingali, J. Jagarlamudi and V. Varma. 2007. Experiments in cross language query focused multi-document summarization. In *Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies in IJCAI2007*.

C. Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proceedings of LREC2004*.

D. R. Radev, H. Y. Jing, M. Stys and D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40: 919-938.

A. Siddharthan and K. McKeown. 2005. Improving multilingual summarization: using redundancy in the input to correct MT errors. In *Proceedings of HLT/EMNLP-2005*.

L. Specia, Z. Wang, M. Turchi, J. Shawe-Taylor, C. Saunders. 2009. Improving the Confidence of Machine Translation Quality Estimates. In *MT Summit 2009 (Machine Translation Summit XII)*.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

X. Wan, H. Li and J. Xiao. 2010. EUSUM: extracting easy-to-understand English summaries for non-native readers. In *Proceedings of SIGIR2010*.

X. Wan, J. Yang and J. Xiao. 2006. Using cross-document random walks for topic-focused multi-documetn summarization. In *Proceedings of WI2006*.

X. Wan and J. Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of SIGIR-08*.

X. Wan, J. Yang and J. Xiao. 2007. Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. In *Proceedings of ACL2007*.

K.-F. Wong, M. Wu and W. Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of COLING-08*.

H. Y. Zha. 2002. Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. In *Proceedings of SIGIR2002*.