Abstractive Cross-Language Summarization via Translation Model Enhanced Predicate Argument Structure Fusing

Jiajun Zhang, *Member, IEEE*, Yu Zhou, and Chengqing Zong, *Senior Member, IEEE*

***Abstract*—Cross-language multidocument summarization is the task to generate a summary in a target language (e.g., Chinese) from a collection of documents in a different source language (e.g., English). Previous methods such as the extractive and compressive algorithms focus only on single sentence selection and compression, which cannot make full use of the similar sentences containing complementary information. Furthermore, the translation model knowledge is not fully explored in previous approaches. To address these two problems, we propose in this paper an abstractive cross-language summarization framework. First, the source language documents are translated into target language with a machine translation system. Then, the method constructs a pool of bilingual concepts and facts represented by the bilingual elements of the source-side predicate-argument structures (PAS) and their target-side counterparts. Finally, new summary sentences are produced by fusing bilingual PAS elements with the integer linear programming algorithm to maximize both of the salience and translation quality of the PAS elements. The experimental results on English-to-Chinese cross-language summarization demonstrate that our proposed method outperforms the state-of-the-art extractive systems in both automatic and manual evaluations.**

***Index Terms*—Abstractive cross-language summarization, predicate-argument structure, translation model, integer linear programming..**

I. Introduction

CROSS-LANGUAGE summarization (CLS) is the task of producing a summary in a different target language (e.g. Chinese) from a collection of documents in a source language (e.g. English) [1]. In big data era nowadays, the data especially texts comes from different languages. To understand the content gist of the documents written in an unfamiliar language, CLS becomes very helpful for us. In this work, we focus on English-to-Chinese CLS.

Ideally, CLS is not a problem if machine translation (MT) is good enough. In that case, one can just generate the summary in the source language and then convert it into target language with the MT system. However, the translation quality is still far from satisfactory although the MT technology has made promising progress in the past decades. Thus, it is a crucial issue how to make full use of the translation models (TM) in CLS.

Existing methods for CLS fall in two categories: extraction-based and compression-based. Most of the CLS systems apply the extraction-based approach which in turn can be divided into three classes. Take English-to-Chinese CLS for example, one method directly extracts summary sentences from English documents using only English side features and then automatically translates the English summary into Chinese summary. The second one firstly translates the English documents into Chinese and then extracts summary sentences from the translated Chinese documents. The third one takes both English documents and the translated Chinese counterparts into account and extracts Chinese summary sentences with features in both languages [2].

Recently, [3] proposes a compressive method for CLS. This approach simultaneously performs sentence selection and compression. The sentence scores are calculated based on the aligned bilingual phrases which are obtained using online MT service. Compression is performed by deleting the phrases which are redundant or badly translated.

It should be noted that the above methods concerning solely single sentence selection and compression cannot merge several salient and well-translated facts from different sentences. For example, the two English sentences and their Chinese translations in the top half of Fig. 1 talk about different but relevant facts for the same concept *president bush*. Due to the redundancy requirements, the extraction-based approach can only pick one Chinese sentence from the two and the compression-based one may delete some noisy or badly translated phrases from the selected sentence. Obviously, the sentence combining the facts from both of the two sentences shown in the bottom half in Fig. 1 should be a much better choice. In this study, our goal is to generate Chinese summaries consisting of such sentences.

To achieve this goal, we propose an abstractive CLS (ACLS) framework which is inspired by the abstractive algorithms in

Manuscript received December 02, 2015; revised April 17, 2016; accepted June 20, 2016. Date of publication June 30, 2016; date of current version August 02, 2016. This research work was supported in part by the Natural Science Foundation of China under Grant 61333018 and Grant 61303181, and in part by the Strategic Priority Research Program of the CAS under Grant XDB02070007 and the Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing. The associate editor coordinating the review of this manuscript and approving it for publication was Ani Nenkova.

J. Zhang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jjzhang@nlpr.ia.ac.cn).

Y. Zhou is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China (e-mail: yzhou@nlpr.ia.ac.cn).

C. Zong is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, and the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: cqzong@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2586608

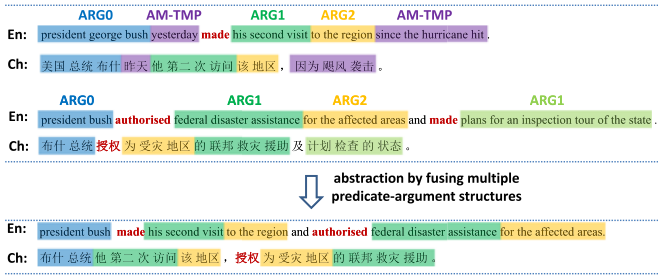


Fig. 1. An example for our ACLS based on PAS fusing. In the top half, two English sentences addressing the same entity and topic are annotated with the PAS. Red bold words are predicates. ARG0, ARG1, ARG2 and AM-TMP denote respectively agent, patient, benefactive and time. The Chinese translations are also given in the example. The bottom half shows the result by merging the elements of different PAS considering both the salience and translation quality of the arguments.

monolingual document summarization [4]–[7]. In our ACLS framework, we assume that all the summary sentences are produced by merging the bilingual concepts and facts[1] which are salient and well-translated. In this work, we define concepts and facts in predicate-argument structures (PAS). PAS for a sentence are usually automatically annotated by semantic role labelling [8], in which each predicate indicates an event and the arguments express the relevant information of this event (as shown in Fig. 1). Thus, concepts are defined by ARG0 that denotes the theme of the event. The facts are defined by two components: core facts represented by predicate+ARG1 (ARG2) and auxiliary facts represented by AM-TMP (AM-LOC). Given the English documents and the automatically translated Chinese counterparts, human writers can generate the Chinese summary sentences by recognizing and merging the concepts and facts (elements of PAS) which are most important and well translated. In this study, we attempt to simulate this process automatically.

Specifically, we use Fig. 1 to illustrate the main idea of our ACLS framework. The input English documents are first translated into Chinese using a MT system. Meanwhile, the English sentences are parsed into PAS. Then, we recognize the bilingual concepts and facts with word alignments which can be automatically estimated. For example, the phrases in the same color in Fig. 1 denote the same concept or fact. The core algorithm lies in two sub-models: one adopts the TM augmented algorithm to measure the salience and translation quality of the bilingual concepts and facts, and the other applies the integer linear programming (ILP) to select and re-organize the concepts and facts into valid and fluent Chinese summary sentences so as to maximize the salience and translation quality of the summary sentences in a global manner. In Fig. 1, the concept *president bush* and the facts *made his visit*, *to the region*, *authorised federal disaster assistance* and *for the affected areas* are selected while others are ignored due to unimportance or low translation quality.

We make the following contributions in this paper:

1) To the best of our knowledge, we are the first to propose a novel ACLS framework.
2) We introduce PAS to define bilingual concepts and facts. Furthermore, TM knowledge is employed to measure the translation quality of the bilingual concepts and facts.
3) Our CLS framework can significantly outperform the state-of-the-art extractive methods in terms of automatic ROUGE metric and manual linguistic quality.

II. ACLS Framework

In the English-to-Chinese summarization, we take the English documents as input and create a Chinese summary as output. This process is completed in five steps in our ACLS framework. First, we translate the English documents into Chinese counterparts with a MT system, and meanwhile parse the English sentences into PAS. Second, a pool of bilingual concepts and facts are recognized and extracted with automatically estimated word alignments. Third, the salience score and translation quality of the bilingual concepts and facts are calculated respectively. Fourth, we employ integer linear optimization model to formalize the sentence generation task with multiple designed constraints. Finally, we perform some post-processing work to improve the readability of the produced Chinese summary sentences. In the following sections, each step will be introduced in detail.

A. Document Translation and PAS Parsing

In order to improve the reproducibility of our method, we apply the state-of-the-art online Google Translate[2] to translate all the sentences in the English documents into Chinese. Note that no pre-processing such as tokenization and lowercasing is performed for English sentences, since we find that Google Translate obtains better translation results for the sentences without lowercasing and tokenization.

It is emphasized in the introduction section that the final summary is composed by the elements of PAS. Therefore, the quality of English PAS parsing is very important. The task of PAS parsing is to recognize every predicate in the sentence and identify all the relevant arguments for each predicate. This task is usually performed through semantic role labelling [8]. In this paper, we employ the approach proposed in [9] which can better handle the sentences containing multiple predicates. Take the second English sentence in Fig. 1 for example. There are two predicates *authorised* and *made* in this sentence, and they share the same agent *president bush*. Few methods provide good solutions to this problem. Fortunately, [9]–[11] designed a smart discriminative model incorporating global features to constrain the shared arguments for different predicates so that the predicted semantic roles are more consistent. This method achieves the state-of-the-art 77.0% F1-score on English PropBank.

[1]The terminology of concept and fact is inspired by the work [7] in which noun phrases are concepts and verb phrases are facts. We will make detailed comparison in the related work.

[2]https://translate.google.com/

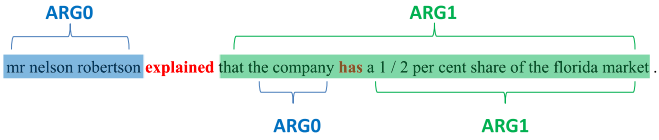


Fig. 2. An example of recursive PAS. *ARG0:*{*the company*}, *predicate: has* and *ARG1:*{*a 1/2 percent share of the Florida market*} server as ARG1 in the outer PAS.

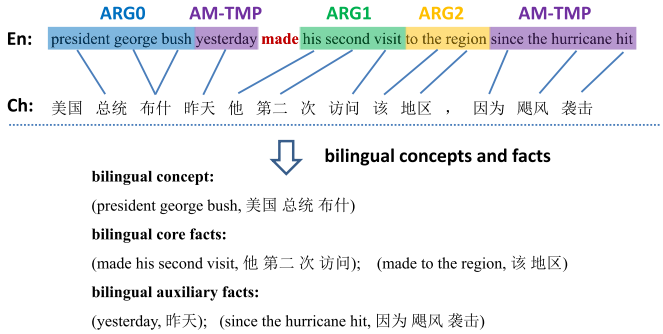


Fig. 3. An example of word alignment and the extraction of bilingual concepts and facts. During extraction, Chinese boundary words will be expanded if they have no alignment and they are not punctuations.

B. Acquisition of Bilingual Concepts and Facts

In our ACLS, the bilingual concepts and facts serve as the basic units for summary sentence generation. This section introduces how to extract the set of bilingual concepts and facts.

We first determine the concepts and facts in the English sentences and then find the corresponding Chinese counterparts. Given an English sentence annotated with PAS, we evaluate each PAS to check whether it is valid for concept and fact extraction. Here, a PAS is valid if and only if there are at least ARG0 and ARG1 in it[3]. Then, for each valid PAS, we define and extract the concepts and facts for each predicate in a monolingual manner.

Concept: ARG0 is regarded as the concept of an event. For example, *president George Bush* and *president bush* are concepts in Fig. 1.

Fact: predicate+ARG1 (ARG2) is defined as the *core fact* of an event. AM-TMP (AM-LOC) is defined as the *auxiliary fact*. The auxiliary fact appears in the final summary only when the corresponding core fact appears. Take Fig. 1 as an example, *made his second visit*, *made to the region*, *authorised federal disaster assistance*, *authorised for the affected areas* and *made plans for an inspection tour of the state* are core facts. *yesterday* and *since the hurricane hit* are auxiliary facts.

Other arguments such as ARG3, ARG4, AM-ADV and AM-MOD are not considered in our framework because they are not core elements of an event. It is worth noting that many PAS are recursive. That is one PAS is part of another, just as Fig. 2 shows. Nevertheless, they are equally treated when we construct the pool of bilingual concepts and facts. When generating the final summary, we will impose constraints to avoid the co-occurrence of the recursive PAS (as shown in Equation (15)).

To obtain the aligned Chinese part for the given English concepts and facts, we need word alignment between the English sentence and its Chinese translation. However, such information is not provided by Google Translate. To solve this problem, we attempt to design an algorithm to automatically align the sentence pair in word level.

The main idea behind this algorithm is that an unsupervised method using latent-variable log-linear models is leveraged first to learn the word alignment model on a large set of bilingual sentence pairs (English and Chinese). Then the model is employed to predict the word alignment between the sentences in the English documents and their Chinese translations.

It is well-known that adequate unlabelled data (without word alignment annotation) would lead to effective models leant by unsupervised approaches. Fortunately, there are a lot of parallel data for Chinese and English in MT community. As we deal with news summarization in this study, we use the news data released by LDC[4] which consists of about 2.1 million Chinese-English sentence pairs.

Given the large-scale parallel sentence pairs, we apply the unsupervised method proposed in [12] which designs a latent-variable log-linear model with the contrastive estimation strategy to learn the word alignment model with non-local features. The method is to optimize the following objective function:

$$\begin{aligned} P(\mathbf{x},\theta) &= \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} P(\mathbf{x},\mathbf{y};\theta) \\ &= \frac{\sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \exp(\theta\cdot\phi(\mathbf{x},\mathbf{y}))}{\mathbf{Z}(\theta)} \end{aligned} \tag{1}$$

$\mathbf{x}$ is sentence pairs without word alignment. $\mathbf{y}$ denotes the latent-variable word alignment. $\phi(\mathbf{x},\mathbf{y})$ can be any local and global feature such as lexical translation probability obtained by the toolkit GIZA++[5]. θ is the feature weight and $\mathbf{Z}(\theta)$ denotes partition function which is difficult to calculate. To avoid the calculation of $\mathbf{Z}(\theta)$, [12] employs the noisy contrastive estimation method to optimize the following objective:

$$J(\theta) = \log \prod_{i=1}^{I} \frac{P(\mathbf{x}^{(i)};\theta)}{P(\widetilde{\mathbf{x}}^{(i)};\theta)}. \tag{2}$$

For each parallel sentence pair $\mathbf{x}$, a noisy example $\widetilde{\mathbf{x}}$ is randomly generated. Then, the algorithm is optimized to guarantee $\mathbf{x}$ have higher probability than $\widetilde{\mathbf{x}}$ and partition function $\mathbf{Z}(\theta)$ can be cancelled out.

With the trained log-linear model, the English document sentences and their Chinese translations can be automatically annotated with word alignment. Top half in Fig. 3 gives an example for an English sentence, its Chinese translation and the word alignment. Using the word alignment, we can determine the

[3]This constraint guarantees the completeness of an event.

[4]https://www.ldc.upenn.edu/. The catalogue number includes LDC2000T50, LDC2002L27, LDC2002T01, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2005T34.

[5]http://www.statmt.org/moses/giza/GIZA++.html

Chinese part[6] of each English arguments and form the bilingual concepts and facts just as shown in the bottom half of Fig. 3.

C. Translation and Salience Score Calculation

Given the pool of extracted bilingual concepts and facts, the next task is to calculate their scores which provide the evidence whether they should be chosen in final summary generation. In our CLS framework, we expect the Chinese summary sentence to be both salient and well translated. Thus, we consider two types of scores: translation score and salience score.

For the translation score, we not only measure the translation confidence of the bilingual concepts and facts, but also measure the fluency of the Chinese translation. Translation confidence estimation is a tough issue in MT community. Here, we adopt the geometric average of the lexical translation probabilities to approximate the translation confidence. Given a bilingual concept or fact (e, c) and a word alignment a between the English word positions $i = 0, \ldots, n$ and the Chinese word positions $j = 0, \ldots, m$, we calculate $p(c|e, a)$ using the following formulae:

$$p(c|e,a) = \left\{ \prod_{0}^{m} \frac{1}{|\{i|(i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(c_j|e_i) \right\}^{\frac{1}{m+1}}. \quad (3)$$

In which $p(c_j|e_i)$ denoting the lexical translation probability of the Chinese word c_j given the English word e_i can be obtained using GIZA++ on the same news bilingual data mentioned in the previous section[7]. The geometric average is employed to eliminate the influence of length variation. $p(e|c, a)$ can be computed in a similar way. Then, the average of $p(c|e, a)$ and $p(e|c, a)$ will be used to approximate the translation confidence. In our future work, we will try other methods, such as measuring the translation confidence in phrase embedding space [13] or in tree structures [14].

We measure the fluency of the Chinese translation with a word-based trigram language model as follows:

$$p(c) = p(c_0) \cdot p(c_1) \cdot \prod_{j=2}^{m} p(c_j|c_{j-2}c_{j-1}). \quad (4)$$

The language model is trained using the toolkit SRILM[8] with Kneser-Ney discounting strategy on the Chinese side of the parallel news data which is introduced in the previous section. The product of translation confidence and language model probability will serve as the TM score.

The salience score indicates the importance of a concept or fact in the documents. Previous methods designed many types of salience, such as position-based method [15], ranking score based method [2], [16], concept-based method [17], [18] and statistical feature based method [19]. Instead of sentences considered in the most of the approaches, bilingual phrases are our focus, which allows us to design more appropriate silence calculation methods.

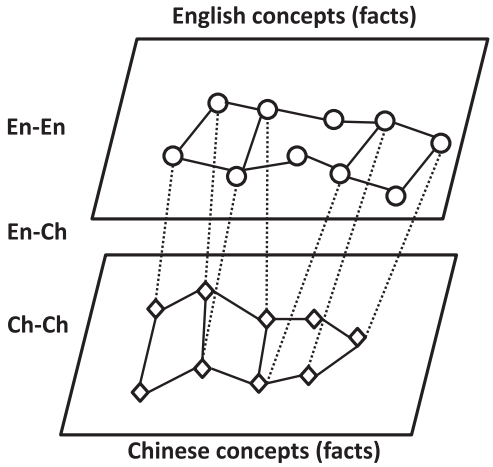


Fig. 4. The simplified illustration of the bilingually-informed CoRank algorithm. Each vertex denotes a English or Chinese concept (fact) and each edge represents the relationship between two vertexes.

On one hand, we adopt the concept-based approach proposed by [17] in which named entity coverage is calculated to denote the salience score. Since named entities should be the same in translation equivalents, we just need to recognize the named entities in the English side for the bilingual concepts and facts. We first apply Stanford Named Entity Recognizer [20] to extract all the English named entities in bilingual concepts and facts. Suppose an English document contains N unique named entities and a bilingual concept (or fact) in this English document has n unique named entities. Then we use n/N as the NE salience score.

On the other hand, we attempt to take full advantage of the bilingual information and try to adapt the bilingually-informed sentence CoRank algorithm [2] to calculate ranking salience score for the concepts and facts. The main idea behind the CoRank algorithm is that the source English concepts (facts) and the translated Chinese parts are simultaneously ranked using a unified graph-based algorithm (as shown in Fig. 4).

The CoRank algorithm first needs to calculate three similarity matrices $M^{\mathrm{en}} = (M^{\mathrm{en}}_{\mathrm{ij}})_{n\times n}$, $M^{\mathrm{ch}} = (M^{\mathrm{ch}}_{\mathrm{ij}})_{n\times n}$ and $M^{\mathrm{ench}} = (M^{\mathrm{ench}}_{\mathrm{ij}})_{n\times n}$. $(M^{\mathrm{en}}_{\mathrm{ij}})_{n\times n}$ is computed as follows:

$$(M^{\mathrm{en}}_{\mathrm{ij}})_{n\times n} = \begin{cases} sim_{\mathrm{cosine}}(ph^{\mathrm{en}}_i, ph^{\mathrm{en}}_j), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

In which $sim_{\mathrm{cosine}}(ph^{\mathrm{en}}_i, ph^{\mathrm{en}}_j)$ is calculated with standard cosine measure and each term weight is computed using TF-IDF formulae. $M^{\mathrm{ch}} = (M^{\mathrm{ch}}_{\mathrm{ij}})_{n\times n}$ is calculated in a similar way on the word level. And $M^{\mathrm{ench}}_{\mathrm{ij}}$ is computed using the following equation

$$M^{\mathrm{ench}}_{\mathrm{ij}} = \sqrt{sim_{\mathrm{cosine}}(ph^{\mathrm{en}}_i, ph^{\mathrm{en}}_j) \times sim_{\mathrm{cosine}}(ph^{\mathrm{ch}}_i, ph^{\mathrm{ch}}_j)}. \quad (6)$$

Note that these three matrices are symmetric and each row will be normalized. The ranking salience scores for English and

[6]We assume that the aligned Chinese part is contiguous. We will finally delete the unrelated words if the Chinese part is discontinuous.

[7]It should be noted that lexical translation probability is obtained using a different training data from that used for Google Translate since we do not know the training data adopted by Google.

[8]http://www.speech.sri.com/projects/srilm/

Chinese are denoted by $\mathbf{v} = [v(ph_i^{\text{en}})]_{n\times 1}$ and $\mathbf{u} = [u(ph_i^{\text{ch}})]_{n\times 1}$ and they are calculated iteratively using the following equations:

$$v(ph_i^{\text{en}}) = \alpha \sum_j M_{\text{ji}}^{\text{en}} v(ph_j^{\text{en}}) + \beta \sum_j M_{\text{ji}}^{\text{ench}} u(ph_j^{\text{ch}}) \quad (7)$$

$$u(ph_i^{\text{ch}}) = \alpha \sum_j M_{\text{ji}}^{\text{ch}} u(ph_j^{\text{ch}}) + \beta \sum_j M_{\text{ji}}^{\text{ench}} v(ph_j^{\text{en}}). \quad (8)$$

We then employ the sum $rs(ph_i) = v(ph_i^{\text{en}}) + u(ph_i^{\text{ch}})$ to represent the ranking score of the i_{th} concept or fact.

Finally, the weighted sum of TM score, NE score and ranking score is employed to denote the overall score of a bilingual concept or fact.

D. Summary Sentence Construction

With the pool of bilingual concepts (facts) and their scores, our problem becomes how to choose the desirable subset of the bilingual concepts and facts in order to construct the final Chinese summary. A summary sentence is composed by a concept and at least one core fact. We formalize this problem as an integer linear optimization task which generates all the summary sentences at the same time. We first introduce the objective function and then detail all the constraints.

1) Objective Function: Our CLS framework aims at producing a Chinese summary which can maximize both of the salience score and the translation score. Each summary sentence is generated by composing the best concepts and facts which optimize the following objective function:

$$\begin{aligned}\max: \quad & \sum_i \alpha_i S_i^C - \sum_{i<j} \alpha_{\text{ij}} (S_i^C + S_j^C) Sim_{\text{ij}}^C \\ & + \sum_i \beta_i S_i^{\text{CF}} - \sum_{i<j} \beta_{\text{ij}} (S_i^{\text{CF}} + S_j^{\text{CF}}) Sim_{\text{ij}}^{\text{CF}} \\ & + \sum_i \gamma_i S_i^{\text{AF}} - \sum_{i<j} \gamma_{\text{ij}} (S_i^{\text{AF}} + S_j^{\text{AF}}) Sim_{\text{ij}}^{\text{AF}}. \quad (9)\end{aligned}$$

In which S_i^C, S_i^{CF} and S_i^{AF} denote the score of concept C_i, core fact CF_i and auxiliary fact AF_i respectively. α_i, β_i and γ_i are the selection indicators about whether the concept C_i (core fact CF_i and auxiliary fact AF_i) should be chosen. α_{ij}, β_{ij} and γ_{ij} are the co-occurrence indicators for the pairs of concepts (core facts and auxiliary facts). Sim_{ij} indicates the similarity score between C_i (CF_i, AF_i) and C_j (CF_j, AF_j), and it is calculated using $sim_{\text{cosine}}(C_i, C_j)$ which is introduced in the previous section.

This objective is similar to that used by [7] in which the authors try to generate English summary sentences by merging noun phrases and verb phrases in the monolingual multi-document summarization task. In contrast, we adopt a more natural way that defines an event by using PAS and design a new TM augmented framework for CLS. Equation (9) attempts to pick the concepts and facts having highest scores with the first, third and fifth terms, and meanwhile discourage the selection of similar concepts and facts by using the second, fourth and last terms. Note that the selected concepts and facts should come from the same PAS in the original sentence or satisfy the compatibility constraint. Next, we first introduce the compatibility constraint between concepts and facts, and then we detail other important constraints which guarantee the validity of the final summary sentences.

2) Compatibility Constraint: A concept and a fact can be chosen to make up a summary sentence only when they meet the compatibility constraint. Given the concept set $\mathbf{C}$, core fact set $\mathbf{CF}$ and auxiliary fact set $\mathbf{AF}$, we define two matrices $\mathbf{I}_{|\mathbf{C}||\mathbf{CF}|}$ and $\mathbf{I}_{|\mathbf{CF}||\mathbf{AF}|}$ indicating whether two elements from different sets are compatible. We define each matrix as follows:

$$\mathbf{I}_{\mathbf{C}_i\mathbf{CF}_j} = \begin{cases} 1, & \text{if } sim(\mathbf{CF}_{i'}, \mathbf{CF}_j) > \text{threshold} \\ 1, & \text{if } sim(\mathbf{C}_i, \mathbf{C}_{j'}) > \text{threshold} \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where $\mathbf{C}_i$ and $\mathbf{CF}_{i'}$ ($\mathbf{C}_{j'}$ and $\mathbf{CF}_j$) belong to the same PAS. It also includes the case that $\mathbf{C}_{i'} = \mathbf{C}_i$ and $\mathbf{CF}_{j'} = \mathbf{CF}_j$. $sim(\mathbf{CF}_{i'}, \mathbf{CF}_j)$ is calculated by cosine similarity in which we further require that the headwords must be the same. For example, we require the predicates are identical in two core facts during similarity computation. If the similarity is bigger than the threshold, we assume that $\mathbf{CF}_{i'}$ is an alternative to $\mathbf{CF}_j$

$$\mathbf{I}_{\mathbf{CF}_i\mathbf{AF}_j} = \begin{cases} 1, & \text{if } \mathbf{CF}_i \text{ and } \mathbf{AF}_j \text{ in the same PAS} \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

By defining this matrix, we require that a core fact and an auxiliary fact are compatible only if they come from the same PAS.

If $\mathbf{I}_{\mathbf{C}_i\mathbf{CF}_j} = 1$, both the concept $\mathbf{C}_i$ and the fact $\mathbf{CF}_j$ can be selected simultaneously to form a summary sentence. We then define an indicator $\theta_{i,j}$ to indicate whether $\mathbf{C}_i$ and $\mathbf{CF}_j$ co-occur in the final summary sentences. In a similar way, $\delta_{i,j}$ is defined to indicate if $\mathbf{CF}_i$ and $\mathbf{AF}_j$ co-appear in the final summary sentences.

Three compatibility constraints are designed for concepts, core facts and auxiliary facts respectively.

Concept Compatibility Constraint: It enforces the selection of the concepts to be consistent with the compatibility matrix $\mathbf{I}_{|\mathbf{C}||\mathbf{CF}|}$:

$$\forall i, j, \alpha_i \geq \theta_{\text{ij}}; \forall i, \sum_j \theta_{\text{ij}} \geq \alpha_i. \quad (12)$$

Core Fact Compatibility Constraints: These constraints maintain the consistency between the selection of the core facts and the concepts (between the core facts and auxiliary facts):

$$\begin{aligned} \forall j, \sum_i \theta_{i,j} &= \beta_j \\ \forall i, j, \beta_i \geq \delta_{\text{ij}}; \forall i, \sum_j \delta_{\text{ij}} &\geq \beta_i. \end{aligned} \quad (13)$$

Auxiliary Fact Compatibility Constraint: It makes sure that the selection of the auxiliary facts to be consistent with the compatibility matrix $\mathbf{I}_{|\mathbf{CF}||\mathbf{AF}|}$:

$$\forall j, \sum_i \delta_{i,j} = \gamma_j. \quad (14)$$

3) Other Constraints: We define some constraints which are similar to those used in the abstractive monolingual summarization [7].

Not-Within Constraints: Concepts (or facts) from two recursive PAS structures (illustrated by Fig. 2) cannot be selected simultaneously since they are obviously redundant

$$\begin{aligned}\alpha_i + \alpha_j &\leq 1, \text{if } \exists\, PAS_{\mathbf{C}_i} \subset PAS_{\mathbf{C}_j} \\ \beta_i + \beta_j &\leq 1, \text{if } \exists\, PAS_{\mathbf{CF}_i} \subset PAS_{\mathbf{CF}_j} \\ \gamma_i + \gamma_j &\leq 1, \text{if } \exists\, PAS_{\mathbf{AF}_i} \subset PAS_{\mathbf{AF}_j}\end{aligned} \tag{15}$$

where $PAS_{\mathbf{C}_i} \subset PAS_{\mathbf{C}_j}$ says that the PAS containing $\mathbf{C}_i$ is nested within the PAS containing $\mathbf{C}_j$.

Co-Occurrence Constraints: We know the fact that $\alpha_{\text{ij}} = 1$ indicating the concepts $\mathbf{C}_i$ and $\mathbf{C}_j$ are both selected ($\alpha_i = 1$ and $\alpha_j = 1$). It is also correct for the inverse case. Thus, the following constraints must be satisfied.

$$\begin{aligned}\alpha_{\text{ij}} - \alpha_i &\leq 0 \\ \alpha_{\text{ij}} - \alpha_j &\leq 0 \\ \alpha_i + \alpha_j - \alpha_{\text{ij}} &\leq 1.\end{aligned} \tag{16}$$

The similar constraints for core facts $\mathbf{CF}$ and auxiliary facts $\mathbf{AF}$ are listed as follows:

$$\begin{aligned}\beta_{\text{ij}} - \beta_i &\leq 0 \\ \beta_{\text{ij}} - \beta_j &\leq 0 \\ \beta_i + \beta_j - \beta_{\text{ij}} &\leq 1\end{aligned} \tag{17}$$

$$\begin{aligned}\gamma_{\text{ij}} - \gamma_i &\leq 0 \\ \gamma_{\text{ij}} - \gamma_j &\leq 0 \\ \gamma_i + \gamma_j - \gamma_{\text{ij}} &\leq 1.\end{aligned} \tag{18}$$

Pronoun Constraints: As suggested in [7], [19], pronouns are normally not used by human summary writers due to the length limitation and content adequacy of the summary. Thus, the concepts which are pronouns should be excluded in our CLS framework:

$$\alpha_i = 0, \text{if } \mathbf{C}_i \text{ is pronoun.} \tag{19}$$

Sentence Number Constraints: We control the sentence number of the final summary to make the summary as concise as possible

$$\sum_i \alpha_i \leq K. \tag{20}$$

In which K is a predefined maximum sentence number.

Summary Length Constraints: Sentence number constraint cannot guarantee the overall length of the final summary. Therefore, we control the overall length with the following constraint

$$\sum_i \alpha_i l(\mathbf{C}_i) + \sum_i \beta_i l(\mathbf{CF}_i) + \sum_i \gamma_i l(\mathbf{AF}_i) \leq L. \tag{21}$$

In which L is a predefined maximum summary length (e.g. 100 words). The above two constraints should be employed at the same time as suggested in [7] because we find in experiments that only one constraint cannot guarantee the overall length of the summary.

In addition to the above constraints adopted, we further design several constraints to control the length of concepts and facts respectively.

Concept Length Constraints: The concepts indicating the topic focus vary in their lengths. In order to maintain the diversity of the topics, we control the overall length of concepts in the final summary:

$$\sum_i \alpha_i l(\mathbf{C}_i) \geq L_{\mathbf{C}} \tag{22}$$

where $L_{\mathbf{C}}$ is a predefined minimum concept length.

Core Fact Length Constraints: It is suggested in [19] that short core fact cannot convey the complete information of an event. Thus, we avoid the selection of core fact whose length is too short

$$\forall i, \beta_i \geq L_{\mathbf{CF}}^{\min} \tag{23}$$

where $L_{\mathbf{CF}}^{\min}$ is the predefined minimum core fact length.

Note that the above objective function and all the constraints are linear. Thus, ILP can solve such kind of problem. In this work, we employ *lp_solve* package[9] to get the exact solution to our problem.

E. Postprocessing

After running the ILP toolkit, we can obtain the best concept subset $\hat{\mathbf{C}}$, core fact subset $\hat{\mathbf{CF}}$ and auxiliary fact subset $\hat{\mathbf{AF}}$. Now, we need to combine the elements from different sets into valid sentences. This task is performed in two steps. First, we determine which ones belong to the same summary sentence. Second, we reorganize the elements to make them in good order.

In the first step, we know that a summary sentence contains one and only one concept. Therefore, we pick a concept $\hat{\mathbf{C}}_i$ from $\hat{\mathbf{C}}$. Then, we find related core and auxiliary facts according to the compatibility relations θ and δ.

In the second step, we determine the final order of concept $\hat{\mathbf{C}}_i$ and its related facts. We put the concept at the beginning and then rearrange the facts according to their natural order in the original documents. If two facts are located in different documents, we use the timestamps of the documents.

It is worth noting that the concepts and facts are bilingual and include English and Chinese parts respectively. Therefore, we have two choices to obtain the final Chinese summary. On one hand, we can directly use the Chinese part to serve as the summary sentences. On the other hand, we can first get the English summary and then automatically translate it to Chinese with Google Translate again. We will compare the performance between these two alternatives.

[9]http://lpsolve.sourceforge.net/

III. Experiments

A. Experimental Setup

Most of the available summarization datasets, such as Document Understanding Conferences (DUC[10]) and TAC[11], are all prepared for multi-document summarization in monolingual language. In contrast, there are few special datasets for CLS. In this study, we employ the same dataset[12] used by [2], [3] to evaluate the performance of our method. This dataset is just an extension to the original DUC 2001 evaluation data for English multi-document summarization, in which there are 30 English document sets. Each set containing a bunch of documents talks about the same topic. NIST organizers provide three human written summaries in English for each document set (about 100 words in each summary). These reference summaries are then manually translated into Chinese so that the English documents and the Chinese summaries become the test data for CLS.

In our ACLS framework, there are several hyper-parameters. We empirically set these hyper-parameters as follows. Through manual evaluation in 200 instances, we find that the best similarity threshold in Equation (10) for finding alternative concepts and facts is $\frac{2}{3}$. Since the reference summary is about 100 words, we set the sentence number K in Equation (20) and summary length L in Equation (21) to $K = 5$ and $L = 100$. We use $L_C = 18$ and $L_{CF}^{\min} = 4$[13] to avoid too short concepts and core facts.

Next, we first evaluate our method using automatic evaluation metric ROUGE [21] and then give the results of manual linguistic quality evaluation.

B. Experimental Results

We will compare our ACLS system with other four baselines:

1) *Baseline-EN*: This system first performs extractive summarization on English documents and then automatically translate the English summary into Chinese. It is called summarization-translation scheme.
2) *Baseline-CH*: This baseline first automatically translates all the English documents into Chinese and then extracts summary from the Chinese translations. It is named translation-summarization scheme.
3) *CoRank*: It is a state-of-the-art extractive CLS system that designs a graph-based CoRank algorithm to extract the Chinese summary sentences using both English and Chinese information [2].
4) *PBCS*: This a phrase-based compressive CLS system which performs summary sentence extraction and redundant phrase deletion simultaneously [3].

In our ACLS framework, we investigate four variations according to the features used and summary generation styles:

1) *ACLS-EN*: This is our ACLS system that employs only the salience score (NE score and ranking score) to indicate the importance of a concept or fact. And it first generates the English summary sentences by composing English parts of the selected bilingual concepts and facts, and the English summary is then automatically translated into Chinese summary.
2) *ACLS-EN-TM*: This system is similar to ACLS-EN except that it adopts both the salience score and the TM score as the significance indicator of the concepts and facts.
3) *ACLS-CH*: This system uses the same features as ACLS-EN. The difference is that ACLS-CH directly generates the Chinese summary from the bilingual concepts and facts. ACLS-EN assumes that the composed English sentence is very good and will lead to better Chinese translations. ACLS-CH assumes that the English summary sentences may contain some noise and will lead to bad Chinese translations.
4) *ACLS-CH-TM*: It applies the same Chinese summary generation style as ACLS-CH does and it utilizes the same feature as that used in ACLS-EN-TM.

To verify the superiority of our PAS-based abstractive summarization method, we further compare our approach to another abstractive summarization method which generates summary by selecting and merging the salient and well-translated noun and verb phrases. It extends the monolingual summarization approach [7] and adds the TM score when calculating the significance of the noun and verb phrases. We call this phrase-based abstractive summarization method *PBAS*.

To avoid the inconsistency problem of Chinese word segmentation, we just evaluate all the systems on the Chinese character level. ROUGE-1.5.5 package[14] is employed for automatic evaluation. This evaluation metric measures the summary quality by counting the matched units such as ngrams and skip ngrams between system summary and the reference summary. We report results using the following scores: ROUGE-1 (unigrams), ROUGE-2 (bigrams) and ROUGE-SU4 (skip bigram with a maximum gap of 4).

Table I gives the ROUGE scores of our proposed methods and the baselines. We find the similar phenomenon reported in [2], [3] that translation-summarization scheme (Baseline-CH) performs better than summarization-translation scheme (Baseline-EN). This indicates that Chinese-side information is more helpful than English-side information if our target is to produce Chinese summary. The other two baselines using both-side knowledge (CoRank and PBCS) can obtain better performance. Specifically, the compression-based method PBCS outperforms the sentence extraction based method CoRank. The reason is that PBCS can find and delete some redundant phrases for each sentence, while CoRank just maintains the original sentences. It is interesting that our reimplemented CoRank method differs slightly from the reported CoRank approach in [3]. It is mainly due to different outputs of Google Translate in different time periods.

The last four rows in Table I demonstrate that our ACLS framework can outperform all the extraction-based baselines in all metrics. When comparing our methods to the compression-

[10] http://duc.nist.gov/

[11] http://www.nist.gov/tac/

[12] Many thanks to Wan's research group for providing us this dataset.

[13] We also test $L_C = 16, 17, 19, 20$ and $L_{CF}^{\min} = 3, 5$. And we observe no significant difference for these settings.

[14] http://www.berouge.com/Pages/default.aspx

TABLE I
EXPERIMENTAL RESULTS OF ROUGE EVALUATION FOR DIFFERENT CLS SYSTEMS

System	ROUGE-1			ROUGE-2			ROUGE-SU4		
	P	R	F1	P	R	F1	P	R	F1
Baseline-EN	0.35666	0.37072	0.35093	0.10313	0.11179	0.10375	0.11943	0.12599	0.11829
Baseline-CH	0.36367	0.36903	0.35822	0.10749	0.11215	0.10740	0.12256	0.12576	0.12133
CoRank (reported)	N/A	N/A	0.37601	N/A	N/A	0.12570	N/A	N/A	0.13352
PBCS (reported)	N/A	N/A	0.37890	N/A	N/A	**0.13549**	N/A	N/A	0.14098
CoRank (reimplemented)	0.36852	0.39647	0.37551	0.11328	0.12349	0.11617	0.12649	0.13668	0.12935
PBAS-EN-TM	0.36768	0.40887	0.38430	0.11652	0.12696	0.12008	0.13020	0.14073	0.13351
PBAS-CH-TM	0.37725	0.40633	0.38470	0.11917	0.12923	0.12196	0.13350	0.14348	0.13597
ACLS-EN	0.37074	0.41321	0.38848	0.11859	0.13122	0.12379	0.13133	0.14619	0.13748
ACLS-EN-TM	0.36613	**0.42896**	0.39299	0.11783	0.12604	0.12633	0.13067	**0.15310**	0.14022
ACLS-CH	0.37516	0.40550	0.38810	0.11966	0.12931	0.12376	0.13333	0.14431	0.13800
ACLS-CH-TM	**0.37554**	0.42354	**0.39393**	**0.12121**	**0.13649**	0.12770	**0.13455**	0.15167	**0.14181**

In ACLS-EN-TM and ACLS-CH-TM, the weights of the NE score, ranking score and TM score are empirically set by 0.2, 0.5 and 0.3.

based system PBCS, we find that our performance is much better in ROUGE-1 and slightly better in ROUGE-SU4, while PBCS wins in ROUGE-2. This phenomenon is easy to understand. The summary sentence obtained by PBCS comes from single original sentence and it can guarantee the performance of bi-gram matching. In contrast, our ACSL framework produces the summary sentence by fusing several original sentences and this may violate the correct word ordering. The results indicate that our method can obtain correct content words and it still remains room to improve the content word reordering. Note that we also conduct significance test on F1-measure using the approximate randomization[36]. We find that the difference between ACLS and other baselines are statistically significant with $p < 0.01$ except for the ROUGE-SU4 metric when compared to PBCS.

Among the four variations of our method, we can see that integrating the TM scores is much beneficial to the performance improvement (ACLS-EN-TM vs. ACLS-EN and ACLS-CH-TM vs. ACSL-CH). This manifests that TM features are indispensable in selecting the concepts and facts with high translation confidence. Overall, the system ACLS-CH-TM obtains the best performance in six out of nine metrics. When it is comparing with ACLS-EN-TM, we can find that it is better to generate Chinese summary by directly composing the Chinese parts of the bilingual concepts and facts rather than getting the English summary first followed by translation again.

We notice from the above experiments that ACLS-CH-TM performs best compared to other variations. In ACLS-CH-TM, the ranking score is the most important, and then the NE score and TM. We first empirically set the weights of the ranking score, TM and NE score 0.5, 0.3 and 0.2. Then, we plan to figure out the question whether the performance is very sensitive to the weight settings. Accordingly, we conduct various experiments by changing the weight settings using a line search strategy, in which the weight of the TM score varies from 0.1 to 0.5 while others from 0.1 to 0.8[15]. Table II gives the detailed results in ROUGE F1 scores. It is interesting to see that most weight settings can lead to good performance. Specifically, the ranking score seems to be the most important and high weights can guarantee the summarization quality. It is in line with our cognition since the first and most important task of summarization is to extract the salient parts from original documents. Table II also demonstrates that TM score is helpful when its weight is not too big (not bigger than 0.4). The best overall performance is obtained when setting the weights to be 0.4 for the NE score and ranking score, and 0.2 for the TM score.

TABLE II
EXPERIMENTAL RESULTS (ROUGE F1) FOR DIFFERENT WEIGHTS OF NE SCORE, RANKING SCORE AND THE TM SCORE

weights	ROUGE-1	ROUGE-2	ROUGE-SU4
0.1 0.8 0.1	0.39287	0.12506	0.13985
0.2 0.7 0.1	0.39289	0.12591	0.14019
0.3 0.6 0.1	0.39251	0.12599	0.13991
0.4 0.5 0.1	0.39179	0.12593	0.13981
0.5 0.4 0.1	0.39173	0.12549	0.13959
0.6 0.3 0.1	0.38800	0.12439	0.13810
0.7 0.2 0.1	0.38725	0.12177	0.13729
0.8 0.1 0.1	0.38190	0.11390	0.13207
0.1 0.7 0.2	0.39758	0.12922	0.14329
0.2 0.6 0.2	0.39223	0.12584	0.13974
0.3 0.5 0.2	0.39687	0.12803	0.14265
0.4 0.4 0.2	0.39570	0.13063	0.14406
0.5 0.3 0.2	0.39086	0.12309	0.13872
0.6 0.2 0.2	0.38586	0.11879	0.13565
0.7 0.1 0.2	0.37576	0.10978	0.12915
0.1 0.6 0.3	0.39475	0.12705	0.14128
0.2 0.5 0.3	0.39393	0.12770	0.14181
0.3 0.4 0.3	0.39156	0.12421	0.13987
0.4 0.3 0.3	0.38729	0.11914	0.13616
0.5 0.2 0.3	0.38050	0.11558	0.13256
0.6 0.1 0.3	0.37436	0.10945	0.12828
0.1 0.5 0.4	0.39432	0.12817	0.14198
0.2 0.4 0.4	0.39002	0.12153	0.13886
0.3 0.3 0.4	0.38527	0.11809	0.13524
0.4 0.2 0.4	0.37813	0.11327	0.13091
0.5 0.1 0.4	0.37142	0.10859	0.12697
0.1 0.4 0.5	0.38853	0.11820	0.13625
0.2 0.3 0.5	0.38091	0.11568	0.13285
0.3 0.2 0.5	0.37520	0.11209	0.12942
0.4 0.1 0.5	0.36776	0.10630	0.12522

To demonstrate the superiority of our abstractive summarization by merging PAS from different sentences, we further compare our method to the (noun-verb) phrase-based abstractive

[15]The intervals are chosen according to their importance.

approach PBAS (line 7-8 in Table I). PBAS-EN-TM is similar to ACLS-EN-TM except that PBAS-EN-TM generates final summary by fusing noun-verb phrases from different sentences. PBAS-CH-TM is similar to ACLS-CH-TM. In our experiments, we also perform a line search strategy to determine the feature weights of PBAS. In PBAS-EN-TM, the best weights are 0.6, 0.3 and 0.1 for the ranking score, NE score and TM score respectively. In PBAS-CH-TM, the best weights are 0.6, 0.2 and 0.2. The results show that our PAS-based method significantly outperforms the (noun-verb) phrase-based method.

We then analyze the results and find that the difference is mainly due to the granularity problem of the noun and verb phrases. The noun and verb phrases usually ignore the temporal and locational information as shown in Fig. 1, and these phrases are either too long or too short. If a long phrase contains noise or parts of it are badly translated, the phrase will be discarded even if some parts of this phrase are salient and well translated. For example, PBAS can only extract the long verb phrase *made his second visit to the region since the hurricane hit* from the original sentence *NP(President George Bush) NP(yesterday) VP(made his second visit to the region since the hurricane hit)* in Fig. 1. The long verb phrase does not get the high salience and translation score, and the whole phrase is not chosen in the final summary. However, in our method, the sentence in PAS is split into six parts *A0(President George Bush)*, *AM-TMP(yesterday)*, *predicate(made)*, *A1(his second visit)*, *A2(to the region)* and *AM-TMP(since the hurricane hit)*. We then successfully select three salient and well-translated parts, and obtain the summary sentence *President George Bush made his second visit*.

C. Manual Linguistic Quality Evaluation

To test whether our method indeed improves the summary quality compared to the state-of-the-art CoRank method, we ask ten native Chinese annotators to measure the summary quality of different systems in different linguistic aspects. In DUC, the linguistic quality of summary is evaluated in five aspects including: Grammaticality (GR), Non-Redundancy (NR), Referential Clarity (RC), Topical Focus (TF) and Structural Coherence (SC).

Following [3], we calculate the average score and standard deviation[16] for each linguistic aspect and the results are displayed in Table III. Overall, our methods (ACLS-CH and ACLS-CH-TM) achieve higher scores in most of the aspects. Specifically, ACLS-CH is slightly worse than CoRank in GR. It is because that summary sentences in ACLS-CH are generated by combining different parts from several original sentences and some grammar mistakes would happen. By adding the TM features (including language model scores), the system ACLS-CH-TM can perform as well as CoRank does. It indicates that TM scores lead to more fluent summary sentences. For example, a phrase *are more potent spreaders of the infection than are blacks* and its badly translated Chinese part are chosen as a part of summary sentence in ACLS-CH although the translation and language model probability of the bilingual phrase is very low. In contrast, ACLS-CH-TM selects an alternative phrase *are about six times as common among blacks as whites* in the final summary since its Chinese translation has a high translation and language model probability. Using the bilingual information provides the chance for our algorithm to find the key phrase which is both salient and well-translated.

For NR scores, we notice that the three systems perform similarly and we find that the concepts and facts extracted from original sentences still contain unimportant parts. Take the second summary sentence in Fig. 5 as an example, the phrase *face a bill in respect of such properties, Lloyd 's exposure there* contains uninformative part *Loyd's exposure there* which our method fails to recognize.

The scores of RC show that our methods obtain much clearer referential relations than CoRank does. It is due to the constraint that disallows the selection of pronouns such as semantically unimportant *he* or *they*. For TF, we are happy to see that our methods can obtain much higher scores, which reveals that the summary generated by our methods is more cohesive and better fits the theme of the original documents. For example, we can focus on the storm *Hurricane Andrew* in the final summary as shown in Fig. 5. In Table III, we also observe some gains in SC.

D. Case Study

To have a better understanding, we choose the first document set D04 in DUC 2001 dataset to compare the summary results between our proposed abstraction-based method ACLS-CH-TM and the extraction-based model CoRank. Fig. 5 shows the results in which CoRank's output is given on the top half and ACLS-CH-TM's output is listed on the bottom half. The original English parts are also displayed for reference.

The top half in Fig. 5 shows that the extraction-based system CoRank can mainly extract the theme-related summary sentences. The Chinese sentences are relatively fluent due to whole sentence extraction. However, the disadvantages are also very obvious. For one thing, the sentence-based extraction contains much noise and redundancy. For example, *Although there had already been some preliminary guesses at the level of insurance claims* in the second sentence is not important and should be deleted. For another thing, this system is apt to generate Chinese summary sentences which are badly translated. For instance, *watch* in the third sentence is incorrectly translated. Finally and the most importantly, the extraction-based models cannot merge the key information from different original sentences.

In contrast, our abstraction-based method ACLS-CH-TM has the capability to address these issues. As shown in the bottom half of Fig. 5, the summary sentences are generated in different ways. The fourth sentence is a compressed one produced by deleting the unimportant parts of an original sentence. The other four summary sentences are created by merging the key informative phrases (facts) from different original sentences. The second sentence is derived from three original sentences while the remaining three come from two original sentences. Table IV reports the number distribution of original sentences used to generate a new abstractive sentence. We can see from the table that more than 60% summary sentences are constructed by merging two or more original

[16]Each time we randomly choose 20 summaries out of 30.

TABLE III
MANUAL LINGUISTIC QUALITY EVALUATION RESULTS OF DIFFERENT CLS SYSTEMS

System	GR	NR	RC	TF	SC
CoRank	3.02 ± 0.21	3.57 ± 0.19	3.08 ± 0.19	2.97 ± 0.18	2.98 ± 0.19
ACLS-CH	2.95 ± 0.19	3.63 ± 0.20	3.26 ± 0.18	3.16 ± 0.20	3.15 ± 0.21
ACLS-CH-TM	3.17 ± 0.22	3.75 ± 0.16	3.47 ± 0.19	3.34 ± 0.19	3.32 ± 0.22

The last serious US hurricane, Hugo, which struck South Carolina in 1989, cost the industry Dollars 4.2bn from insured losses, though estimates of the total damage caused ranged between Dollars 6bn and Dollars 10bn.
最后一个严重的美国飓风雨果袭击南卡罗来纳州在1989年，从费用保险损失该行业42亿美元，但造成的伤害总量的估计60亿美元和100亿美元之间不等。
Although there had already been some preliminary guesses at the level of insurance claims, yesterday's figure comes from the Property Claims Services division of the American Insurance Services Group, the property-casualty insurers' trade association.
虽然当时已经在保险理赔的水平一些初步猜测，昨天的数字来自美国保险服务集团，财产险保险公司行业协会的财产索赔服务部门。
US CITIES along the Gulf of Mexico from Alabama to eastern Texas were on storm watch last night as Hurricane Andrew headed west after sweeping across southern Florida, causing at least eight deaths and severe property damage.
美国城市沿墨西哥湾的阿拉巴马州到得克萨斯州东部是在风暴手表昨晚安德鲁飓风向西横跨佛罗里达州南部席卷后，造成至少八人死亡和严重的财产损失。
GENERAL ACCIDENT, the leading British insurer , said yesterday that insurance claims arising from Hurricane Andrew could cost it as much as Dollars 40m.
一般事故，英国着名保险公司昨日表示，从安德鲁飓风所带来的保险索赔可能“成本就高达4000万美元。
The town of Homestead, near the centre of the storm, was largely flattened, including a local air force base.
家园镇，靠近风暴的中心，主要是夷为平地，其中包括当地的空军基地。

Hurricane Andrew hit the eastern coast of Florida, causing billions of dollars of property damage and at least 12 deaths.
安德鲁飓风所袭击佛罗里达州的东海岸，造成数十亿美元的财产损失和至少12人死亡。
US insurers an estimated dollars 7.3bn (pounds 3.7bn) , facing their worst-ever year for catastrophe losses, face a bill in respect of such properties, Lloyd 's exposure there.
美国的保险公司估计73亿美元（37亿英镑），他们面临有史以来最严重的一年，巨灾损失，的保险公司将面临一项法案，英国劳氏曝光那里。
President Bush made his second visit, authorised federal disaster assistance for the affected areas.
布什总统做了他的第二次访问，授权给受灾地区的联邦救灾援助。
One of the costliest US storms this century threatened a further devastating landfall near the city of New Orleans.
最昂贵的美国的风暴之一本世纪威胁靠近新奥尔良市的又一个毁灭性的登陆。
The storm to Louisiana's important oil-refining industry, caused more than dollars 20bn of damage.
风暴路易斯安那州重要的炼油行业，造成超过200亿美元的损失。

Fig. 5. The comparison of summary results between ACLS-CH-TM and CoRank on D04 of DUC 2001. Both the original English sentences and Chinese sentences are given. The top half shows the summary result of CoRank and the bottom half displays the summary result generated by our method ACLS-CH-TM. The phrases in the same color denote that they are from the same original sentence.

TABLE IV
THE NUMBER DISTRIBUTION OF THE ORIGINAL SENTENCES USED FOR GENERATING A NEW SENTENCE BY OUR ABSTRACTION-BASED METHOD ACLS-CH-TM

1	2	3	≥ 4
0.36	0.267	0.193	0.18

sentences. In the standard extraction-based and compression-based approaches, several sentences discussing the same topic may be discarded due to the redundancy processing. The summary results shown in Fig. 5 reveals that the proposed abstraction-based framework is able to combine important information from different original sentences by fusing PAS sharing the same concept.

E. Efficiency of the Proposed Algorithm

The automatic and human evaluations have shown that our proposed algorithm performs better than other methods in summarization quality. In this subsection, we discuss the efficiency of the proposed algorithm. The most time-consuming part is the module applying ILP to select the salient and well-translated arguments. The run-time varies in different document sets and ranges from 2 seconds to 4 minutes. According to our analysis on the 30 document sets, more than 25 document sets only need less than 10 seconds. Since we run the algorithm in parallel, the overall demanding run-time is the same as the slowest one which requires about 4 minutes. We will further improve the efficiency of the proposed algorithm in our future work.

IV. RELATED WORK

Research works in summarization can be grouped into three categories: extraction-based methods, compression-based methods and abstraction-based methods. In CLS which we focus in this paper, researchers have investigated only the methods of the first two categories.

Extraction-based approaches dominate the CLS community. Pilot studies investigate this task by utilizing only one language side information (source or target language side) [1], [22]–[26]. Two main schemes are employed in extraction-based

methods: summarization-translation scheme and translation-summarization scheme. The former one first produces the summary in the source language and then automatically translate it into target language. While the latter one first gets the target translations of source language documents and then extract summary sentences in target language. [2] conducts a comparison and finds that translation-summarization scheme performs better. Furthermore, [2] proposes a graph-based CoRank algorithm using bilingual knowledge to extract target language summary sentences. The drawback is that the extraction-based approaches would retain much noise and redundancy due to the whole sentence selection.

Compression-based approaches are recently studied in CLS by [3], in which sentence selection and compression are performed simultaneously. The phrase alignment information of the bilingual sentences are exploited to calculate the sentence scores. Meanwhile, the unimportant phrases which are redundant or badly translated will be discarded to keep the summary sentence as compact as possible. However, the compression-based methods cannot merge the complementary information from different original sentences.

Abstraction-based approaches are capable to solve all the above problems in theory but they are not well studied in CLS. In monolingual multi-document summarization, abstraction-based methods [7], [27]–[35] are deeply explored in recent years. For example, [28] and [29] propose the summary revision method that revises the summary sentence by rewriting the key phrases in it. [35] introduces an attention-based neural network model in order to generate new sentence summary from original sentences. The most relevant work to ours extracts *informative items* [31], *abstraction schemes* [32] or *noun and verb phrases* [7] as the basic units for producing the summary sentences. The difference between our method and these approaches lies in two aspects. On one hand, the event definition is not so clear in previous approaches. For instance, [7] assumes that a noun phrase followed by a verb phrase becomes an event. However, the noun phrase and verb phrase are extracted heuristically and contain much noise. In contrast, we use the well-defined PAS as the events and the arguments can be extracted without any heuristics. On the other hand, these studies concern only the monolingual summarization task. It is well-known that CLS faces more challenges. It requires that the summary should be not only concise and informative, but only well translated. In our proposed method, we integrate the TM scores together with the salience scores, which leads to better target language summaries.

V. Conclusion and Future Work

In this paper, we have presented a novel ACLS framework. We first assume that an event is composed of the bilingual concepts and facts which are represented by the source language PAS and their target language translations. We finally generate the target language summary by merging multiple bilingual PAS structures using integer linear optimization that attempts to maximize the salience score and the translation quality simultaneously. The automatic and manual evaluations show that our proposed framework performs much better than the state-of-the-art CLS methods.

In the future work, we plan to design a better algorithm for translation confidence estimation. Furthermore, we are going to annotate more CLS dataset so that we are able to automatically tune the hyper-parameters with a development set.

Acknowledgment

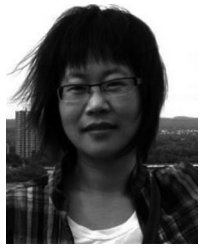
The authors would like to thank Prof. X. Wan for the great help in preparing the draft and the reviewers for their valuable suggestions.

References

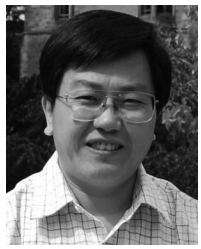
[1] X. Wan, H. Li, and J. Xiao, "Cross-language document summarization based on machine translation quality prediction," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 912–926.
[2] X. Wan, "Using bilingual information for cross-language document summarization," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, 2011, pp. 118–127.
[3] J. Yao, X. Wan, and J. Xiao, "Phrase-based compressive cross-language summarization," in *Proc. Empirical Methods Natural Lang. Process.*, 2015, pp. 388–395.
[4] R. Barzilay and K. McKeown, "Sentence fusion for multidocument news summarization," *Comput. Linguistic*, vol. 31, no. 3, pp. 297–328, 2005.
[5] K. Filippova and M. Strube, "Sentence fusion via dependency graph compression," in *Proc. Empirical Methods Natural Lang. Process.*, 2008, pp. 177–185.
[6] K. Filippova, "Multi-sentence compression: Finding shortest paths in word graphs," *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 322–330.
[7] L. Bing, P. Li, Y. Liao, W. Lam, W. Guo, and R. Passonneau, "Abstractive multi-document summarization via phrase selection and merging," in *Proc. Assoc. Comput. Linguistics*, 2015, pp. 1587–1597.
[8] D. Gildea and D. Jurafsky, "Automatic labeling of semantic roles," *Comput. Linguistics*, vol. 28, no. 3, pp. 245–288, 2002.
[9] H. Yang and C. Zong, "Multi-predicate semantic role labeling," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 363–373.
[10] H. Yang, Y. Zhou and C. Zong, "Bilingual semantic role labeling inference via dual decomposition," *ACM Trans. Asian Lang. Low-Resour. Lang. Inform. Process*, vol. 15, 2015, Art. no. 15.
[11] T. Zhuang and C. Zong, "Joint inference for bilingual semantic role labeling," in *Proc. Empirical Methods Natural Lang. Process.*, 2010, pp. 304–314.
[12] Y. Liu and M. Sun, "Contrastive unsupervised word alignment with non-local features," in *Proc. Assoc. Adv. Artif. Intell.*, 2015, pp. 2295–2301.
[13] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, "Bilingually-constrained phrase embeddings for machine translation," in *Proc. Assoc. Comput. Linguistics*, 2014, pp. 111–121.
[14] J. Zhang, F. Zhai, and C. Zong, "Augmenting string-to-tree translation models with fuzzy use of source-side syntax," in *Proc. Empirical Methods Natural Lang. Process.*, 2011, pp. 204–215.
[15] W. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multi-document summarization by maximizing informative content-words," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 1776–1782.
[16] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic-focused multi-document summarization," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, 2007, pp. 2903–2908.
[17] H. Li, Y. Hu, Z. Li, X. Wan, and J. Xiao, "Pkutm participation in tac2011," in *Proc. Text Anal. Conf.*, 2011.
[18] F. Boudin, H. Mougard1, and B. Favre, "Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions," in *Proc. Empirical Methods Natural Lang. Process.*, 2015, pp. 1914–1918.
[19] K. Woodsend and M. Lapata, "Multiple aspect summarization using integer linear programming," in *Proc. Empirical Methods Natural Lang. Process./Comput. Natural Lang. Learn.*, 2012, pp. 233–243.
[20] J. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. Assoc. Comput. Linguistics*, 2005, pp. 363–370.

[21] C. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram cooccurrence statistics," in *Proc. North Am. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2003, pp. 71–78.
[22] G. Chalendar, R. Besancon, O. Ferret, G. Grefenstette, and O. Mesnard, "Crosslingual summarization with thematic extraction, syntactic sentence simplification, and bilingual generation," presented at the Workshop Crossing Barriers Text Summarization Res./5th Int. Conf. Recent Advances Natural Language Processing, 2005.
[23] P. Pingali, J. Jagarlamudi, and V. Varma, "Experiments in cross language query focused multi-document summarization," presented at the Int. Joint Conf. Artificial Intelligence Workshop Cross Lingual Information Access Addressing Information Need Multilingual Societies, Hyderabad, India, 2007.
[24] C. Orasan and O. Chiorean, "Evaluation of a cross-lingual romanian-english multi-document summariser," presented at the 6th Int. Language Resources Evaluation Conf., Marrakesh, Morocco, 2008.
[25] J. Lim, I. Kang, and J. Lee, "Multidocument summarization using cross-language texts," in *Proc. NTCIR*, 2004, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/TSC/NTCIR4-TSC-LimJM.pdf.
[26] M. Litvak, M. Last, and M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 927–936.
[27] Y. Mehdad, G. Carenini, and R. Ng, "Abstractive summarization of spoken and written conversations based on phrasal queries," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 1220–1230.
[28] A. Nenkova, "Entity-driven rewrite for multidocument summarization," in *Proc. 3rd Int. Joint Conf. Natural Lang. Process.*, 2008, pp. 118–125.
[29] A. Siddharthan, A. Nenkova, and K. McKeown, "Information status distinctions and referring expressions: An empirical study of references to people in news summaries," *Comput. Linguistics*, vol. 37, no. 4, pp. 811–842, 2011.
[30] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 340–348.
[31] P. Genest and G. Lapalme, "Framework for abstractive summarization using text-to-text generation," in *Proc. Workshop Monolingual Text-To-Text Gener.*, 2011, pp. 64–73.
[32] P. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 354–358.
[33] S. Gerani, Y. Mehbad, G. Carennini, R. Ng, and B. Nejat, "Abstractive summarization of product reviews using discourse structure," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 1602–1613.
[34] L. Wang and C. Cardie, "Domain-independent abstract generation for focused meeting summarization," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistic*, 2013, pp. 1395–1405.
[35] A. Rush, S. Chopra, and J. Weston, "A neural attention model for sentence summarization," in *Proc. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389.
[36] A. Yeh, "More accurate tests for the statistical significance of result differences," in *Proc. 18th Int. Conf. Comput. Linguistics*, 2000, pp. 947–953.

Jiajun Zhang received the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, in 2011, Beijing, China. He is currently an Associate Professor at the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences. His research interests include machine translation, multilingual natural language processing and deep learning.



Yu Zhou is an Associate Professor at the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China. Her research interests include machine translation and the key techniques of natural language processing.



Chengqing Zong received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1998. He is a Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include natural language processing, machine translation, and sentiment analysis. He is the Director in the Chinese Association of Artificial Intelligence and the Society of Chinese Information Processing, and a Member of International Committee on Computational Linguistics. He is an Associate Editor of the *ACM Transactions on Asian and Low-Resource Language Information Processing* and an Editorial Board Member of the IEEE INTELLIGENT SYSTEMS, the *Machine Translation*, and the *Journal of Computer Science and Technology*. He served ACL-IJCNLP2015 as a PC co-Chair.