A Multilingual News Summarizer

Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
hh_chen@csie.ntu.edu.tw

Chuan-Jie Lin
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN, R.O.C.
cjlin@nlg2.csie.ntu.edu.tw

Abstract

Huge multilingual news articles are reported and disseminated on the Internet. How to extract the key information and save the reading time is a crucial issue. This paper proposes architecture of multilingual news summarizer, including monolingual and multilingual clustering, similarity measure among meaningful units, and presentation of summarization results. Translation among news stories, idiosyncrasy among languages, implicit information, and user preference are addressed.

Introduction

Today many web sites on the Internet provide online news services. Multilingual news articles are reported periodically, and across geographic barrier to disseminate to readers. Readers can access the news stories conveniently, but it takes much time for people to read all the news. This paper will present a personal news secretariat that helps on-line readers absorb news information from multiple sources in different languages. Such a news secretariat eliminates the redundant information in the news articles, reorganizes the news for readers, and helps them resolve the language barriers.

Reorganization of news is some sort of document summarization, which creates a short version of original document. Recently, many papers touch on single document summarization (Hovy and Marcu, 1998a). Only a few touch on multiple document summarization (Chen and Huang, 1999; Mani and Bloedorn, 1997; Radev and McKeown, 1998) and multilingual document summarization (Hovy and Marcu, 1998b). For multilingual multiple news summarization, several issues have to be addressed:

(1) Translation among news stories in different languages

The basic idea in multiple document summarizations is to identify which parts of news articles present similar reports. Because the news stories are in different languages, some kind of translation is required, e.g., term translation. Besides the problem of translation ambiguity, different news sites often use different names to refer the same entity. The translation of named entities, which are usually unknown words, is another problem.

(2) Idiosyncrasy among languages

Different languages have their own specific features. For example, a Chinese sentence is composed of characters without word boundary. Word segmentation is indispensable for Chinese. Besides, Chinese writers often assign punctuation marks at random, how to determine a meaningful unit for similarity checking is a crucial issue. Thus some tasks may be done for specific languages during summarization.

(3) Implicit information in news reports

Some information is implicit in news stories. For example, the name of a country is usually not mentioned in a news article reporting an event that happened in that country. On the contrary, the country name is important in foreign news. Besides, time zone is used to specify date/time implicitly in the news.

(4) User preference

When users want to read documents in their familiar languages, news fragments in some

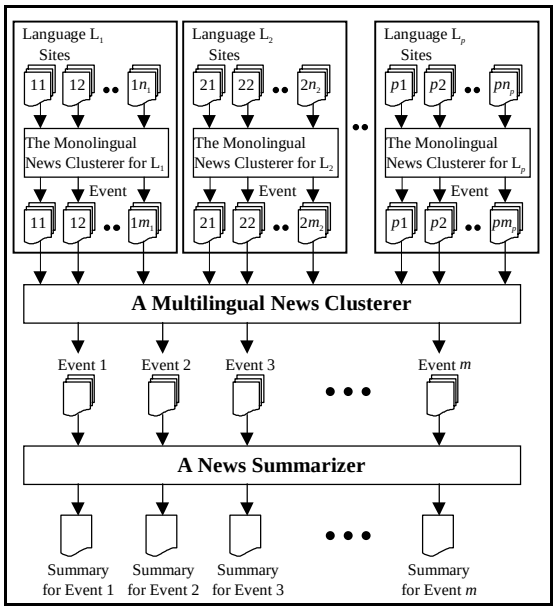

Figure 1. Architecture of Our Multilingual Summarization System

languages are preferred to those in other languages. Even machine translation should be introduced to translate news fragments. Besides, if a user prefers the news from the view of his country, or more precisely, of some news sites, we should meet his need.

Figure 1 shows the architecture of a multilingual summarization system, which is used to summarize the news from multiple sources in different languages. It is composed of three major components: several monolingual news clusterers, a multilingual news clusterer, and a news summarizer. The monolingual news clusterer receives a news stream from multiple on-line newspapers in its respective language, and directs them into several output news streams by using events. The multilingual news clusterer then matches and merges the news streams of the same event but in different languages in a cluster. The news summarizer summarizes the news stories for each event.

The possible tasks for each component depend on the languages used. Some major tasks of a monolingual clusterer are listed below.

(1) Identifying word boundaries for Chinese and Japanese sentences,

(2) Extracting named entities like people, place, organization, time, date and monetary expressions,

(3) Clustering news streams based on predefined topic set and named entities.

The task for the multilingual clusterer is to align the news clusters in the same topic set, but in different languages. It is similar to document alignment in comparable corpus. Named entities are also useful cues.

The major tasks for the news summarizer are shown as follows.

(1) Partitioning a news story into several meaningful units (MUs),

(2) Linking the meaningful units, denoting the same thing, from different news reports,

(3) Displaying the summarization results under the consideration of language type users prefer, information decay and views of reporters.

1. Clustering

1.1 Monolingual Clustering

We adopt a two-level approach to cluster the news from multiple sources. At first, news is classified on the basis of a predefined topic set. Then, the news articles in the same topic set are partitioned into several clusters according to named entities. Classification is necessary. On the one hand, a famous person may appear in many kinds of news stories. For example, President Clinton may make a public speech (political news), join an international meeting (international news), or even just show up in the opening of a baseball game (sports news). On the other hand, a common name is frequently seen but denotes different persons. Classification reduces the ambiguity introduced by famous persons and/or common names.

An event in a news story is characterized by five basic entities such as people, affairs, time, places and things. These entities form important cues during clustering. Systems for named entity extraction in a famous message understanding competition (MUC, 1998) demonstrate promising performances for English, Japanese and Chinese. In our multilingual summarization system, we focus on English and Chinese. Gazetteer approach is adopted to deal with English news articles. Comparatively, Chinese news articles are segmented at first. Then, several types of information from character, sentence and text levels are employed to extract Chinese named

entities. These tasks are similar to the approaches in the papers (Chen and Lee, 1996; Chen, *et al.*, 1998a).

1.2 Multilingual Clustering

The multilingual clusterer takes input from the monolingual clusterers, and determines which news clusters in which languages talk about the same story. Recall that a news cluster consists of several news articles reporting the same event, and one news cluster exists for one event after monolingual clustering. In this way, there is at most one corresponding news cluster in another language. Therefore, the main task of the multilingual news clusterer is to find the matchings among the clusters in different languages. Figure 2 shows an example. In Topic 1, cluster c_{i11} is aligned to $c_{j1\gamma}$, and cluster c_{i12} is aligned to c_{j11}. Clusters $c_{i1\alpha}$ and c_{j12} are left unaligned. That means the denoted events are reported in only one language.

Similarity of two clusters is measured based on verbs, named entities, and the other nouns. Because Chinese words are less ambiguous than English ones (Chen, Bian and Lin, 1999), we translate nouns and verbs in the Chinese news articles into English. If a word has more than one translation, we select high frequent English translation. For the named entities not listed in the lexicon, name transliteration similar to the algorithm (Chen, *et al.*, 1998b) is introduced for matching in non-alphabetic (e.g., Chinese) and alphabetic languages (e.g., English).

Alignment is made under the same topic. A news cluster c_i is aligned to another cluster c_j if their similarity is above a threshold, and is the highest between c_i and the other clusters. If the similarity of c_i and the other clusters is less than a given threshold, c_i is not aligned. It is possible because local news is reported only in the restricted areas.

2. Similarity Analysis

2.1 Meaningful Units

The basic idea during summarization is to tell which parts of the news articles are similar in the same event. The basic unit for similarity measure may be a paragraph or a sentence. For the former, text segmentation is necessary for documents without paragraph markers (Chen and Chen, 1995). For the latter, text segmentation is necessary for languages like Chinese. Unlike English writers, Chinese writers often assign punctuation marks at random (Chen, 1994). Thus the sentence boundary is not clear. Consider the following Chinese example (C1):

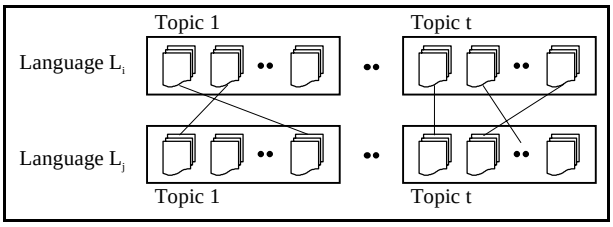

Figure 2. Matching among the Clusters in Two Languages

(C1) 儘管警方大肆鎮壓與逮捕，無數反對自由貿易的示威群眾今天仍繼續向西雅圖市中心前進，他們發動和平集會，以抗議世界各國貿易部長即將在西雅圖舉行討論全球貿易自由化的會議。

(Central News Agency, 1999.12.02)

(Although they were undeterred by mass arrests and a police crackdown, anti free-trade protesters still marched on downtown Seattle today. The protesters, carrying signs and chanting, opposed the global trade liberalization being worked on at a meeting of trade ministers from the World Trade Organization.)

It is composed of four sentence segments separated by commas. If a sentence segment is regarded as a unit for similarity checking, it may contain too little information. On the contrary, if a sentence is regarded as a unit, it may contain too much information. Here we consider a meaningful unit (MU) as a basic unit for measurement. A MU is composed of several sentence segments and denotes a complete meaning. We will find two MUs shown as follows for (C1):

(C2) 儘管警方大肆鎮壓與逮捕，無數反對自由貿易的示威群眾今天仍繼續向西雅圖市中心前進

(Although they were undeterred by mass arrests and a police crackdown, anti free-trade protesters still marched on downtown Seattle today.)

(C3) 他們發動和平集會，以抗議世界各國貿易

部長即將在西雅圖舉行討論全球貿易自由化的會議

(The protesters, carrying signs and chanting, opposed the global trade liberalization being worked on at a meeting of trade ministers from the World Trade Organization.)

In our summarization system, an English sentence itself is an MU. Comparatively, it is a little harder to identify Chinese MUs. Three kinds of linguistic knowledge – punctuation marks, linking elements and topic chains, are proposed.

(1) Punctuation marks

There are fourteen marks in Chinese (Yang, 1981). Only period, question mark, exclamation mark, comma, semicolon and caesura mark are employed. The former three are sentence terminators, and the latter three are segment separators.

(2) Linking elements

There are three kinds of linking elements (Li and Thompson, 1981): forward-linking elements, backward-linking elements, and couple-linking elements. A segment with a forward-linking (backward-linking) element is linked with its next (previous) segment. A couple-linking element is a pair of words that exist in two segments. Apparently, these two segments are joined together. Examples (C4)-(C6) show each kind of linkings.

(C4) 下課之後，我要去看電影。
(After school, I wanted to see a movie.)

(C5) 我本來想去看電影，可是我沒有買到票
(I wanted to see a movie, but I couldn't get a ticket.)

(C6) 因為我沒有買到票，所以我沒有去看電影。
(Because I couldn't get a ticket, (so) I didn't see a movie.)

(3) Topic chains

The topic of a clausal segment is usually deleted under the identity with a topic in its preceding segment. The result of such a deleting process is a *topic chain.* We employ part of speech information to predict if a subject of a verb is missing. If it does, we postulate that it must appear in the previous segment and the two segments are connected to form a larger unit.

Consider example (C1). The word "儘管" (although) is a forward linking element. Thus the first two segments are connected together (C2). The last segment does not have any subject, so that it is connected to the previous one by topic chain (C3). In summary, two MUs are formed.

2.2 Similarity Model

The next step is to find the similarity among MUs in the news articles reporting the same event, and to link the similar MUs together. We analyze the news stories within the same language, and then the news stories among different languages. The key idea is similar at these two steps. That is, predicate argument structure forms the kernel of a sentence, thus verbs and nouns are regarded as important cues for similarity measures. The difference between these two steps is that we have to translate nouns and verbs in one language into another language. The approach of select-high-frequent translation and name transliteration shown in Section 1.2 is adopted here too. Consider (MU1) – (MU3). The former two are in Chinese and the last one is in English. They denote a similar event "Seattle's Curfew Hours". Each noun (verb) is enclosed by parentheses and assigned an index. There are 9 common terms between (MU1) and (MU2); 10 common terms between (MU1) and (MU3); and 8 common terms between (MU2) and (MU3). Note the time zones used in (MU2) and (MU1) are different, so are (MU2) and (MU3).

(MU1) 在($_{1}$西雅圖)($_{2}$市長)($_{3}$謝爾)($_{4}$宣布)該($_{5}$市)($_{6}$進入)($_{7}$緊急)($_{8}$狀態)， 並在($_{9}$昨晚)($_{10}$七時)至($_{11}$今晨)($_{12}$七時)($_{13}$實施)($_{14}$宵禁)後，($_{15}$警方)已將($_{16}$示威者)($_{17}$驅離)了($_{18}$市中心)。

(Chinatimes, 1999.12.02)

(MU2) ($_{1}$西雅圖)($_{2}$市長)($_{4}$宣布)，全($_{5}$市)($_{6}$進入)($_{7}$緊急)($_{8}$狀態)，並在($_{19}$台北時間)($_{20}$一號)($_{21}$上午)($_{22}$十一點)、到($_{23}$晚上)($_{24}$十一點半)，($_{13}$實施)($_{14}$宵禁)。

(Formosa Television, 1999.12.02)

(MU3) ($_{1}$Seattle) ($_{2}$Mayor) ($_{25}$Paul) ($_{3}$Schell) has ($_{4}$declared) a ($_{8}$state) of ($_{5}$civil) ($_{7}$emergency) and ($_{13}$imposed) a ($_{10}$7 p.m.) to ($_{26}$7:30 a.m) (($_{27}$10 p.m.) EST – ($_{28}$10:30 a.m.) EST) ($_{14}$curfew) on ($_{28}$downtown) ($_{29}$areas) of the ($_{30}$city).

(Reuters)

Several strategies may be considered in similarity measure:
(S1) Nouns in one MU are matched to nouns in another MU, so are verbs.
(S2) The operations in (1) are exact matches.
(S3) Thesauri are employed during matching.
(S4) Each term specified in (S1) is matched only once.
(S5) The order of nouns and verbs in MU is not considered.
(S6) The order of nouns and verbs in MU is critical, but it is relaxed within a window.
(S7) When continuous terms are matched, an extra score is added.
(S8) When the object of transitive verbs are not matched, a score is subtracted.
(S9) When date/time expressions and monetary and percentage expressions are matched, an extra score is added.

Five models shown below are constructed under different combinations of the strategies specified in the above.
(M1) (S1)+(S3)+(S4)+(S5)
(M2) (S1)+(S3)+(S4)+(S6)
(M3) (S1)+(S3)+(S4)+(S5)+(S7)+(S8)
(M4) (S1)+(S3)+(S4)+(S5)+(S7)+(S8)+(S9)
(M5) (S1)+(S2)+(S4)+(S5)+(S7)+(S8)+(S9)

3. Experiments

3.1 Preparation of Testing Corpus

Six events selected from Central Daily News, China Daily Newspaper, China Times Interactive, and FTV News Online in Taiwan are used to measure the performance of each model. They are shown as follows:

(1) military service: 6 articles
(2) construction permit: 4 articles
(3) landslide in Shan Jr: 6 articles
(4) Bush's sons: 4 articles
(5) Typhoon Babis: 3 articles
(6) stabilization fund: 5 articles

The news events are selected from different editions, including social edition, economic edition, international edition, political edition, *etc*. An annotator reads all the news articles, and connects the MUs that discuss the same story. Because each MU is assigned a unique ID, the links among MUs form the answer keys for the performance evaluation.

Table 1. Performance of Similarity of MUs

Model	Precision Rate	Recall Rate
M1	0.5000	0.5434
M2	0.4871	0.3905
M3	0.5080	0.5888
M4	0.5164	0.6198
M5	0.5243	0.5579

3.2 Results

Traditional precision and recall are computed. Table 1 lists the performance of these five models. M1 is regarded as a baseline model. M2 is different from M1 in that the matching order of nouns and verbs are kept conditionally. It tries to consider the subject-verb-object sequence. The experiment shows that the performance is worse. The major reason is that we can express the same meaning using different syntactic structures. Movement transformation may affect the order of subject-verb-object. Thus in M3 we give up the order criterion, but we add an extra score when continuous terms are matched, and subtract some score when the object of a transitive verb is not matched. Compared with M1, the precision is a little higher, and the recall is improved about 4.5%. If we further consider some special named entities such as date/time expressions and monetary and percentage expressions in M4, the recall is increased about 7.6% at no expense of precision. M5 tries to estimate the function of the thesauri. It uses exact matching. The precision is a little higher, but the recall is decreased about 6% compared with M4.

Several major errors affect the overall performance. Using nouns and verbs to find the similar MUs is not always workable. The same meaning may not be expressed in terms of the same words or synonymous words. Besides, we can use different format to express monetary and percentage expressions. Word segmentation is another source of errors. Two sentences denoting the similar meaning may be segmented differently due to the segmentation strategies. Unknown words generate many single-character words. After tagging, these words tend to be nouns and verbs, which are used in computing the scores for similarity measure. Thus errors may be introduced.

4. Presentation Model

Two models, i.e., focusing model and browsing model, are proposed to display the summarization results. In the focusing model, a summarization is presented by voting from reporters. For each event, a reporter records a news story from his own viewpoint. Recall that a news article is composed of several MUs. Those MUs that are similar in a specific event are common focuses of different reporters. In other words, they are worthy of reading. In the current implementation, the MUs that are reported more than once are our target. For readability, the original sentences that cover the MUs are selected. For each set of similar MUs, the longest sentence in user-preferred language is displayed. The display order of the selected sentences is determined by relative position in the original news articles.

In the browsing model, the news articles are listed by information decay. The first news article is shown to the user in its whole content. In the latter shown news articles, the MUs denoting the information mentioned before are shadowed (or eliminated), so that the reader can focus on the new information. The amount of information in a news article is measured in terms of the number of MUs, so that the article that contains more MUs is displayed before the others. For readability, a sentence is a display unit. In this model, users can read both the common views and different views of reporters. It saves the reading time by listing the common view only once.

5. Evaluation of Summarization Results

The same six events specified in Section 3.1 are used to measure the performance of the two summarization models. Three kinds of metrics are considered – say, the document reduction rate, the reading-time reduction rate, and the information carried. The higher the document reduction rate is, the more time the reader may save, but the higher possibility the important information may be lost. Tables 2 and 3 list the document reduction rates for focusing and browsing summarization, respectively. Only focuses are displayed in focusing summarization,

Table 2. Reduction Rates for Focusing Summarization

Event Name	Doc Len	Sum Len	Sum/Doc	Reduction
military service	7658	2402	0.3137	68.63%
construction permit	4182	1226	0.2932	70.68%
landslide in Shan Jr	5491	1823	0.3320	66.80%
Bush's sons	6186	924	0.1494	85.06%
Typhoon Babis	4068	1460	0.3589	64.11%
stabilization fund	8434	2243	0.2659	73.41%
Average	36019	10078	0.2798	72.02%

Table 3. Reduction Rates for Browsing Summarization

Event Name	Doc Len	Sum Len	Sum/Doc	Reduction
military service	7658	2716	0.3547	64.53%
construction permit	4182	2916	0.6973	30.27%
landslide in Shan Jr	5491	2946	0.5365	46.35%
Bush's sons	6186	5098	0.8241	17.59%
Typhoon Babis	4068	2270	0.5580	44.20%
stabilization fund	8434	4299	0.5097	49.03%
Average	36019	20245	0.5621	43.79%

Table 4. Assessors' Evaluation

Event Name	Document Reduction Rate	Question-Answering Correct Rate	Reading-Time Reduction Rate
military service	64.53%	100%	45.24%
construction permit	30.27%	33.33%	33.54%
landslide in Shan Jr	46.35%	80%	10.28%
Bush's sons	17.59%	100%	36.49%
Typhoon Babis	44.20%	100%	35.10%
stabilization fund	49.03%	100%	18.49%
Average	43.79%	88.46%	30.86%

so that the average document reduction rate is higher than that of browsing summarization.

Besides the document reduction rate, we also measure the correct rate of question-answering, and reading-time reduction rate. Assessors read the highlight parts only in the browsing summarization, and answer 3 to 5 questions. Table 4 lists the evaluation results of the six events. The average document reduction rate is 43.79%. On the average, the summary saves 30.86% of reading time. While reading the summary only, the correct rate of question-answering task is 88.46%.

Conclusion

This paper sketches architecture for multilingual news summarizer. In multilingual clustering, matching all pairs of news clusters in all languages is time-exhaustive. Because only English and Chinese news articles are considered in this paper, it is not a problem. In general, an

effective way is to predefine a sequence of language pairs according to the degree of translation ambiguity. The language pair of less ambiguity is tried first.

To discuss which fragments of multilingual news stories denote the same things, this paper defines the concept of MUs. Punctuation marks, linking elements and topic chains are cues to identify MUs for Chinese. Select-high-frequent English translation and name transliteration are adopted to translate Chinese MUs into English. Five models are proposed to link the similar MUs together. Different formats used in time, date and monetary expressions, e.g., implicit time zone, affect the performance of linking. It should be studied in the future.

In presentation of summarization results, the information decay strategy helps reduce the redundancy, and the user can get all the information provided by the news sites. However, the news sequence is not presented according to the importance. The user may quit reading and miss the information not shown yet. The voting strategy from reporters gives a shorter summarization in terms of user-preferred languages. However, it also misses some unique information reported only by one site. A hybrid strategy should be developed in the future to meet all the requirements.

References

Chen, H.H. (1994) "The Contextual Analysis of Chinese Sentences with Punctuation Marks," *Literal and Linguistic Computing*, Oxford University Press, **9**(4), 1994, pp. 281-289.

Chen, H.H; *et al*. (1998a) "Description of the NTU System Used for MET2." *Proceedings of 7th Message Understanding Conference*, 1998.

Chen, H.H.; *et al*. (1998b) "Proper Name Translation in Cross-Language Information Retrieval," *Proceedings of COLING-ACL98*, 1998, pp. 232-236.

Chen, H.H.; Bian, G.W. and Lin, W.C. (1999) "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval," *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 215-222.

Chen, K.H. and Chen, H.H. (1995) "A Corpus-Based Approach to Text Partition," *Proceedings of International Conference of Recent Advances on Natural Language Processing*, Tzigov Chark, Bulgaria, 1995, pp. 152-160.

Chen, H.H. and Huang, S.J. (1999) "A Summarization System for Chinese News from Multiple Sources," *Proceedings of 4th International Workshop on Information Retrieval with Asia Languages*, 1999, pp. 1-7.

Chen, H.H. and Lee, J.C. (1996) "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 222-229.

Hovy, E. and Marcu, D. (1998a) *Automated Text Summarization*, Tutorial in *17th ACL and 36th COLING*, Montreal, Quebec, Canada, 1998.

Hovy, E. and Marcu, D. (1998b) *Multilingual Text Summarization*, Tutorial in *AMTA-98*, 1998.

Li, C.N. and Thompson, S.A. (1981) *Mandarin Chinese: A Functional Reference Grammar*, University of California Press, 1981.

Mani, I. and Bloedorn, E. (1997) "Multi-document Summarization by Graph Search and Matching," *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Providence, RI, pp. 622-628.

MUC (1998) *Proceedings of 7th Message Understanding Conference*, http://www.muc.saic.com /proceedings/proceedings_index.html.

Radev, D.R. and McKeown, K.R. (1998) "Generating Natural Language Summaries from Multiple On-Line Sources," *Computational Linguistics*, Vol. 24, No. 3, pp. 469-500.

Yang, Y. (1981) *The Research on Punctuation Marks*, Tian-jian Publishing Company, Hong Kong, 1981.