Clustering and Visualization in a Multi-lingual Multi-document Summarization System

Hsin-Hsi Chen, June-Jei Kuo, and Tsei-Chun Su

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
hh_chen@csie.ntu.edu.tw
{jjkuo,tcsu}@nlg.csie.ntu.edu.tw

Abstract. To measure the similarity of words, sentences, and documents is one of the major issues in multi-lingual multi-document summarization. This paper presents five strategies to compute the multilingual sentence similarity. The experimental results show that sentence alignment without considering the word position or order in a sentence obtains the best performance. Besides, two strategies are proposed for multilingual document clustering. The two-phase strategy (translation after clustering) is better than one-phase strategy (translation before clustering). Translation deferred to sentence clustering, which reduces the propagation of translation errors, is most promising. Moreover, three strategies are proposed to tackle the sentence clustering. Complete link within a cluster has the best performance, however, the subsumption-based clustering has the advantage of lower computation complexity and similar performance. Finally, two visualization models (i.e., focusing and browsing), which consider the users' language preference, are proposed.

1 Introduction

In a basic multi-document summarization system (Chen and Huang, 1999; Mckeown, Klavans, Hatzivassiloglou, Barzilay and Eskin, 1999; Goldstein, Mittal, Carbonell and Callan, 2000; Hatzivassiloglou, Klavans, Holcombe, Barzilay, Kan and Mckeown 2001), how to decide which documents deal with the same topic, and which sentences touch on the same event are indispensable. Because a document is composed of sentences and a sentence consists of words, how to measure the similarity on different levels (i.e., words, sentences and documents), is one of the major issues in multi-document summarization (Barzilay and Elhadad, 1997; Mani and Bloedorn, 1999; Goldstein, Mittal, Carbonell and Callan, 2000; Radev, Jing and Budzikowska, 2000). In multi-lingual multi-document summarization, we have to face one more issue, i.e., the multilinguality problem (Chen and Lin, 2000). However, most of the previous works did not touch this issue. The same concepts, themes and topics may be in terms

F. Sebastiani (Ed.): ECIR 2003, LNCS 2633, pp. 266-280, 2003.

of different languages. Translation among words (sentences, documents) in different languages, idiosyncrasy among languages, implicit information in documents, and user preference should be tackled.

Clustering puts together those words/sentences/documents that denote the same concepts/themes/topics. The granuality of clustering units and the features used in the clustering should be considered. Because sentences contain less information than documents, i.e., fewer features can be employed in sentence clustering, similarity computation among sentences is more challenging than that among documents. In multilingual clustering, three possible ways may be adopted. That is, (1) merge the documents from different language sources, do the document and sentence clustering; (2) do the document clustering for each language source, merge the documents clusters denoting the same topic in different languages, and do the sentence clustering; (3) do the document and sentence clustering for each language source, and merge the sentence clusters denoting the same event in different languages.

This paper presents methods for event clustering on different levels, and show how to summarize the results from event clusters. Section 2 depicts the basic architecture of a multi-lingual multi-document summarization system. Section 3 touches on similarity measurement. Section 4 proposes clustering models for multi-lingual documents. Section 5 deals with multi-lingual sentence clustering. After linking the sentences denoting the same event, Section 6 addresses the visualization issue, e.g., which sentence in which language will be selected, and the preference. Section 7 concludes the remarks.

2 Basic Architecture

Figure 1 shows a multi-lingual multi-document summarization system. We receive documents from multi-lingual sources and send them for document pre-processing. Different languages have their own specific features. Document pre-processing module deals with idiosyncrasy among languages. For example, a Chinese sentence is composed of characters without word boundary. Word segmentation is indispensable for Chinese. Document clustering partitions documents into event clusters. Document content analysis module analyzes document in each event cluster, and links together those sentences denoting the same themes. Finally, summaries are generated.

The major issues behind such a system are how to represent documents in different languages; how to measure the similarity among document representations of different languages; the granularity of similarity computation; and visualization of summaries. The following sections will discuss each issue in detail.

3 Similarity Measurement

3.1 Methods

Word exact matching cannot resolve paraphrase problem. Relaxation with WordNet-like resources (Fellbaum, 1998) postulates that words in the same synset are similar.

EuroWordNet (Vossen, 1998) and Chinese-English WordNet (Chen, Lin and Lin, 2002) facilitate the inexact matching among different languages.

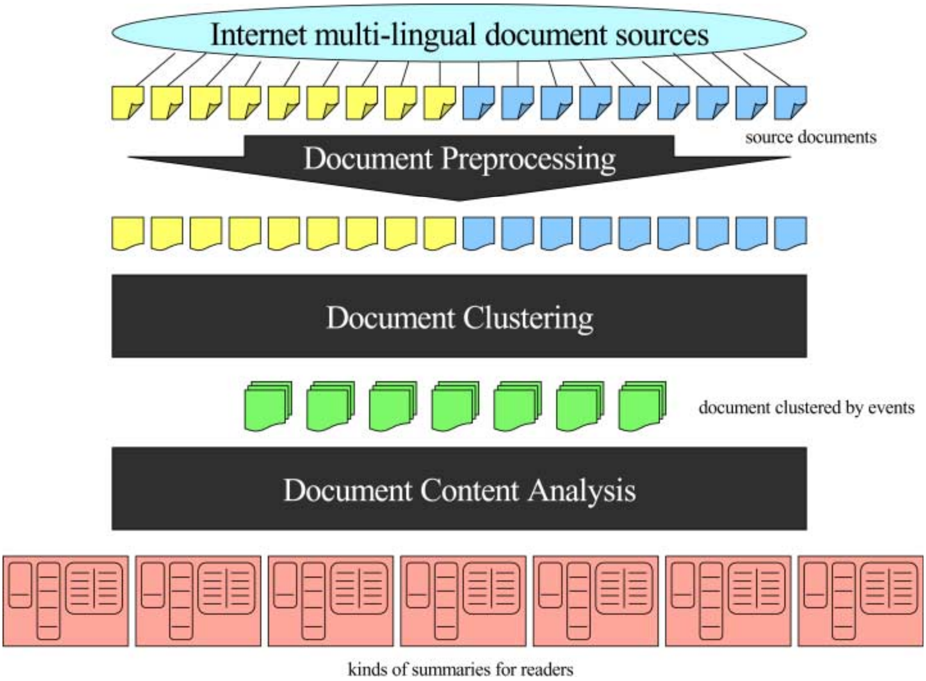
Internet multi-lingual document sources
source documents
Document Preprocessing
Document Clustering
document clustered by events
Document Content Analysis
kinds of summaries for readers

Fig. 1. A Multi-lingual Multi-document Summarization System

Predicate and the surrounding arguments form the basic skeleton in a sentence, so that verbs and nouns are considered as the basic features for similarity measurement. The similarity of two monolingual sentences is defined as follows.

$$SIM(S_i, S_j) = \frac{|S_i \cap S_j|}{\sqrt{|S_i|}\sqrt{|S_j|}} \tag{1}$$

where S_i and S_j are two sets denoting two sentences, $S_i \cap S_j$ denotes the common occurrences of two sentences by inexact matching, and $|S_i|$, $|S_j|$ and $|S_i \cap S_j|$ denote the number of elements in the sets S_i, S_j, and $S_i \cap S_j$, respectively.

For computing the similarity of two sentences in different languages, the ambiguity problem floats up. That is, a word may have more than one translation equivalent in a bilingual dictionary. Five strategies are proposed.

1. position-free

   This strategy is similar to the above method. For each word in S_i, find its translation equivalents by a bilingual dictionary. Then, merge all the equivalents. Let the set be S_i'. Formula 1 is modified as follows.

$$SIM(S_i, S_j) = \frac{|S_i^{'} \cap S_j|}{\sqrt{|S_i|}\sqrt{|S_j|}} \tag{2}$$

2. first-match-first-occupy

   Compare the translation of each word in S_i with the words in S_j. When a word in S_j is matched, it is removed from S_j and the similarity score (SC) is added by 1. In other words, the word is occupied, and will not be considered in the later comparison. Formula (3) shows the revision.

$$SIM(S_i, S_j) = \frac{SC}{\sqrt{|S_i|}\sqrt{|S_j|}} \tag{3}$$

3. first-match-first-occupy and position-dependent within a window

   This method is similar to Strategy (2) except that the latter comparison is restricted by the results of the previous matching. The range of comparison is limited within a window size of the previous matching. Figure 2 shows an example. Assume C2 has been matched by E1 and the window size is 3. The candidates for E2 in the later comparisons are C1 and C3.

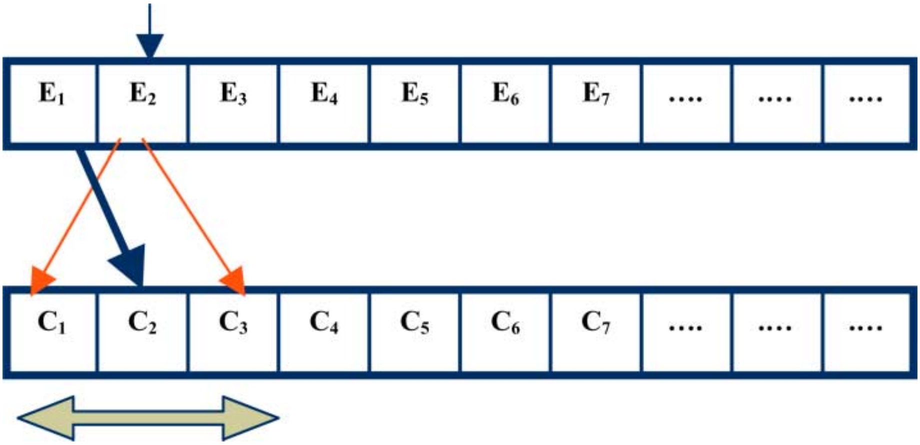

Fig. 2. First-Match-First-Occupy and Position-Dependent

4. unambiguous-word-first and position-dependent within a window

   This strategy links those pairs without ambiguity first, and then performs the similar operation as strategy (3).

5. unambiguous-word-first and position-dependent within a range
   This strategy does not set the window size beforehand. The range for matching is restricted by the decided pairs.

We adopt the same five strategies to compute the document similarity except that the window size is changed. Formula (4) defines the document similarity.

$$SIM(D_i, D_j) = \frac{|D_i \cap D_j|}{\sqrt{|D_i|}\sqrt{|D_j|}} \tag{4}$$

Here, D_i and D_j are two sets denoting two documents.

3.2 Experiments

We selected 81 pairs of English and Chinese news stories from the web site of United Daily News in Taiwan. Another 80 unrelated news stories, i.e., 40 English ones and 40 Chinese ones, were added and mixed together. For each English news story, we try to find the best matching from the remaining 241 candidates. Besides, from the above 81 pairs of English and Chinese news stories we extracted 43 pairs of English and Chinese sentences at random and regarded them as an answer set to evaluate the performance of sentence similarity computation. For each English news sentences, we try to find the best matching from the remaining 85 candidates. Correct rate is defined as follows.

$$CorrectRate = \frac{CorrectPairsSystemFind}{TotalCorrectPairs} \tag{5}$$

Table 1. Performance of Document Alignment

	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Best 1	0.951	0.839	0.506	0.320	0.320
Best 2	0.987	0.925	0.604	0.432	0.444
Best 3	1.000	0.925	0.666	0.469	0.469
Best 4	1.000	0.950	0.740	0.518	0.518
Best 5	1.000	0.975	0.740	0.530	0.530

Table 2. Performance of Sentence Alignment

	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
Best 1	0.883	0.767	0.441	0.255	0.255

Best 2	0.930	0.813	0.674	0.279	0.279
Best 3	0.976	0.860	0.697	0.325	0.325
Best 4	1.000	0.930	0.790	0.372	0.372
Best 5	1.000	0.930	0.790	0.372	0.372

Tables 1 and 2 summarize the experimental results for document and sentence alignments, respectively. Best *n* means *n* documents should be proposed to cover the correct matching. The experimental results show that Strategies 1 and 2 are better than the other three strategies. Moreover, Strategy 1 is also superior to Strategy 2. The position-dependent seems not to be useful in both first-match-first-occupy and unambiguous-word-first models. This is due to the difference of word order between Chinese and English sentences, e.g., the arguments in relative clause may be extra-posed to different positions in Chinese and in English. To fix the non-ambiguous word first does not have a clear effect in the experiments. After analyzing the results, we find that there are 172,734 lexical items in our bilingual dictionary. Of these, 111,120 lexical items have only one translation. The average number of translation equivalents per lexical item is 2.17. However, only 841 of 9,636 words in the test corpus are unambiguous. On the average, each lexical item has 10.84 translation equivalents. The experimental results also reveal that the performance of document alignment is better than that of sentence alignment. The amount of information affects the similarity computation.

4 Event Clustering

4.1 Clustering Models

Translation is indispensable for multi-lingual multi-document clustering. Three possible models are proposed as follows. They deal with when translation is performed.

1. translation BEFORE document clustering
   This model clusters the multi-lingual multi-documents directly. Figure 3 shows this model, which is a one-phase model. The similarity computation among documents in Section 3 belongs to this type.
2. translation AFTER document clustering
   This model clusters documents in each language separately, and merges the clustering results. Figure 4 shows this model, which is a two-phase model.
3. translation DEFERRED to sentence clustering
   In this model, multilingual problem is dealt with on the sentence level. Figure 5 shows this model which will be discussed further in Section 5.

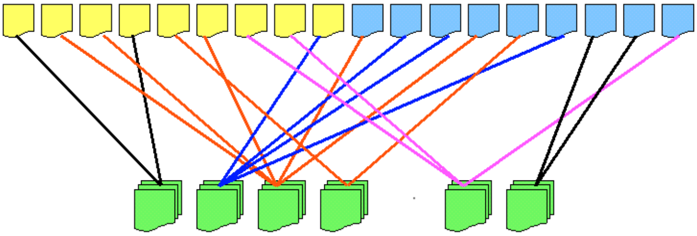

Fig. 3. Translation before Document Clustering

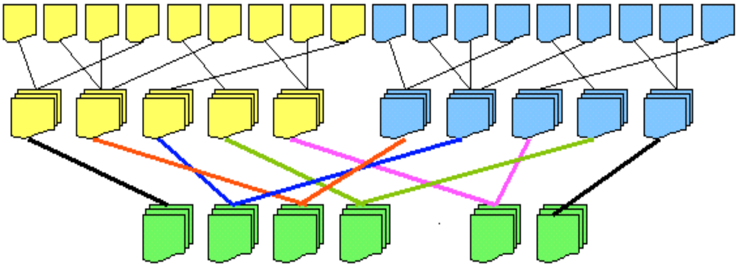

Fig. 4. Translation after Document Clustering

4.2 Experiments

We collected English and Chinese news articles reported on May 8, 2001 from the following news sites in Taiwan. There are 460 news articles in the test corpus.

1. English: Central News Agency, China Post, China Times and United Daily News
2. Chinese: Central News Agency, Central Daily News, China Times and United Daily News

First, we cluster those news articles manually and the result is shown as in Table 3. Tables 4 and 5 show the experimental results of one-phase model (i.e., translation before clustering) and two-phase model (i.e., translation after clustering), respectively. In the one-phase scheme, only one threshold is used. Table 4 lists three sets of results under three different threshold assignments. Comparatively, due to the different document features, e.g. document numbers, three different thresholds are used in the two-phase scheme (see Table 5), including one for Chinese document clustering (i.e., 0.3), one for English document clustering (i.e., 0.5), and one for the final cluster merging (i.e., 0.2). The performance of two-phase scheme is better than that of one-phase scheme. The major reason is translation is performed after monolingual clustering. That reduces not only the translation errors, but also the computation complexity. This concept leads to the Model 3 (translation deferred to sentence clustering).

Table 3. Manual Clustering Result

	Article Number	Cluster Number	Cluster Number = 1	Cluster Number > 1
Chinese	360	265	230	35
English	91	75	65	10
CE	460	318	276	42

Table 4. Experimental Result Using One-Phase Model

Threshold	Number of Articles in a Cluster			Exact Match	Precision	Recall
	1	1<N<5	>5			
0.1	156	37	50	154	0.633	0.484
0.2	250	16	36	223	0.738	0.701
0.4	430	0	1	264	0.612	0.830

Table 5. Experimental Results Using Two-Phase Model

	Threshold	Number of Articles in a Cluster			Exact Match	Precision	Recall
		1	1<N<5	>5			
C	0.3	240	33	21	253	0.860	0.954
E	0.5	52	13	6	60	0.845	0.800
CE	0.2	281	29	44	296	0.841	0.931

5 Sentence Clustering

5.1 Clustering Models

Figure 5 shows that after monolingual document clustering, those documents in a cluster denote the same event in a specific language. To generate the extract summary of an event, we must cluster the similar sentences among documents and then choose a representative sentence from each cluster. Position-free strategy proposed in Section 2 has the best performance, thus it is employed to compute the similarity between two bilingual sentences.

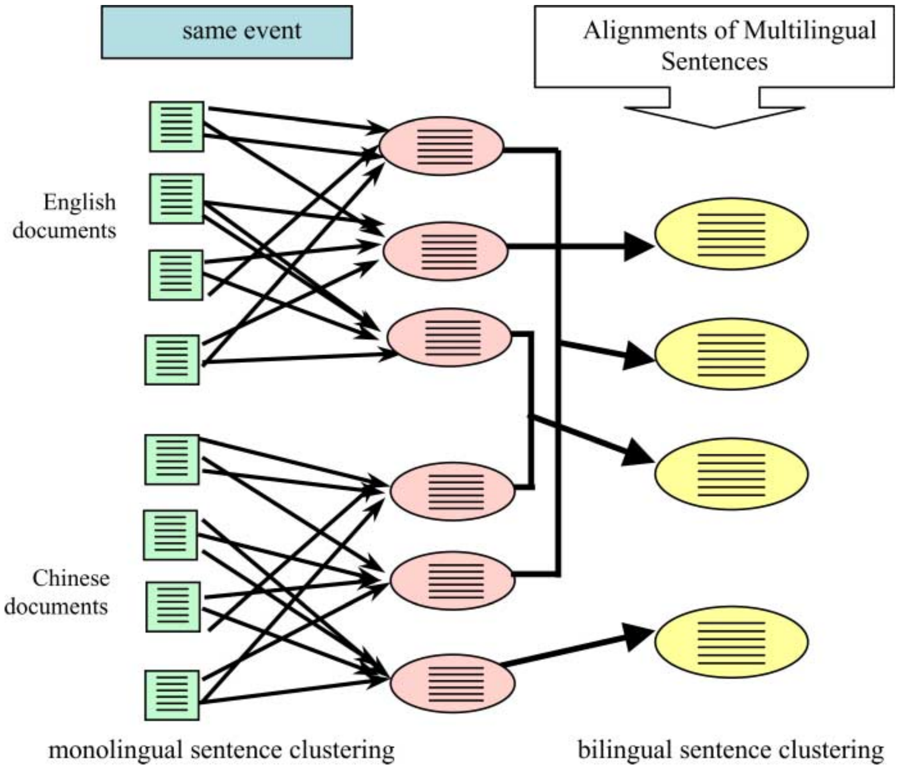
same event
Alignments of Multilingual Sentences
English documents
Chinese documents
monolingual sentence clustering
bilingual sentence clustering

Fig. 5. Translation Deferred to Sentence Clustering

There are three alternatives shown as follows for sentence clustering.

1. complete link using all sentence
   We compute the similarity between any two sentences in the same event cluster, and employ complete link strategy to cluster the sentences.
2. complete link within a cluster
   To tackle the computational issue, we read each sentence in the same event cluster in sequence. The first sentence s_1 is assigned to a cluster c_1. Assume there already are k clusters when a new sentence s_i is considered. The sentence s_i may belong to one of k clusters, or it may form a new cluster c_{k+1}. The determination depends on the similarity between s_i and all the sentences in each cluster. If all the sentence similarities in a specific cluster are greater than the threshold, it is added into that cluster. If there is no such a cluster, s_i becomes a new cluster.
3. subsumption-based clustering
   The basic idea is similar to Model 2 except that in this model a centroid is determined for each cluster and a subsumption test is used to tell if a sentence be-

longs to a specific cluster. The following formula (6) defines an information score of a sentence in a cluster. Total 25 words of higher document frequency in a cluster are considered as topic words of this cluster.

$$\inf(S) = \left(|S_n| + |S_v| + |S_t|\right) \tag{6}$$

where S denotes a sentence,
$|S_n|$ is the number of nouns in S,
$|S_v|$ is the number of verbs in S, and
$|S_t|$ is the number of topic words in S.

The sentence of the highest information score in a cluster is selected as the centroid of this cluster. We only compute sentence similarity between a sentence and a centroid, and the sentence similarity is in terms of a subsumption score shown as follows.

$$SIM\ (S_i, S_j) = \frac{|S_i \cap S_j|}{\min\left(|S_i|, |S_j|\right)} \tag{7}$$

whereS$_i$ and S$_j$ are two sets representing two sentences,
S$_i$ $\cap$ S$_j$ denotes the common occurrences[1] of two sentences, and | S$_i$ |, | S$_j$ | and | S$_i$ $\cap$ S$_j$ | denotes the number of elements in the sets S$_i$, S$_j$, and S$_i$ $\cap$ S$_j$, respectively.
The larger the score is, the more the subsumption is.

5.2 Experiments

We used the same materials specified in Section 4.2. After manual clustering, we selected five events shown below and the related numbers of English and Chinese articles are listed in Table 5.

1. Investment for bioinformatics
2. The relation between President Chen and Vice President Lu
3. Mr. Hsiao Wuan-Chang visited mainland China
4. Can the management of Kaoshong harbor return to city government?
5. The court rejected the application from the Journalist Magazin

Besides, we also cluster the related sentences munually for each event. There are 662 correct links. The following shows sample of answer keys used in evaluation. Each sentence is denoted by NewsAgencyType_DocumentID_Sentence_ID. For example, ChinaEng_022_001 is an English sentence (ID : 001) in a news (ID : 022) published by China Times. This sentence is related to CnaEng_021_001, UdnBI_e_003_001, and UdnBI_c_003_001.

Sentence	Link1	Link2	Link3
ChinaEng_022_001	CnaEng_021_001	UdnBI_e_003_001	UdnBI_c_003_01
ChinaEng_022_007	CnaEng_021_007		
UdnBI_e_007_001	ChinaEng_002_001	CpostEng_004_017	
UdnBI_e_007_002	CpostEng_004_003		

[1] The synonym matching and position-free method specified in Section 2 are adopted.

Tables 7-9 list experimental results using the three sentence clustering methods.

Table 6. Test Data for Sentence Clustering

	Total Chinese Documents	Total English Documents	Total Chinese Sentences	Total English Sentences
Event 1	4	3	69	25
Event 2	5	2	87	39
Event 3	5	3	92	40
Event 4	5	2	82	16
Event 5	2	3	23	46

Table 7. Performance of Complete Link Using All Sentences

Threshold	Total Links Proposed	Number of Correct Links	Precision	Recall
0.20	852	436	0.511	0.658
0.25	702	408	0.581	0.616
0.30	668	384	0.574	0.580

Table 8. Performance of Complete Link within a Cluster

Threshold	Total Links Proposed	Number of Correct Links	Precision	Recall
0.50	892	478	0.536	0.722
0.55	718	420	0.585	0.634
0.60	622	376	0.604	0.567

Table 9. Performance Using Subsumption-based Links

Threshold	Total Links Proposed	Number of Correct Links	Precision	Recall
0.50	874	462	0.529	0.698
0.55	708	418	0.590	0.631
0.60	602	358	0.595	0.540

By observing Tables 7 and 8, the performance of Strategy 2 is better than that of Strategy 1. Although the performance of Strategy 3 is a little worse than that of Strategy 2, its time complexity is decreased very much. If the score function can be further improved to obtain the more representative sentence, this strategy is competible.

6 Visualization

In multi-lingual multi-document summarization, how to display the results to readers is an important issue. Two models, i.e., focusing model and browsing model, are proposed. The readers' preference is also taken into consideration. For example, a Chinese reader prefers to read more Chinese summarization than English one.

6.1 Focusing Model

A summarization is presented by voting from reporters. For each event, reporter records a news story from his own viewpoint. Recall that a news story is composed of several sentences. Those sentences that are similar in a specific event are common focus of different reporters. In other words, they are worthy of reading. For each set of similar sentences, only the longest sentence is displayed. The display order of the extracted sentences is determined by the related position in the original news articles. The following formula defines a position score function.

$$PositionScore = \frac{position(S,D)}{sizeof(D)} \tag{8}$$

where *position*(*S,D*) denotes the position of sentence S in document *D*,
sizeof(*D*) is the size of document *D*.

The extracted sentences are sorted in the ascending order of position scores. When users' language preference is considered, the sentences are selected by languages and voting of reporters, and displayed by position scores. That is, they are grouped by languages. Figure 6 sketches the concepts of focusing model. Chinese *i-j* means the *j-th* sentence in *i-th* Chinese news article.

6.2 Browsing Model

The news articles are listed by information decay and chronological order. The first article is shown to the user in its whole content. In the latter news articles, those sentences, that have higher similarity scores with the sentences in former news articles, are shadowed (or eliminated), so that the reader can only focus on the novel information. We also consider the readers' preference in multi-lingual multi-document summarization. The news articles in the preferred language are shown before those in other languages. Figure 7 sketches the concepts of browsing mode with Chinese preference. Chinese news article 1 is the first article, so the whole content is shown to the reader. However, due to the high similarity with sentences in Chinese news article 1, sentences 3, 4 and 5 in English news article 1 are shadowed (underlined). Similarly, sentence 3 in Chinese news article 3 has high similarity score with some sentences in Chinese news article 1, 2 or English news article 1, 2.

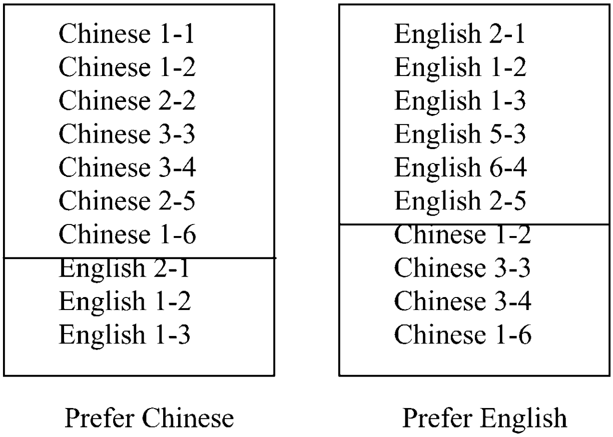

Fig. 6. Visualization in Focusing Model

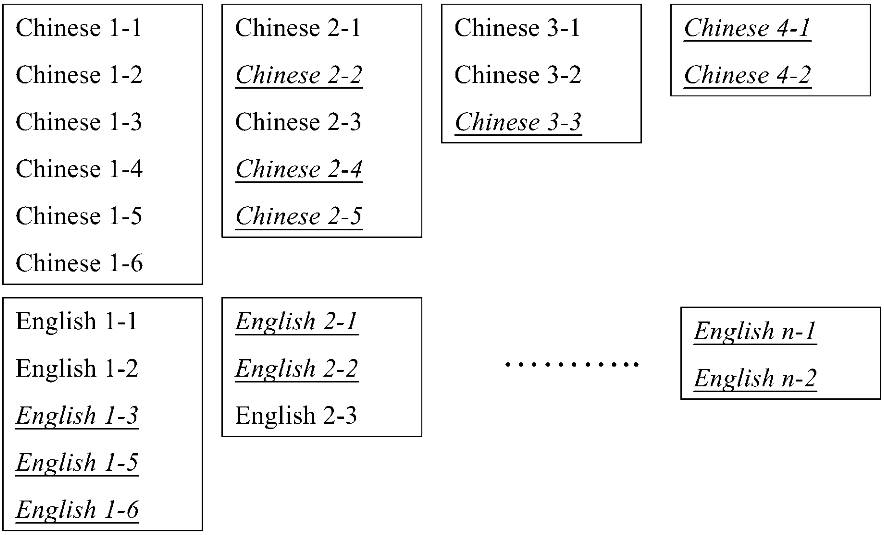

Fig. 7. Visualization in Browsing Model

7 Concluding Remarks

This paper presents a multi-lingual multi-document summarization system. Five strategies are proposed to measure the similarities between two bilingual sentences. The position-free strategy is better than the position-dependent strategy. Besides, two strategies are proposed for multi-lingual document clustering. The two-phase strategy (translation after clustering) is better than one-phase strategy (translation before clustering). Translation deferred to sentence clustering, which reduces the propagation of translation errors, is most promising. Moreover, three strategies are proposed to tackle the sentence clustering. Complete link within a cluster has the best performance, however, the subsumption-based clustering has the advantage of lower computation complexity and similar performance. Finally, two visualization models (i.e., focusing and browsing), which considers the users' language preference, are proposed.

Acknowledgements

Part of the research results of this paper was supported by National Science Council, Taiwan, under the contract NSC90-2213-E-002-045.

References

[1] Barzilay, Regina and Elhadad, Michael (1997) "Using Lexical Chains for Text Summarization," *Proceedings of ACL/EACL 1997 Workshop on The Intelligent Scalable Text Summarization*, pp. 10-16.

[2] Chen, Hsin-Hsi and Huang, Sheng-Jie (1999) "A Summarization System for Chinese News from Multiple Sources," *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, pp. 1-7.

[3] Chen, Hsin-Hsi and Lin, Chuan-Jie (2000) "A Multilingual News Summarizer," Proceedings of 18th International Conference on Computational Linguistics, pp. 159-165.

[4] Chen, Hsin-Hsi, Lin, Chi-Ching and Lin, Wen-Cheng (2002) "Building a Chinese-English WordNet for Translingual Applications," *ACM Transactions on Asian Language Information Processing*, **1**(2), pp. 103-122.

[5] Fellbaum, Christinae., Ed. (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

[6] Goldstein, Jade, Mittal, Vibhu, Carbonell, Jaime and Callan, Jamie (2000) "Creating and Evaluating Multi-Document Sentence Extract Summaries," *Proceedings of the 2000 ACM International Conference on Information and Knowledge Management*, pp. 165-172.

[7] Hatzivassiloglou, Vasileios, Klavans, Judith L., Holcombe, Melissa L. Barzilay, Regina, Kan, Min-Yen and Mckeown, Kathleen R. (2001) "SIMFINDER: A Flexible Clustering Tool for Summarization," *Proceedings of NAACL2001 Workshop on Automation Summarization*, pp. 41-49.

[8] Mani, Inderjeet and Bloedorn, Eric (1999) "Summarizing Similarities and Difference among Related Documents," *Information Retrieval*, 1(1-2), pp. 35-67.

[9] Mckeown, Kathleen, Klavans, Judith L., Hatzivassiloglou, Vasileios, Barzilay, Regina and Eskin, Eleazar (1999) "Towards Multi-document Summarization by Reformulation," *Proceedings of AAAI-99*, pp. 453-460.

[10] Radev, Dragomir.R., Jing, Hongyan and Budzikowska, Malgorzata (2000) "Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies," *Proceedings of Workshop on Summarization*, ANLP/NAACL, 2000.

[11] Vossen, Piek (1998) "EuroWordNet: Building a Multilingual Database with Wordnets for European languages, " *The ELRA Newsletter*, 3(1), 7-10.