Computational Linguistics and Chinese Language Processing
Vol. 13, No. 1, March 2008, pp. 41-60 


# Cross-Lingual News Group Recommendation Using Cluster-Based Cross-Training

**Cheng-Zen Yang*, Ing-Xiang Chen*, and Ping-Jung Wu***

**Abstract**

Many Web news portals have provided clustered news categories for readers to browse many related news articles. However, to the best of our knowledge, they only provide monolingual services. For readers who want to find related news articles in different languages, the search process is very cumbersome. In this paper, we propose a cross-lingual news group recommendation framework using the cross-training technique to help readers find related cross-lingual news groups. The framework is studied with different implementations of SVM and Maximum Entropy models. We have conducted several experiments with news articles from Google News as the experimental data sets. From the experimental results, we find that the proposed cross-training framework can achieve accuracy improvement in most cases.

**Keywords:** Cross-Lingual News Group Mapping, Cross-Training, Semantic Overlapping, Mapping Recommendation

## 1. Introduction

As the Web becomes an abundant source of news information, it also becomes an important medium for people to learn recent tidings. To provide readers a convenient way of viewing a news event described by different news agencies, many Web news portals, such as AltaVista News and Google News, cluster news articles according to their relevance with consistent user interfaces. With such news clustering services, readers could easily acquire more details of an interesting news event from numerous reports. Ideally, they can simply click through an entry link to browse many related news reports without need of a cumbersome searching procedure. Nevertheless, if the news event is originally reported by foreign news agencies, the readers usually find that there are only few translated news articles and can only acquire an overview

* Dept. of Computer Sci. and Eng., Yuan Ze University, 135 Yuan-Tung Rd., Chungli, 320, Taiwan.
Tel.: +886-3-4638800 ext: 2361 Fax: +886-3-4638850.
E-mail: {czyang,sean,pjwu}@syslab.cse.yzu.edu.tw

[Received June 30, 2007; Revised October 8, 2007; Accepted January 12, 2008]

of the news event. If they want to find more related foreign news stories, they may generally get frustrated due to the following two reasons. First, the translated news articles seldom provide as much information as the original news articles. Second, the translation may add more interpretations that can mislead in the searching direction. The following example illustrates these situations.

This news story, reported in BBC News [2006], is a good example to show these problems. The title of its English version is "First impressions count for web" and the article contains 15 paragraphs mainly focused on the impressions in a 20th of a second after first sight [BBC News 2006]. However, the title of its Chinese news story is "好網頁還需要讓讀者一見鍾情" and may be translated into "Good web pages need to let readers fall in love at first sight", which includes additional semantic information related to love. In addition, the Chinese news article has only 7 paragraphs. When readers read the Chinese news article (the source document) and want to find more information from (for example) English news articles (target documents), they will most likely search for the news article entitled with "fall in love at first sight" and find nothing related. Apparently, the readers cannot easily find the English news article. Additionally, the amount of information of the source news article may not be equal to that of the corresponding target news article. In this example, the amount of information of the translated Chinese article is much less than that of the original English article. The scant amount of translated information will perplex the readers in other searching operations. These observations suggest the need of a cross-lingual news recommendation framework for readers to get a broader view to a news event.

To address the recommendation issue for cross-lingual news groups, the simplest approach is to directly translate the source news article and find the related news group in another language. Unfortunately, the quality of translation and the amount of news information highly influence the recommendation results. Readers may get translated results of poor quality. For instance, using Google Translation (http://www.google.com/translate_t?hl=zh-TW) to translate the Chinese news title of the above example gets "Readers need to make a good website was love at first sight". As many Web news portals have provided monolingual cluster-based news browsing interfaces, the quality of cross-lingual news group recommendation can be improved if the cluster information of the source documents is exploited. Such exploration of cluster information has been studied recently in many applications, such as Web catalog integration [Agrawal and Srikan 2001; Tsay *et al*. 2003; Sarawagi *et al*. 2003; Zhang and Lee 2004a; Zhang and Lee 2004b; Chen 2005] and title generation [Tseng *et al*. 2006].

In this paper, we propose a cross-lingual news group recommendation framework using the cross-training approach from recent Web taxonomy integration techniques [Sarawagi *et al*. 2003] to find the possible semantic corresponding relationships between news groups of

different languages. With the cross-training approach, the framework explores the implicit clustering information from the source news groups and the target news groups by learning the group features alternately. Then, the framework utilizes the implicit clustering information to improve the mapping accuracy between news groups of different languages.

Such a framework has two major advantages. First, it will save considerable news searching effort resulting from the cumbersome searching procedure in which readers need to query different monolingual news portals in a trial-and-error manner. Second, it mitigates the translation inaccuracy to provide readers a broader panorama of news events from different aspects.

The cross-training framework has been implemented in Support Vector Machines (SVM) and Maximum Entropy (ME) classifiers. We have also conducted experiments to investigate the accuracy improvement of the cross-training approach with a 21-day data set containing English and Chinese news articles collected from Google News. In the experiments, we measured the accuracy performance for different approaches. The experimental results show that the cross-training approach can benefit the mapping accuracy in most cases.

The rest of the paper is organized as follows. In Section 2, we present the problem definitions and briefly review previous related research on Web catalog integration. Section 3 elaborates the proposed cross-training framework. Section 4 describes our experiments in which English news and Chinese news articles from Google News were used as the data sets. Section 5 concludes the paper and discusses future directions.

## 2. Problem Statement and Related Research

For the recommendation problem of clustered news groups in different languages, we assume that the recommendation process deals with two Web news catalogs in two different languages to find the best semantically correlated relationships between the two news catalogs. We also assume that readers browse one news catalog and want to find related news articles in another news catalog of another language for the sake of simplicity. The catalog browsed by readers is the source $S$ in which the news articles (source documents) are written in language $L_s$ and have been classified into $m$ event clusters $S_1, S_2,\ldots, S_m$. The other is the target catalog $T$ in which the news articles (target documents) are written in language $L_t$ and have been also classified into $n$ clusters $T_1, T_2,\ldots, T_n$. The terms of the documents of each cluster comprise the feature space of the corresponding news event.

In the recommendation process, therefore, the objective of the framework is to discover all possible cluster-to-cluster mapping relationships between $S$ and $T$, and report these relationships to the readers for recommendation. For the sake of simplicity in discussion, we only consider the best mapping relationships in this paper, *i.e.*, given a source catalog $S_i$, the

best corresponding target catalog $T_j$ ($S_i \rightarrow T_j$) is identified in this study. Ideally, if both news clusters $S_i$ and $T_j$ focus on the same news event, the news articles in both clusters should have semantic overlap, as shown in Figure 1. Generally, the mapping relationships are one-to-one and symmetric. However, in our observations, one-to-many situations indeed have occurred because more than one target cluster is overlapped by the same source cluster. Furthermore, source documents will be translated in $L_t$ first, and the quality of the feature space of the translated source documents may be hindered due to the poor translation process. These factors may make the symmetric relationships asymmetric. Therefore, the reverse mappings ($T_j \rightarrow S_i$) are separately considered.

Generally, the cluster-to-cluster mapping discovery problem can be viewed as a generalization of the Web catalog integration problem on a coarse-grained basis. In the Web catalog integration problem, the objective of the integration process is to classify the documents in the source catalog into the target catalog with the enhancement of the implicit source information. In recent years, there have been many approaches proposed for the general catalog integration problem. For example, the Naïve Bayes approaches [Agrawal and Srikan 2001; Tsay *et al*. 2003], the SVM-based approaches [Sarawagi *et al*. 2003; Zhang and Lee 2004a; Zhang and Lee 2004b; Chen 2005], and the Maximum Entropy approach [Wu *et al*. 2005] have shown that the integration improvement can be effectively achieved.

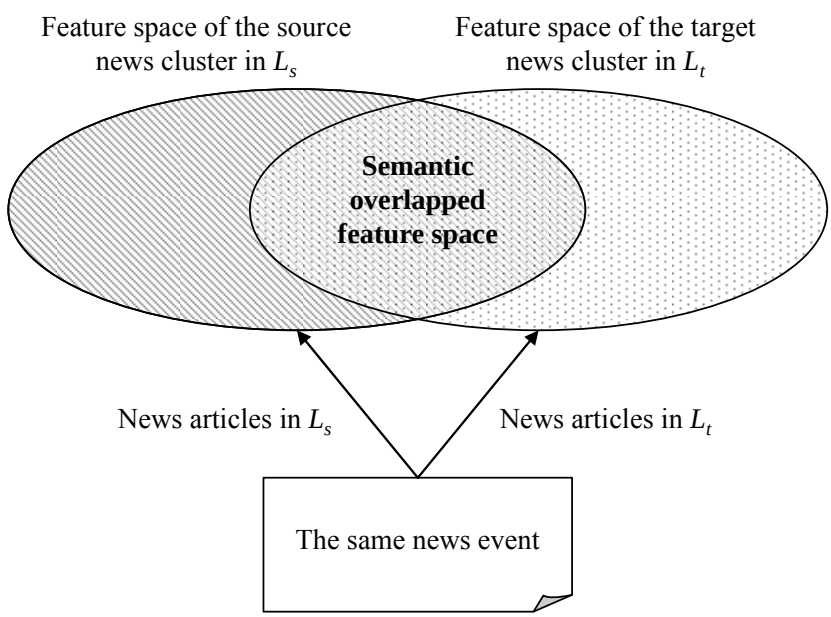


***Figure 1. The relation of the news event and the correspondent news clusters in $L_s$ and $L_t$.***

Some enhancement approaches, however, may not be suitable for the cross-lingual cluster-to-cluster mapping discovery problem. For example, the topic restriction approach proposed in Tsay *et al*. [2003] requires that the testing target clusters are the clusters containing common documents from the source cluster. Nonetheless, in the cross-lingual cluster-to-cluster mapping discovery problem, there cannot be such a common subset. The enhanced Naïve Bayes (ENB) approach proposed in [Agrawal and Srikan 2001] exploits the implicit source catalog information to enhance the integration accuracy performance. However, due to the diversity of news articles and the translation variety, the iterative algorithm may introduce many false-positive mappings to twist the overlapped space into a larger one. The shrinkage approach adopted in Wu *et al*. [2005] also needs to be adapted because the news clusters are usually not hierarchically organized.

Our recommendation framework uses the cross-training approach adapted from the cross-training (CT) approach proposed in [Sarawagi *et al*. 2003]. The CT approach is a semi-supervised learning strategy. The idea behind CT is that a better classifier can be built with the assistance of another catalog that has semantic overlap. The overlapped document set is fully-labeled and partitioned into a development set and a test set where the development set is used to tune the system performance and the test set is used to evaluate the system. Through the cross-training process, the implicit information in the source taxonomy is learnt, and more source documents can be accurately integrated into the target taxonomy.

The proposed framework utilizes the CT approach to first obtain the potential mapping relationships from the reverse mappings ($T_j \rightarrow S_i$) through a learning process. The extracted information then is used to augment the feature space in the next learning phase. Finally, the mappings from $S_i$ to $T_j$ are explored in a classification process.

## 3. Cross-Training for Mapping Discovery

The main design principle of the cross-training framework is that the implicit mapping relationships are extracted through the first learning phase on reverse mappings. In this phase, the strength of each possible mapping is identified and ranked. For each $S_i$, the framework can find the most possibly corresponding $T_j$. Before the second learning phase, the feature space of each $T_j$ is expanded with the discovered mapping information. Then, the augmented classifiers are used to identify the mapping relationships from $S_i$ to $T_j$, and give the recommendations.

### 3.1 The Processing Flow

Figure 2 depicts the processing flow in the cross-training framework. Without loss of generality, we use English and Chinese here as two language representatives for $L_s$ and $L_t$ to explain our bilingual recommendation process in this paper.

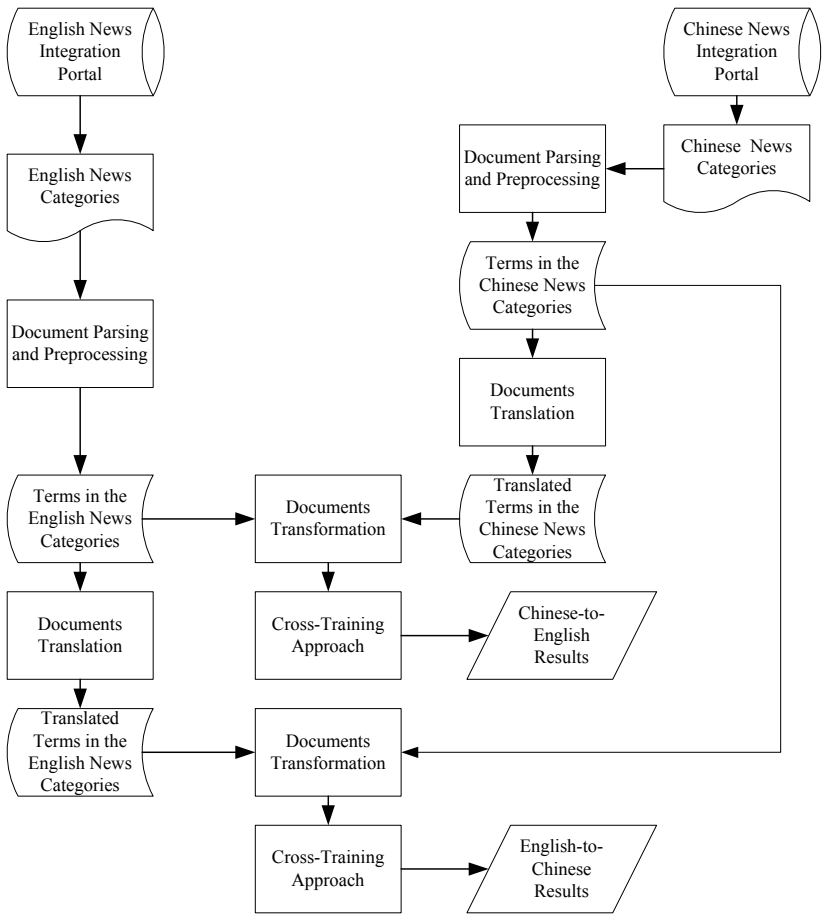


***Figure 2. The processing flow for bilingual news group recommendation in the cross-training framework.***

In the framework, the classification system first retrieves English and Chinese news articles from news portals, say Google News or Yahoo! News. These news articles have been usually clustered well in the news portals. The framework then performs parsing and preprocessing on each news cluster to get its feature space. The preprocessor parses the Web news, and eliminates stopwords [Fox 1992] and HTML tags. After the preprocessing, the source news groups are translated into the target language. For example, if a reader wants to find the possible English news groups for a designated Chinese news group, the English news articles are in the source news groups and will be translated into Chinese. After the translation process, all the source and target news groups are prepared as the data sets for further cross-training operations.

A debate may arise about whether the framework should re-cluster the news articles after the translation process. Since the translation process may introduce semantic variety into the news clusters, re-clustering the news articles may produce clusters with better semantic integrity for the following recommendation process. Nonetheless, the observations in Chen *et al*. [2003] show that the re-clustering process can contrarily reduce the quality of the original semantic integrity. Therefore, the proposed framework will not re-cluster the news articles.

## 3.2 Parsing and Preprocessing

As each Web news article is composed of plain text and HTML tags, it needs to be parsed first to extract useful information. For simplicity sake, the document parsing procedure is currently designed in a conservative manner by ignoring the HTML tags and extracting only the plain text.

Both Chinese and English news articles are then preprocessed. There are four steps for English news articles: (1) tokenization, (2) stopword removal, (3) stemming, and (4) generation of term-frequency vectors. As there is no word boundary in Chinese sentences, the Chinese articles need to be segmented first [Nie and Ren 1999; Nie *et al*. 2000; Foo and Li 2004]. We use a hybrid approach proposed by Tseng [2002], which can achieve a high precision rate and a considerably good recall rate by considering unknown words. The hybrid approach combines the longest match dictionary-based segmentation method and a statistical-based approach which is a fast keyword/key-phrase extraction algorithm. With this hybrid approach, each sentence is scanned sequentially and the longest matched words based on the dictionary entries are extracted. This process is repeated until all characters are scanned.

## 3.3 Translation and Transformation

After preprocessing, the Chinese and English news articles in each category are tokenized. Then, the Chinese news documents are translated. The translation can be based on a bilingual dictionary or a well-trained machine translation system. In the translation, we adopt a straightforward word expansion method. Each Chinese word is simply translated to a set of English terms listed in a bilingual dictionary or derived from a machine translation system. The same procedure is also applied to the English news articles. Currently, the translation process does not consider the word choice disambiguation problem when there are several candidates for each word. The translation quality is not further addressed using different translation technologies. Nonetheless, it can be found that the proposed cross-training approach achieves around 90% accuracy performance in top-1 ranking.

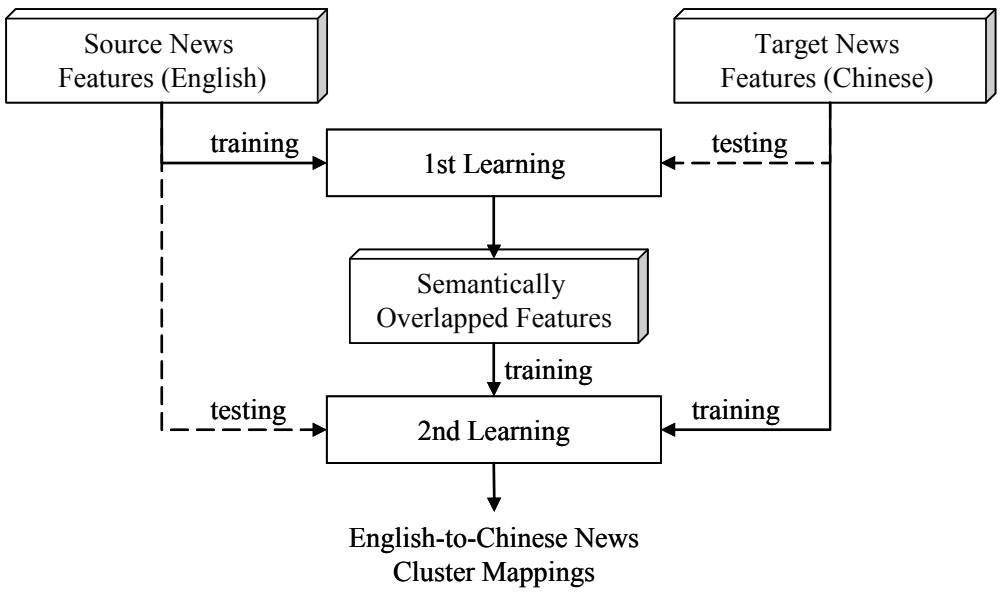


***Figure 3. The basic concept of the cross-training process.***

Finally, each news article is converted to a feature vector. For each index term in the feature vector, a weight is associated with the term to express the importance. In the current design, the weight of each term is calculated by $TF_x / \sum TF_i$, where *i* denotes the number of the stemmed terms in each news article.

## 3.4 The Cross-Training Process

Previous studies on the general Web catalog integration problem show that, if a source document can be integrated into a target category, there must be a sufficiently large semantic overlap between them [Agrawal and Srikan 2001; Sarawagi *et al*. 2003; Tsay *et al*. 2003; Zhang and Lee 2004a; Zhang and Lee 2004b; Wu *et al*. 2005; Yang 2006]. For the cluster-to-cluster mapping discovery problem, this observation is also an important basis. If an English news category can be associated with a Chinese news category, this mapping must be concluded from a situation in which the semantically overlapped feature space is sufficiently large.

### 3.4.1 Learning to Extract the Implicit Information

The cross-training process is incorporated mainly for exploring the overlapped feature space. Figure 3 illustrates a cross-training process in which there are two learning phases. In the first phase, the source news clusters are used as the training data sets to train *m* classifiers, and the target news clusters are used as the testing data sets to extract the implicit mapping

information. The $m$ classifiers then calculate the mapping scores ($Sc_{ij}$) for $n$ target news clusters to predict the strengths of the semantic overlaps.

Since SVM and ME are studied in the framework implementations, the mapping score $Sc_{ij}$ of $T_j \rightarrow S_i$ can be defined as either the ratio at which the target documents in $T_j$ are classified into the source news cluster $S_i$ or the average weight derived from the classifier. For example, if the classification scheme used in the framework is SVM, the mapping score $Sc_{ij}$ can be calculated by either Eq. (1) where $N_{Tj}$ is the news documents of the target cluster $T_j$ or Eq. (2) which is the average of the distance from each document to the hyperplane. This average can be viewed as the discriminative characteristic of all documents to the classifier.

$$Sc_{ij} = \frac{\#\text{of } N_{Tj} \text{ classified in } S_i}{\#\text{of } N_{Tj}} \tag{1}$$

$$Sc_{ij} = \frac{\sum w_i x_i + b}{\#\text{of } N_{Tj}} \tag{2}$$

Basically, Equation (1) represents a voting scheme in which the predicted rank of a target cluster $T_j$ depends on the number of the positively classified news articles in $T_j$. Equation (2) represents a weighting scheme in which the predicted rank of $T_j$ depends on the average of the total distance to the hyperplane. For each source cluster, the target cluster with the highest mapping score is qualified as the potential candidate that may have the accurate $S_i \rightarrow T_j$ mapping relationship in the second learning phase. The reason the mapping scores are considered in an asymmetric way is that the cross-training approach will adjust the feature vectors back and forth in each learning iteration. Other mapping discovery approaches may provide efficient schemes to consider both mapping scores of $S_i \rightarrow T_j$ and $T_j \rightarrow S_i$ as an integrated scoring method. This has been left for our future study.

### 3.4.2 Learning to Find the Corresponding Mappings

The implicit information explored in the first learning phase is then used as the prediction information in the second learning phase. The cross training process can be continued until the results converge. The category information of the corresponding source cluster, say $S_i$, for the previously discovered candidate target cluster, say $T_j$, is inserted into the feature space of $T_j$. The category information can be category identifiers or the category title words. For example, we used identifiers starting from 1000001 to 1000040 for categories in the current experiments.

Figure 4 depicts the detailed process of concatenating the predicted implicit information to the ordinary feature vectors of the target cluster in the cross-training approach. In the figure, $F_T$ is a feature vector for the term features of the target news articles, $L_T$ is a feature vector for the label features (category information) of the target cluster, and the *test output* contains the

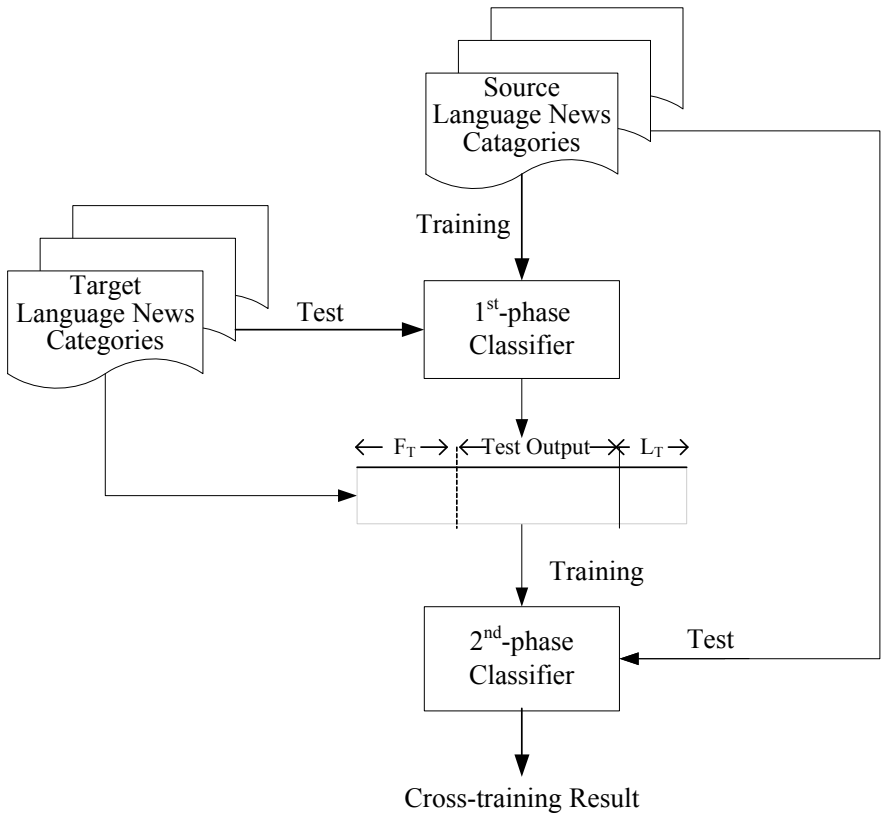


***Figure 4. Adding the predicted implicit information in the cross-training process.***

label features of the predicted source clusters. With the predicted mapping information, the discriminative power of the classifiers of the second phase can be enhanced.

For controlling the discriminative power of the added semantically-overlapped implicit label information, as in Sarawagi *et al*. [2003], the ordinary feature weights in the augmented target vectors are scaled by a factor of $f$, and the weight of each label attribute by a factor of $1 - f$. The parameter $f$ is used to decide the relative weights of the label and term features and can be tuned for different application environments. In the current experiments, the results show that the best $f$ value ranges from 0.02 to 0.05. The small $f$ values show that the augmented information should not be overemphasized in the cross-training process. This observation for factoring is consistent with previous studies [Sarawagi *et al*. 2003; Chen *et al*. 2004].

Finally, the second-phase classifiers are trained with the augmented target vectors. The recommended source news groups of the target news groups are calculated using the same mapping scoring method.

## 4. Experiments

We have implemented the cross-training framework in SVM and ME classifiers. To rank the predictive corresponding target clusters, we implemented the voting scheme in the cross-training framework of SVM (SVM-VCT) and ME (ME-VCT), and the weighting scheme with SVM (SVM-WCT). As stated in Section 3.4.1, Equation (1) was used to rank the target clusters in SVM-VCT and ME-VCT. Equation (2) was used in the weighting scheme SVM-WCT. We also implemented the voting scheme and the weighting scheme in SVM (SVM-V and SVM-W) for comparison. In the experiments, an English news catalog and a Chinese news catalog from Google News were used as the representatives to demonstrate the classification performance of the proposed cross-training framework. We measured the accuracy performance at top-1, top-3, and top-5 ranks. The details of the experiments are presented as follows.

### 4.1 The Experimental Environment

The framework is currently implemented in Java. The segmentation corpus is based on the Academia Sinica Bilingual Wordnet 1.0 published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) [Sinica BOW 2005]. We used $SVM^{light}$ (version 5.00) [Joachims 2002] as the SVM tool with a linear kernel, and the maximum-entropy toolkit (version 20041229) [Zhang 2004] as the Maximum Entropy model kernel.

The bilingual word lists published by Linguistic Data Consortium (LDC) were used as the bilingual dictionaries. The Chinese-to-English dictionary ver. 2 (ldc2ce) has about 120,000 records, and the English-to-Chinese dictionary (ldc2ec) has about 110,000 records. In ldc2ce and ldc2ec, each entry is composed of a single word and several translated words separated by slashes without any indication of the importance. Therefore, the translated words are treated equally in our experiments. In the translation, each word in the source document was replaced with these translated words. The translation quality issue is not addressed in depth because we want to follow the normal reader behaviors. Furthermore, the implicit semantic information embedded in each category of news articles may mitigate the poorness of the translation.

### 4.2 Data Sets

In our experiments, two news portals were chosen as the bilingual news sources: Google News U.S. version for English news and Google News Taiwan version for Chinese news. Both the Chinese and English news articles were retrieved from the world news category from May 10 to May 23, 2005 and from October 21 to October 27, 2007. The experiments were performed on the data set of each day. Twenty news categories were collected per day. All the English

news articles were translated into Chinese with the bilingual dictionaries. The size of the English-to-Chinese data set is 454.5 Mbytes. All the Chinese news articles were also translated. The size of the Chinese-to-English data set is 341.5 Mbytes. The 21-day data sets contain 36,548 English news articles and 8,224 Chinese news articles.

In the experiments, the mapping relations between the Chinese and English news reports were first identified by three graduate students manually and independently. The mapping between an English news category and a Chinese news category is recognized if at least two students have the same mapping identification. These manually-identified mapping relations were used to evaluate the accuracy performance of the bilingual classification systems. We found that there were 122 identified mappings in the Chinese-to-English recommendation task and 123 identified mappings in the English-to-Chinese recommendation task. The difference existed because an English category was identified that was to be mapped to two Chinese categories. The data sets collected currently cannot significantly reveal the influences of one-to-many situations. In our future work plan, more news categories need to be collected to verify our scheme for one-to-many cases.

The experiments were conducted in two ways: finding the related Chinese news groups from the English news groups (Chinese-to-English) and finding the related English news groups from the Chinese news groups (English-to-Chinese). Here, we take the Chinese-to-English recommendation process as the example to present the experimental details. The English-to-Chinese recommendation process was conducted in a similar manner.

In the Chinese-to-English experiments, each Chinese news catalog was first used as the training set in the first learning phase. To find a corresponding Chinese category ($S_i$) of an English target category ($T_j$ ), the news articles in $S_i$ were all used as the positive training examples, and the news articles in the other Chinese news categories ($S_k$, $k \neq i$) were randomly selected as the negative training examples. Then, all mapping scores between English categories and Chinese categories were measured based on the first-phase classification results. The English category with the highest mapping score was considered as the possibly mapped category.

In the second learning phase of the Chinese-to-English experiments, the category information of the previously identified English cluster was concatenated to the corresponding Chinese cluster. Then, the English categories were used as the training set to train the second-phase classifiers. The augmented source Chinese categories were classified to calculate the mapping scores for each English news category. Finally, we measured the accuracy performance for each day using the correct mappings at the top-1, top-3, and top-5 recommendation ranks by the following equation:

$$\text{Accuracy} = \frac{\text{Number of the correctly discovered mapping in } S \to T}{\text{Total number of the correct mapping in } S \to T}, \quad (3)$$

which is similar to Agrawal and Srikan [2001]. Accuracy, rather than precision or recall, is used because the recommendation process is performed on a cluster-to-cluster basis. The error rate is the complement of the accuracy. In the English-to-Chinese experiments, the roles of two catalogs were switched.

## 4.3 Results and Discussion

***Table 1. Experimental results of the correctly discovered Chinese-to-English mappings in the top-1 recommendation lists.***

| Day | Tagged Mappings | SVM-V | SVM-VCT | SVM-W | SVM-WCT | ME | ME-VCT |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 6 | 5 | 5 | 4 | 7 |
| 2 | 6 | 1 | 6 | 6 | 4 | 3 | 6 |
| 3 | 6 | 0 | 5 | 4 | 5 | 3 | 6 |
| 4 | 6 | 1 | 6 | 6 | 5 | 5 | 6 |
| 5 | 6 | 0 | 6 | 6 | 5 | 3 | 5 |
| 6 | 6 | 2 | 3 | 6 | 6 | 3 | 3 |
| 7 | 3 | 1 | 3 | 1 | 0 | 1 | 3 |
| 8 | 6 | 2 | 5 | 6 | 3 | 2 | 6 |
| 9 | 5 | 1 | 5 | 3 | 4 | 1 | 5 |
| 10 | 5 | 1 | 5 | 4 | 4 | 2 | 3 |
| 11 | 4 | 0 | 4 | 2 | 2 | 2 | 3 |
| 12 | 7 | 3 | 7 | 4 | 5 | 2 | 7 |
| 13 | 4 | 3 | 4 | 4 | 3 | 4 | 4 |
| 14 | 8 | 5 | 7 | 7 | 6 | 4 | 5 |
| 15 | 10 | 5 | 10 | 10 | 9 | 9 | 9 |
| 16 | 8 | 2 | 8 | 7 | 7 | 6 | 7 |
| 17 | 5 | 1 | 5 | 4 | 4 | 4 | 5 |
| 18 | 6 | 1 | 6 | 4 | 6 | 5 | 6 |
| 19 | 2 | 0 | 2 | 1 | 1 | 2 | 2 |
| 20 | 5 | 2 | 5 | 5 | 4 | 5 | 5 |
| 21 | 7 | 1 | 7 | 7 | 5 | 5 | 7 |
| Total | 122 | 32 | 115 | 102 | 93 | 75 | 110 |
| Avg. Acc. | | 26.23% | 94.26% | 83.61% | 76.23% | 61.48% | 90.16% |

Table 1 lists the experimental results of the correctly discovered Chinese-to-English mappings at the top-1 recommendation lists identified by different approaches. Table 2 lists the correctly discovered Chinese-to-English mappings at the top-3 recommendation lists. From these tables, we can notice that the cross-training approach significantly improves the voting approaches in SVM-V and ME to find correct mappings in the top-1 recommendation results. In addition, it improves SVM-V, SVM-W, and ME entirely to find correct mappings in the top-3 recommendation results. Here, the scaling factor $f$ is 0.05. When $f$ ranged from 0.02 to 0.05, we attained similar results.

***Table 2. Experimental results of the correctly discovered Chinese-to-English mappings in the top-3 recommendation lists.***

| Day | Tagged Mappings | SVM-V | SVM-VCT | SVM-W | SVM-WCT | ME | ME-VCT |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 6 | 7 | 7 | 5 | 7 |
| 2 | 6 | 1 | 6 | 6 | 6 | 6 | 6 |
| 3 | 6 | 0 | 5 | 6 | 6 | 5 | 6 |
| 4 | 6 | 1 | 6 | 6 | 6 | 5 | 6 |
| 5 | 6 | 0 | 6 | 6 | 6 | 4 | 6 |
| 6 | 6 | 3 | 3 | 6 | 6 | 4 | 4 |
| 7 | 3 | 1 | 3 | 3 | 3 | 1 | 3 |
| 8 | 6 | 2 | 5 | 6 | 6 | 3 | 6 |
| 9 | 5 | 2 | 5 | 4 | 5 | 3 | 5 |
| 10 | 5 | 1 | 5 | 5 | 5 | 3 | 3 |
| 11 | 4 | 1 | 4 | 3 | 3 | 3 | 3 |
| 12 | 7 | 3 | 7 | 7 | 7 | 5 | 7 |
| 13 | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| 14 | 8 | 5 | 7 | 8 | 8 | 7 | 7 |
| 15 | 10 | 5 | 10 | 10 | 10 | 9 | 9 |
| 16 | 8 | 2 | 8 | 7 | 7 | 6 | 7 |
| 17 | 5 | 1 | 5 | 5 | 5 | 4 | 5 |
| 18 | 6 | 1 | 6 | 5 | 6 | 5 | 6 |
| 19 | 2 | 0 | 2 | 2 | 2 | 2 | 2 |
| 20 | 5 | 2 | 5 | 5 | 5 | 5 | 5 |
| 21 | 7 | 1 | 7 | 7 | 7 | 6 | 7 |
| Total | 122 | 35 | 115 | 118 | 120 | 95 | 114 |
| Avg. Acc. | | 28.69% | 94.26% | 96.72% | 98.36% | 77.87% | 93.44% |

From Table 1, it is noticeable that SVM-W outperformed SVM-WCT. The reason the cross-training approach cannot benefit the accuracy performance is because adding more features changes the characteristics of the hyperplanes learned by SVM, thereby affecting the distance summation results in Eq. (2). Therefore, some correct mappings were ranked at the second rank in the recommendation lists for SVM-WCT but at the top rank for SVM-W. For the top-3 recommendation lists as shown in Table 2, SVM-WCT outperformed SVM-W and got the best accuracy performance.

***Table 3. Experimental results of the correctly discovered English-to-Chinese mappings in the top-1 recommendation lists.***

| Day | Tagged Mappings | SVM-V | SVM-VCT | SVM-W | SVM-WCT | ME | ME-VCT |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 4 | 7 | 5 | 6 | 5 |
| 2 | 6 | 3 | 6 | 6 | 4 | 6 | 6 |
| 3 | 6 | 0 | 5 | 6 | 6 | 6 | 5 |
| 4 | 6 | 0 | 4 | 6 | 5 | 5 | 6 |
| 5 | 6 | 0 | 2 | 5 | 5 | 6 | 6 |
| 6 | 6 | 2 | 5 | 5 | 5 | 4 | 6 |
| 7 | 3 | 0 | 2 | 2 | 2 | 3 | 2 |
| 8 | 6 | 1 | 5 | 6 | 3 | 5 | 5 |
| 9 | 5 | 1 | 3 | 4 | 3 | 3 | 4 |
| 10 | 5 | 1 | 4 | 5 | 4 | 2 | 5 |
| 11 | 4 | 1 | 4 | 2 | 1 | 1 | 2 |
| 12 | 7 | 2 | 7 | 6 | 5 | 5 | 6 |
| 13 | 4 | 0 | 3 | 3 | 3 | 3 | 4 |
| 14 | 8 | 1 | 8 | 7 | 6 | 7 | 7 |
| 15 | 10 | 5 | 10 | 9 | 10 | 8 | 10 |
| 16 | 8 | 0 | 6 | 6 | 4 | 5 | 5 |
| 17 | 5 | 2 | 5 | 4 | 5 | 2 | 2 |
| 18 | 7 | 6 | 7 | 7 | 7 | 6 | 6 |
| 19 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| 20 | 5 | 2 | 5 | 5 | 5 | 3 | 5 |
| 21 | 7 | 3 | 7 | 7 | 6 | 5 | 7 |
| Total | 123 | 31 | 104 | 110 | 95 | 93 | 106 |
| Avg. Acc. | | 25.20% | 84.55% | 89.43% | 77.24% | 75.61% | 86.18% |

Table 3 and Table 4 list the experimental results of the correct English-to-Chinese mappings in the top-1 and top-3 recommendation lists, respectively. From these two tables, we can see that the cross-training approach significantly improved SVM-V and ME in finding the correct mappings in the top-1 recommendation results. From the top-3 recommendation results, we can observe that SVM-V is highly improved by the cross-training approach. SVM-W and SVM-WCT has the same results and both achieve the best performance. Although the cross-training approach cannot benefit ME more in the top-3 results as in the Chinese-to-English experiments, the performance of ME-VCT is comparable to ME.

***Table 4. Experimental results of the correctly discovered English-to-Chinese mappings in the top-3 recommendation lists.***

| Day | Tagged Mappings | SVM-V | SVM-VCT | SVM-W | SVM-WCT | ME | ME-VCT |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | 4 | 7 | 7 | 7 | 6 |
| 2 | 6 | 3 | 6 | 6 | 6 | 6 | 6 |
| 3 | 6 | 0 | 5 | 6 | 6 | 6 | 6 |
| 4 | 6 | 0 | 4 | 6 | 5 | 6 | 6 |
| 5 | 6 | 0 | 2 | 5 | 6 | 6 | 6 |
| 6 | 6 | 2 | 5 | 6 | 6 | 5 | 6 |
| 7 | 3 | 0 | 2 | 3 | 3 | 3 | 2 |
| 8 | 6 | 2 | 6 | 6 | 6 | 6 | 5 |
| 9 | 5 | 1 | 3 | 4 | 4 | 3 | 4 |
| 10 | 5 | 1 | 4 | 5 | 5 | 5 | 5 |
| 11 | 4 | 1 | 4 | 3 | 3 | 2 | 3 |
| 12 | 7 | 2 | 7 | 6 | 6 | 6 | 6 |
| 13 | 4 | 1 | 3 | 4 | 4 | 4 | 4 |
| 14 | 8 | 1 | 8 | 8 | 8 | 8 | 8 |
| 15 | 10 | 5 | 10 | 10 | 10 | 10 | 10 |
| 16 | 8 | 0 | 6 | 7 | 7 | 7 | 6 |
| 17 | 5 | 2 | 5 | 5 | 5 | 5 | 3 |
| 18 | 7 | 6 | 7 | 7 | 7 | 6 | 6 |
| 19 | 2 | 1 | 2 | 2 | 2 | 2 | 2 |
| 20 | 5 | 2 | 5 | 5 | 5 | 5 | 5 |
| 21 | 7 | 6 | 7 | 7 | 7 | 7 | 7 |
| Total | 123 | 36 | 105 | 118 | 118 | 115 | 112 |
| Avg. Acc. | | 29.27% | 85.37% | 95.93% | 95.93% | 93.50% | 91.06% |

Table 5 lists the experimental results of the correct Chinese-to-English and English-to-Chinese mappings in the top-5 recommendation lists. Here, we omit the details of the correct mappings of each day and only show the total results. The top-5 results are very similar to the top-3 results.

***Table 5. Experimental results of the correctly discovered Chinese-to-English and English-to-Chinese mappings in the top-5 recommendation lists.***

(a) Results of Chinese-to-English mappings

| Day | Tagged Mappings | SVM-V | SVM-VCT | SVM-W | SVM-WCT | ME | ME-VCT |
|---|---|---|---|---|---|---|---|
| Total | 122 | 37 | 115 | 119 | 120 | 103 | 117 |
| Avg. Acc. | | 30.33% | 94.26% | 97.54% | 98.36% | 84.43% | 95.90% |

(b) Results of English-to-Chinese mappings

| Day | Tagged Mappings | SVM-V | SVM-VCT | SVM-W | SVM-WCT | ME | ME-VCT |
|---|---|---|---|---|---|---|---|
| Total | 123 | 36 | 105 | 120 | 120 | 117 | 113 |
| Avg. Acc. | | 29.27% | 85.37% | 97.56% | 97.56% | 95.12% | 91.87% |

Other improvements can still be introduced in the recommendation framework. For example, unknown name entity recognition (NER) and transliteration processing are two important issues for cross-lingual processing. Improvements to the quality of machine translation in the framework should further enhance the accuracy performance.

## 5. Conclusion

As the amount of news information explosively grows over the Internet, on-line Web news services have played an important role in delivering news information to people. Although these Web news portals have provided readers with clustered monolingual news services, cross-lingual news clustering services are still in great demand.

In this paper, we propose a cross-lingual news group recommendation framework with the cross-training approach to get high accuracy performance in finding the mapping relationships between two news catalogs in different languages. From the experimental results, we can find that the proposed cross-training recommendation framework comprehensively has the superior accuracy performance. Among all approaches, SVM-WCT can achieve the best accuracy in the top-3 and top-5 recommendation lists for both Chinese-to-English and English-to-Chinese.

There are still many research issues left for our future study. For example, feature weighting plays an important role in system performance. Meaningful features should be explored and employed for integration. In addition, we only consider the accuracy rate of correct mappings in current experiments. The correct rejection rate needs to be further studied for independent source/target categories. Furthermore, the scoring method can be discussed to find whether there are other better approaches to discover the correct mapping. In addition, a filtering scheme needs be discussed to screen out incorrect mapping recommendations (negative mappings) for practical use. One of the most challenging issues is how to translate new words which are created daily due to the rapidly changing Web. A better automatic bilingual translation system is needed to fulfill the requirements of effective term translation for the NER problem and the transliteration problem.

### Acknowledgement

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for partially supporting this research under Contract No. NSC 95-2745-E-155-008. The authors would also like to express many thanks to the anonymous reviewers for their precious suggestions for this paper.

## References

Agrawal, R. and R. Srikan, "On Integrating Catalogs," in *Proceedings of the 10th International Conference on World Wide Web*, 2001, pp. 603-612.

Chen, H.-H., J.-J. Kuo, and T.-C. Su, "Clustering and Visualization in a Multi-lingual Multidocument Summarization System," in *Proceedings of 25th European Conference on Information Retrieval Research*, 2003, pp. 266-280.

Chen, I.-X., C.-H. Shih, and C.-Z. Yang, "Web Catalog Integration using Support Vector Machines," in *Proceedings of the 1st Workshop on Intelligent Web Technology (IWT 2004)*, Taipei, Taiwan, 2004, pp. 7-13.

Chen, I.-X., J.-C. Ho, and C.-Z. Yang, "An Iterative Approach for Web Catalog Integration with Support Vector Machines," in *Proceedings of 2nd Asia Information Retrieval Symposium (AIRS 2005)*, 2005, pp. 703-708.

Foo, S. and H. Li, "Chinese Word Segmentation and Its Effect on Information Retrieval," *Information Processing and Management*, 40(1), 2004, pp. 161-190.

Fox, C., "Lexical Analysis and Stop Lists", *Information Retrieval: Data Structures and Algorithms*, Chapter 7, Frakes, W. and Baeza-Yates, R., (eds.), Prentice-Hall, 1992, pp. 102-130.

Nie, J.Y. and F. Ren, "Chinese Information Retrieval: Using Characters or Words," *Information Processing and Management*, 35(4), 1999, pp. 443-162.

Nie, J.Y., J. Gao, J. Zhang, and M. Zhou, "On the Use of Words and N-grams for Chinese Information Retrieval," in *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, 2000, pp. 141-148.

Sarawagi, S., S. Chakrabarti, and S. Godbole, "Cross-training: Learning Probabilistic Mappings between Topics," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 177-186.

Sinica BOW, The Academia Sinica Bilingual Wordnet. Ver. 1.0, The Association for Computational Linguistics and Chinese Language Processing, 2005.

Tsay, J.-J., H.-Y. Chen, C.-F. Chang, and C.-H. Lin, "Enhancing Techniques for Efficient Topic Hierarchy Integration," in *Proceedings of the 3rd International Conference on Data Mining (ICDM'03)*, 2003, pp. 657-660.

Tseng, Y.-H., "Automatic Thesaurus Generation for Chinese Documents," *Journal of the American Society for Information Science and Technology*, 53(13), 2002, pp. 1130-1138.

Tseng, Y.-H., C.-J. Lin, H.-H. Chen, and Y.-I Lin, "Toward Generic Title Generation for Clustered Documents," in *Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS 2006)*, 2006, Singapore, pp. 145-157.

Wu, C. W., T. H. Tsai, and W. L. Hsu, "Learning to Integrate Web Taxonomies with Fine-Grained Relations: A Case Study Using Maximum Entropy Model," in *Proceedings of 2nd Asia Information Retrieval Symposium (AIRS 2005)*, 2005, pp. 190-205.

Yang, C.-Z., C.-M. Chen, and I.-X. Chen, "A Cross-Lingual Framework for Web News Taxonomy Integration," in *Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS 2006)*, 2006, Singapore, pp. 270-283.

Zhang, D. and W. S. Lee, "Web Taxonomy Integration using Support Vector Machines," in *Proceedings of the 13th International Conference on World Wide Web*, 2004a, pp. 472-481.

Zhang, D. and W. S. Lee, "Web Taxonomy Integration through Co-Bootstrapping," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004b, pp. 410-417.

## Online Resources

Altavista News, http://www.altavista.com/news/default.

BBC News, "First impressions count for web." English version is available at http://news.bbc.co.uk/2/hi/technology/4616700.stm; Chinese version is available at http://news.bbc.co.uk/chinese/trad/hi/newsid_4610000/newsid_4618500/4618552.stm, 2006.

Google News, http://news.google.com/.

Google Translation, http://www.google.com/translate_t?hl=zh-TW.

Joachims, T., SVM$^{light}$, version 5.0, http://svmlight.joachims.org/, 2002.

Linguistic Data Consortium, http://projects.ldc.upenn.edu/Chinese/LDCch.htm.

Yahoo! News, http://news.yahoo.com/.

Zhang, L., The Maximum Entropy model toolkit, version 20041229, http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html, 2004.