English and Chinese Bilingual Topic Aspect Classification: Exploring Similarity Measures, Optimal LSA Dimensions, and Centroid Correction of Translated Training Examples

Yejun Wu
School of Library and Information Science,
Louisiana State University
267 Coates Hall, Baton Rouge, LA 70803
wuyj@lsu.edu

Douglas W. Oard
College of Information Studies and UMIACS,
University of Maryland
College Park, MD 20742
oard@umd.edu

ABSTRACT

This paper explores topic aspect (i.e., subtopic or facet) classification for collections that contain more than one language (in this case, English and Chinese), and investigates several key technical issues that may affect the classification effectiveness. The evaluation model assumes a bilingual user who has found some documents on a topic and identified a few passages in each language on specific aspects of that topic that are of interest. Additional passages are then automatically labeled using a k-Nearest-Neighbor classifier and local (i.e., result set) Latent Semantic Analysis (LSA). Experiments show that when few manually annotated passages are available in either language, a classification system trained using passages from both languages can often achieve higher effectiveness than a similar system trained using passages from just one language. Using this experimental framework, this paper answers three technical research questions: whether the normalized cosine similarity measure is better than the more common unnormalized cosine similarity measure (yes), whether the number of retained LSA dimensions (which was heuristically chosen) is appropriate (yes), and whether partial corrections of the translated training examples in the LSA space can yield an improvement over no correction (no).

Keywords

Bilingual classification, English, Chinese, topic, aspect, similarity measure, LSA dimension, translation, training examples.

INTRODUCTION

The fundamental problem that we address is the wish to identify contiguous passages of text that address a specific aspect of a topic. By *aspect*, we mean a sub-topic or facet of the topic.[1] We wish to do this for several distinct aspects of the same topic, and we are particularly interested in associating passages from documents written in different languages with the same set of aspects. Our interest in this problem is motivated by sentiment analysis applications in which both positive and negative sentiment about different aspects of a topic are expressed. Rather than annotating the sentiment at document-scale as "mixed," we would prefer to identify which parts of the document address specific aspects of the topic addressed by the document, and then associate positive or negative sentiment with those specific passages. We have been able to do sentence-scale Chinese sentiment classification with moderate success, and English sentiment classification has been studied for more than a decade at the scale of words, sentences, passages, and entire documents (Wu & Oard, 2009). Our interest in cross-language comparative sentiment analysis thus leads directly to a need for bilingual topic aspect classification as a prerequisite task. It is that prerequisite task on which we focus in this paper. Being able to perform this task would also allow us to analyze the aspects of a topic for many other purposes, such as passage retrieval, question answering, summarization, and discourse analysis.

To frame this challenge in a manner amenable to evaluation, we assume that the results of a topic-based search are already available in two languages (e.g., from Cross-Language Information Retrieval (CLIR)). We have chosen to focus for this work on English and Chinese since they are widely spoken languages with quite different characteristics. For our evaluation, we model the bilingual

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

[1] Our choice of "aspect" rather than facet results from common use in information retrieval evaluation, where aspect recall is used as a measure of how comprehensively the different aspects of a topic are covered by a result set. In linguistics, "aspect" is often read as "grammatical aspect." That is not our intended meaning; throughout this paper, aspect should be read as "aspect of a topic."

search result set by using documents that have been manually annotated for topical relevance.

Of course, we need some way of specifying which aspects will interest the user. For this work, we have adopted a fairly straightforward approach: we assume that the user is able to read both languages, and that they will hand-annotate a few passages in each language for each aspect that is of interest. This leads directly to an evaluation design that resembles the example-based classification task of the information filtering track of the Text Retrieval Conferences (TREC), or the topic tracking task of the Topic Detection and Tracking (TDT) evaluations (NIST, 2000), but at passage scale. The TDT collections include ground truth topic annotations (for event-based topics), so we have chosen those collections as a starting point, adding automatic passage segmentation and (for some passages) ground truth topic aspect annotations.

It is already well known that document-scale training examples in one language can be used to build a topic classifier for documents in another language (Bel et al., 2003; Gliozzo & Strapparava, 2006; Olsson et al., 2005; Rigutini et al., 2005; Prettenhofer & Stein, 2010; Shi et al., 2010; Ni et al., 2011). Our problem is different in three ways, however: (1) we seek to classify passages (which are typically shorter than full documents), (2) we seek to classify those passages into aspects (which are typically more closely related to each other than topics are), and (3) we seek to use training examples in two languages (rather than just one). The research question in our previous study (Wu & Oard, 2008) was focused on the third of those differences: can examples in two languages be used together to improve classification effectiveness over what could be achieved with training examples only in one language. Our previous results indicate that balancing the investment in annotation of training examples across languages can be helpful when seeking to simultaneously optimize classification effectiveness for more than one language.

In this paper, we investigate several key technical issues that may affect classification effectiveness: the optimal number of dimensions for local Latent Semantic Analysis, whether vector length normalization after dimensionality reduction is beneficial, and the effect of partial corrections of the positions of the translated training examples in the LSA space.

The remainder of this paper is organized as follows. We first summarize our experimental framework, providing some additional details that for space reasons could not be included in (Wu & Oard, 2008). We then pose three research questions and present experiment results. Finally, we conclude the paper with a few remarks about future work.

EXPERIMENTAL FRAMEWORK

Perhaps surprisingly, prior work indicates that while cross-language topic classification is a somewhat harder problem than same-language classification, it is in general a fairly tractable problem. The reason for this is that reasonably accurate statistical term translation models can be learned from (translation-equivalent) parallel text, and that robust ways of using those term translation mappings. Monolingual classification remains more accurate than cross-language classification, however, which motivates the question that we asked in (Wu & Oard, 2008): can monolingual and cross-language training be used together in a way that will result in more accurate classification than in a single language? We explored that question in the context of classification of short passages, an underexplored condition that is important for focused analysis (e.g., for associating topic aspects with sentiment expressed by an author in that part of the text). In this section we briefly summarize those earlier experiments.

Our goal is to classify English and Chinese document segments (or passages) in documents that are already known to be relevant to a topic, based on their relevance to the aspects of that topic. For our intended application, we assume that the user provides only a limited number of training examples for each aspect, so the classification methods here employ what might be called weakly supervised learning.

Before starting, it is useful to define some terminology. We define *bilingual* aspect classification as a topic aspect classification task in which aspect-annotated training passages are available in both languages, and the evaluation passages are in only one of those languages. We consistently refer to the language of the evaluation passages (which might be Chinese or English) as the *evaluation* language and the other language (English or Chinese, respectively) simply as the *other* language. This terminology makes it easy to compare bilingual classification with same-language classification (training examples are in the *evaluation* language) or cross-language classification (training examples are in the *other* language).

In this section, we start by describing our method for same-language aspect classification; then extend that method, first to cross-language aspect classification, and then to bilingual aspect classification; and finally we introduce our test collection and then present our experiment results.

Same-Language Aspect Classification

Our same-language aspect classification system serves both as a baseline and as a point of departure for adding cross-language and bilingual aspect classification capabilities. Figure 1 illustrates the key stages in the process.

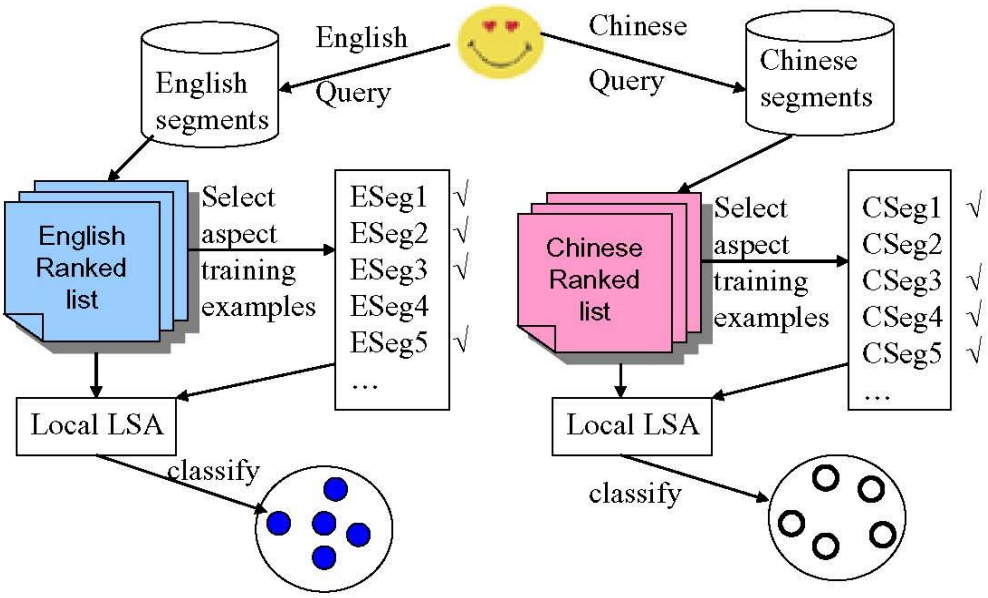

Figure 1. The procedure for same-language aspect classification.

First, all documents are automatically partitioned into segments[2] of sufficient length to support reasonably accurate term-based classification, doing so in such a way as to (hopefully) have each resulting segment address at most one aspect of the topic. Our documents are news stories from the Topic Detection and Tracking (TDT) collection. TextTiling, a process for automatically subdividing a text document into word sequences ("tiles") that are topically coherent (Hearst, 1997), was used to perform automatic segmentation, after some tuning and adaptation (Wu & Oard, 2008).

This preprocessing step abstracts away for our ultimate application scenario (in which users would most likely actually retrieve documents and then manually annotate user-designated passages as training instances) in a way that simplifies our experiment design.

Second, we model a user who can search effectively in both English and Chinese who retrieves two sets of segments for some topic, one set in English and one set in Chinese. We do this by indexing all segments in the document collection for a language (either English or Chinese) using Indri,[3] creating an Indri query in that language that is appropriate to the topic (based on the TDT topic description), and then selecting as our set some fixed number of top-ranked documents from an Indri search using that query. The user examines both sets of retrieved segments (English and Chinese) and, for each aspect, selects a few of the highly ranked segments (in our experiments, between one and seven) from each set as training examples for one aspect. She/he then repeats this examination and marking process for additional aspects. The number of different aspects that are of interest to the user will vary by topic; for our experiments we required that there be at least two aspects (i.e., single-aspect topics were removed).

To identify appropriate sets of English and Chinese segments for each topic we indexed the segments, formulated a query, performed a search, and then selected some fixed number of the highest ranked segments. In order to decide on a fixed size for our result sets, we experimented with a range of options. We had to balance two concerns: (1) we wanted enough segments so that most of the training examples would usually be in our result set, and (2) taking many more segments than we needed could yield a less focused Local Latent Semantic Analysis (LSA) model in our next processing stage. After some preliminary experimentation, we chose to select the top 1,500 Chinese segments and the top 2,500 English segments.

Third, each segment in a set is represented as a dense fixed-length vector using Local LSA. Local LSA, introduced by Hull (1994), is a variant of the LSA feature transformation and selection technique in which Singular Value Decomposition (SVD) is performed on the term-document matrix constructed from a result set rather than from the entire collection. The effect of this is to emphasize the

[2] For clarity, we consistently use "segment" to refer to an automatically partitioned span of text, and "passage" to refer to any arbitrary span of text that is designated by the user.

[3] http://www.lemurproject.org/indri (accessed Apr 9, 2013).

effect of differences within the result set (which we would expect to preserve differences resulting from different aspects of a topic) while suppressing the effect of patterns of term usage which are shared by most or all elements in the result set (which tends to reduce the influence of terms used in the query). This technique has previously been used to emphasize differences among top-ranked documents that might better distinguish documents which are topically relevant from those which are not (Schütze, et al, 1995), but we are not aware of any prior use of the technique for topic aspect classification. In our experiments, we compute Local LSA on the term-segment matrix, since our result set contains segments rather than documents. We build two reduced-dimensional spaces, one for English as the evaluation language and one for Chinese as the evaluation language.

Dumais (1991) has shown that improved effectiveness can result when the elements of the term-document matrix are term weights rather than the raw term counts used by Deerwester et al. (1990). We have therefore chosen the widely used Okapi BM25 term weighting function (Spärck Jones, et al., 1998; Olsson, 2006), which has been shown to be robust and to achieve retrieval effectiveness that is on par with other known techniques.

A previous study of the relationship between the number of retained dimensions and retrieval effectiveness (as measured by mean average precision) for the Cranfield collection of 1,398 aerospace abstracts indicated that retaining 100 dimensions yielded good results (Oard, 1996). Both the number of abstracts and the length of the abstracts in that experiment were close to our number of segments and our typical segment length, so we had decided to set *q=100* for our earlier experiments. A sensitivity analysis, reported below, indicates that this was a reasonable choice.

Finally, the vector representations of the training segments in a language are used to train an aspect classifier for that language. Since a topic can (and, in our experiments, will) have multiple aspects, our classification problem is naturally modeled as an *m*-way multiple-class problem. Early experiments with a linear kernel SVM yielded disappointing results, perhaps because the small number of training instances is not sufficient to learn an appropriate separating hyperplane in the reduced dimensional space.
Instance-based classification techniques such as the k-Nearest-Neighbor (kNN) classifier are well suited to multi-class problems, and a kNN classifer yielded better results in those early experiments, so we focused exclusively on three kNN classifer variants that we explored.

The classic kNN algorithm is quite simple: to classify a segment, consult the *k* most similar training examples, where k is some integer, *k≥1*. Each of the *k* labeled neighbors "votes" for its aspect, and the aspect with the largest number of votes wins (Manning & Schütze, 2000).

We compute the similarity of two segments using the cosine (i.e., the normalized inner product) of the two Local LSA vectors:

$$sim(d_1, d_2) = \frac{\sum_{j=1}^{q}(T_{1j} * T_{2j})}{\sum_{j=1}^{q} T_{1j}^2 \sum_{j=1}^{q} T_{2j}^2}$$

where T_i is the right singular vector for segment d_i, T_{ij} is therefore the value of the j^{th} Local LSA feature for segment d_i, and q is the number of retained dimensions (and thus the rank of the Local LSA feature space). In earlier work it has been more common to use the unnormalized inner product rather than first normalizing the length of each vector to the unit hypersphere (as is effectively done in the cosine computation). However, our new experiment yields better results when using the normalized similarity in most cases.

Generally, larger values of *q* reduce the effect of noise on classification, but make boundaries between classes less distinct. The optimal *q* will vary depending on the difficulty of the classification problem and the amount of available training data. For a topic with *m* aspects, we set *q=2m+1* (always an odd integer, to minimize cases of ties). The only exception to this rule was when we conducted experiments with only 1 or 2 training instances for each aspect; in those cases, we set *q* to the number of training instances. For example, for a topic with 3 aspects, but only 2 training instances for each aspect, we would automatically set *q=2* rather than *q=7*. This heuristic approach might be improved upon (e.g., by learning optimal values for *q* using held-out data), but because we applied it consistently it serves as a useful basis for system comparisons.

In the classic kNN algorithm, every training instance in the top *k* gets an equal vote. This tends to make the results quite sensitive to the choice of *k* because too large a value for *k* will bring in many confounding examples. This effect can be minimized by using some form of weighted kNN. The idea of "distance weighted" kNN was originally introduced by Dudani (1976). Dudani's simplest proposed implementation, simply summing the similarity values for each training instance of a class among the *k* nearest neighbors, seems to work well in text classification applications (see, for example, Olsson & Oard (2007)). The category with the largest sum of similarity scores is assigned to the test instance. We refer to this as the *similarity-weighted* kNN algorithm.

An alternative approach to m-way classification proposed by Yang et al. (2000) is to build a suite of binary classifiers, one for each aspect of a topic, and then to select the classifier with the maximum margin. In what we refer to as the *maximum-margin* approach, one kNN classifier is built for each aspect to perform a binary classification problem in which each training instance is labeled either as a positive or negative instance for that aspect. The score (i.e., the margin) $r_a(d_1)$ assigned to segment d_1 for aspect a is

defined as the difference in the arithmetic means of the similarity values of the positive and negative examples:

$$r_a(d_1) = \frac{1}{k_p}\sum_{j=1}^{k_p} sim(d_1, dp_j) - \frac{1}{k_n}\sum_{j=1}^{k_n} sim(d_1, dn_j)$$

Where the k_p positive instances among the nearest neighbors are dp_j, the k_n negative instances are dn_j. This process is repeated for each aspect, and the aspect a with the highest score r_a is then assigned to segment d_1.

Cross-Language Aspect Classification

The key to cross-language aspect classification is to define a similarity measure that can match segments in one language with segments in another language. Because the similarity computation will need to be performed for every segment in the evaluation language, but only for the training instances in the other language, it is convenient to use the evaluation language as the language from which the LSA representation is built.[4] This leads naturally to a process in which each training segment is first represented using a term vector in the evaluation language, and then that term vector is re-represented in the LSA space. For clarity, we focus here on the case in which English as the evaluation language and Chinese as the other language, but the opposite direction is handled identically (with Chinese and English substituted for English and Chinese, respectively).

Although it is tempting to think of the process of developing an English term vector for a Chinese segment as "translation," simply replacing each term with its most likely translation is well known to often yield a suboptimal result (see, for example, Darwish & Oard (2003)). A better approach is to account for the uncertainty in the selection of appropriate translations by representing the translation function as a probability distribution. One way to do this would be to first compute term weights and then map those weights through the translation probability distribution; we call this approach *C-TrW* (for "cross-language with translated term weights," although "translated" is used loosely in this sense), an alternative approach that has been reported to be better is to map the term statistics (specifically, tf_{ij} and df_j) from Chinese into English (Wang & Oard, 2006). Essentially, this approach of mapping term statistics yields a maximum likelihood estimate of the term statistics that would have observed had the Chinese segment been instead originally written in English. Okapi BM25 weights are then computed from the resulting mapped term statistics in the usual way. The segment length (*dl*) statistic does not require any mapping because relative segment lengths are (to a sufficiently close approximation) preserved by translation, and the effect of *dl* in the BM25 formula relies only on relative segment lengths. We call this approach *C-TrTD* (for "cross-language with translated term frequency and document frequency").

For our experiments, we used the same translation probability matrices as (Wang, 2005) and (Wang & Oard, 2006). These were automatically built from the Foreign Broadcast Information Service (FBIS) parallel corpus[5] using a word alignment procedure implemented in the freely available GIZA++ toolkit[6] (Och & Ney, 2003).

Once an English vector has been generated for the Chinese segment (by multiplying the tf_{ij} and df_j term statistic vectors by the translation probability matrix and then using the resulting values to compute Okapi BM25 weights for each term), we can then further map the resulting English vector of Okapi BM25 term weights into the LSA space using well known techniques (Deerwester, 1990). This is done by multiplying the English term vector by the t x q matrix of left singular vectors T_0.

The foregoing description is somewhat idealized because we actually built our term statistics vectors using our Indri index, and for English our Indri index actually contains stems as terms, whereas the translation probability tables built by Wang (2005) had used English words. Although we could have trained a translation model for English stems, we instead simply mapped the word statistics to stem statistics. An English stem could have resulted from multiple word forms, so we conflated the probabilities associated with translation of English words into probabilities for the corresponding stems.

Bilingual Aspect Classification

For bilingual aspect classification, we assume the existence of training examples in both the evaluation language and the other language. The simplest way to perform bilingual aspect classification would be to first create evaluation-language term vectors and then Local LSA vectors for the other-language training segments, and then to use all of the examples together without adjustments. This can be done following *C-TrW* or *C-TrTD*, in which case we call the resulting bilingual classification process *B-TrW* or *B-TrTD*.

Systematic translation errors might, however, result in systematic mispositioning of the other-language training segments in the Local LSA space, and thus suboptimal

[4] There has been a considerable amount of prior work on constructing LSA representations using topically paired training instances (e.g., (Landauer and Littman, 1990)), but that approach is not practical for Local LSI because we have no *a priori* way of pairing a substantial fraction of the segments in a result set with topically related segments in the other language.

[5] LDC catalog: LDC2003E14. http://projects.ldc.upenn.edu/TIDES/mt2003.html (accessed April 10, 2013).

[6] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html (accessed April 10, 2013).

classification results. For cross-language classification, we have little choice but to accept these errors because we have no independent evidence for what the correct positioning would be. For bilingual classification, by contrast, we can compare the Local LSA vectors for the evaluation-language segments and the other-language segments. One simple way of doing this would be to compute the differences in the centroids (i.e., the arithmetic means) of the two sets; we could then correct the position of each other-language Local LSA vector by that difference. This technique can be applied following either *C-TrW* or *C-TrTD*, in which case we call the resulting process *B-TrWΔC* or *B-TrTDΔC*, respectively.

Because we are working with only a small number of training samples, the observed difference in the centroid positions will result from some combination of systematic errors (e.g., from translation) and random errors (e.g., from sampling), so applying the full difference could be harmful if it were to bring in more noise (effects of random variation) than signal (measurable effects from systematic errors). If the direction of the correction vector is informative but the magnitude is infelicitously scaled, we might benefit more from partial than from full correction. We therefore experiment with no correction (*B-TrW* and *B-TrTD*), partial correction (sweeping a fixed scaling parameter, which we refer to as *B-TrWδC* and *B-TrTDδC*), and full correction (*B-TrWΔC* and *B-TrTDΔC*) in our new experiments.

Test Collection

We used the TDT3 collection (Graff et al., 1999) and the TDT4 collection (LDC, 2004) to develop our reusable test collections. We selected the three English news sources and the three Chinese news sources that contain the largest number of documents that had been marked as relevant (to some topic). This selection results in 33,388 Chinese documents and 37,083 English documents for the union of the TDT3 and TDT4 collections.

Two bilingual annotators were recruited to annotate a group of consecutive sentences as an aspect (or subtopic, facet). They annotated all 50 of the TDT3 and TDT4 topics for which at least 15 relevant documents were known to exist (in the TDT relevance judgments). They were asked to identify between two and five aspects for each topic (in addition to the optional "all others" category which was defined for each topic for use by annotators wishing to provide negative training instances), and to try to finish annotating each topic within 4 hours (to limit annotation costs). They were asked to focus exclusively on aspects that could be found in both English and Chinese documents. If an aspect appeared in only one language, they were asked not to annotate that aspect. We asked that they try to find at least 8 passages per aspect and to choose passages to be annotated for an aspect in as many different documents as possible. They were allowed to designate overlapping passages and to assign overlapped passages to different aspects. The annotation was performed in two phases. In the first phase, each person annotated 25 topics. In the second phase, each ("re-")annotator re-annotated every aspect for five topics that had been annotated by the other ("first") annotator.

The first annotators annotated a total of 176 bilingual aspects for the 50 topics, and we used the annotations from both annotators to build our test collections. The annotated passage-level annotations were mapped onto the automatically-generated segments that were used in our experiments. We created two test collections from the resulting segment-level annotations. In our first test collection (*Test Collection 1*), we retained aspects that had at least 5 annotated segments in one language and at least 4 annotated segments in the other language, and then deleted any topics for which this reduced the number of aspects below two. There were a total of 106 bilingual aspects for 36 topics that met this requirement. To simplify our experiments, we also deleted any segment that was not in the set of retrieved segments from which the Local LSA space was constructed.[7] Finally, *Test Collection 1* includes a total of 92 bilingual aspects for 33 topics, 3 of which could only be used with English as the evaluation language (because those 3 aspects had each been assigned to only 4 Chinese segments).

Our second test collection was designed to support ablation studies with as many as 6 training segments per aspect. For this test collection (*Test Collection 2*), we therefore retained only aspects for which at least 7 aspects in one language and at least 6 aspects in the other language were available and then deleted any topics for which this reduced the number of aspects below two. *Test Collection 2* includes a total of 40 aspects for 17 topics, all 40 of which can be used with either evaluation language. The full set of 50 topics, a list of which topics are in each of the two test collections, and more details are provided in (Wu, 2008) and (Wu & Oard, 2008).

We performed an inter-annotator agreement study on a third test collection built in the same way as the others, but in which no minimum number of segments per aspect was enforced and in which overlapping passages were allowed.[8] This test collection (*Test Collection 3*, used only for computing inter-annotator agreement) included 36 aspects for 10 topics. The unit on which agreement was assessed

[7] These segments were deleted simply for convenience; we could have kept them by folding them into the LSA space in the same manner as the evaluation-language term vectors that had been automatically constructed from other-language segments.

[8] Removing overlapping passages was not necessary in this case because in reality a sentence is not an atomic unit and it might contain clauses that are properly assigned to two or more aspects.

was an automatically generated segment that was mapped onto the annotated passages. The average value (across all aspects) of Cohen's kappa was 0.57 for Chinese and 0.29 for English. We expect that the agreement for Chinese was higher because Chinese was the native language of our annotators. Because the annotators had chosen which topics would be re-annotated, there may be some risk that these results are somewhat higher than would have been the case had random selection been used.

Evaluation Metric

We chose precision, recall, and the *F* measure to compare the effectiveness of our aspect classifiers. Precision measures the fraction of the segments that are assigned to an aspect that are correctly assigned. Recall measures the fraction of the segments that should be assigned to an aspect that actually were assigned to that aspect by the classifier. Both precision and recall are clearly important in our intended application (low precision would adversely affect correctness; low recall would adversely affect comprehensiveness), so we want a measure that rewards both. The harmonic mean of recall and precision (the F measure) is a natural way to produce a single-valued effectiveness measure (van Rijsbergen, 1973). As a mean, the value of *F* for any single aspect will always be between its precision and recall values. The F measure is typically parameterized as F_β, where β specifies the ratio between precision and recall at which F is maximized. In this paper, we report $F_{\beta=1}$. Because we are interested not in the particular aspects in our test collection, but rather in the effectiveness of our classifiers on future (as yet unseen) aspects, we use the arithmetic mean of $F_{\beta=1}$ over the aspects as our primary measure of effectiveness for a classifier design; this way of aggregating results is referred to as *macro-averaging*. For brevity, we consistently refer to macro-averaged $F_{\beta=1}$ simply as F_1 . Because we have a particular interest in comparing systems, we must pay attention to whether differences in the arithmetic mean are likely the result of real differences between the systems or are likely to have resulted from the chance effects in our sampling. For F_1, we report the results from a two-tailed paired-sample t-test as *statistically significant* when $p<0.05$.

Previous Findings

We performed two sets of segment classification experiments. Our first set of experiments was designed to compare our three kNN classifier designs in combination with different ways of exploiting other-language training examples and different values for some key parameters. For the second set of experiments, we used one of the best configurations from those first experiments as a basis for an ablation study to investigate the effects of varying the number of evaluation-language and other-language training examples.

Our first set of experiments was designed to identify effective ways of using other-language training segments. Test collection 1 was used for these experiments. The segments for each aspect were partitioned into training and test sets using cross-validation.

We performed our experiments by trying all three kNN classifiers (voting, similarity-weighted, and maximum margin) with four classification techniques defined above (*B-TrW*, *B-TrWΔC*, *B-TrTD* and *B-TrTDΔC*) and a monolingual classification baseline (*M*) in which only evaluation-language training segments were used. Note that the *M* condition used half as many training examples as the other four (bilingual) conditions. Our previous experiments found that both *B-TrWΔC* and *B-TrTDΔC* consistently yielded lower mean precision, recall, and F_1. *B-TrW* and *B-TrTD* consistently improved classification effectiveness over the baseline (M), indicating that other-language training examples were useful. *B-TrTD* was better than *B-TrW*, confirming that translating TF and DF vectors then computing Okapi term weights was better than translating a vector of pre-computed term weights. *Similarity-Weighted B-TrTD* outperformed an unweighted contrastive condition, and was therefore selected for use in our second experiment.

In the second experiment, we used 1-6 segments for training and the reminder for test. Test Collection 2 was used, in which each aspect has at least 7 segments in both languages. Our previous experiments found that other-language training examples were useful, and in particular that when equal numbers of other-language and evaluation-language training examples were used classification effectiveness usually increased. When adding other-language training examples to a fixed number of same-language training examples, however, a point of diminishing returns was reached. More details of the previous findings can be found in (Wu & Oard, 2008).

RESEARCH QUESTIONS

Our previous findings were constrained by specific parameter settings, some of which were set heuristically. We hope those findings are not sensitive to those particular parameters, therefore some important research questions remain to be answered. The first issue to be examined is the cosine similarity measure, which is the normalized inner product of the two local LSA vectors. In earlier work it has been more common to use the unnormalized inner product after dimensionality reduction rather than renormalizing using the cosine measure. The similarity measures are critical to the classification algorithm, so our research question here is: is cosine normalization better than using an unnormalized inner product?

The second issue to be examined is the number of LSA dimensions retained in the reduced-dimensional space. In our earlier experiments we took 100 dimensions

	Voting			Similarity-Weighted			Maximum-Margin		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Monolingual	0.506	0.551	0.495	0.536	0.576	0.523	0.552	0.592	0.536
B-TrW	0.562	0.593	0.536	0.596	0.637	0.582	0.576	0.624	0.563
B-TrΔW	0.511	0.511	0.464	0.501	0.506	0.451	0.491	0.492	0.442
B-TrTD	0.572	0.595	**0.539**	0.614	0.647	**0.590**	0.589	0.638	**0.576**
B-TrTDΔC	0.501	0.521	0.469	0.507	0.525	0.473	0.516	0.530	0.477

Table 1: English as evaluation language: 4 English and 4 Chinese training examples; arithmetic mean over 92 aspects from 33 topics; bold indicates best F1. **Normalized cosine similarity.**

	Voting			Similarity-Weighted			Maximum-Margin		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Monolingual	0.511	0.552	**0.494**	0.581	0.601	**0.556**	0.559	0.603	**0.544**
B-TrW	0.517	0.548	0.490	0.554	0.562	0.511	0.584	0.593	0.534
B-TrΔW	0.500	0.517	0.467	0.511	0.531	0.471	0.464	0.496	0.438
B-TrTD	0.521	0.525	0.478	0.558	0.562	0.513	0.544	0.559	0.501
B-TrTDΔC	0.499	0.521	0.463	0.516	0.535	0.475	0.491	0.507	0.444

Table 2: English as evaluation language: 4 English and 4 Chinese training examples; arithmetic mean over 92 aspects from 33 topics; bold indicates best F1. **Unnormalized cosine similarity.**

heuristically. The number of dimensions retained affects the number of concepts that are used to represent the segments, and directly affects the vector space where the segments are represented and their positions in the vector space. Since the correct choice of dimensionality is important to success (Landauer and Dumais, 1997), a sensitivity analysis should be done to validate that choice.

In an effort to partially mitigate systematic translation errors, in earlier experiments we moved the other-language training examples toward the evaluation-language training examples in the LSA vector space until their centroids met. However, the full correction (*B-TrWΔC* and *B-TrTDΔC*) had not proven to be effective, so our third research question is: what about partial correction?

EXPERIMENTS

We use the same experimental framework described above, but with one important difference. In our earlier work, when translating an English stem to Chinese, we had incorrectly normalized translation probabilities after conflating the probabilities of English words; that normalization is appropriate only when translating from Chinese to English (Wang, 2006).

Test Collection 1 Experiment: Examining Unnormalized Cosine Similarity

We ran the same first set of experiments as described above, but for both normalized and unnormalized cosine similarity. Tables 1 and 2 show the classification effectiveness of three kNN classifiers with five classification techniques, using English as the evaluation language. The general pattern is that for most cases normalized cosine similarity is better than unnormalized cosine similarity (with two exceptions: *Similarity-Weighted M* with normalized similarity is statistically significantly worse than that with unnormalized similarity, and *Maximum-Margin M* with normalized cosine similarity is numerically worse than that with unnormalized similarity, but that difference is not statistically significant. When using Chinese as evaluation language, in all cases normalized cosine similarity is either numerically or statistically significantly better than unnormalized cosine similarity. We are also able to replicate our earlier results in the current experiment settings, so we can take *Similarity-Weighted B-TrTD* as the best technique for the remaining experiments.

Exploring the Number of Retained LSA Dimensions

The results in Tables 1 and 2 are informative only to the extent that we used a reasonable value of *q* (the number of retained dimensions) when building the Local LSA space. Working with the Cranfield collection of 1,398 aerospace abstracts, Oard (1996) illustrated that the optimal number of LSA dimensions was associated with the departure from linear decay of the singular values on a log-log plot. To get a sense for whether this effect is evident in our setting, we ran Local LSA on the top 300 English segments for each query in Test Collection 1 and generated a log-log plot for each topic. Figure 2 shows one example, in which the divergence from linearity begins around *q*=168.

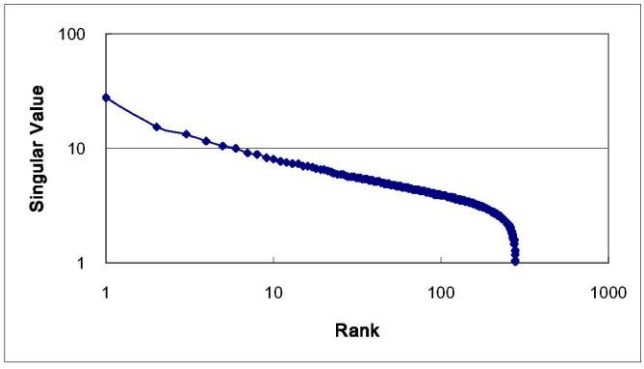

Figure 2. Singular values for the top 300 English document segments for Topic 30001

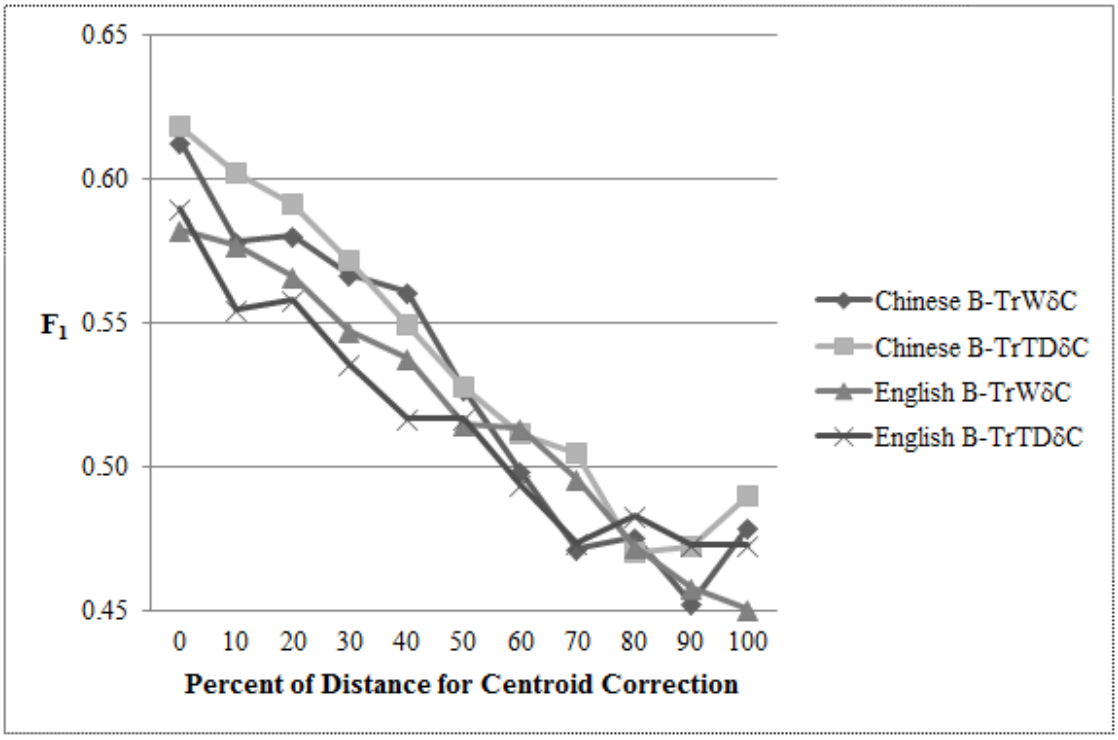

Figure 3. Effect of partial centroid moving of other-language training examples on classification effectiveness.

Of course, with 33 such plots, it is not clear how best to select a single value for *q*. Moreover, the SVD implementation that we used (SVDPACKC) limited us to about 335 dimensions (because of memory limitations), which was not sufficient to see the divergence from linearity for some topics. Since this analysis suggested that a larger number of dimensions might be useful, we reran our *Similarity-Weighted B-TrTD* condition with 150 and 250 dimensions; Tables 3 and 4 show the results. Although a modest numerical improvement in F_1 is evident when moving from 100 to 150 dimensions for each evaluation language, that difference is not statistically significant in either case. We therefore conclude that 100 dimensions was a reasonable choice for a topic-independent value of *q* in our experiments (and that choosing 150 dimensions would also have been reasonable). We leave the question of whether further improvements might be obtained from some way of setting topic-specific values of *q* for future work.

q	*P*	*R*	F_1
100	0.614	0.647	0.590
150	0.630	0.646	**0.597**
250	0.585	0.632	0.574

Table 3. Effect of varying the number of Local LSA dimensions (*q*) on classification effectiveness: English evaluation language, similarity-weighted *B-TrTD*, Test Collection 1.

q	*P*	*R*	F_1
100	0.647	0.645	0.618
150	0.681	0.670	**0.647**
250	0.647	0.653	0.617

Table 4. Effect of varying the number of Local LSA dimensions (*q*) on classification effectiveness: Chinese evaluation language, similarity-weighted *B-TrTD*, Test Collection 1.

Partial Position Correction for Other-Language Training

Because full correction (*B-TrWΔC* and *B-TrTDΔC*) had not proven to be effective (see Tables 1 and 2), we tried sweeping through values for partial correction (B-TrWδC and B-TrTDδC) with step size 0.1, that is, moving the other-language training examples toward the evaluation-language training examples in the LSA space built with the evaluation-language training examples by 10% at a step. Figure 3 shows the effect on classification effectiveness using *Similarity-Weighted kNN*. As Figure 3 shows, partial correction generally results in lower F_1 values than no correction (i.e., B-TrW and B-TrTD, plotted at the far left), and increasing the amount of partial correction generally results in greater degradation. This indicates correcting the positions of the other-language training segments in the Local LSA space brings in more noise (effects from random

variation) than signal (measurable effects from systematic translation errors).

This first set of experiments led us to conclude that *Similarity-Weighted B-TrTD* with *q=100* dimensions was a reasonable basis for ablation studies, so we consistently used that configuration for our experiments with Test Collection 2.

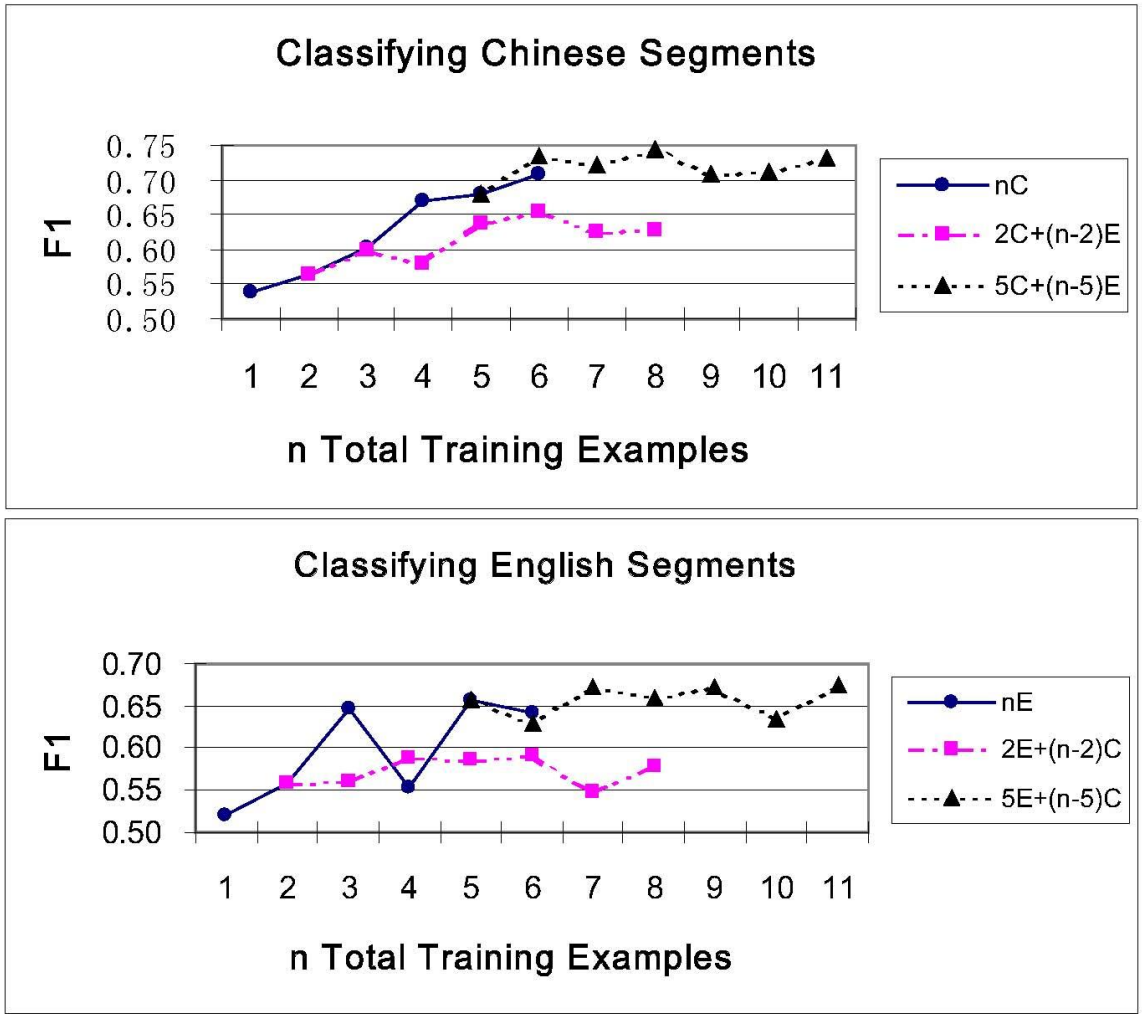

Figure 4. Effect of adding other language training segments to some fixed number of evaluation-language training instances. Top graph: evaluation language Chinese, bottom graph: Evaluation language English. C=Chinese, E-English, *similarity-weighted B-TrTD*, Test Collection 2.

Test Collection 2 Experiments: Ablation Studies

A set of ablation studies was designed to explore the effect of varying the number evaluation-language and other-language training examples. Test Collection 2, with at least 7 annotated segments in each language for each of 40 aspects, was used. As in our experiments with Test Collection 1, we performed at most 70 rounds of cross-validation.

Figure 4 helps to answer the question of whether evaluation-language training examples are more useful than other-language training examples. Each line in the upper plot connects cases in which some fixed number (e.g., two, five) of Chinese training examples are used, with the total number of training examples shown on the horizontal axis. For example, the line labeled with squares starts at F_1=0.566 for the 2C condition, then increases to F_1=0.601 for the 2C+1E condition, and F_1=0.579 for the 2C+2E condition, and then stays in that range as more English training examples are added (i.e., 2C+2E, 2C+3E, …). They generally show that a point of diminishing returns is reached beyond which fluctuations appear random.

CONCLUSION AND FUTURE WORK

From prior work, we know that same-language classification generally yields better results than cross-language classification. In our experiments we have viewed this as a continuum that we call bilingual

classification, with results for various combinations of evaluation-language and other-language training examples. When evaluation instances in both languages must be classified, we have shown that it can be useful to have some annotated training examples in each language, and to use the training examples from both languages to train the classifiers for each of the evaluation languages. Although we have only shown this in one setting (English and Chinese news stories, event-oriented document-scale topical relevance, segment-scale topic aspect classification, kNN classifiers, cosine similarity, probability translation of term vectors, Local LSA, F_1 measure), our techniques are broadly applicable to other settings, and they have been shown (in work by others) to be relatively robust. We therefore believe that this result should be of interest to anyone building example-based classifiers for more than one language. We have also augmented the existing TDT3 and TDT4 test collections with aspect annotations for English and Chinese in ways that other researchers may find useful.

We used our experimental framework to answer three research questions: whether the normalized cosine similarity measure is better than the more common unnormalized cosine similarity measure (yes), whether the number of retained LSA dimensions (which was heuristically chosen) is appropriate (yes), and whether partial corrections of the mapping of the translated other-language training examples into the evaluation-language LSA space can yield an improvement over no correction (no).

Our results also suggest several potentially productive directions for future research. Our inter-annotator agreement results for English are somewhat disappointing, so further work on test collection development is certainly called for. The exploratory analysis of a heuristic for selecting a suitable value for the number of dimensions to retain seems particularly well matched to the topic-specific nature of Local LSA, and definitely merits further investigation using a computing environment that can support larger SVD computations. Topic-specific rank cutoffs for defining the document space from which the Local LSA is computed might also be explored. And, of course, we will ultimately want to apply the classifiers that we have built to the aspect-specific sentiment analysis task that originally motivated this work. Although there is now a substantial body of work on cross-language text classification and related topics (most notably, cross-language information retrieval), we are only beginning to explore issues like this one addressed in this paper that arise when integrating those technologies into more comprehensive applications.

ACKNOWLEDGEMENTS

Thanks to Jianqiang Wang for providing us with the bi-directional English-Chinese translation probability tables, and Gina-Anne Levow for proving us with the Chinese stopword list. This work has been supported in part by DARPA contract HR-0011-06-2-0001 and NSF award DHB-0729459.

REFERENCES

Bel, N., Koster, C., & Villegas, M. (2003). Cross-lingual text categorization. European Conference on Digital Libraries (ECDL), 18(11), 613-620.

Darwish, K. & Oard, D. (2003). Probabilistic Structured Query Methods, Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, July 2003, Toronto, Canada.

Deerwester, S. et al. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science (JASIS), 41(6), 391-407.

Dudani, S. (1976). The Distance-Weighted k-Nearest-Neighbor Rule, IEEE Transactions on Systems, Man, and Cybernetics, 6(4)325-327.

Dumais, S. (1991). Improving the Retrieval of Information from External Sources, Behavior Research Methods, Instruments, and Computers, 23(2)229-236.

Gliozzo, A. & Strapparava, C. (2006). Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, July 2006, Sydney. 553-560.

Graff, D., Cieri, C., Strassel, S. and Martei, N. (1999). The TDT-3 Text and Speech Corpus, Broadcast News Workshop '99 Proceedings, February 1999, Herndon, Virginia, pp. 57-60.

Hearst, M. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1), 33-64.

Hull, D. (1994). Information Retrieval Using Statistical Classification. Ph.D. dissertation, Stanford University.

Landauer, T. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T. & Littman, M. (1990). Fully Automatic Cross-Language Retrieval Using Latent Semantic Indexing, Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, October, 1990, Waterloo, Ontario, Canada, pp. 31-38.

LDC (2004). Annotation Guide -- TDT3: 2003 Evaluation. Linguistic Data Consortium, http://projects.ldc.upenn.edu/TDT4/Annotation/.

Manning, C. & Schütze, H. (2000). Foundations of Statistical Natural Language Processing, MIT Press.

Ni, X., Sun, J., Hu, J., & Chen, Z. (2011). Cross lingual text classification by mining multilingual topics from wikipedia. The 4th ACM International Conference on Web Search and Data Mining. February 2011. Hong Kong, China. 375-383.

NIST (2000). The Year 2000 Topic Detection and Tracking Task Definition and Evaluation Plan, version 1.4, National Institute of Standards and Technology, pp. 7-8. ftp://jaguar.ncsl.nist.gov//tdt/tdt2000/evalplans/TDT00.Eval.Plan.v1.4.doc

Oard, D. (1996). Adaptive vector space text filtering for same-language and cross-language applications. Ph.D. Dissertation, University of Maryland, College Park.

Och, F. & Ney, H. (2000). Improved statistical alignment models. The 38th Annual Meeting of the Association for Computational Linguistics. October 2000, Hong Kong, 440-447.

Olsson, S. (2006). An analysis of the coupling between training set and neighborhood sizes for the kNN classifier. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. August 2006, Seattle, Washington. 685-686.

Olsson, S. & Oard, D. (2007). Improving text classification for oral history archives with temporal domain knowledge. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. July 2007, Amsterdam, The Netherlands, 623-630.

Olsson, S, Oard, D., & Hajic, J. (2005). Cross-language text classification. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. August 2005, Salvador, Bahia, Brazil. 645-646.

Prettenhofer, P. & Benno Stein, B. (2010). Cross-language text classification using structural correspondence learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 1118-1127.

Rigutini, L., Maggini, M., & Liu, B. (2005). An EM based training algorithm for cross-language text classification. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), Compiegne, France. 529-535.

Schütze, H., Hull, D. and Pedersen, J. (1995). A Comparison of Classifiers and Document Representations for the Routing Problem, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 1995, Philadelphia, Pennsylvania, USA, 229-237.

Shi, L., Mihalcea, R., & Tian, M. (2010). Cross language text classification by model translation and semi-supervised learning. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Massachusetts. USA, October 2010. 1057-1067.

Spärck Jones, K., Walker S., and Robertson, S. (2000). A probabilistic model of information retrieval: Development and comparative experiments. Information Processing & Management, 36 (6): 779-840.

Van Rijsbergen, C. (1979), Information Retrieval, 2nd edition, Butterworths, London.

Wang, J. (2005). Matching Meaning for Cross-Language Information Retrieval. Ph.D. thesis, University of Maryland, College Park.

Wang, J. & Oard, D. (2006). Combining bidirectional translation and synonym forcross-language information retrieval. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. August 2006, Seattle, Washington, USA, 202-209.

Yang, Y., Ault, T., et al. (2000). Improving text categorization methods for event tracking. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. July 2000. Athens, Greece, 65-72.

Wu, Y. (2008). Classifying Attitude by Topic Aspect for Chinese and English Document Collections. Ph.D. dissertation, University of Maryland, College Park.

Wu, Y. & Oard, W. (2008). Bilingual topic aspect classification with a few training examples. Proceedings of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, 2008, Singapore, 203-210.

Wu, Y. & Oard, W. (2009). Beyond topicality: finding opinionated Chinese documents. Proceedings of the 2009 Annual Meeting of the Association of the American Society for Information Science & Technology (ASIS&T), November 6-11, 2009, Vancouver, British Columbia, Canada.