Cross-Language Query Classification using Web Search for Exogenous Knowledge

Xuerui Wang*
Univ. of Massachusetts
140 Governors Drive
Amherst, MA 01003
xuerui@cs.umass.edu

Andrei Broder Evgeniy Gabrilovich Vanja Josifovski Bo Pang
Yahoo! Research
2821 Mission College Blvd.
Santa Clara, CA 95054
{broder, gabr, vanjaj, bopang}@yahoo-inc.com

ABSTRACT

The non-English Web is growing at phenomenal speed, but available language processing tools and resources are predominantly English-based. Taxonomies are a case in point: while there are plenty of commercial and non-commercial taxonomies for the English Web, taxonomies for other languages are either not available or of arguable quality. Given that building comprehensive taxonomies for each language is prohibitively expensive, it is natural to ask whether existing English taxonomies can be leveraged, possibly via machine translation, to enable text processing tasks in other languages. Our experimental results confirm that the answer is affirmative with respect to at least one task. In this study we focus on query classification, which is essential for understanding the user intent both in Web search and in online advertising. We propose a robust method for classifying non-English queries into an English taxonomy, using an existing English text classifier and off-the-shelf machine translation systems. In particular, we show that by considering the Web search results in the query's original language as additional sources of information, we can alleviate the effect of erroneous machine translation. Empirical evaluation on query sets in languages as diverse as Chinese and Russian yields very encouraging results; consequently, we believe that our approach is also applicable to many additional languages.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*

General Terms

Algorithms, Experimentation, Language, Measurement, Verification

*The research described herein was conducted while the first author was a summer intern at Yahoo! Research.

WSDM'09, February 9-11, 2009, Barcelona, Catalunya, Spain.

Keywords

Query classification, Web search, relevance feedback, machine translation

1. INTRODUCTION

Since its inception in 1992, the Web has grown at nearly exponential rate. Initially, most of the Web content was in English; however, as more users go online worldwide, the importance of the non-English part of the Web increases steadily. While English still dominates the Web, in 2002 as much as 43.6% of the Web content was in languages other than English, and the percentage of non-English queries submitted to Google was reported to have increased from 36% to 43% over a 6-month period.[1] Web usage in non-English-speaking countries has also exploded in the last decade. For example, according to *Internet World Stats*,[2] as of March 2008, China had about 210 million Internet users, second only to the United States that had 218 million.[3]

Despite the increasing importance of the non-English Web, significantly fewer tools and resources are available in other languages. Notably, developing taxonomies and annotating training examples for automatic classification is an extremely labor-intensive and expertise-intensive task. Even if high-quality inexpensive translation (either human or automatic) were easily available, one would need considerable expertise in both languages and cultures, as not every concept in the source language corresponds to exactly one concept in the target language, and vice versa. Instead of developing a new taxonomy for each language, the approach adopted in the *Open Directory Project*,[4] we propose to use resources already available in English to help process text in other languages.

One intuitive route to achieving this aim is to use automatic machine translation systems. While the field of machine translation (MT) has advanced significantly over the recent years, it is still not feasible to depend on MT systems to reliably translate training examples (let alone entire taxonomies) into the target language, owing to the less-than-perfect quality of MT output. Instead, we use MT systems to provide an admittedly *imperfect* mapping between English and other languages, and use MT output as an intermediate step that undergoes further processing. It is this

[1] http://www.netz-tipp.de/languages.html
[2] http://www.internetworldstats.com/top20.htm
[3] In July 2008, numerous news sources reported the number of Chinese Internet users to surpass that in the US.
[4] http://www.dmoz.org/

Table 1: An example of machine-translated text (Chinese to English) by Google Translate and its predicted classes from a text classifier trained in English.

Chinese text
视频 专辑 影视 会员 全网 上传视频	录制视频	制作相册	帮助 [电视剧] 电视剧龙游天下38b大结局 与朋友分享视频《电视剧龙游天下38b大结局》 《电视剧龙游天下38b大结局》的评论(共条) 对《电视剧龙游天下38b大结局》发表评论 下载更多时尚表情 当前视频标题： 电视剧龙游天下38b大结局 标签：电视剧龙游天下大结局38b 简介：电视剧龙游天下大结局38b * 发布者其他视频 更多..
Translated text by Google Translate
Television video album Member entire network Upload video | video recording | produced albums | Help TV Yongyu the outcome of the world 38 b Sharing video with friends, "the world drama Yongyu 38 b great outcome" "TV dramas Yongyu 38 b major outcome of the world," the commentary (total) "TV dramas Yongyu 38 b major outcome of the world," comments Download more fashionable expression Current Video Title: TV Yongyu the outcome of the world 38 b Tags: the outcome of world drama Yongyu 38 b Introduction: the outcome of world drama Yongyu 38 b * Publishers other video More ...
Top 3 predicted classes for the translated text
Class 1: Entertainment and Social Event Services/Television/TV Programs/Soap Opera TV Shows Class 2: Entertainment and Social Event Services/Television/TV Programs Class 3: Computing/Computer Software/Internet Software/Internet Downloads/Video Downloads

indirect use of machine translation systems that allows our system to more robustly tolerate occasional translation errors.

In this paper, we focus on query classification, where most of the previous work was conducted for the English Web. Query classification proved to be effective for better understanding query intent and improving user experience, as well as for boosting the relevance of online advertising [4, 5]. For instance, knowing that the query "TI-83" is about graphical calculators while "E248WFP" is about LCD monitors can obviously lead to more relevant ads even though no advertiser has specifically bid on these particular queries. In our previous work [5], we developed a commercial taxonomy for classifying English texts, which had approximately 6000 nodes and where each node was populated with hundreds of manually labeled examples. Translating such a taxonomy into numerous other languages and populating the translated taxonomy with labeled examples in the target language can be prohibitively labor-intensive. Instead, we propose a methodology for classifying non-English queries with respect to the original English taxonomy by using classifiers trained solely on English text.

A straightforward way to classify a non-English query is to directly machine translate the query into English, and use existing techniques for English query classification. However, while state-of-the-art machine translation tools work reasonably well on longer text fragments, they can be quite inaccurate on very short text such as typical Web queries. Consequently, inaccurate translation at this early stage, which can not be corrected via additional Web evidence on the English side, can be cascaded and wildly exaggerated, and cause subsequent classification to go completely astray. See Section 4 (Table 3) for anecdotal examples.

We propose a more robust method for classifying non-English queries. Instead of directly translating a query into English, we first submit the query in its native language to a search engine. We then collect top-scoring Web search results and use MT tools to translate the result pages into English. We classify the translated pages into the English taxonomy, and finally perform voting to determine the best overall class labels for the original query. It should be noted that state-of-the-art English-language query classification systems also use Web search results for more robust classification [5]. In the case of cross-language classification, however,

using English search results for the translated query is simply "too late". By using Web search results in the query's native language, in contrast to doing so on the English side, we effectively move the imperfect translation from high information density area (query) to low information density area (search results). Table 1 shows an example of text translated by a machine translation system from Chinese to English, along with the automatically predicted classes. The Chinese text is extracted from a Chinese Web page (top search result) found for the highlighted query; the English text was produced by Google Translate. Observe that although the query itself is not translated correctly, and most of the translated text is hardly human readable, a classifier operating on the entire translated text can still robustly predict the page's classes.

The contributions of this paper are two-fold. First, we develop a robust methodology for cross-language query classification using a mainstream Web search engine paired with off-the-shelf machine translation. Second, we present experimental results on Chinese and Russian query logs, and show that using our approach leads to significantly better classification accuracy. We also experimented with multiple different machine translation packages, and the results imply that as the quality of machine translation improves over time, so will the accuracy of query classification that uses machine translation.

2. RELATED WORK

Two areas of research are most closely related to our work: cross-language text classification (CLTC) and query classification (QC).

Recent years have seen increasing interest in cross-language text classification. Classification results have been reported for various language pairs: e.g., English-Italian [18], English-Czech [16], English-Spanish [13], English-Japanese [10], and English-Chinese [14]. This body of work typically falls into one of the two main approaches discussed in Bel et. al [2]: *poly-lingual training*, where a classifier is trained on labeled training documents in multiple languages, and *cross-lingual training*, where a classifier is trained in one source language, and documents in other languages are completely or selectively translated into the source language for classification. Our method in this paper bares more resemblance to the second approach.

Query classification can be considered as a special case of text classification in general, but it is in a sense much more difficult due to the brevity of queries. On the other hand, in many cases a human looking at a search query and its search results do remarkably well in making sense of it. Unfortunately, the sheer volume of search queries does not lend itself to human supervision. The state-of-the-art method [5] uses a blind relevance feedback technique: given a query, the class label is determined by classifying the Web search results retrieved for the query. Empirical evaluation confirms that this procedure yields a considerably higher classification accuracy than previous methods, particularly for rare queries.

In this paper, we approach the task of non-English query classification by taking advantage of advances in both cross-language classification and query classification. To the best of our knowledge, none of previously published work has addressed this important problem.

Another related research topic is that of cross-lingual information retrieval (CLIR), which deals with retrieving information written in a language different from the language in which the query is issued [12, 6, 8, 19, 9, 20]. Note that the goal of a CLIR system is very different from that of ours. A cross-lingual information retrieval system seeks to identify the most relevant documents in a language other than the query's original language, therefore when a MT system is involved, a good translation system is one that preserves the order of relevancy. In contrast, in our proposed approach for cross-language query classification, we assume top results in the query's native language to be mostly relevant, and only need the translation system to provide partially correct translations that retain the class of the original query. It is therefore not surprising that techniques that were shown to work well for CLIR may not necessarily be as effective for our task. For instance, in CLIR literature, a hybrid system that combined query translation and document translation was reported to outperform any non-hybrid systems for CLIR [15]; our experiments with a similar hybrid strategy, however, did not result in the best performing system for our task (see Section 4 for more detail).

There is also a connection between our approach and work in statistical machine learning that addresses domain adaptation or transfer learning [3, 7], which in turn has a connection to earlier work in multi-task learning [1][5]. If we consider different languages as different domains, our task can be viewed as one that seeks to adapt an English query classifier to non-English languages. This, at the surface level, mirrors work in domain adaptation that focuses on bridging the vocabulary differences in different domains, where one line of solutions focuses on developing a mapping that projects unknown tokens in the new domain with little resources to an observed token in the domain with ample training data. In contrast, we use the output of machine translation systems as a noisy projection provided to us, and focus on studying the best way of utilizing this projection and the effect of the quality of such a mapping in terms of the performance on the query classification task. We are certainly interested in further exploring this connection in our future work.

3. METHOD

We present a method for classifying non-English queries with respect to an English taxonomy with the help of external knowledge in the query's native language. Given a query, we first dispatch it to a Web search engine and retain a few dozen top-scoring search results. Then we translate these search results into English using a machine translation system. The translated results are subsequently classified using an existing classifier trained on English data. Finally, we perform voting among the predicted classes of individual search results to determine the class(es) for the original query.

Suppose a document written in the source (non-English) language, denoted as d_s, is translated into the target (English) language, denoted as d_t, by a machine translation system. Since the down-stream text classifier we plan to use is based on the bag-of-words representation of the document, we focus on the unigram precision of the translation system for simplicity, although the analysis should hold for n-gram

[5]Note that the usage of these terms are not always consistent within the machine learning literature. We refer to the collective body of work.

based classification in general. Note that unigram precision is an important component of BLEU score [17], a widely-used measure for automatic evaluation of machine translation systems. Let N be the total number of words in d_t, and l be the number of correctly translated words in d_t, we quantify the quality of the translation system by a quality factor $\alpha = l/N$. This notion is very similar to the unigram precision discussed in [17], where a unigram precision of 0.3 to 0.5 was reported for example machine translation systems on sample Chinese to English translations.

Again for simplicity, we consider a basic voting mechanism as our text classifier, where each word casts a vote for one of the classes and the class with the majority votes is predicted for the text document d_t. In addition, we assume there is only one correct class for each query, which can be relaxed if necessary, and all search results d_s preserve the class information of the query. While for most practical classifiers this assumption might be overly strong, we can approximate the imperfect classification with and an effective document length $N' < N$ to account for the fact that not all words cast a vote at all, and an effective quality factor $\alpha' < \alpha$ to account for the fact that a correctly translated word casts the right vote with (a non-trivial) probability $p < 1$ (in the following analysis, we assume $p = 1$ for simplicity). Note that the following analysis still holds for the adjusted α' and N'.

Let the number of classes in the taxonomy be K (again, for simplicity, we ignore the hierarchical structure in the taxonomy). For now, we also assume all correctly translated words cast one vote on the correct class c^*, and all incorrectly translated words cast a vote on one of the K classes uniformly at random. Thus, the correct class will receive a total of αN votes, and in order for d_t to receive the incorrect label, at least $\alpha N + 1$ out of the other $(1-\alpha)N$ votes need to aggregate over a class other than the ground truth c^*.

In this simplified setting, if $\alpha > 0.5$, it is impossible to classify the document incorrectly. If $\alpha \leq 0.5$, we consider the chance of at least $\alpha N+1$ of the random votes aggregating into one of the $K-1$ incorrect classes. Out of $K^{(1-\alpha)N}$ possible voting configurations, at most

$$(K-1)\binom{(1-\alpha)N}{\alpha N+1}K^{(1-2\alpha)N-1}$$

of them result in at least $\alpha N + 1$ votes in a class other than c^*. That is, the chance of d_t getting an incorrect label is bounded by

$$(K-1)\binom{(1-\alpha)N}{\alpha N+1}(\frac{1}{K})^{\alpha N+1}$$

.

Clearly, with a fixed N, the higher α is, the lower the chance of getting an incorrect class label induced by incorrect translation is. This also explains why the proposed method is better than classifying a translated query directly. First, as we mentioned earlier, translation of short queries directly are likely to be of lower quality since the MT system has less context information to resolve ambiguity. In addition, as queries are short, it is much more likely to have the entire query translated incorrectly, since K is typically quite high (over 6000 in our case), a completely irrelevant query in the target language is very unlikely to lead to the correct label by chance. But even if we assume multi-words queries

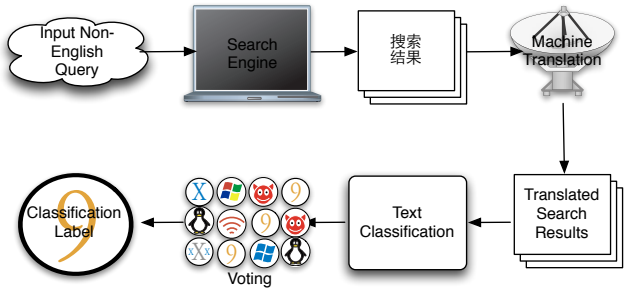

Figure 1: Robust classification of non-English queries with external Web evidence.

are partially correctly translated with the same translation quality, that is, the same α, as translating documents, the fact that queries are much shorter (i.e., much smaller N) leads to much higher chance of incorrect labels based on our model. This is in line with our intuition: if we have a query translated into three words in English, with one of the words being correct, then with high probability the two incorrectly translated words will vote for incorrect classes; on the other hand, if we have a 300-word document, 100 of which are correct translations, the chance of at least 100 of the random votes from the 200 incorrectly translated words aggregated into one class is significantly lower.

In what follows, we discuss each step in the overall procedure of our proposed method (as depicted in Figure 1) in more detail:

Web Search.

First, we dispatch a given non-English query to one or more major search engines to retrieve top search results in the query's native language.

In this study, queries are dispatched to Google to retrieve up to 32 search results (due to the limit imposed by Google AJAX Search API). The top search results are crawled from the Web using the returned URLs. When a fresh copy is not available, Google's cached page is retrieved with Google's cache header removed to ensure that these pages are comparable to the original pages.

All crawled Web pages are processed to remove all the tags, Java scripts, and other non-content information. If the returned results are not HTML files (e.g., PDF files, MS Word documents, etc.), they are simply removed from consideration. The resulting non-English textual content is re-encoded into UTF-8 regardless of what the original encoding was.

Machine Translation.

The crawled Web pages are translated into English via an off-the-shelf machine translation system. To study the impact of using different machine translation systems, we experiment with several different systems that are easily accessible over the Web:

- *Babelfish*[6] is powered by the technology of SYSTRAN, one of the oldest machine translation companies, that was known to be a rule-based MT system.[7]

[6]http://babelfish.yahoo.com

[7]http://www.systran.co.uk/translation/systran/corporate-

- *Google Translate*[8] is said to be based on statistical machine translation techniques[9] with which the system is trained on large-scale parallel corpora.
- *Dictionary-based Translation* is a very simple MT system based on a bilingual dictionary, where we translate a given document word by word, without considering the context in which the words appear. In the case of a language like Chinese where the input text needs to be first segmented into words, we segment the character sequences based on the entries in the bilingual dictionary. The translation quality should be clearly inferior to either of the above systems, in particular, it does not take advantage of the contextual information when translating documents. Still, there should be a non-trivial value of α associated with this system. As a practical question, it is interesting to see how well we can do for languages with no existing full-fledged MT systems; at the same time, it is also interesting to study whether our proposed system still improves over translating the queries directly even when the translation quality α is the same for short text snippets and documents with full sentences. (We expect both Babelfish and Google Translate to work better on normal texts than on short queries, i.e., with different α for query translation). We use the publicly available CEDICT dictionary[10] as our bilingual dictionary, the Chinese entries of which are also used to segment incoming Chinese text.

Note that as the public interface of both *Babelfish* and *Google Translate* impose limits on query length, we break long text into parts, send them in separately, and merge the translations afterwards.

Text Classification.

The translated pages are classified into an English taxonomy by a centroid-based classifier [11] trained on English data, which has been shown to be efficient and effective for large-scale experiments, and up to 5 ranked labels are returned for each page.

Label Voting.

Finally, we infer the query class from the page classes. More specifically, we take the majority vote from the page classes as the class label of the original query, with each translated page contributing up to 5 votes equally. Our choices here are motivated by our previous work on English query classification [5].

Compared to the baseline approach of direct query translation, our method has three advantages. First, by dispatching the original query to a search engine, we expand the query with exogenous knowledge that would not be available otherwise. In particular, while the query itself might be difficult to translate (e.g., the name of a popular Chinese TV series), the search results will likely contain additional pertinent keywords indicative of the correct class label that are easier to translate. Second, state-of-the-art machine translation systems are much better at translating long Web pages than short queries, thus considerably reducing the amount of erroneous translations introduced by the MT system. Even though the translated Web pages might not be easily readable by human readers, a machine-learned classifier can still reliably classify MT output [14], which is also demonstrated in Table 1. Finally, the voting mechanism further increases the robustness of our method as it alleviates the impacts of irrelevant search results or partially incorrect translations. The ranking of search results also gives us the flexibility to experiment with some weighted voting procedures in the future.

profile/translation-documentation/white-papers

[8]http://translate.google.com

[9]http://googleresearch.blogspot.com/2006/04/statistical-machine-translation-live.html

[10]http://www.mandarintools.com/cedict.html

4. EXPERIMENTAL RESULTS

We implemented our methodology described earlier for cross-language query classification in a software system called **Jasper**[11]. In what follows, we first describe the data sets used in this study, the baseline system, and the evaluation procedure; we then present the results of experimental evaluation of JASPER.

4.1 Data Sets

The volume of queries in today's search engines follows the familiar power law, where a few queries appear very often while most queries appear only a few times. In order to comprehensively evaluate our approach on queries of different frequency, we employ a stratified sampling procedure. To this end, we divide the query log into ten deciles by query frequency (in log scale), and randomly sample the same number of queries from each decile.

We sampled 1000 queries from a large-scale Chinese query log (100 queries per decile); we call this data set **C1000**. To conduct our pilot study, we also define a subset of it by taking 200 queries (20 per decile) from this large set, which we call **C200**. In order to assess the applicability of our method to another language, we also sampled a large-scale Russian query log, and selected 100 queries (10 per decile), which we call **R100**.

4.2 Baseline System

Recall that our hypothesis is JASPER will benefit from using Web search results in the query's native language, so that machine translation is applied to considerably longer input where culture-specific references are better resolved. To validate this hypothesis, we compare against a baseline system where machine translation is applied to queries directly. Our experiments show that directly classifying machine translated queries yields extremely poor results. As prior studies showed that using Web search results is beneficial in monolingual query classification [5], we further strengthen our baseline with Web evidence on the English side. More specifically, our baseline system first uses machine translation to translate the query into English, and then classifies the output (as if it were a regular English query) using English Web search results with the technique presented in [5]. By comparing JASPER with this baseline system, we address the following research question: is it indeed critical to use Web search results in the query's native language rather than search results for the translated query in English?

[11]Jasper is a semi-precious gem popular in the ancient world; its name can be traced back in Hebrew, Assyrian, Persian, Greek and Latin. (Source: `http://www.gemstone.org/gem-by-gem/english/jasper.html`)

Table 2: Average accuracy of Chinese query classification using different systems on C200.

Precision @ n		Jasper			Baseline			No search results
		GoogleTranslate	Babelfish	Dictionary	GoogleTranslate	Babelfish	Dictionary	
AND	1	$0.590^{\star}$	$0.545^{\star\diamond}$	$0.420^{\star}$	0.365	0.330	0.320	0.125
	2	$0.530^{\star+}$	$0.475^{\star\diamond}$	$0.385^{\star}$	0.335	0.297	0.273	0.108
	3	$0.483^{\star+}$	$0.428^{\star\diamond}$	$0.337^{\star}$	0.297^{+}	$0.268^{\diamond}$	0.222	0.083
	4	$0.429^{\star+}$	$0.381^{\star\diamond}$	$0.305^{\star}$	0.278	$0.254^{\diamond}$	0.195	0.074
	5	$0.388^{\star+}$	$0.359^{\star\diamond}$	$0.280^{\star}$	0.250^{+}	$0.221^{\diamond}$	0.174	0.065
OR	1	$0.715^{\star}$	$0.670^{\star\diamond}$	$0.520^{\star}$	0.485^{+}	0.425	0.420	0.180
	2	$0.680^{\star+}$	$0.615^{\star\diamond}$	$0.510^{\star}$	0.472^{+}	0.395	0.400	0.155
	3	$0.645^{\star+}$	$0.570^{\star\diamond}$	$0.465^{\star}$	0.443^{+}	0.370	0.352	0.122
	4	$0.595^{\star+}$	$0.531^{\star\diamond}$	$0.439^{\star}$	0.421^{+}	$0.357^{\diamond}$	0.319	0.109
	5	$0.566^{\star+}$	$0.513^{\star\diamond}$	$0.414^{\star}$	0.390^{+}	0.321	0.294	0.098

4.3 Evaluation Procedure

In order to assess the performance of the different query classification techniques, we had their output judged by native Chinese and Russian speakers, all of whom are native speakers in the respective language and are also highly proficient in English. Since human judgments are very expensive, we first conducted a pilot study with the smaller C200 data set, where each system produced up to 5 classes per query, and each query-class pair was judged by two Chinese speakers. We performed the rest of experiments using the entire C1000 data set and the R100 data set, where each system produced up to 3 classes per query, and each automatic classification was judged by a single native speaker. In all cases, human judgments were binary, where the predicted class was judged either *correct* (1) or *incorrect* (0). To avoid possible bias towards any particular approach, the results from different systems were mixed and presented for judgment in a randomized order.

For the pilot experiment with C200, we define the correctness of automatic classification in two ways, by combining the two human judgments using the logical AND (both editors consider the label as correct) or the logical OR (one of the judges considers the label as correct). The performance is then measured by the percentage of correct predictions among the top $n = 1, \ldots, 5$ predicted labels, that is, *Precision @ n*.[12] For the other two data sets, C1000 and R100, we report *Precision @ 3*.

We evaluate the performance of six different systems, which are three variants of our methodology (JASPER) and of the baseline. In each variant, JASPER or baseline is paired with a different translation system, namely, Google Translate, Babelfish, or CEDICT Dictionary.

4.4 Results

4.4.1 Performance on the C200 data set

Table 2 reports the performance of the different methods on the C200 data set. The upper part of the table reports the results of using logical AND to combine the editorial judgments, while the lower part of the table uses logical OR. We used one-tail paired t-test with p-value< 0.05 to assess the statistical significance of the results. The following superscripts are used to denote statistical significance. We first compare the performance of JASPER and the baseline using the same machine translation system: "$\star$" denotes that the performance of JASPER is significantly better than the corresponding performance of the baseline. We then consider the effect of using different MT systems for either JASPER or baseline: "+" represents that *GoogleTranslate* significantly edges out *Babelfish* and "$\diamond$" shows that *Babelfish* significantly outperforms *Dictionary*.

JASPER *vs. baseline.*

Several conclusions can be drawn from Table 2. First, with the same machine translation system, JASPER consistently and significantly outperforms the baseline (denoted by $\star$) across the board for any of the performance measures we considered. Note that this improvement holds regardless of the MT system used. In particular, it holds even when we use the simple dictionary-based MT system. This is interesting on two accounts: (a) as the unigram precision of the translation (α) gets lower, JASPER still has clear lead over the baseline; and (b) as predicted by our analysis in Section 3, even when the translation of the query has the same α as the translation of search result pages in the query's native language, which is very likely the case for our simple dictionary-based MT since it does not attempt to use any contextual information in the text, JASPER still clearly beats the baseline.

To further understand why it is critical to seek Web knowledge in the query's native language, we show some anecdotal results for two sample queries in Table 3. The first query is about a Chinese singer whose name contains two Chinese characters corresponding to "wheat" in English. Translating this name alone with any MT system gets erroneous results: all translations are wheat-related. Since the baseline uses Web search results on the English side, it runs completely astray due to the literal translation of "wheat", and produces food and health-related classes. Indeed, when the MT system incurs translation errors in the baseline system, additional search results on the English side can not compensate for these errors. In contrast, since JASPER uses search results in the original query language, it can robustly collect evi-

[12]Note that although we refer to this measure as "accuracy" in what follows, it is not accuracy in the traditional sense: a query may have only one or two correct labels, in which case even a perfect classifier is bounded by 40% accuracy when predicting top 5 classes. Still, this measure is often used in information retrieval literature and it is able to demonstrate the relative effectiveness of the different approaches under consideration.

Table 3: Automatic classification of two sample Chinese queries, which are translated into English using GoogleTranslate (G), Babelfish (B), and CEDICT Dictionary (D).

Query: 麦子杰 (A Chinese singer's name; the first two characters translate into "wheat" literally)	
JASPER: translating the search results of 麦子杰	· Computing/Computer Software/Internet Software/Internet Downloads · Computing/Computer Software/Internet Software/Internet Downloads/Audio Downloads · Computing/Computer Software/Internet Software/Internet Downloads/Audio Downloads/ Audio Downloads-Free
Baseline using the translated query: *Wheat* - (G) *The wheat is outstanding* (B) *wheat hero* (D)	 · Health and Beauty/Medical Conditions/Allergies/Allergy Treatments and Therapies/ Allergy and Immunology · Mass Merchants/Food · Computing/Computer Software/Home and Personal Planning Software/Recipe Software
Query: 跑跑卡丁车 (A popular kart racing video game)	
JASPER: translating the search results of 跑跑卡丁车	· Automotive/Parts and Accessories/Go-Kart Parts and Accessories · Toys and Hobbies/Toys/Games/Video Games · Toys and Hobbies/Toys/Games/Video Games/PC Video Games
Baseline using the translated query: Paopao Karting (G) Runs the kart (B) to run to run kart racing (D)	 · Automotive/Parts and Accessories/Go-Kart Parts and Accessories · Automotive/Powersport Vehicles/Go-Karts · Entertainment and Social Event Services/Events/Sports Events/Motor Sports

dence in the Chinese Web to substantiate the music-related interpretation of the query. As long as enough keywords are correctly translated, correct class labels can still be predicted accordingly.

The second query in Table 3 is about a very popular video game on kart racing. At first glance, machine translation of the query produces acceptable results. But by translating the query alone, all MT systems "overlook" the video game aspect of the query intent, which is absent from Web evidence available in English as it is a video game popular "locally" in China. This results in prediction of classes on the general topic of racing. While the Chinese query itself does not contain the phrase video game explicitly either, top search results in Chinese provide enough evidence to reflect the correct query intent, thus enabling the correct classifications that are related to video games. This is the other main reason why it is advantageous to seek web evidence in the query's native language: to capture "local" search intent, or resolve culture-specific jargons.

Impact of different MT systems.

The results in Table 2 also show that the quality of different machine translation systems impacts the performance of both JASPER and the baseline. In most scenarios, statistical *GoogleTranslate* is significantly (denoted by $+$) better than rule-based *Babelfish*, while *Babelfish* in turn significantly (denoted by $\diamond$) outperforms the simple *Dictionary*-based translation.

We note that the worst performance of JASPER (with *Dictionary*) is still better than the best baseline version. This supports the hypothesis that additional Web evidence should be sought in the original language of the query and that it is better to classify partially incorrectly translated documents rather than partially incorrectly translated short queries, even when the unigram precision α for the query translation is higher than that of document translation. Although using the dictionary-based MT system does not perform as well as using more advanced ones, the dictionary-based JASPER implementation still exhibits solid performance. This is a particularly encouraging result for rare languages that do not have full-fledged machine translation systems readily available. Paired with a bilingual dictionary, which is often available even for rare languages, JASPER can deliver decent query classification performance, which indicates the broad applicability of JASPER.

Other observations.

Finally, the rightmost column of Table 2 shows the result of query classification using MT alone, without the help of Web search results in either the original query language or in English. Predictably, although we use the best performing MT system, Google Translate, the results are consistently inferior, reinforcing the value of augmenting query classification with exogenous knowledge using Web search.

The above results show that the relative ordering of the examined methods is the same whether we use logical OR or logical AND to combine two editorial judgments, i.e., it is not affected by how strict the editorial judgment is. Therefore, as we are more interested in studying the relative performance of different systems, to reduce the amount of editorial work, we performed experiments with R100 and the rest of the C1000 data set using only one editorial judgment for each query-class pair. To further reduce the number of expensive human judgments, in the following experiments we also reduce the number of predicted classes to 3, as the relative ordering is not sensitive to how many classes are predicted, as is shown in Table 2.

4.4.2 Performance on additional data sets

Table 4 presents the results of query classification on C1000 and R100 data sets. Again, the superscripts denote statistical significance under one-tail paired t-test with p-value$<$ 0.05. For the same machine translation system, "$\star$" denotes that the average accuracy of JASPER is significantly better than the corresponding accuracy of the baseline. In either JASPER or baseline, "$+$" denotes that *GoogleTranslate* significantly outperforms *Babelfish*.

On the much larger C1000 data set, we again find that JASPER significantly outperforms the baseline with either one of the MT systems. The lead of Google Translate still

Table 4: Average precision@3 on C1000 and R100 data sets.

Data set	Jasper		Baseline	
	GoogleTranslate	Babelfish	GoogleTranslate	Babelfish
C1000	$0.535^{\star+}$	$0.465^{\star}$	0.332	0.297
R100	$0.613^{\star+}$	$0.543^{\star}$	0.530	0.417

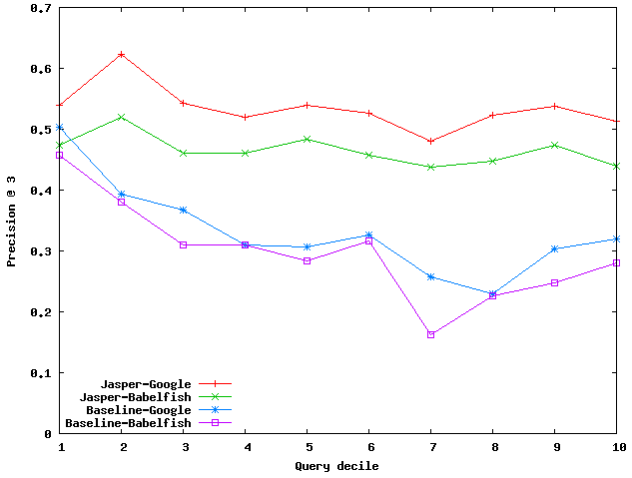

Figure 2: Performance by query deciles

holds for JASPER.

Our results also show that most of the conclusions drawn for Chinese are also valid for Russian. The performance numbers reported for Russian in Table 4 are much larger than those for Chinese. We speculate that this reflects the fact that Russian to English translation is easier because both languages belong to the same family of Indo-European languages, and are therefore closer to each other than English and Chinese. Consequently, better quality of page translation yields better query classification results.

4.4.3 *Stratified analysis of query frequency*

We also analyze the results according to query frequency deciles. For this experiment we used all 1000 queries in C1000 with 100 queries from each decile.

As can be seen in Figure 2 and Table 5, the performance of JASPER does not fluctuate very much across different frequency deciles, showing more robust performance than the baseline. As a result, for less frequent queries the performance gap between JASPER and the baseline becomes progressively larger, reflecting the increased difficulty of translating rare queries directly.

The detailed results presented in Table 5 show that JASPER significantly (denoted by $\star$) outperforms the baseline at all deciles, with the exception of the decile with highest frequency, where the difference is not statistically significant. We explain this by the fact that the most frequent queries are so common that they are simply easier for the MT system to translate correctly. Similarly to the C200 data set (Table 2), JASPER with *GoogleTranslate* significantly outperforms JASPER with *Babelfish* (denoted by $+$).

Furthermore, with the C1000 data set, we show that overall JASPER paired with *Babelfish* performs statistically significantly (denoted by $\diamond$) better than baseline paired with *GoogleTranslate*, providing more convincing evidence that JASPER outperforms the baseline even when it is paired with an MT system that is weaker for our task.

It is interesting to note that, the above conclusion holds for all deciles except the first one (= most frequent queries), where the baseline method with *GoogleTranslate* outperforms JASPER with *Babelfish*. With deeper analysis, we found that the most frequent queries can often be translated perfectly by *GoogleTranslate*, even when they are proper nouns not included in typical bi-lingual dictionaries. For example, 网易 (the name of a popular Chinese Internet portal Website[13]) is recognized by the Web-friendly *GoogleTranslate* as a proper noun "Netease", while the rule-based *Babelfish* translates it into "The net is easy". By collecting knowledge about "Netease" in English, most of which is high-quality information about the portal, the baseline method can assign accurate class labels to the query. On the other hand, the evidence collected in Chinese mostly consists of the *content* of the portal (such as news, blogs, and games), rather than the information *about* the portal itself, and is thus less useful for class predictions.

4.5 Further Analysis: What did not Work

4.5.1 *Combining* JASPER *with the baseline*

In cross-lingual information retrieval (CLIR) literature, a hybrid approach of translating both queries and documents was shown to be significantly superior to translating either queries or documents alone [15]. As we noted in Section 2, cross-language query classification is very different from CLIR in its main goal and what it requires from a MT system. We examine whether a similar hybrid approach of JASPER and the baseline can yield better results. From experiments conducted thus far, we clearly show that Web evidence in the native language of a query provides useful information for query classification. In the following experiment we assess if the evidence collected from English pages provide additional information that further help improving classification accuracy.

We construct a hybrid system of JASPER and the baseline by allowing search results from both the English and native language side to vote for the query classes. Based on our experiments with *GoogleTranslate* reported in Table 6, the performance of hybrid approach is right between that of JASPER and the baseline (precision@3). This suggests that on average incorporating English search results does not provide much additional information, and performs worse than JASPER as it brings down the average α.

4.5.2 *Combining different MT systems*

A combination of different machine translation systems could potentially be beneficial, especially when the underly-

[13]http://www.163.com

Table 5: Average precision@3 on C1000 for individual deciles of query volume.

Decile	Jasper		Baseline	
	GoogleTranslate	Babelfish	GoogleTranslate	Babelfish
1	0.540^{+}	0.473	0.503^{+}	0.457
2	$0.623^{\star+}$	$0.520^{\star\diamond}$	0.393	0.380
3	$0.543^{\star+}$	$0.460^{\star\diamond}$	0.367^{+}	0.310
4	$0.520^{\star+}$	$0.460^{\star\diamond}$	0.310	0.310
5	$0.540^{\star+}$	$0.483^{\star\diamond}$	0.307	0.283
6	$0.527^{\star+}$	$0.457^{\star\diamond}$	0.327	0.317
7	$0.480^{\star+}$	$0.437^{\star\diamond}$	0.257^{+}	0.163
8	$0.523^{\star+}$	$0.447^{\star\diamond}$	0.230	0.227
9	$0.537^{\star+}$	$0.473^{\star\diamond}$	0.303^{+}	0.247
10	$0.513^{\star+}$	$0.440^{\star\diamond}$	0.320	0.280
Overall	$0.535^{\star+}$	$0.465^{\star\diamond}$	0.332^{+}	0.297

Table 6: Average precision@3 on C200 for Jasper, baseline, and a combined voting scheme that uses search results on both the English side and the native language side (*GoogleTranslate* is used for all methods).

Judgment	Jasper	Baseline	Combined
Logical AND	0.483	0.297	0.457
Logical OR	0.645	0.443	0.618

Table 7: Average precision@3 on C200 for different combinations of voting of the machine translation systems for Jasper.

System	Average Accuracy
GoogleTranslate	0.533
Babelfish	0.475
Dictionary	0.372
GoogleTranslate+Babelfish	0.510
GoogleTranslate+Dictionary	0.475
Babelfish+Dictionary	0.440
All Three Combined	0.490

ing systems based on MT techniques as different as statistical MT (*GoogleTranslate*) and rule-based MT (*Babelfish*). If the different MT techniques excel on different input texts, combining them could potentially get "the best of both worlds". On the other hand, a simple combination mechanism might result in a system with middling α, in which case we should not expect it to outperform the better-performing system of the two.

We consider two options to combine different machine translation systems: a macro-combination where we combine the votes obtained from JASPER paired with each of the MT systems individually; or a micro-combination where we use a combined translation system with the results from different MT combined into one document for each Chinese result page. The experimental results on the C200 data set for these two options are reported in Table 7 and Table 8, respectively. Note that the numbers reported here are again with single judgment, and should be expected to be between the corresponding performances based on two sets of editorial judgments reported in the previous sections.

Table 8: Average precision@3 on C200 for different combinations of MT systems with Jasper.

System	Average Accuracy
GoogleTranslate	0.533
Babelfish	0.475
GoogleTranslate+Babelfish	0.487

As can be seen from the tables, we do not observe the benefit of combining translation results. The performance of a combined system is always between that of the two base systems, indicating the second conjecture is more likely to be the dominant factor here.

5. DISCUSSION

Query classification is the cornerstone of many Web applications such as search and online advertising. However, constructing taxonomies and collecting training data for every language requires substantial effort and can be extremely expensive. In this paper, we have presented a robust method for classifying non-English queries with respect to an English taxonomy. We submit a non-English query to a general purpose search engine, and retrieve the top search results in the query's native language. The search result pages are then translated into English automatically using publicly available MT systems, and the translated pages are subsequently classified using a classifier trained on English data. Finally, we determine the query class by performing voting on the individual classes of the translated pages.

Experimental results with queries sampled from Chinese and Russian query logs show that our method significantly outperforms a baseline method that directly translates the query and then uses Web knowledge in English. By employing blind relevance feedback in the query's native language we significantly reduce the impact of erroneous machine translation.

The experiments show that our method is very robust against query frequency fluctuation, and can be successfully used on a very diverse set of queries including tail queries. Tail queries are important since together they account for

a significant fraction of the query volume. However, taken individually they make one-off events for which it is virtually impossible to collect enough training statistics, hence it is important to develop query classification techniques that are robust enough to handle tail queries. The superior performance of our method on rare queries provides a substantial opportunity for applications such as cross-language Web search and online advertising.

Our experiments with dictionary-based translation show that when a full-fledged machine translation system is not available, as is the case for many rare languages, combining our approach with a simple dictionary lookup can still deliver decent performance.

In our future work we plan to examine additional ways to improve JASPER performance. We are experimenting with varying the number of Web search results obtained for each query, weighting different results according to the search rank and score, and using auxiliary information produced by machine translation systems.

6. ACKNOWLEDGMENTS

We thank Ravi Kumar, Donald Metzler, and Kishore Papineni for fruitful discussions and pointers. We thank Pavel Braslavski and his colleagues at Yandex for providing us with information about popular Russian queries.

7. REFERENCES

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] N. Bel, C. H. A. Koster, and M. Villegas. Cross-lingual text categorization. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*, pages 126–139, 2003.

[3] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006.

[4] A. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using Web relevance feedback. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008.

[5] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 231–238, 2007.

[6] C. Buckley, M. Mitra, J. Walz, and C. Cardie. Using clustering and superconcepts within smart: Trec 6. *Information Processing Management*, 36(1):109–131, 2000.

[7] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

[8] S. T. Dumais, T. K. Landauer, and M. L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Workshop on Cross-Linguistic Information Retrieval*, pages 16–23, 1996.

[9] F. C. Gey, N. Kando, and C. Peters. Cross-language information retrieval: the way ahead. *Information Processing and Management*, 41(3):415–431, 2005.

[10] A. Gliozzo and C. Strapparava. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 553–560, 2006.

[11] E.-H. Han and G. Karypis. Centroid-based document classification: Analysis and experimental results. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 424–431, 2000.

[12] K. Kishida. Technical issues of cross-language information retrieval: a review. *Information Processing and Management*, 41(3):433–455, 2005.

[13] Y. Li and J. Shawe-Taylor. Advanced learning algorithms for cross-language patent retrieval and classification. *Information Processing and Management*, 43(5):1183–1199, 2007.

[14] X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can chinese web pages be classified with english data source? In *Proceeding of the 17th international conference on World Wide Web*, pages 969–978, 2008.

[15] J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214, 1999.

[16] J. S. Olsson, D. W. Oard, and J. Hajič. Cross-language text classification. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–646, 2005.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2001.

[18] L. Rigutini, M. Maggini, and B. Liu. An EM based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 529–535, 2005.

[19] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual information retrieval: learning from bilingual corpora. *Artifical Intelligence*, 103(1-2):323–345, 1998.

[20] D. Zhou, M. Truran, T. Brailsford, and H. Ashman. A hybrid technique for english-chinese cross language information retrieval. *ACM Transactions on Asian Language Information Processing*, 7(2):1–35, 2008.