Cross-Lingual Query Classification: a Preliminary Study

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

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

ABSTRACT

The non-English Web is growing at breakneck speed, but available language processing tools are mostly 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 very limited quality. Given that building taxonomies in all non-English languages is prohibitively expensive, it is natural to ask whether existing English taxonomies can be leveraged, possibly via machine translation, to enable information processing tasks in other languages. Preliminary results presented in this paper indicate that the answer is affirmative with respect to query classification, a task which is essential both for understanding the user intent and thus providing better search results, and for better targeting of search-based advertising, the economic underpinning of commercial Web search engines. We propose a robust method for classifying non-English queries against an English taxonomy using widely available, off-the-shelf machine translation systems. In particular, we show that by viewing the search results in the query's original language as independent sources of information, we can alleviate the impact of poor quality or erroneous machine translations. Empirical results for Chinese queries show that we achieve remarkably encouraging results.

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, Measurement, Experimentation

1. INTRODUCTION

The Web has grown rapidly since its inception in 1992 at an approximately exponential growth 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 many 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] Furthermore, Web usage in non-English-speaking countries has 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.

Despite the increasing importance of the non-English Web, significantly fewer technical resources are available in those languages. Developing such technical resources for each language of interest to us can be an extremely labor-intensive and expertise-intensive task. It is therefore of great interest to apply resources already available in English to processing other languages instead of developing such resources anew for each language.

One natural direction to achieve 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 develop 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 non-English languages, and use MT output as an intermediate step that undergoes further processing. It is this indirect use of machine translation that allows our system to tolerate translation errors.

In this paper, we focus on query classification, where most of the previous work was conducted for the English Web. Query classification has proven to be effective for better understanding query intent and improving user experience, as well as for boosting the relevance of online advertising [2, 3]. For instance, knowing that the query "TI-83" is about graphical calculators while "E248WFP" is about LCD monitors can obviously lead to more focused advertisements even though no advertiser has specifically bid on these particular queries. A commercial English taxonomy of Web queries with approximately 6000 nodes where each node was populated with example queries has been developed in previous work [3]. Translating this taxonomy into each non-English language of interest to us and re-populating the translated

[*] This work was performed during the author's summer internship at Yahoo! Research in 2008.

iNEWS'08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558-253-5/08/10 ...$5.00.

[1] http://www.netz-tipp.de/languages.html

[2] http://www.internetworldstats.com/top20.htm

taxonomy with example queries in that non-English language can be very labor-intensive. Instead, we classify non-English queries with respect to the original English taxonomy by utilizing classifiers built for English text directly. The labels can be used to improve the algorithmic results for non-English Web search as well as the quality of advertisement placement in non-English languages.

A straight-forward way to classify a non-English query is to directly translate the query into English, and use existing techniques for English query classification. However, while 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 can cause the subsequent classification to go completely astray, which can no longer be corrected even with additional resource on the English side.

In this paper we propose a more robust method for classifying non-English queries. Instead of directly translating a query into English, we submit the query to a search engine, machine translate and classify the resulting pages, and then infer the query class from the page classes. We present preliminary experimental results on queries sampled from a Chinese query log. We show that significantly better classification accuracy can be obtained via our approach compared to directly translating queries.

2. RELATED WORK

Recently, there has been a surge of interest in cross-language text classification. Classification results over various language pairs have been reported, including, but not limited to, English-Italian [9], English-Czech [8], English-Spanish [6], English-Japanese [4], and English-Chinese [7]. Bel et. al [1] discuss two main approaches to cross-language text classification: *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 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. Observe, however, that in many cases a human looking at a search query and the search results does remarkably well in making sense of it. Unfortunately, the sheer volume of search queries does not lend itself to human supervision, and alternative sources of knowledge about the world are needed. The state-of-the-art method [3] 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.

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, and extremely difficult problem.

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 submit it to a Web search engine and retain a few top-scoring search results; we then 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. The overall procedure of our method is shown in Figure 1:

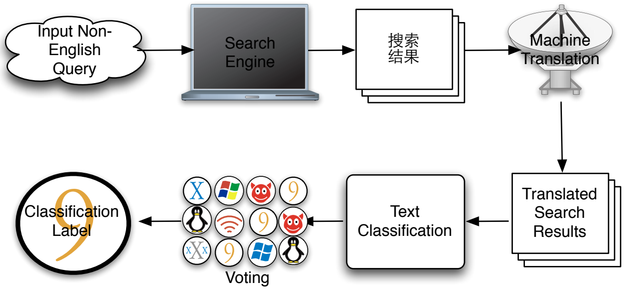

Figure 1: Robust classification of non-English queries.

Web Search.

First, we dispatch a given non-English query to one or more major search engines to retrieve top N 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 the original encoding.

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 MT systems, we experiment with two different systems that are easily accessible over the Web:

- *Babel Fish*[3] is powered by the technology of SYSTRAN, one of the oldest machine translation companies, that at least at one point was known to be a rule-based MT system.
- *Google Translate*[4] is said to be based on statistical machine translation techniques with which the system is trained on large-scale parallel corpora.

The public interfaces of both systems have limits on query length. We break long texts into parts, translate them separately, and merge the translations afterwards.

[3] http://babelfish.yahoo.com
[4] http://translate.google.com

Text Classification.

The translated pages are classified into an English taxonomy by a centroid-based classifier [5] trained on English data. This classifier has been shown to be efficient and effective for large-scale experiments. 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 top 5 classes with the most votes from the page class predictions as the most likely class labels of the original query, with each translated page contributing up to 5 votes with equal weights.

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 [7]. 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 weighted voting procedures.

4. EXPERIMENTAL RESULTS

In this section, we first describe our data set; we then evaluate our approach using two different machine translation systems and compare the results to the baseline approach that directly translates queries.

4.1 Data Set

We apply our method to 200 queries sampled from a large-scale Chinese query log. It is well known that the volume of queries in today's search engines roughly follows a power law. To make the 200 samples representative of the overall traffic, we divide the query log into ten deciles with respect to the logarithm of query frequency, and sample 20 queries uniformly from each decile. This way, we ensure both popular and rare queries are represented in our sample.

4.2 Evaluation Mechanism

Preliminary pilot studies show that directly classifying machine translated queries yields extremely poor results due to the poor quality of machine translation of short queries. To further strengthen the baseline, we employ the blind relevance feedback procedure by expanding the translated English query with search results in the English Web, similar to what was done in state-of-the-art English query classification systems [3]. Note that this enhanced baseline approach is quite powerful in itself with the help of external evidence from the Web. Thus, both the proposed approach and the baseline system take advantage of machine translation as well as the blind relevance feedback techniques. However, by doing Web search first in the original language, we significantly increase the robustness of our approach since machine translation errors are reduced or partially compensated for as more relevant content is available in the native language.

For each system under comparison, we take up to 5 predicted labels for each query. Since there is no existing ground truth of Chinese query labels, two native Chinese speakers were asked to make editorial judgments over each predicted label into *correct* (1) or *incorrect* (0). To remove possible bias towards any particular approach, all predicted labels from different systems are mixed and presented to the human editors in random order.

As different editors can have different interpretations of the original query intent, their judgments can slightly differ from time to time. We define the correctness of a prediction in two ways: the logical AND (both judges consider the label as correct) or the logical OR (one of the judges considers the label as correct) of the two judgments. For each query, the performance of a particular method is measured by the percentage of correct predictions among the top 5 predicted labels. Note that although we refer to this performance measure as accuracy in the text to follow, it is not accuracy per se: a query might have only two correct labels, in which case even a perfect classifier is bounded by 40% accuracy with this measure. Still, this measure demonstrates the relative effectiveness of different approaches under consideration, as more comprehensive comparisons using different metrics are not included due to space limit.

4.3 Results

We report performance of four different systems: the proposed method and the enhanced baseline method paired up with Google Translate or Babel Fish, denoted as *Method+ Google*, *Method+Babelfish*, *Baseline+Google*, and *Baseline+ Babelfish*, respectively.

The average accuracies over queries sampled from different deciles are shown in Table 1, where Decile 1 corresponds to the most frequent queries, and Decile 10 corresponds to the least frequent queries. Performances measured by using logical AND (OR) to combine editorial judgments are presented in the top (bottom) part of the table.

We first compare the performance of the proposed method against that of the corresponding baseline system using the same MT system, and mark the corresponding number with a superscript when our method significantly outperforms the baseline under one-tail paired t-test with p-value< 0.05: "+" for *Method+Google* vs. *Baseline+Google*; and "$\diamond$" for *Method+Babelfish* vs. *Baseline+Babelfish.* As shown in Table 1, regardless of which translation system is used, the proposed method outperforms the baseline approach most of the time. We conjecture that, with more queries, the performance difference will be much more significant. Note that for less frequent queries, the performance gap between our method and the baseline method becomes larger, which probably reflects the difficulty of translating rare queries. Given the queries are sampled from different volume deciles and is therefore somewhat representative for the overall traffic, we conjecture that this improvement will reasonably carry over to larger sets of sample queries. In the future, a more comprehensive larger-scale experiment will enable us to draw stronger conclusions.

Table 1: The average accuracy over queries, with the top part corresponding to logical AND of editorial judgments, and the bottom part corresponding to logical OR of editorial judgments. The superscripts denote significance under one-tail paired t-test with p-value< 0.05.

Decile	**Method+ Google**	**Method+ Babelfish**	**Baseline+ Google**	**Baseline+ Babelfish**
1	0.470	$0.480^{\diamond}$	0.440	0.370
2	0.470^{+}	$0.440^{\diamond}$	0.290	0.190
3	0.350^{+}	$0.340^{\diamond}$	0.180	0.180
4	0.320	0.290	0.270	0.300
5	0.380^{+}	$0.390^{\diamond}$	0.170	0.160
6	$0.410^{\star}$	0.340	0.310	0.250
7	0.410^{+}	$0.350^{\diamond}$	0.080	0.100
8	0.320^{+}	$0.290^{\diamond}$	0.220	0.190
9	0.420^{+}	$0.360^{\diamond}$	0.210	0.180
10	0.270	0.250	0.240	0.220
Overall	$0.382^{\star+}$	$0.353^{\diamond}$	0.241	0.214
1	0.620	0.610	0.620	0.560
2	0.620^{+}	$0.610^{\diamond}$	0.440	0.260
3	0.520^{+}	$0.480^{\diamond}$	0.310	0.250
4	0.550^{+}	0.510	0.400	0.380
5	0.570^{+}	$0.530^{\diamond}$	0.310	0.250
6	0.610	$0.530^{\diamond}$	0.480	0.380
7	$0.550^{\star+}$	$0.440^{\diamond}$	0.170	0.130
8	$0.440^{\star}$	0.370	0.350	0.290
9	0.610^{+}	$0.560^{\diamond}$	0.340	0.300
10	0.470^{+}	0.430	0.380	0.350
Overall	$0.556^{\star+}$	$0.507^{\diamond}$	0.380	0.315

We also examine the effect of using different MT systems. "$\star$" denotes significant difference between *Method+Google* and *Method+Babelfish*. We observe that, overall, *Method+Google* significantly outperforms *Method+Babelfish* on the 200 queries. In the future, we plan to apply our approach with a simple bilingual-dictionary-based tranlsation module to further investigate how the quality of machine translation affects the performance of our system.

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we presented a robust method to classify non-English queries against an English taxonomy. We dispatch a non-English query to a general purpose search engine, and retrieve top search results in the query's native language. These non-English pages are then translated into English via publicly available MT systems. The translated pages are classified using a classifier trained on English data, and the label of the given query is inferred from the classes of the translated pages by a voting mechanism.

Preliminary experiments with queries sampled from a Chinese query log show that our method almost always significantly outperforms a strong baseline method, and the conclusion holds consistently for two different machine translation systems. By employing the blind relevance feedback techniques in the query's native language, rather than in the English Web with the translated query, the impact of erroneous translation is significantly reduced.

It is also important to note that the performance of our method, as we expect, seems to be less sensitive to the volume of the query, that is, less query dependent, with an overall smaller variance in average accuracy over different deciles, and relatively better performance on rare queries compared to the baseline approach. The rare queries, in aggregation accounting for a considerable mass of the search engine traffic, simply do not have enough occurrences to allow statistical learning on a per-query basis. The superior performance of our method on rare queries provides a substantial opportunity for down-stream applications such as online advertising.

While a relatively small query set is used in our preliminary study, the results are quite promising. In the future, we plan to further investigate the robustness of our method with larger data sets in multiple non-English languages.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.