Multilingual Search with Machine Translation in the Intel Communities

Ryan Martin ryan.c.martin@intel.com

Abstract

This paper describes an experiment performed at Intel to assess the viability of using machine translation for cross-language information retrieval within the Intel Communities (public user forums). Many of the Intel Communities are mixed-language, with a large majority of content being posted in English. In order to make this information available to non-English speakers, our team researched the effectiveness of using machine translation to translate search queries to English using general domain machine translation systems.

1 Introduction

Many of the Intel Communities are multilingual by necessity; a single forum supports users worldwide. Although it is not uncommon for visitors to post content in other languages, the majority of content in these forums is English. In order to improve the experience for non-English speaking visitors, a real-time translation feature was added to the forums in 2012. This feature gives site visitors the ability to translate individual posts to one of 10 languages. While this feature is used often and gets positive feedback from site visitors, a notable shortcoming has been the lack of cross-language search.

It is believed that users would be more likely to find information relevant to their visit if the search results from non-English search terms also included English content within the forums. In order to test this hypothesis, it was first necessary to understand if using machine translation for cross-language search[1] could be used with good results in the selected user forum.

Previous studies have looked at using dictionary-based techniques in cross-language information access (Levow et al., 2005). For this experiment, we evaluated the use of machine translation (MT) without any special query *pre*- or *post*-processing. This decision was largely influenced by the fact that a third-party collaboration platform is used to host the user forums, and we are limited to the amount of customization that can be implemented.

To better understand the viability of using machine translation for this purpose, the Intel team used machine translation to translate non-English search queries written in Spanish and Simplified Chinese[2], and compared the results with searches performed using roughly equivalent English queries.

The following sections describe the experiment set up, scoring methods, and results.

2 Platform and User Description

The *Intel Support Community* was used as the test platform. This forum is hosted using a third-party collaboration platform. The Support Community serves a large number of active users

[1] Or *cross-language information retrieval* (CLIR), using the more common description.

[2] https://communities.intel.com/community/tech

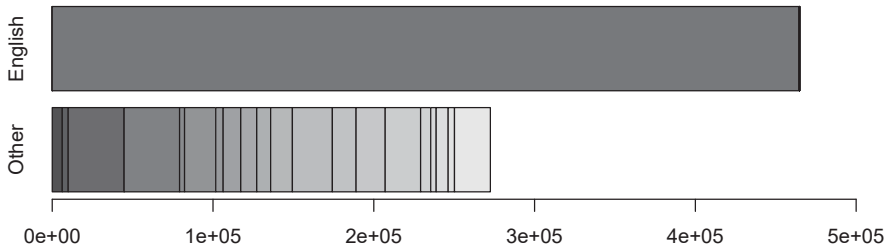

Figure 1: Unique sessions by browser language (based on 1 month of data)

and contains discussions on a diverse set of Intel products and services. There are currently over 70K searchable items (blog posts, discussions, documents), focused on computer hardware and software. Of this content, approximately 82% is from discussion posts[3]. The variety and quantity of data made this particular community a good choice for the study.

2.0.1 User Language Profile

Based on one month of data from 2016, approximately 37 percent of sessions have a browser language other than English. This ratio of non-English browsers is illustrated in Figure 1.

While this is a relatively high percentage of users with non-English browser settings, an informal review of actual user searches shows that most visitors are searching in English[4].

2.0.2 Search Engine

The experiment was carried out using the collaboration platform's native search engine. The search engine supports the following features:

- Basic word search (ignores case and order)
- Exact phrase search using quotation marks to delimit phrases
- A simple one-or-more wildcard (*)
- Compound expressions using AND, OR, NOT, and grouping ( )
- No removal of stop words
- Content must contain *all* search terms

All of the searches performed for this experiment were basic word searches, and didn't make use of any special query syntax.

[3] Data collected on August 29, 2016

[4] User data does not record search language. The actual ratio of non-English search queries requires further investigation.

LANG	N	TER	EQUIV.
ES	59	0.300	45.8%
ZH	59	0.318	44.1%
ES*	37	0.324	45.9%
ZH*	37	0.117	82.9%

Table 1: Search Data (* represent actual user data)

2.1 Experiment

Spanish and Simplified Chinese were selected as the source query languages. Both languages are spoken in locales with a relatively high percentage of site visitors. Additionally, these languages both had sufficient examples of non-English queries collected from real user sessions.

2.1.1 Search Data

The search queries were a combination of both real-world searches collected from the site, and artificial searches that were developed by the team based on existing site content. It was necessary to develop our own search examples for a couple of reasons. For one, many searches fail to return results regardless of the language used. It was desirable to have a reference set that was guaranteed to return some meaningful results – there is little reason to translate a search that is known to return few or no results. Additionally, the number of non-English searches collected from the site was relatively small after being filtered for quality.

A total of 59 samples were developed in English by the team as a reference set. These samples were then human translated into Spanish and Simplified Chinese in order to achieve a close approximation of the English source query. The translators were given instructions that each line of the source text was composed of independent search terms – this was done so that translators did not attempt to fix the input by producing more grammatical output.

Additionally, 37 Spanish and 37 Simplified Chinese searches were collected from real-world sessions and human translated to English.

The median length of the searches was 4 terms, although the real user searches for Chinese were somewhat shorter with a median of 2 (See Figure 2).

2.1.2 MT systems

Generally available commercial MT systems were used for the translation of the non-English queries to English. In this experiment, we chose to focus on the use of general domain MT alone. This decision was largely influenced by the fact that we have limited control over the collaboration platform and search engine; the later being more-or-less a black box. Although the domain is relatively well-defined (hardware and software), we have typically relied upon general domain systems for user-generated content since the quality and style of the source content is considered to be less predictable. Improving the system using domain-specific MT systems, or by including in-domain dictionaries in the query translation process (Jones et al., 2008) certainly deserves further investigation.

The selected MT systems are also used to support the real-time translation feature that has been integrated into the platform *Discussions*. TER scores were calculated between the English reference searches, and the English output from machine translating the non-English source (See Table 1).

The EQUIV column of Table 1 shows the percentage of translations that were functionally equivalent to the reference query. In other words, the percentage of searches where the the translation and reference contained the same terms when ignoring both letter case and order. These

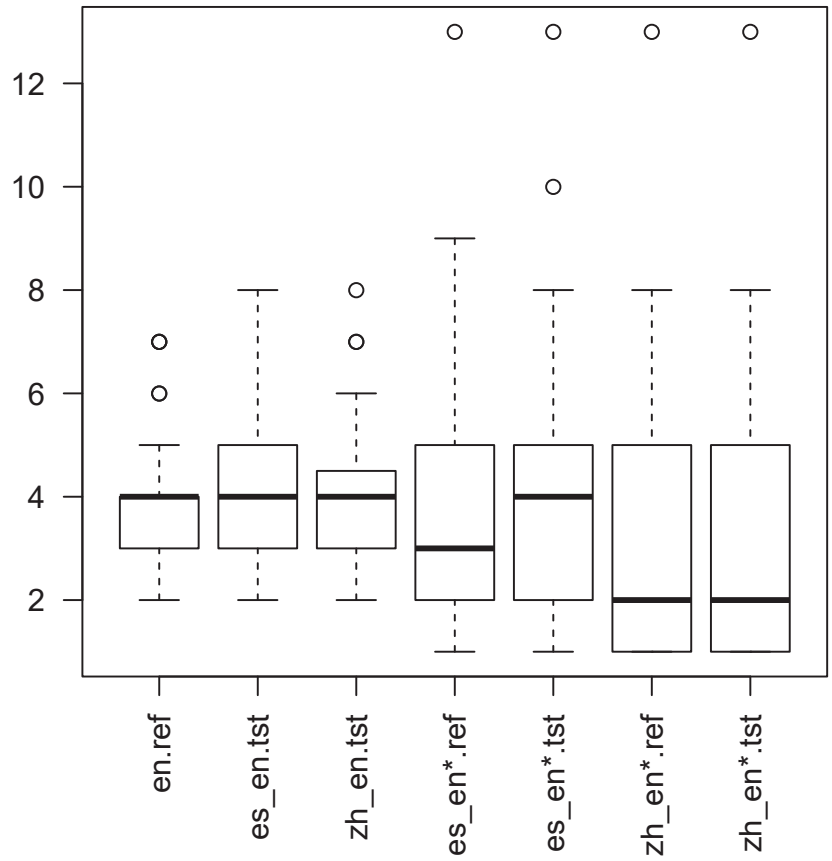

Figure 2: Search query length (words)

translated queries will return the same search results as the reference based on the behavior of the search engine (Section 2.0.2).

2.1.3 Data Collection

An automated script was used to perform searches of each of the English reference queries. The same process was used to collect results from the equivalent machine translated versions of each search (the 'test' set).

For each search performed, the top 5 search result URLs were parsed from the result page[5]. When fewer than 5 results were returned, then all URLs from the results page were collected.

3 Scoring Methods and Results

Each search query was given a point for each search result that matched between the reference (English) and the translated search. The maximum score for any search was 5 points given that only the top five results for any search were recorded. Note that the total number of possible points for a particular search may be less than 5 if the reference search returned less than 5 results.

A final calculation was made by summing the points, and then dividing by the total number of search results in the reference set. These results are displayed in the SCORE column of Table 2.

In addition to calculating the simple ratios shown in Table 2, we also calculated adjusted scores that scaled the points for each search to the range 0 - 5. In the resulting frequency distribution, 0 represents no matches between the reference and test search, while 5 means that the search results were identical between the reference and test (e.g. This was done so that 3/3 and 5/5 would both get a score of 5). The results are shown in Figure 3.

Finally, the TOP RESULT column of Table 2 shows the percentage of searches where the top (first) search result matched between the test and reference searches.

3.1 Analysis

The scores for the four test sets displayed in Figure 3 all share a very similar distribution. There was also very similar distributions between the test sets developed by the team and those harvested from site users.

It is important not to interpret the scores displayed in Table 2 as a measure of search effectiveness since the experiment used a diverse set of search queries that were known to return reasonable results. To actually measure the total search accuracy, one would also need to include failed or low quality searches. In other words, the final score does not imply that actual users would have a similar success rate in finding relevant information.

These results show that given a *good* search query where the equivalent English would return meaningful search results, that the machine translated queries provided identical results most of the time. In our experiment, a 4 or 5 could be expected at least 70% of the time.

When the translated search fails, it usually fails completely. This is shown by the spikes at 0 in Figure 3. This particular feature of the distributions could be explained a number of ways. One likely explanations is that the basic search fails if there are any out of vocabulary (OOV) terms. This is true regardless of language, or whether MT was used to produce the search terms. For the basic search to succeed, all of the terms in the search query must appear in a target document. It is reasonable to conclude that a translation error in a single term could cause a search to completely fail – this hypothesis would need to be confirmed by a complete review of the translated search terms.

[5]Each URL contains a unique numeric identifier for the target content.

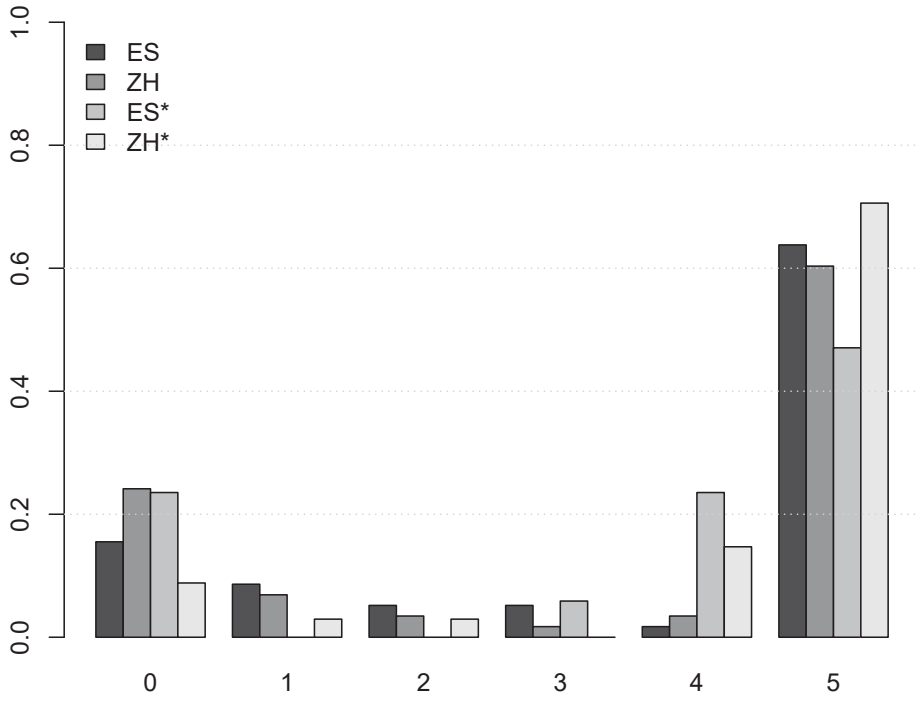

Figure 3: Distribution of scores

LANG	N	SCORE	TOP RESULT
ES	37	71.3%	71.2%
ZH	37	67.9%	64.4%
ES*	59	71.1%	64.9%
ZH*	59	79.3%	78.3%

Table 2: Total ratio of successful (intersecting) results.

Although Spanish and Simplified Chinese had very similar results, it should be noted that these are some of our best performing MT languages, especially when the target language is English. Similar results should not be assumed for other language pairs. Additionally, the relevance of the search results was not considered in the scoring process; only the intersection count of reference and test search results are used in the scoring.

4 Conclusions

Our study looked at the viability of using machine translation to translate non-English search terms to English within the Intel Communities. The ability for users to search cross-language across discussion forums is one possible method to connect users with relevant content in multi-language user forums. Once users are able to find relevant content, they can use the real-time translation features already available on the site. The median score of the four test sets (based on the scoring method described in Section 3) was approximately 71%, and we consider this to be a good baseline when evaluating methods to improve the system. Further research may include evaluating custom MT system developed for the Intel domain.

Acknowledgments

I would like to thank Julie Chang for her assistance coordinating translations and providing actual user search queries from the Communities.

References

Jones, G. J. F., Fantino, F., Newman, E., and Zhang, Y. (2008). Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from wikipedia. In *Proceedings of the 2 nd International Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies*, pages 34–41.

Levow, G.-A., Oard, D. W., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41:523–547.