



Simultaneous Multilingual Search for Translingual Information Retrieval

Kristen Parton¹
kristen@cs.columbia.edu

Kathleen R. McKeown¹
kathy@cs.columbia.edu

James Allan²
allan@cs.umass.edu

Enrique Henestroza¹
eh2348@columbia.edu

¹ Department of Computer Science
Columbia University

² Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst

ABSTRACT

We consider the problem of *translingual* information retrieval, where monolingual searchers issue queries in a different language than the document language(s) and the results must be returned in the language they know, the query language. We present a framework for translingual IR that integrates document translation and query translation into the retrieval model. The corpus is represented as an aligned, jointly indexed “pseudo-parallel” corpus, where each document contains the text of the document along with its translation into the query language. The queries are formulated as multilingual structured queries, where each query term and its translations into the document language(s) are treated as synonym sets. This model leverages simultaneous search in multiple languages against jointly indexed documents to improve the accuracy of results over search using document translation or query translation alone. For query translation, we compared a statistical machine translation (SMT) approach to a dictionary-based approach. We found that using a Wikipedia-derived dictionary for named entities combined with an SMT-based dictionary worked better than SMT alone. Simultaneous multilingual search also has other important features suited to translingual search, since it can provide an indication of poor document translation when a match with the source document is found. We show how close integration of CLIR and SMT allows us to improve result translation in addition to IR results.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models, Search Process.

General Terms

Design, Experimentation.

Keywords

Cross-lingual IR, query translation, document translation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-59593-991-3/08/10...\$5.00.

1. INTRODUCTION

Cross-Lingual Information Retrieval (CLIR) typically focuses on the task of retrieving documents in language f (or a set of M languages F) that are relevant to a query in language e . However, in many real-life scenarios, such as intelligence work, international business, or tourism, users may not be able to read language f and therefore need results in the query language, e . This requires searching documents in another language, but returning results in the query language; we call this task *translingual information retrieval*. In CLIR, result translation has usually been considered a separate, post-processing step that should be evaluated separately. In contrast, we show that by viewing CLIR and result translation as parts of the same translingual IR task, we can leverage an integrated, simultaneous search over source and translation to improve the accuracy of search results, and, at the same time, use search results to improve document translations.

Many CLIR systems have used a query translation approach, in which the documents are indexed in their source language(s), the query is translated into each of the F document languages, and the retrieval is done completely in languages F . Another approach to CLIR is to do retrieval in the query language e by indexing the document translations in language e and searching using the original query. Both query translation (QT) and document translation (DT) attempt to map the query and the documents into a common language, but they use different translation strategies. In DT, each document is a large, coherent context with full, grammatical sentences, whereas a query may be short and non-grammatical, with little or no context. On the other hand, once a document is translated, any mistakes or deletions in the translation cannot be remedied, whereas translating a query allows for more flexibility in incorporating multiple possible translations using synonyms and related terms.

In this paper, we present a framework for translingual IR that integrates document translation and query translation in a novel way that is particularly suited for translingual applications. Our approach uses representations in the query language and the document languages simultaneously. We take advantage of having a corpus that has been automatically translated to develop a hybrid model that integrates QT and DT into the indexing and searching, rather than as a post-processing step. We define a pseudo-parallel document as a single document containing both the source (in $f \in F$) and machine translation (in e) of the

document. The query is translated into each $f \in F$, and then a multilingual structured query is created, which represents the original query terms as well as their translations into all languages. Finally, the multilingual structured query is run over the pseudo-parallel indexed corpus to retrieve the results. In this Simultaneous Multilingual IR (SMLIR) model, we simultaneously search each document in both the document language, $f \in F$, and the query language e , thus allowing the relative advantages of the QT and DT approaches to complement each other. Our comparison of SMLIR against another state-of-the-art hybrid approach using both query and document translation shows that this advantage yields better performance for SMLIR.

We use the SMLIR framework to experiment with different knowledge sources for query translation. We present a novel approach using Wikipedia to obtain both name translations and name variants, including both spelling variations and slang. Given the large difference between the English, Arabic and Chinese languages, differences in name spellings pose significant problems that variants help to address. Our experiments demonstrate that using Wikipedia for query translation yields better results than using statistical machine translation (SMT) alone for query term translation.

The SMLIR approach has additional benefits in a translanguing setting. When returning documents in translation, a relevant document that is unreadable to the user is just as useless as an irrelevant document. As part of our system, we show how to use SMLIR to detect translation errors in the documents at query time and subsequently correct them. This approach to automatic post-editing allows us to use query-time information to improve the translated responses, and therefore present the user with a more readable document. Both approaches seek to improve translanguing IR: SMLIR improves result relevance, and through post-editing improves the translation quality of returned results.

2. PREVIOUS WORK

Many CLIR systems use some form of QT to search documents in languages other than the query language. Since a term frequently has more than one translation in the target language, previous research has focused on methods for dealing with this kind of ambiguity. Pirkola's [98] approach features "structured queries" which use the term frequencies of all possible translations of a query term; document frequencies are computed based on all documents which contain any of the possible translations. This method has been extended in various ways, addressing computational issues in computing frequencies (e.g., [Kwok 00; Oard and Ertrunc 02]) and augmenting the query translation with translation probabilities that are used as weights for each translated term [Darwish and Oard 03]. Language modeling has also been used as a basis for weighting term translations appropriately (e.g., [Xu and Weischedel 00; Lavrenko and Croft 01; Kraaij 04]). Our research critically differs from all of these approaches in our joint use of query and document translation.

Recent research examines the joint use of query and document term translation [Wang and Oard 06]. Wang and Oard use bidirectional term alignments derived using Giza++ [Och and Ney 00] to translate both query terms and document terms. A key characteristic of their work is the mapping of translated terms to language specific synsets from WordNet [Miller et al. 90] for English and other languages. When a WordNet for a specific

language does not exist, they compute the synsets for that language automatically from the Giza++ word alignments. This approach thus attempts to capture matches between query terms and document terms based on meaning. They experiment with a number of different methods for matching query-language synsets against document-language synsets. While closer to the work we report here, our use of full document translation rather than term translation alone allows us to carry out simultaneous search in document and query languages. We also demonstrate different ways of doing the query translation using a combination of statistical phrase tables and Wikipedia.

Early work by Oard [98] compares CLIR search using QT and DT separately against more standard approaches of dictionary based look-up for query term translation, and finds the use of DT promising. The approaches of McCarley [99] and Chen and Gey [04] for the joint use of QT and DT are perhaps most similar to our own. Chen and Gey do an "approximate" fast document translation by replacing each word in a document with the single most likely translation and subsequently build a query language index. McCarley uses a full machine translation system to translate his corpus, as we do in our system. Then, Chen and Gey as well as McCarley translate the query using either a 1-best machine translation or the 1-best translation from a statistical translation lexicon, and then do pseudo-relevance feedback for query expansion. In both systems, a document-language search is done with the query translation over the indexed source documents (QT), and then a second query-language search is done with the original query over the separately indexed translations (DT). Finally, the scores from the QT and DT runs are merged to get a score for the hybrid system, and the result documents are reranked. McCarley merges results using arithmetic mean; Chen and Gey sum the scores, which is rank equivalent to mean. Both find that their hybrid systems outperform the QT and DT systems; McCarley's hybrid system even outperforms his monolingual system, where human translations are used to search the source documents.

Rather than build two indexes and run two separate searches, we build a single index where each document is indexed bilingually, as both the original document and its translation into the query language. Then we create a single multilingual query which combines the original query with the query translation(s) into the document language(s). Since we are not merging output from different systems, no parameter tuning is required to determine whether to weight QT or DT higher or how to merge them. This approach is similar to the approach taken by [Nie and Jin 03] for multilingual information retrieval, where documents in multiple languages were combined in a multilingual index that could be searched with a single multilingual query which was constructed via query translation. In their system, the multilingual approach outperformed the "separate indexing then merging" approach. However, their goal was different than ours, since the documents were already multilingual and the purpose was to return relevant documents in multiple languages.

We also explore various improvements to query translation, beyond the single-best machine translation used by [McCarley] and [Chen and Gey]. We derive synonym and translation dictionaries from Wikipedia. Since Wikipedia is created and edited by humans, we hypothesize that it will be better at translating than machine translation. In particular, Wikipedia contains the translations of many named entities, which may be

أدلى شوارزنججر بهذه الملاحظة هنا اليوم / الثلاثاء / في عشاء أقامته شبكة أعمال كاليفورنيا الأمريكية - الصينية
 He SwArznjr by these pointed out here today in a dinner banquet held by the network of California American.

Actual translation: "Schwarzenegger made this statement here today, Tuesday, at a dinner held by the Chinese-American business association of California."

Figure 1. A pseudo-parallel document, with Arabic source and English machine translation. A reference translation is shown below. Note that the word "Chinese" is deleted in the machine translation, and "Schwarzenegger" is mistranslated.

mistranslated by statistical machine translation. [Ferrandez et al. 07] demonstrate that Wikipedia is an excellent source of named entity translations for cross-lingual question answering. For their Spanish-English cross-lingual question-answering application, a full 59% of the named entities should not be translated. Wikipedia helps them detect which ones to translate, as well as providing translations for many of the named entities. In contrast, in our case, we are dealing with languages that use non-overlapping alphabets (English to Chinese and Arabic) and thus all names must be translated. Furthermore, given the differences, name misspellings abound. Our use of re-directs to build a set of name variants addresses this problem.

3. SIMULTANEOUS MULTILINGUAL IR

The problem we seek to address is the following. We are given a corpus with documents in language **F**, where all the documents have been translated into a common language **e**. Given a query in language **e**, what is the best way to retrieve relevant documents from this corpus and return the documents in language **e**?

3.1 Approach

Our solution is to index the translation together with the source document, and then to search them both simultaneously using multilingual structured queries. Figure 1 shows a pseudo-parallel document, that is, a document aligned with its machine translation. The actual translation of the example is also shown in the figure. This translation exhibits problems typical to statistical machine translation (SMT): Schwarzenegger's name could not be translated correctly, the sentence is ungrammatical, and an important word, "Chinese," has been deleted. This demonstrates a problem with a pure document translation (DT) approach to CLIR: although many queries contain named entities, names are especially difficult for MT to handle. We expect that MT will improve in the future, but it will never be perfect. Foreign and rare names will always be problematic, especially since there are often several acceptable spellings of these names (our corpus contains at least three versions of Arnold Schwarzenegger's name in Arabic). Searching in less formal genres also requires handling name variations, nicknames and misspellings: various English documents in our corpus refer to Schwarzenegger as Schwartzenegger, Arnold, and the Governorator. Unless we have perfect name translation, and uniform spelling of name variations, we will always miss some documents by doing document translation only. Document translation using SMT also suffers from deleted tokens and so-called "hallucinated" insertions, both of which can hurt retrieval accuracy.

On the other hand, a pure query translation (QT) approach faces different challenges. The best approach would combine MT, manual dictionaries, and language-pair-specific transliteration strategies. Even then, many query terms are likely to be out-of-vocabulary or translated incorrectly. In addition, query terms that are unambiguous in language **e** may be translated to ambiguous or polysemous terms in languages **F**. For example, the Arabic words for "Brad Pitt" mean refrigerator or teapot, and house, respectively; the Arabic word for Dean (as in Howard Dean) is a

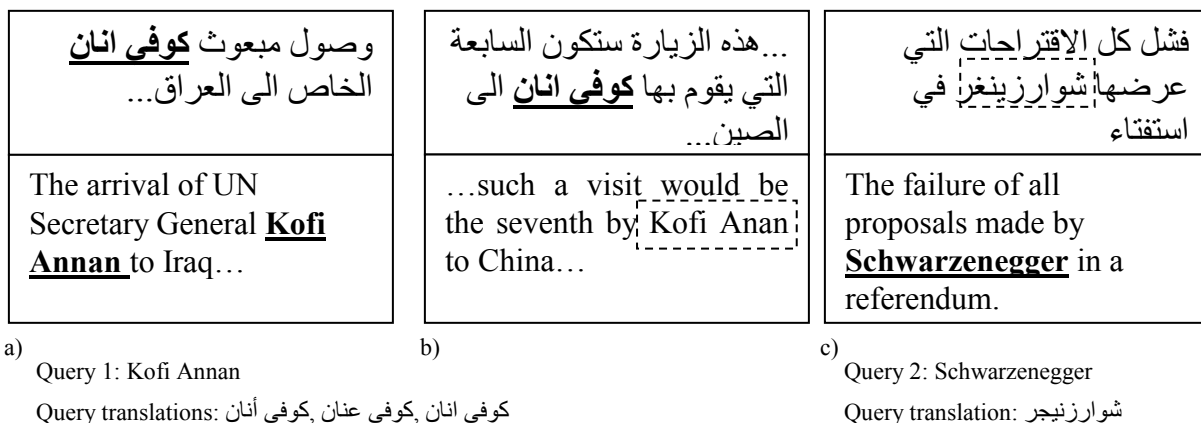


Figure 2. Examples of indexed "pseudo-parallel" documents and multilingual queries. Each indexed document contains the Arabic source and the English machine translation. The query is made up of the original query terms combined with query translations. In document a) both the query translation and document translation match. In document b) only the query translation matches, because the name is mistranslated in the document. In document c) only the document translation matches because our query translation did not come up with this spelling variation, although the document translation system was able to translate it.

very common noun that can mean religion, loan, and devoutness, among other things. While relevance feedback and structured queries can alleviate these problems to an extent, there is still room for improvement with a QT approach.

We attempt to mitigate the problems of DT and QT, and benefit from the advantages of each, by using both approaches simultaneously. Consider the pseudo-parallel documents and queries in Figure 2. For document a, the document translation and query translation succeed, and query 1 matches the document twice, once in the query language and once in the document language. However, document b is translated wrong, so the match is only on the document-language side. A system using only document translation would not return this match. For document c and query 2, we see the opposite result. Our query translation system did not return this Arabic spelling of Schwarzenegger, but the MT system was still able to translate it correctly, so the SMLIR system matches on the query-language side only. Query translation alone would not return this match. By combining both methods, we are able to retrieve all three matches, and weight the first one higher. For translational applications, returning a bad translation can make a relevant document look irrelevant. It is desirable to be able to detect translations we believe to be correct, and rank them higher in the results.

A crucial feature of our hybrid system is that QT and DT are integrated into the retrieval model, rather than merged via parameter tuning. To see how this is done, consider INDRI’s approach to retrieval which is done using a query likelihood model; Indri estimates $\log P(Q|D)$ for each document and ranks the documents according to the log probability. The model assumes term independence so that $\log P(Q|D) = \sum_{\{q \in Q\}} \log P(q|D)$. Individual term probabilities are estimated using a maximum likelihood estimate, so $P(q|D) = \text{tf}(q, D) / |D|$, the proportion of terms in D that are q . (To compensate for varying document length and to avoid the zero-probability problem, that estimate is adjusted with Dirichlet smoothing from the full corpus’ statistics.)

For cross-lingual approaches, let D_f be the original document and D_e be its translation (similarly for Q_e and Q_f). For QT, q is replaced by its (weighted) set of possible translations, meaning that the $P(q|D) = \sum_{\{w \in \text{trans}(q)\}} \text{tf}(w, D_f) / |D_f|$. For DT, the estimate is done in the translation of the document: $P(q|D) = \text{tf}(q, D_e) / |D_e|$.

For SMLIR, however, D_f and D_e are blended into a single representation D_{fe} and counts are estimated there, so

$$P(q | D) = \sum_{w \in \text{trans}(q)} \frac{\text{tf}(w, D_f)}{|D_f|} + \frac{\text{tf}(q, D_e)}{|D_e|}$$

Hybrid models used in earlier work (e.g., [McCarley 99]) average the QT and DT values. Ignoring the Dirichlet smoothing parameter, the merged approach is equivalent to

$$P(q | D) = \sum_{w \in \text{trans}(q)} \frac{\text{tf}(w, D_f)}{|D_f|} + \frac{\text{tf}(q, D_e)}{|D_e|}$$

Note that if $|D_e| = |D_f|$ (or they differ only by a constant) and if no words in e occur in D_f and no words in f occur in D_e , then the two estimates differ only by a constant (since $|D_{fe}| = 2|D_e|$). That is, ignoring smoothing, the main theoretical difference between

the two models arises when the source or translation document contains terms in the other language.

The ability to handle mixed-language text may be useful. We noticed that Chinese news articles tended to contain translations for Western names in parentheses after the first mention. Also, in our corpus, many blogs quoted English sources verbatim, but contained comments and discussion in Chinese or Arabic. A pure query translation approach would miss these matches, since the query would not contain the original query terms.

In addition, translating from language e to language f may be easier than translating in the reverse direction. Therefore, one approach may perform better for a given query than another. Using the SMLIR model described above, the relative importance of QT and DT changes per query, depending on the relative quality of the query and document translations.

3.2 Practical Considerations

Although combining query translation with document translation has been shown to improve relevance [McCarley 99], CLIR systems typically do not use document translation since corpus translation is very resource-intensive. Alternatively, some systems use fast “approximate” document translation [Chen 04, Oard 00], where the result may be unreadable to humans but useful for IR. For the CLIR task, where the goal is to return documents in their source language, full machine translation for a corpus is a large cost with low reward. However, for the translational IR task that we have defined, where the goal is to return documents translated into the query language, corpus translation proves to be very useful. Since machine translation is typically resource-intensive, doing result translation at search time results in a trade-off between time and translation quality. Translating the documents ahead of time may allow for better translations, and also enables further offline analysis in the query language (for example, named entity recognition).

From a practical point of view, joint indexing may also be simpler than a separate-indexing-then-merge approach. In the merged approach, two indexes are built and each query results in two separate queries, which then have to be merged. Using SMLIR, only a single index has to be managed, and each query results in one IR query, whose results can be returned without further processing. The query time for both methods is comparable.

A further consideration for translational applications is how the system scales with additional languages. Our model considers the case where documents may be in multiple languages, $f \in \mathbf{F}$, but there is only a single query/response language, e . Then we need $|\mathbf{F}|$ document translation resources from \mathbf{F} to e , and $|\mathbf{F}|$ query translation resources from e to \mathbf{F} . We consider the translation resources an essential part of the translational task as we have defined it; without a document translation system, an application would not be able to return results in the query language.

3.3 Implementation

Our research on simultaneous multilingual IR is done as part of the DARPA GALE (Global Autonomous Language Exploitation) program. The end user of the GALE system may ask a variety of open-ended questions, defined by a set of templates. End users are English speaking only and need responses in English. Each of the DARPA GALE teams is provided with a multilingual corpus, including text and speech, consisting of English, Arabic and Chinese documents. We handled documents from both the

“formal” genre, such as newswire or broadcast news, and the “informal” genre, such as blogs, newsgroups and broadcast conversation. In this context, it is the task of IR to find all documents that are potentially relevant to a given query; INDRI V2.5 [Metzler and Croft 2004] was used for retrieval. These relevant documents are then passed to response generators which filter out irrelevant sentences to produce the final response. Results must be returned to the user in English, even if they come from Chinese or Arabic sources.

Here we describe the nature of GALE queries, the use of Indri and our method for query translation using translation dictionaries and synonym dictionaries. Then we describe how to represent a query in the SMLIR framework, and the settings we used to experiment with query translation.

3.3.1 Queries

GALE queries are significantly different than typical TREC, NTCIR or CLEF-like queries, because they are based on 17 pre-defined templates with argument slots – e.g., “Describe the connection between [event/topic X] and [event/topic Y]” (the recent TREC ciQA task [Lin and Kelly 2006] used templated queries of the same style). For each GALE template, there is a specific set of relevance guidelines specifying what kind of information is relevant and what is not. For example, template 7 is “Describe involvement of [person/organization/country] in [event/topic].” An excerpt of the GALE Relevance Guidelines states:

For a country to be involved in an event/topic, there must exist an official state action regarding the event. The involvement of ordinary citizens (of the country) in the event does not constitute that country’s involvement in the event...Background information about the event or the involved people, organizations, and countries, is irrelevant if it does not connect explicitly with some involvement in the event.

Due to the specific nature of the guidelines, after the IR results are returned, a template-specific response generator does further filtering for relevant content.

English Query	English Redirects	Cross-Language Links	Arabic Redirects
mahmoud abbas	mahmoud abbas abu mazen mahmud 'abbas mahmud abbas abbas, mahmoud	محمود عباس	محمود عباس أبو مازن
kofi annan	annan, kofi kofi kofi a annan kofi annan kofi atta annan kofi bo bofi nana maria annan	كوفي عنان	كوفي عنان كوفي انان كوفي اتان

Figure 3. Deriving translations and synonyms from Wikipedia. Redirects are used for common misspellings as well as name variants. Using the redirects can add noise, but also increases the likelihood of finding a translation.

Overall, out of 17 templates, 10 have named entities as arguments, 3 have events or topics as arguments, and 4 contain both named entities and non-named entities as arguments. The arguments are typically short and name-centric: in GALE’s second year, the average argument length was 3.4 words, and less than 10% of arguments contained no named entities.

The argument length is interesting, because the query arguments are more similar to short web queries than long TREC-like query narration/descriptions. Unlike web queries, GALE query arguments are well-formed grammatical units – they are always noun phrases. Given these characteristics, query translation is in some ways easier for GALE than for other tasks because we do not have to translate long sentences or analyze semantic roles. However, the context of a longer narrative is not available to guide query translation.

3.3.2 Query Translation

In order to build a high-precision name-focused translation dictionary for English to Arabic and English to Mandarin, we took advantage of the user-created content in Wikipedia. In Wikipedia, each article may have links to the same (or similar) articles in other languages. We extracted these links to create a simple translation dictionary. (The links are not always bidirectional, so we extracted in both directions.) Users often add name translations in the first sentence of an article (e.g., “Mahmoud Abbas (Arabic: محمود عباس)...”), so we extract those as well. Since this dictionary is derived from an encyclopedia, it contains many nouns and noun phrases, including many named entities. The name entries are biased towards famous people, and in particular, people that are somehow notable in both languages. Unlike typical MT dictionaries, these translations are not exact, word-for-word translations; for instance, “Hillary Rodham Clinton” is a headword in English, which links to the Arabic headword for “Hillary Clinton.” As noted in [Ferrandez et al. 07], Wikipedia is a particularly suitable source for name translation because new names are constantly being updated by users.

In addition to translation links, Wikipedia users can also add redirect links. These links match a name variation with the canonical form of an article title. For example, there is an English link that redirects the common misspelling “schwarzenegger” to “Arnold Schwarzenegger”, and another for the slang term “governator”. By aggregating all the redirects for a certain article, we can create sets of name variations from each version of Wikipedia; for our purposes, we extracted redirect sets from Arabic, Chinese, and English. It is important to note that these sets are not always synonyms, but may be related words, common misspellings, or even intentional spam. Since our corpus contained blogs and newsgroups, misspellings and slang were useful to us. Relying on user-generated content was important, since these variations would not normally be found in a standard dictionary; however, it may add noise to the dictionary.

As of January, 2008, the number of articles for each language was [Wikipedia]:

English: 2,153,891
 Chinese: 159,392
 Arabic: 50,098

To compensate for the small Arabic Wikipedia, we combined it with a translation dictionary extracted from other Arabic dictionaries.

Using the extracted translation dictionaries and redirect sets, we are able to look up a query term and get a set of variants in both languages. Figure 3 shows the results of looking up two queries. First, each English query is expanded by the English redirects list. Then, we use the translation dictionary to try to translate each redirect variant. Finally, we expand each translation in the Arabic redirects list. For “mahmoud abbas”, the English list contains valid spelling variants and a common nickname (“abu mazen”), and the Arabic list contains “mahmoud abbas” and “abu mazen”. However, for Kofi Annan, we have two errors in the English list: “Nana Maria Annan” is his wife, and “Kofi Bo Bofi” is the punchline to a joke. The Arabic list contains a translation of “Kofi Annan” and two spelling variants.

As is typical, there is a trade-off between precision and recall using the name variations (expansions). For example, the term “William Jefferson Clinton” is not present in our English-Arabic dictionary, but if we expand it to “Bill Clinton”, we can find a translation. In Figure 3, both the Arabic Wikipedia and the English Wikipedia list “Abu Mazen”, but there is no explicit link between them, so they are only available through the redirect lists. However, famous people may have tens or hundreds of redirect terms, so it may hurt precision to include them all.

Although our queries are name-focused, there are many non-names as well, such as topics and events. To translate non-names, we used a probabilistic word translation table derived from bidirectional word alignments extracted from GIZA++ [Och and Ney 00] by the MT members of our DARPA GALE team.

3.3.3 Indri

We used the v2.5 of the open-source Indri retrieval system for all retrieval experiments. Indri is a powerful retrieval system that combines inference networks and statistical language modeling approaches to retrieval [Metzler and Croft 2004]. We chose Indri primarily for two reasons: (1) it provides a convenient mechanism for restricting queries to XML elements and (2) it provides a weighted synonym operator as part of the query language.

We use the XML operators to restrict query terms to matching in a single language. For example, the French query term “are” should not match an English document containing “are”. Although each document contains text in two languages, the source and translation are in separate XML elements, and we can query them separately using Indri. For example, it may be useful to restrict common nouns to matching in one language only, but match proper names in both languages.

Our queries used the following operators of the Indri query language:

- #1(a b c) indicates that terms or features *a*, *b*, and *c* are a phrase and must appear in order and adjacent to each other.
- #combine(a b) indicates that the score associated with a document should be a combination of the scores of its operands. It is the default multi-term operator of Indri.
- #wsyn(w_a a w_b b w_c c) indicates that terms or features *a*, *b*, and *c* should be treated by Indri as if they are the same term, and counts of the terms are weighted as indicated. We use this operator to incorporate probabilistic translations of words into the queries: alternate translations are listed with weights reflecting the estimated chance that the words are actually translations of the query term. That is, when

calculating document or collection probabilities, the counts of all terms are added together and treated as one. Note that the synonym operator allows other operators such as #1() as a feature.

- #OP().field forces Indri to evaluate the operator only within the indicated field name. This feature allows us to run a query within the source text, the translated text., or both.

3.3.4 Query Construction

The GALE queries are non-factoid, template-based questions, for example, “WHERE HAS [Tony Blair] BEEN AND WHEN?” The filler text (in capitals) is used to frame the template, but does not supply useful query terms for querying the corpus. The argument is indicated by brackets, and is sometimes marked as a named entity of a specific type. We ran a named entity recognizer on the query to get more fine-grained markup. There are also optional slots for related words, name variants, and locations. From the whole query, we extracted a set Q_e of English query phrases/words which contained all arguments, related words, locations and phrases marked as named entities. For example, the query argument “Osama bin Laden in Iraq” would generate three terms based on named entity markup: “Osama bin Laden” “Iraq” and “Osama bin Laden in Iraq”.

3.4 Query-Directed SMT Post-Editing

Consider a query such as “Provide information on [Arnold Schwarzenegger]”. Assuming our dictionary contains the correct translation, SMLIR could find the document in Figure 1. This document is relevant in Arabic, but a user is unlikely to rate it relevant in English given the poor translation. In a translanguag setting, we would like to return documents where the relevant parts are readable to the user in translation.

Using the SMLIR approach, we can detect potentially incorrect translations in a document when a document-language match is found for a query but a query-language match is not found. In the example from Figure 1, a good query translation would match the query “Schwarzenegger” in Arabic (شوارزنجر), but the document translation does not match the query due to a translation error. Since we have word alignments from the document translation system, after detecting an incorrect translation, we can replace the incorrect translation (“\$warznr”) with the original query (“Schwarzenegger”). The user then gets a response with the correctly spelled name, and the translated document is perceived relevant. Figure 4 illustrates the post-editing process.

Our algorithm for query-directed SMT post-editing is:

1. Use SMLIR system to detect potential mistranslation. If a result contains a match in the foreign source but not in the English, consider it a potential error.
2. Using word alignments, extract the MT hypothesis: the English words that correspond to the foreign source match.
3. If the MT hypothesis matches a name variation in our dictionary, do not rewrite this translation.
4. Use word alignments to decide which translation tokens to replace. Name translations are not necessarily contiguous, so it is important not to insert the name multiple times. If the match is part of a larger phrase match, there may be links to other words, and it is important not to replace other tokens in the sentence.

WHERE HAS [Former United Nations Secretary General Kofi Annan]
BEEN AND WHEN?

Arabic match:

...هذه الزيارة ستكون السابعة التي يقوم بها كوفي انان الى الصين...

Corresponding English translation:

...such a visit would be the seventh by Kofi Anan to China...

Rewritten English translation:

...such a visit would be the seventh by **Kofi Annan** to China...

Figure 4. Query-directed statistical machine translation post-editing. SMLIR returned a match on the Arabic part of the document for “Kofi Annan”, but not on the English side. We use word alignments from SMT to extract “Kofi Anan” from the English translation. Since “Kofi Anan” is not a known name variation for “Kofi Annan”, we rewrite the sentence with the term as it appeared in the query.

In general, a find-and-replace approach to machine translation is too simplistic and likely to be problematic. We restrict our post-editing to proper names, which are more amenable to rewriting than arbitrary words or phrases. Names are particularly hard for MT systems to translate, but translating them correctly is especially important for question-answering.

There are several reasons that SMT post-editing could improve over SMT. First, statistical machine translation takes a sentence-by-sentence approach to translation, ignoring issues of coherence and consistency. In one document in our corpus, the same name was translated three different ways. A name that co-referred in the source language then appears to be three different people in the translation. Second, we have more information at query-time than at document translation time. In our application, we get name-tagged queries, and we can try to match them to documents that are name-tagged in two languages. Therefore, we are using information from multiple sources to make an informed translation decision. This is similar to the approach of [Ji and Grishman 07] for using joint inference over information extraction and entity translation to improve name translation.

4. EXPERIMENT DESIGN

4.1 Data

The data we are using is the GALE Y2 corpus, which includes text and speech in English, Chinese, and Arabic, with both formal (newswire and broadcast news) and informal (blogs and broadcast conversation) genres.

For all experiments, the entire corpus was processed as follows. Speech was automatically transcribed into text and algorithmically divided into documents based on story segmentation. Foreign text was marked up with named entities, and then translated using SMT into English. The English text was then marked up with named entities. The final corpus included 133,695 Arabic documents and 102,859 Chinese documents; each non-English source was associated with the translated version.

For each foreign-source document, we created a pseudo-document that contained both the source, with source named-entity markup, as well as the English MT, with translation named-entity markup. The source and translation are separate XML elements, so query terms can be restricted to match either in the source or in the target. For Chinese text, we used character segmentation. For Arabic text, we used query-side stemming, meaning we indexed the unstemmed text and expanded the query terms with known morphological prefixes [Larkey et al.].

For evaluating our system, we used queries produced for GALE’s year two evaluation, but we restricted the queries to only Chinese or only Arabic documents. We eliminated queries that had no known relevant documents in a given language. Therefore, some queries had both Arabic and Chinese versions, others had only one or the other, and a few had no foreign source matches.

4.2 Corpus Translation

The Chinese and Arabic documents in our corpus were translated by our DARPA GALE MT team. They use two pass, phrase-based statistical MT. In the first pass, the N best translations are generated, using phrase count features to smooth phrase probabilities. In the second pass, the system uses sentence mixture language models to rescore the N best results.

Our document translation system is sophisticated, but since we are translating an entire corpus, we cannot use the full MT system. The current best SMT systems spend about a week translating about 35,000 Chinese words (NIST MT evaluation 2006). At that rate, translating our corpus would take over 30 years. Despite the trade-off in quality, it is still useful for us to translate the corpus ahead of time in order to do further annotation for question answering.

4.3 Baselines

We implemented two baseline systems in order to compare our hybrid system with query translation only (QT) and document translation only (DT).

4.4 Comparisons with Previous Approaches

Following McCarley [99], we implemented a simple hybrid system that reranks results from the QT and DT baseline systems by their averaged, normalized Indri scores.

We also implemented a straightforward probabilistic structured query approach [Darwish & Oard 03], where we translated all words (including names) using the probabilistic translation described in section 3.3.2.

4.5 Evaluation

We evaluated the results by asking native language readers to judge the returned documents in the source language, in order to eliminate the effects of poor MT. For this paper, we limited the IR evaluation to Mandarin Chinese due to the difficulty in finding an adequate number of Arabic speakers. The judgments were done by 12 native speakers of Mandarin Chinese. Each judge was assigned one or more templates, and given the full GALE relevance guidelines for each template, which include examples of relevant and irrelevant sentences. After reviewing the relevance guidelines for a certain template, the judge used a web-based interface which presented a query along with a set of documents to judge for each query. In this manner, judges consistently worked on the same template and the same queries under different experimental settings, so as to avoid confusing relevant guidelines for different templates.

The documents could be judged Relevant or Not Relevant. The judges were told to judge based on the source document (in Chinese), but they could also see the machine translation. We found that judging based on the source language was important, as sometimes a document was found to be relevant whose garbled translation did not appear relevant.

We used 39 queries as training data and tested on 96 queries. We included the top 10 documents for each query for SMLIR as well as each baseline and previous approach. A total of 13,942 Chinese documents were judged.

4.6 Metrics

We evaluated our IR model using Normalized Discounted Cumulative Gain (NDCG) [Jarvelin & Kekalainen 00], which takes into account the relative ranking that each system gives to the returned documents. The NDCG at n for query Q is defined by the following formula, where $rel(i)$ is the relevance judgment of

the document at rank i , and Z is a normalization factor that makes it so the perfect ranking gets an NDCG score of 1.

$$NDCG(Q) = Z \sum_{i=1}^n \frac{2^{rel(i)} - 1}{\log(1 + i)}$$

In addition, the normalization factor correctly adjusts for queries with less than 10 relevant documents in the corpus, whereas precision at 10 would penalize a perfect system.

5. RESULTS

5.1 Simultaneous Multilingual IR

We compared document translation (DT) and query translation (QT) baselines against two hybrid approaches which combine DT and QT: merged and SMLIR. Merged is similar to the hybrid system in [McCarley 99]: two separate searches are run, one in the query language and one in the document language, and then the scores are averaged and the result list is reranked. SMLIR is our simultaneous multilingual IR approach, in which the documents are jointly indexed and the query is multilingual. (All experiments in this section used both the Wikipedia and SMT dictionary for query translation. Significance is from a two-tailed t-test.)

The results are shown in Table 1. Document translation (DT) does significantly better than query translation (QT). The poor performance of QT is due to two problems: the prevalence of rare names in the queries, which were not covered by the translation dictionary, as well as some issues in translation of non-name phrases. A query translation module that included transliteration might improve performance of QT for names. For non-name arguments (such as “the cigarette smoking ban”), using full SMT rather than the typical approach of word-by-word translation might lead to better QT.

Surprisingly, the merged hybrid system does worse than DT, though at a lower level of significance (97.5%). It seems that for the merged approach, the poor performance of the QT system just serves to degrade the performance of the DT system. However, the SMLIR system does significantly better than QT and DT as well as the merged system.

One of our initial criticisms of the merged system was that document scores are not comparable across queries, so combining them in any way is ad-hoc. We found numerous examples of this

What are we translating	Hybrid Approach	NDCG at 10
Queries only (QT)	-	0.4156
Documents and Queries	Separate searches, then merge results	0.5245*
Documents only (DT)	-	0.5345+
Documents and Queries	Joint indexing (SMLIR)	0.5517*

Table 1. Overall results. For all settings, we used a combination of Wikipedia for name translation and a probabilistic translation dictionary to translate query terms (* indicates 99% confidence, + indicates 97.5% confidence).

Query Translation Method	NDCG at 10
Probabilistic dictionary (all)	0.5136
Wikipedia (names) + Probabilistic dictionary	0.5517*
Wikipedia (names)	0.5572

Table 2. Results for various query translation strategies using the SMLIR hybrid approach. Using Wikipedia for name translation performed significantly better than just using a probabilistic dictionary (* indicates 99% confidence). Combining the methods did not appear to improve the results.

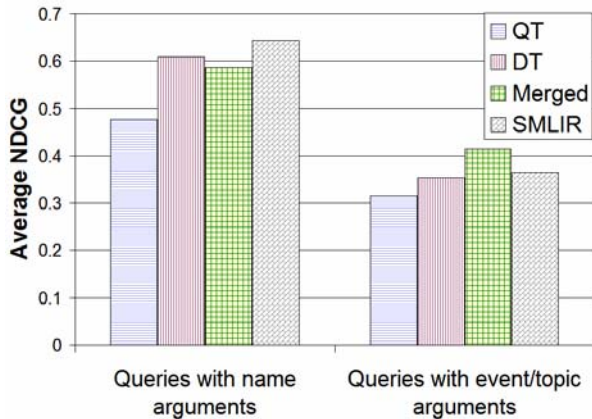


Figure 5. Merged outperforms SMLIR on queries with event/topic arguments, but does poorly on queries with only name arguments.

in our error analysis. Sometimes scores from a QT query were orders of magnitude smaller or larger than scores from a DT query, in which case merged would end up favoring the system with the larger scores, rather than combining DT and QT in a more principled way, as we hoped SMLIR would do.

On the other hand, merged outperformed SMLIR on queries with only event/topic arguments. (The arguments are marked as events or topics, although the events may contain named entities in them.) In Figure 5, we can see that all systems performed worse on these queries than on queries with name arguments only. Whereas merged made significant gains over both QT and DT on event/topic queries, SMLIR barely outperformed them. On the other hand, for queries with name arguments, merged does worse than DT, and SMLIR does best.

5.2 Query Translation

We also performed experiments with various query translation methods. Since Wikipedia is an encyclopedia, we used it for translated named entities, which are problematic for machine translation. We compared translating the names only using Wikipedia against translating all terms using a probabilistic MT dictionary. For all settings that used Wikipedia, we expanded names in the query using the synonym lists derived from Wikipedia, and then translated all synonyms. Surprisingly, just translating the names with Wikipedia did better than using the MT dictionary to do translation of all terms. This may be due to the fact that Wikipedia translations are typically high precision but low recall, whereas an MT dictionary typically contains many (weighted) translations, not all of which are appropriate for a given context. We expected that combining high-precision named entity translation with a probabilistic translation dictionary would perform best, but combining the dictionaries did not improve the results. In any case, the success of using Wikipedia highlights the importance of named entity translation for cross-lingual IR.

5.3 Query-Directed Post-Editing

The goal of our automatic machine translation post-editing system is to use query-time information to improve the translation quality of returned results. Our preliminary results indicate that, despite its simplicity, this approach is able to improve our MT output. We ran the post-editing system on 127 Arabic GALE queries, using

the top 10 document results from our SMLIR system. Of those, 28 (22%) of the queries returned documents that required post-editing. For the queries that were post-edited, 15% of the IR name matches were rewritten. For each query, up to the first 5 post-edits were examined by a student of Arabic. The annotator decided whether the replacement was Acceptable, Not Acceptable or Ambiguous. Of the 101 rewrites examined, our replacements were Acceptable 93% of the time. 6% were Not Acceptable and 1% were Ambiguous. Our post-editing algorithm was especially conservative, so it aimed for precision rather than recall. A more in-depth evaluation is required to explore this issue further.

Improved name translation is essential for good translangual applications, such as question answering. However, MT metrics such as the BLEU score [Papineni et al. 02] do not take into account the relative importance of various words in the sentence. Producing an incorrect translation of a name such as “Zarqawi” has the same effect on BLEU score as producing an incorrect determiner (“a” instead of “the”), though the latter is unlikely to diminish a reader’s comprehension of the text. For translangual question answering, a relevant result with a poor name translation can seem irrelevant to the end-user. By using query-directed post-editing, we can improve result translation for translangual applications.

6. CONCLUSIONS AND FUTURE WORK

Our novel approach for translangual IR features the use of multilingual structured queries which are issued over a pseudo-parallel indexed corpus to retrieve results. This hybrid model integrates QT and DT into the indexing and searching, rather than as a post-processing step, and thus allows proper treatment of multilingual documents. Our experimental results show that this approach significantly outperforms a previous hybrid approach, which merges the results of separate queries issued over separately indexed source and English documents. Our experiments evaluated results for English queries and Chinese documents, but our implementation of SMLIR currently includes three languages, English, Chinese and Arabic, demonstrating the ability to seamlessly integrate multiple languages into one framework.

We also experimented with different approaches to query translation. We introduced a method to generate name variants from Wikipedia, an approach that is critical to capturing the variety of name translations between languages as different as Chinese, English and Arabic. Our experimental results show that query translation based on Wikipedia with expansion using name variants outperforms query translation using only a probabilistic phrase table from statistical machine translation. While we expected query translation using a combination of Wikipedia for names and probabilistic translation for topic terms to outperform either approach alone, our results show that using Wikipedia for names alone is not significantly different from the combined approach.

Our research was motivated by a need to provide monolingual speakers the ability to query a multilingual corpus in their own language and receive documents in that same language. The SMLIR framework meets this need and furthermore, provides clues about the quality of document translation. We showed how the framework allows us to detect mis-translated names when query names match against the document but not the translation. Our preliminary method for post-editing names, correcting the translation, shows that we can dramatically improve the quality of translated names in the relevant documents. This is important in

the context of translingual IR, since it is difficult to tell whether a document is relevant when it is poorly translated.

We are currently working on improving our techniques for post-editing of names, incorporating better methods for handling morphological differences and extending it to handle both Chinese and Arabic names. Another direction for future work is the inclusion of name transliteration in query translation. Finally, we will also explore how we can modify our implementation to better exploit a combination of probabilistic term translation with Wikipedia translation of names.

7. ACKNOWLEDGMENTS

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, in part by an NSF Graduate Research Fellowship, and in part by the Center for Intelligent Information Retrieval at the University of Massachusetts. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors. We would like to thank Bob Armstrong for making the annotation happen, Mark Smucker and Giridhar Kumaran for help with INDRI and corpus issues. We would also like to thank Ben Carterette for help with discussions about evaluation. We would also like to thank the members of the NIGHTINGALE machine translation team for translation data, especially Nizar Habash.

8. REFERENCES

- [1] Chen, A. and F. Gey: Combining Query Translation and Document Translation in Cross-Language Retrieval. CLEF 2003: 108-121.
- [2] Darwish, K. And D.W. Oard, "Probabilistic Structured Query Methods," In *Proceedings of the 26th Annual SIGIR Conference, 2003*, pp. 338-44.
- [3] Ferrández, S., Toral, A., Ferrández, O., Ferrández, A., Muñoz, R. Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering. In *Proceedings of the 12th International Conference on applications of Natural Language to Information Systems*. Paris (France). pp. 352-363. June 2007.
- [4] Habash, N. and Ghoneim, Personal communication, 2007.
- [5] Jarvelin, K. and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of SIGIR 2000*, 2000.
- [6] Ji, H. and R. Grishman. Collaborative Entity Extraction and Translation. *Proc. International Conference on Recent Advances in Natural Language Processing 2007*. Borovets, Bulgaria. Sept 2007.
- [7] Kelly, D. and Lin, J. 2007. Overview of the TREC 2006 ciQA task. *SIGIR Forum* 41, 1 (Jun. 2007), 107-116.
- [8] Kraaij, W., *Variations on Language Modeling on Information Retrieval*, Ph.D. thesis, University of Twente, 2004.
- [9] Larkey, L., L. Ballesteros, and M. Connell, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis," In *Proceedings of the 25th Annual SIGIR Conference 2002*, 275-82.
- [10] Lavrenko, V. and W.B. Croft, Relevance-based Language Models, In *Proceedings of the 24th Annual SIGIR Conference, 2001*, pp. 69-74.
- [11] McCarley, J.S., "Should we Translate the Documents or the Queries in Cross-language Information Retrieval?," in *Proc. of the 37th Annual Conference of the Association for Computational Linguistics*, 1999, pp. 208-214.
- [12] Metzler, D., and Croft, W.B., "Combining the language model and inference network approaches to retrieval," *Information Processing and Management*, 40(5):735-750, 2004.
- [13] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. "Introduction to WordNet: an on-line lexical database." In: *International Journal of Lexicography* 3 (4), 1990, pp. 235 - 244.
- [14] Nie, J-Y. and Jin, F. A Multilingual Approach to Multilingual Information Retrieval. *Proceedings of the Cross-Language Evaluation Forum*, 2003.
- [15] Oard, D., "A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval," In D. Farwell, E. Hovy, and L. Gerber (eds), *Machine Translation and the Information Soup*, p. 472.
- [16] Och, F.J. and H. Newy, "The Alignment Template Approach to Statistical Machine Translation," *Computational Linguistics*, 30(4), 2004.
- [17] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311-318
- [18] Pirkola, A., "The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval," In *Proceedings of the 21st Annual SIGIR Conference*, 1998, pp. 55-64.
- [19] Wan, J. and D.W., Oard, "Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval," In *Proceedings of the 29th Annual SIGIR Conference*, 2006. 202-9.
- [20] Wikipedia. "Wikipedia: Multilingual Statistics.", January, 2008. http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics
- [21] Xu, J. and R. Weischedel, "TREC-9 Cross-lingual Retrieval at BBN," In *Proceedings of The Ninth Text Retrieval Conference*, National Institutes of Standards and Technology, 2000.