Lost and Found in Translation: The Impact of Machine Translated Results on Translingual Information Retrieval

Kristen Parton Nizar Habash Kathleen McKeown
Columbia University, NY, USA
{kristen, habash, kathy}@cs.columbia.edu

Abstract

In an ideal cross-lingual information retrieval (CLIR) system, a user query would generate a search over documents in a different language and the relevant results would be presented in the user's language. In practice, CLIR systems are typically evaluated by judging result relevance in the document language, to factor out the effects of translating the results using machine translation (MT). In this paper, we investigate the influence of four different approaches for integrating MT and CLIR on both retrieval accuracy and user judgment of relevancy. We create a corpus with relevance judgments for both human and machine translated results, and use it to quantify the effect that MT quality has on end-to-end relevance. We find that MT errors result in a 16-39% decrease in mean average precision over the ground truth system that uses human translations. MT errors also caused relevant sentences to appear irrelevant – 5-19% of sentences were relevant in human translation, but were judged irrelevant in MT. To counter this degradation, we present two hybrid retrieval models and two automatic MT post-editing techniques and show that these approaches substantially mitigate the errors and improve the end-to-end relevance.

1 Introduction

With the increasing quality and availability of machine translation (MT) on the Internet, users can consume content in many languages other than their own. Cross-lingual applications go beyond passive consumption to enable users to find, analyze and use information in other languages. In cross-lingual information retrieval (CLIR), a query in one language is used to retrieve results from documents in another language. In real life, the results would need to be translated back to the query language for the user to read, but CLIR system evaluations are typically done in the document language, to rule out the effects of MT on understanding result relevance.

Just as CLIR evaluations focus on intrinsic measures of retrieval accuracy without taking into account result translation, MT evaluations are based on intrinsic measures of translation quality without considering the usability of MT output in applications such as CLIR. The fact that MT and CLIR are studied in isolation from each other leaves a large, under-studied gap surrounding the role of MT in real-life applications of CLIR. This paper is an attempt to bridge that gap.

We distinguish CLIR from CLIR with result translation, or *translingual* IR (TLIR). The TLIR task is: given a query in language ℓ, and a corpus in language m, return relevant results from the corpus, translated from the document language m into the query language ℓ.[1] A system that carries out the TLIR task consists of a CLIR model, which is responsible for retrieving and ranking results, as well as an MT system, which is needed to translate results back to the query (user's) language.

To study the TLIR task, we created a TLIR evaluation corpus with relevance judgments on human translations as well as the output of two MT systems (MT A and MT B). The corpus is based on a standard Arabic-English MT test set. From an MT perspective, this corpus provides an extrinsic, task-based evaluation of MT output; from the viewpoint of CLIR, it models a real-world application, where the end user can read the results of the CLIR system.

We use the TLIR corpus to compare various CLIR models and MT systems on a shared end-to-end task. We find some results that contradict intrinsic evaluations. Although MT A and MT B have similar BLEU scores, MT B performs better on the TLIR task. We compare two baseline CLIR models, and find that the one that performs better in an intrinsic retrieval evaluation actually performs worse in the TLIR evaluation, where both retrieval and result translation are taken into account.

[1] Note that using MT for CLIR query translation has been studied extensively (Gao et al., 2001; Herbert et al., 2011; Magdy and Jones, 2011) and is *not* the focus of this paper.

In addition to studying the complete TLIR task, we can also use the corpus to analyze two separate aspects of the task: retrieval accuracy and translated result understanding. MT errors can degrade each of these aspects of the TLIR task. When CLIR models retrieve over machine translated documents, MT errors can lead to recall errors, where the CLIR model fails to retrieve relevant results. On the other hand, even when a relevant result is found, if it is translated incorrectly it can appear irrelevant to the user. We refer to these two types of errors as **lost in retrieval** errors and **lost in translation** errors.

We use the TLIR corpus as a novel testbed to evaluate previously proposed solutions to each of these types of errors. A bilingual CLIR model has been shown to outperform baseline models in a CLIR evaluation without result translation (Parton et al., 2008); here we show that it also helps in a TLIR setting because it addresses many lost in retrieval errors as well as some lost in result translation errors. Automatic post-editors (APEs) have been shown to improve intrinsic MT quality (Mareček et al., 2011); here we compare the impact of adequacy-oriented APEs on lost in translation errors in the TLIR task.

In the next section, we motivate our work by explaining how MT errors can impact the TLIR task. Then we describe the TLIR evaluation corpus and how we use it to analyze the effect of different MT systems and CLIR models on TLIR performance (Section 3). After describing the experimental setup (Section 4), we present an analysis of two baseline approaches to TLIR that suffer from lost in retrieval errors and lost in translation errors (Section 5). We experiment with bilingual retrieval models to address the lost in retrieval errors (Section 6), and APEs to address the lost in translation errors (Section 7). By integrating CLIR and MT more closely and evaluating them in an end-to-end task, we are able to improve over the baseline approaches.

Arabic (document language)	وقد افرج عن **فعنونو** فى ابريل (نيسان) 2004 بعد ان امضى 18 سنة فى السجن	
English MT (query language)	And was released in April, 2004 after he had spent 18 years in prison	
Human translation	**Vanunu** was released in April 2004 after spending 18 years in prison	
	Document Translation (DT)	**Query Translation (QT)**
Query	Mordechai Vanunu	مردخاي فعنونو
Indexed Sentence	English MT	Arabic
Retrieved Result	None: **Lost in Retrieval** error	Arabic
Translated Result	-	English MT
Translated Relevance	-	Irrelevant: **Lost in Result Translation** error

Figure 1: The results of running the query "Provide information about Mordechai Vanunu" against two different TLIR pipelines. The NE deletion in MT affects the QT and DT retrieval models differently.

2 Motivation

CLIR enables users to find information in languages they do not know, but CLIR search results are not immediately useful because a separate MT system must be applied before the user can read the results. Shared CLIR tasks have been run by TREC, NTCIR and CLEF across a variety of language pairs and domains, and in nearly all cases, relevance is judged in the document language. These evaluations are sufficient for evaluating retrieval models, but inadequate for assessing the usefulness of an end-to-end TLIR system: even if a system returns all the relevant results, if they are translated poorly, the user will not be able to understand them.

As with CLIR, MT systems are usually evaluated intrinsically with human judgments of adequacy and fluency, or with automatic metrics such as BLEU (Papineni et al., 2002). These evaluations are aimed at open-domain, task-agnostic MT, so they seek to balance fluency and adequacy. In contrast, for *task-embedded MT*, where MT is used to translate the results of a cross-lingual application, adequacy may be more important than fluency (or vice versa).

Evaluating task-embedded MT is more difficult than evaluating either the task or the MT alone. In a CLIR evaluation, Hakkani-Tür et al. (2007) were unable to judge against MT output "because when the translation quality is poor that procedure tends to be too subjective and MT system-specific." In a large-scale translingual evaluation for the GALE distillation (question-answering) task, inter-annotator consistency on relevance judgments over MT ranged from 59-89% (Glenn et al., 2011). The Interactive Track at CLEF has also carried out TLIR evaluations (Oard and Gonzalo, 2003), though the evaluations were relatively small in scale because they involved in-depth user studies.

2.1 MT Adequacy Errors

MT systems make a variety of errors which may degrade translation fluency, adequacy or both. In the TLIR task, adequacy is more important because users are searching for specific pieces of information. The MT errors that most impact the TLIR task include named entity (NE) mistranslation and missing content words.

NEs are crucial for many search and IR tasks: Guo et al. (2009) report that 71% of Bing web queries contain NEs, and in a recent question-answering shared task (DARPA GALE Y2), over 90% of the queries contained NEs. Yet NEs are particularly challenging for MT systems to translate correctly (Hermjakob et al., 2008), which can have a major impact on search result translation. NEs are frequently out-of-vocabulary (OOV), and many MT systems delete OOV words, as in Figure 1, where the NE deletion causes the sentence to appear irrelevant to the query. Alternatively, OOV words may also be left in the source language or transliterated. In this example, the user would not be able to understand the NE if it were left in Arabic or if it were transliterated poorly (e.g., the Buckwalter transliteration is *fEnwnw*).

Another type of adequacy error that impacts TLIR is deleted or missing content words. (We refer to content words that are mistranslated as function words as missing since they appear deleted to the user, but not to the MT system.) Manual error analysis has shown that missing content words produce adequacy errors across different language pairs and different types of SMT systems (Condon et al., 2010; Vilar et al., 2006; Popović and Ney, 2007).

2.2 MT Errors and TLIR

MT errors affect TLIR differently depending on which model is used for retrieval. Two naive TLIR baseline systems are simple pipelines of independent MT and CLIR systems, which are demonstrated in Figure 1. In the document translation (DT) approach, the Arabic corpus is translated offline and then indexed in the query language, English. At query time, a monolingual search in the query language (English) is performed, and the MT sentences are retrieved. In the query translation (QT) approach, the corpus is indexed in the document language (Arabic). At query time, the English query is translated into the document language and Arabic sentences are retrieved. Then MT is run to translate the results into the query language (English).

Figure 1 shows how an MT error can affect each of these retrieval models in different ways. In the DT approach, the indexed MT sentence is missing the NE, so at query time, it cannot be retrieved by the English query. In terms of CLIR, this is a recall error. We refer to this as a *lost in retrieval* error.

In the QT approach, the English query is translated into Arabic using MT and additional resources. In the QT pipeline in Figure 1, the Arabic sentence is retrieved and then translated using MT. However, when the user sees the sentence, it appears irrelevant because there is no mention of the query. In CLIR, this would not be considered an error, since the sentence is relevant in Arabic. In TLIR, it is a precision error because an irrelevant result was returned. But the error is due only to a loss in MT adequacy and not due to the retrieval model. We refer to this as a *lost in translation* error.

When MT errors occur, they cause lost in retrieval errors in the DT model and any retrieval model that relies on translated documents. The QT model avoids lost in retrieval errors, but is still affected by lost in translation errors, where the relevant information is garbled or missing due to MT. In our experiments, we quantify these errors and their impact on the end-to-end TLIR system, and evaluate two approaches to mitigating these errors.

3 TLIR Evaluation Corpus

In the English-Arabic TLIR task, the system is presented with an English query and a set of Arabic documents. The goal is to find all Arabic sentences relevant to the query and return them machine translated into English. Each experimental setting consists of a CLIR model and an MT system; we experiment with four CLIR models and three types of translations (human translations, MT A and MT B). As in a standard IR evaluation, we run a set of English test queries on all systems and pool the top k returned translated sentences. Then we ask annotators to judge each query-sentence pair as Relevant or Not Relevant. We aggregate the judgments using mean average precision (MAP), defined below.

One of the challenges in evaluating TLIR is finding an appropriate evaluation corpus. The ideal corpus would enable us to measure both retrieval relevance and MT quality, so we would like a corpus

that has human translations (HTs) as well as query-result relevance judgments. CLIR corpora are very large datasets with query-document pairs annotated in the document language. It is expensive to translate such large corpora with MT and infeasible to translate manually. In contrast, MT test sets have reference translations, but are small in comparison.

To create the TLIR evaluation corpus, we augment a standard MT test set (NIST MT08 Arabic-English) with queries and sentence-level relevance judgments: we create 94 queries for 813 sentences in the newswire (NW) corpus, and 77 queries for 547 sentences in the web (WB) corpus. This is a tiny corpus by CLIR standards. We are explicitly trading off the size of the corpus in order to have a corpus with reference translations because we are ultimately interested in the impact of MT errors on TLIR. For each query-sentence pair, we collect two types of relevance judgments:

HT relevance: Relevance judgments on the HT of each sentence. Human translation is the upper bound for TLIR result translation quality.

MT relevance: Relevance judgments on each MT version of each sentence. Annotators judge MT relevance without seeing the HT, so sentences that are garbled during MT are judged irrelevant even when the HT version of the sentence is actually relevant.

3.1 TLIR Queries

Our task was inspired by the GALE distillation task, which was an open-ended, template-based cross-lingual question answering task with result translation. Since the queries in that task focused on NEs, we chose to base our queries on NEs. Specifically, annotators were asked "Which sentences contain facts about NE?" The instructions emphasized that merely mentioning the NE was neither necessary nor sufficient for relevance. For instance,"The US President visited France in 2012." is relevant to a query about Barack Obama despite not mentioning his name. It is not relevant to a query about the US, even though the US is mentioned, because there is no fact about the US. This type of query is interesting because it requires sentence-level MT understanding. If a sentence is completely garbled, it does not have a fact, so it is not relevant.

3.2 Evaluation Metric

For results using HT relevance, we report the mean average precision (MAP), which takes into account both recall and precision (Manning et al., 2008). MAP is a standard metric for evaluating ranked results that is commonly used in IR evaluations. This metric summarizes the overall performance of the system by taking the mean over all queries of the average precision across all levels of recall. More formally, for a query q, denote by Rel_q^{HT} the set of all HT sentences that are relevant to q. Then for each result d_q in a set of n ranked results, relevance, precision at k and average precision are defined as:

$$rel^{HT}(d_q) = \begin{cases} 1, & \text{if } d_q \in Rel_q^{HT} \\ 0, & \text{otherwise} \end{cases}$$

$$Prec(k) = \frac{\sum_{i=1}^{k} rel^{HT}(d_q^i)}{k}$$

$$AvePrec = \sum_{k=1}^{n} \frac{(Prec(k) \times rel^{HT}(d_q^k))}{|Rel_q^{HT}|}$$

MAP is the average of $AvePrec$ over all the queries.

When evaluating TLIR, a result is relevant only if it is actually relevant (in HT) *and* is perceived as relevant by the user (in MT). Let Rel_q^{MT} represent the set of all MT sentences that are relevant to q. We will consider a sentence d_q is relevant to a query q only if it is in Rel_q^{MT} in addition to Rel_q^{HT}. Formally, the relevance, precision and average prevision are defined as:

$$rel^{MT}(d_q) = \begin{cases} 1, & \text{if } d_q \in (Rel_q^{MT} \cap Rel_q^{HT}) \\ 0, & \text{otherwise} \end{cases}$$

$$Prec(k) = \frac{\sum_{i=1}^{k} rel^{MT}(d_q^i)}{k}$$

$$AvePrec = \sum_{k=1}^{n} \frac{(Prec(k) \times rel^{MT}(d_q^k))}{|Rel_q^{HT}|}$$

For a given retrieval model and MT test set, HT MAP measures how good the retrieval model is, independent of whether the end-user can understand the results, while MAP using MT relevance measures both retrieval and result understanding. In other words, MT MAP is essentially regular (HT) MAP with a penalty for results that cannot be understood in translation.

3.3 Analysis Methods

Since we have relevance judgments on both HT and MT, we can compare the TLIR performance upper bound (on HT) to the TLIR performance using MT, and quantify the percent lost due to MT. We can also

Experiment Name	Indexed Corpus	Annotated Results
Gold (HT)	Arabic, HT	HT relevance
Lost in Retrieval	Arabic, MT	HT relevance
Lost in Translation	Arabic, HT	MT relevance
End-to-end	Arabic, MT	MT relevance

Table 1: The four different types of analysis we perform on each TLIR system, using different combinations of human translation (HT) and MT.

Retrieval Model	Indexed Language
Query translation (QT)	Arabic
Document translation (DT)	English
SMLIR	Arabic and English
QT-rerank	Arabic

Table 2: QT and DT are the baseline models, and SMLIR and QT-rerank are hybrid models. In DT, SMLIR and QT-rerank, "English" may refer to either HT or MT, depending on the experiment. Note that even though QT-rerank does not index English, it does use English to re-rank the retrieved results.

use these two types of relevance judgments to isolate the effects of MT on retrieval and on result translation separately. In each TLIR setting, the MT is used in two ways: during retrieval, most of the CLIR models use the translated corpus to index and/or rank the results [2]; during relevance annotation, the translated results are presented to the user. By varying retrieval and relevance annotation between HT and MT, we can analyze our results for each setting four different ways:

Gold (HT): HT is used for indexing, ranking and result translation. This setting is an upper bound on TLIR system performance.

Lost in Retrieval (CLIR): MT is used for retrieval only; the results are annotated in HT, to remove the influence of MT errors on result understanding. This setting is similar to standard CLIR evaluations where result translation is ignored.

Lost in Translation: HT is used for retrieval, but results are judged in MT. This setting ignores the impact of MT errors on retrieval accuracy, and focuses on sentences that should be relevant but are judged irrelevant due to errors in result translation.

End-to-end (TLIR): MT is used for both retrieval and relevance annotation. This setting represents the full end-to-end TLIR system, where MT has an impact on both retrieval and result understanding.

[2]Comparing query translation methods is not the focus of this paper, so query translations are identical across all experimental settings. HT is never used for query translation.

4 Experiments

4.1 Query Extraction

Queries were created by running the Stanford NE recognizer (Finkel et al., 2005) on one of the HTs. This list of all possible NE queries for the corpus was filtered to remove near-duplicates and incorrectly tagged phrases. After relevance annotation, queries that had one or zero results were filtered. The average number of relevant sentences per query was 8 for NW and 5 for WB.

4.2 MT systems

We use two pre-existing state-of-the-art Arabic-English SMT systems with widely different implementations MT A was built using HiFST (de Gispert et al., 2010), a hierarchical phrase-based SMT system implemented using finite state transducers. It is trained on all the parallel corpora in the NIST MT08 Arabic Constrained Data track (5.9M parallel sentences). The first-pass 4-gram language model (LM) is trained on the English side of the parallel text and Gigaword 3. The second-pass 5-gram LM is a zero-cutoff stupid-backoff (Brants et al., 2007) estimated using 6.6B words of English newswire text.

MT B was built using Moses (Koehn et al., 2007), and is a non-hierarchical phrase-based system. It is trained on 3.2M sentences of parallel text using several LDC corpora including some available only through the GALE program (e.g., LDC2004T17, LDC2004E72, LDC2005E46 and LDC2004T18). The data includes some sentences from the ISI corpus (LDC2007T08) and UN corpus (LDC2004E13) selected to specifically add vocabulary absent in the other resources. The Arabic text is tokenized and lemmatized using the MADA+TOKAN system (Habash et al., 2009). The system uses a 5-gram LM that was trained on Gigaword 4. Both systems are tuned for BLEU score using MERT.

The BLEU scores for MT A and MT B on NW are 51.32 and 51.23, respectively; and on WB, 36.15 and 37.60, respectively. Based on these scores, MT A and MT B are comparable in quality.

4.3 Relevance Annotation

The same HT that was used to create the queries was also used as the gold standard for collecting relevance judgments. While an ideal CLIR evaluation would have gold relevance measured in the source language, Arabic annotations were difficult to get,

and we opted for a large quantity of reference annotations instead of a small set of source language annotations. The relevance judgments were done on Amazon Mechanical Turk (AMT) using Crowdflower[3] to filter for trusted workers. Relevance judgments on AMT have been shown to have high agreement with TREC raters (Alonso and Baeza-Yates, 2011). We pooled the top-10 results of each TLIR run and crowd-sourced 3 relevance judgments each on three versions of each query-sentence pair: the HT and both MT versions.

4.4 Baseline Systems

Document Translation For DT, the translated English sentences are indexed, and the query is run against the indexed translations. An index is built for each version of the corpus: HT, MT A and MT B.

Query Translation For QT, the Arabic source sentences are indexed, and the English query is translated into an Arabic query in order to do monolingual search in Arabic. The QT method uses a cascaded approach to query translation, and then converts the translated query into a structured query (Pirkola, 1998). The first step in translation is a NE dictionary built from Wikipedia, the CIA world factbook and the NEs from the Buckwalter analyzer dictionary (Buckwalter, 2004). Since all of the queries are NEs, this dictionary is a high-precision, but low recall resource. If the translation is not found, a phrase table from MT B is used as a dictionary. The final back-off searches a large corpus of machine translated documents, which is a lower precision resource. This resource simulates the task context that a large CLIR corpus would provide; prior work has shown that words that are deleted in one sentence are often successfully translated in others (Ma and McKeown, 2009). If the translation is still not found, the original query is expanded using synonyms extracted from Wikipedia, and then the cascaded translation is applied again.

5 Analysis of Baselines

Figure 2 shows results from all experimental settings; in this section we discuss the results for QT and DT. All of the models are evaluated on the newswire and web genres (the top and bottom charts, respectively) and MT A and MT B (left and right, respectively). Each setting is analyzed four different ways, summarized in Table 1.

This analysis demonstrates the strengths and weaknesses of the different models. The QT model is unaffected by retrieval errors, since retrieval is done in the document language only (MAP is the same for Gold and Lost in Retrieval bars), but is significantly degraded by result translation errors. For instance, for QT in the MT A NW setting, the MAP is 19% lower when results are judged in MT instead of HT (Lost in Translation vs. Gold).

On the other hand, lost in retrieval errors have a significant impact on the DT model because it fails to retrieve sentences with MT errors. Across various conditions, MAP for DT decreases by 16-39% when retrieval is done using MT instead of HT (Lost in Retrieval vs. Gold).

In all cases, QT has higher MAP than DT in the Lost in Retrieval setting, and the reverse is true in the Lost in Translation setting. If we ignore result translations, as in a standard CLIR evaluation, the QT model is better than DT. When we ignore MT errors during retrieval but annotate relevance in MT, the DT model is better than QT. In other words, translated sentences that "look" relevant are ranked higher by DT than by QT.

In the End-to-end evaluation, where MT is used for both retrieval and relevance annotation, the results depend on the genre. In the NW genre, DT has higher end-to-end MAP than QT. In the WB genre, even though DT has higher MAP than QT when HT is used for retrieval (the lost in translation setting), its much lower retrieval accuracy (the lost in retrieval setting) ultimately makes the end-to-end DT MAP lower than QT. We attribute the poor performance of DT on WB to the difficulty of the genre: the text is informal and harder to translate, and as a result, there are more MT errors. If we consider the drop from Gold to End-to-end, we can see that it is larger in the WB genre than in NW (across all conditions). The relatively worse MT quality has a measurable impact on the TLIR task.

Across all settings, MT B has higher MAP than MT A. Although both systems have similar BLEU scores on the evaluation corpora, when we looked at the examples, we found that MT B had much better NE translation, which is crucial for this task. The extrinsic TLIR evaluation shows the relative usefulness of the MT systems, which the intrinsic mea-

[3] http://www.crowdflower.com

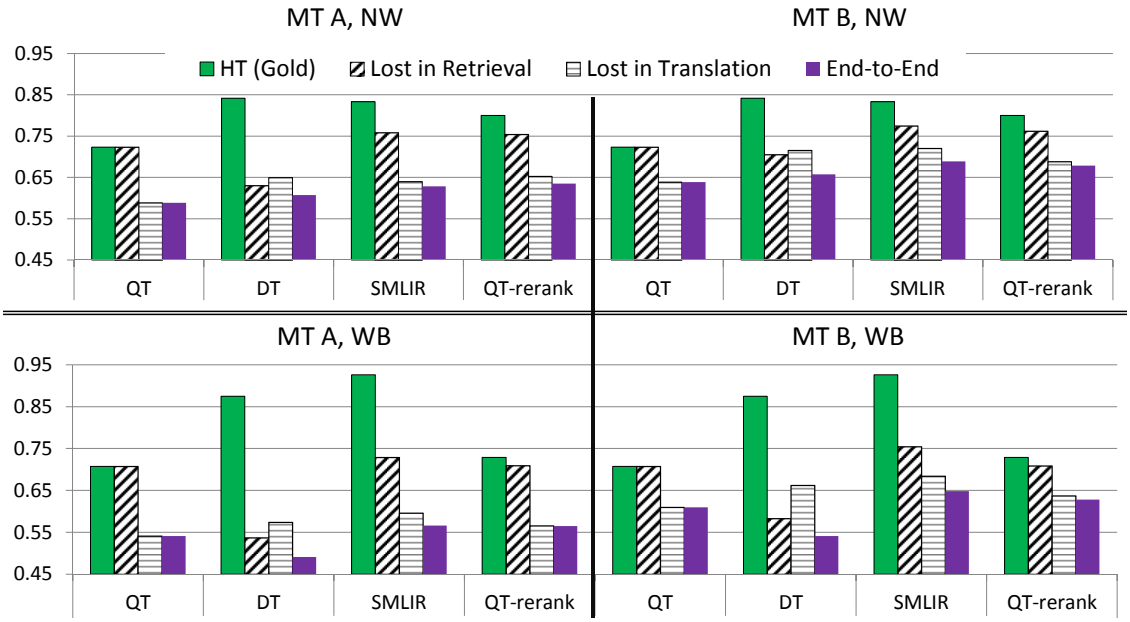

Figure 2: The MAP of different combinations of MT systems and retrieval models using the analysis methods from Table 1.

sures of MT quality could not capture.

Similarly, results of the intrinsic CLIR evaluation do not always match results from the full end-to-end TLIR task. For instance, in the NW genre, QT has higher MAP than DT in the CLIR evaluation (Lost in Retrieval), but DT has higher (translated) MAP than QT in the TLIR evaluation (End-to-End). This highlights the limitations of evaluating CLIR models without taking result translation into account.

6 Lost and Found in Retrieval

The baseline QT and DT models are both severely affected by errors in MT. A better retrieval model would retrieve all the relevant results, but rank the translations that appear relevant the highest. Several hybrid methods for combining QT and DT exist. For instance, McCarley (1999) and Chen and Gey (2003) describe methods for combining the results of separate QT and DT searches using re-ranking. We chose to use the simultaneous multilingual IR (SMLIR) model from (Parton et al., 2008) because it requires only a single index and a single search at runtime, and does not require tuning a re-ranker.

SMLIR: In the SMLIR model, each sentence is indexed as a bilingual sentence with different fields for each language, and the structured query is composed of both query-language and document-language terms. In a CLIR evaluation without result translation, SMLIR outperformed both QT and DT.

QT-rerank: The SMLIR model uses an offline MT system to do a full corpus translation, which maybe infeasible depending on the size of the corpus and the available compute power. As an alternative, we introduce the QT-rerank model, which uses the same intuition as the SMLIR model, but relies on an online MT system to translate search results only rather than the full corpus. The QT-rerank model retrieves results using QT only, then translates the results using an online MT system, and finally re-ranks the translated results using DT.

Table 2 summarizes the different CLIR models we evaluate, and Table 3 compares the effect of an MT error on result ranking in different models.

6.1 Results

The hybrid models use the complementary strengths of QT and DT to their advantage to mitigate lost in

	MT	QT	DT	SMLIR
Correct MT	Vanunu	1	1	1
OOV	fEnwnw	1	-	2
Deletion	(empty)	1	-	2

Table 3: Given a sentence with a single Arabic NE, the rows show three different MT outputs and their ranking according to the different CLIR models. DT retrieves only the correct MT. QT retrieves all of them, but cannot distinguish between them. SMLIR retrieves all of them and ranks the correct MT highest (as would QT-rerank).

retrieval errors. In the lost in retrieval setting, the SMLIR model always does better than either QT or DT alone. The document language part of the query ensures that relevant sentences are not missed during retrieval due to MT errors, and the query language part of the query gives higher rank to retrieved sentences that are translated correctly.

Although the hybrid models were designed to improve retrieval only, SMLIR also helps lost in translation errors in the web genre. This is because sentences that match in both the query language and the document language are ranked higher than sentences that match in only one language, so the highest ranking sentences tend to be those that are actually relevant in translation.

By reducing lost in retrieval errors and some lost in translation errors, the SMLIR model is able to achieve higher end-to-end MAP than either QT or DT alone across all settings.

The QT-rerank model also has better end-to-end performance then either DT or QT alone. QT-rerank does not use the query language during the first-pass retrieval, only during the second pass reranking. This means that sentences ranked low by QT but high by DT may be missed by QT-rerank, but will be returned by SMLIR. For MT A, QT-rerank does as well or better than SMLIR, while for MT B, SMLIR has better end-to-end performance. In other words, QT-rerank has the advantages of SMLIR without the overhead of translating the full corpus.

7 Lost and Found in Result Translation

Figure 3 shows the Lost in Translation errors (where sentences were judged relevant in HT but not in MT) as a percent of query-sentence pairs. In these cases, even a retrieval model with perfect recall would not help, since the problem is in the MT only.

Automatic post-editors (APEs) can be used to correct specific types of errors in MT output, for example: English determiner selection (Knight and Chander, 1994), grammatical agreement in Czech (Mareček et al., 2011), and grammatical errors in English (Doyon et al., 2008) and Swedish (Stymne and Ahrenberg, 2010). APEs are also useful for adapting the output of task-agnostic MT systems to the needs of a particular task. This is important, since application developers often use out-of-the-box MT systems that cannot be re-trained or re-tuned. For the TLIR task, the goal is to use APE to correct MT errors before showing translated results to the user.

7.1 Adequacy-Oriented APEs

Since adequacy is crucial for TLIR, we use two adequacy-oriented APEs that are described in more detail in (Parton et al., 2012). There, manual evaluation showed that 30-56% of the translations edited by the APEs had improved adequacy. In this paper, we use the TLIR evaluation corpus as a novel testbed for these APEs, to see whether improved adequacy leads to an improvement in TLIR relevance.

Both APEs focus on phrase-level adequacy errors that impact TLIR: content words that are not translated at all, content words that are translated to function words, and mistranslated NEs. They use part of speech tags in the source and target language as well as word alignments from the MT system to flag possible errors (e.g., a noun translated into a determiner would get flagged.) Then, the resources from query translation (described in 4.4) are used to find a list of suggestions for each flagged Arabic word or phrase. The APEs differ in how they apply the suggestions:

APE1: Rule-Based APE: The rule-based APE takes the top-ranked translation suggestion and inserts it into the translated sentence. A simple set of rules uses the word alignments of the adjacent Arabic words to determine where to insert the suggested word, and whether to overwrite the existing translation or insert the suggestion next to it. APE1 always edits flagged errors unless the rules fail to determine an insertion point: 85-100% of sentences with errors were post-edited.

APE2: Feedback APE: The feedback APE passes the full list of translation suggestions back to the MT system and re-translates the source sentence using the same MT system, given the new translations. The advantage of APE2 is that the MT system can modify other parts of the sentence to make the post-edited output more fluent, rather than just inserting a word in the middle of an existing translation. The behavior of the feedback APE depends

MT	And was released in April, 2004 after he had spent 18 years in prison
APE1	And was vanunu released in April, 2004 after he had spent 18 years in prison
APE2	Vanunu was released in April, 2004 after he had spent 18 years in prison

Table 4: Adequacy-oriented APE on the MT from Figure 1.

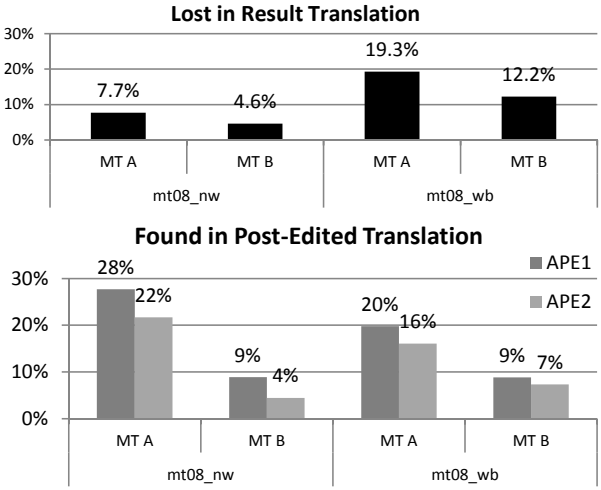

Figure 3: a) The percent of sentence-query pairs with lost in translation errors, where a relevant HT result became irrelevant in MT. b) The percent of lost in translation errors that were corrected by each APE.

on how each MT system handles the feedback: for MT A, it modified 83-90% of sentences with flagged errors, while for MT B, it changed only 63-74%.

Figure 3 shows the differences between the APEs: while both APEs add the deleted NE back into the sentence, APE2 results in a more fluent translation.

7.2 Results

Both APEs were run on all MT sentences. We collected 5 relevance judgments on post-edited sentences that were lost in translation errors (relevant in HT but not in MT), and on a sample of post-edited sentences that were relevant in both MT and HT, to measure the effect of errors made by the APEs.

Figure 3 shows the percent of errors that were corrected by the APEs – sentence-query pairs that were relevant in HT, irrelevant in MT, and relevant in post-edited MT. The APEs had a positive impact on result relevance, with 16-28% of MT A errors and 4-9% of MT B errors corrected. Since APE1 is more aggressive than APE2 in editing more often, it corrects more errors. Both APEs rarely caused a relevant MT sentence to look irrelevant (less than 0.01% of the time). These results show that adequacy improvements in the translated results can lead to improved end-to-end TLIR relevance.

8 Discussion

We presented a TLIR evaluation that quantified the impact of result translation on retrieval and translated relevance, as well as on the end-to-end system. The QT and DT baselines that simply pipelined CLIR and MT were significantly degraded by MT errors compared to HT upper bounds. The TLIR evaluation corpus that we created was a crucial resource for this analysis, and also provided an experimental framework for testing proposed improvements over the baseline CLIR and MT systems. Task-based evaluation of MT highlights specific errors that affect translation usefulness that are not evident from standard MT evaluations: while MT A and MT B had similar BLEU scores on newswire, MT A had worse translated relevance because MT B was better at translating NEs.

Similarly, while the standard QT CLIR model performed better in the intrinsic evaluation (without result translation), DT had better relevance in translation. Results showed that the hybrid model SMLIR exploited these complementary advantages to outperform both QT and DT. We also introduced a new hybrid model, QT-rerank, which performed as well as SMLIR, but uses online result translation instead of translating the full corpus offline.

We also experimented with modifying the MT output to improve result relevance. We applied two adequacy-oriented APEs to sentences that were relevant in HT, but irrelevant in MT. The APEs corrected adequacy errors such as missing content words and mistranslated NEs, which are particularly crucial to our TLIR task. The APEs improved sentence relevance 4-28% of the time.

SMLIR and QT-rerank were successful because they integrated result translations into the retrieval model, while the adequacy-oriented APEs used the CLIR task context to find translation suggestions that ultimately improved MT adequacy. In both cases, by coupling the MT and CLIR more closely, we were able to improve the end-to-end TLIR system. We believe that other cross-lingual applications could also benefit from tighter integration with MT.

Recent work in MT confidence estimation and task-embedded MT offers more opportunities for improving the quality and ranking of translated results for TLIR. TrustRank (Soricut and Echihabi, 2010) ranks the quality of translations from good to bad. Specia et al. (2011) use confidence estimation to predict MT adequacy, with a special emphasis on NEs. In the future, it would be interesting to apply these confidence estimators to the TLIR task, to rerank translations by both quality and relevance.

Acknowledgments

Thanks to Bill Byrne, Gonzalo Iglesias and Adrià de Gispert for letting us use their HiFST system and collaborating with us on the HiFST post-editor. This paper is based on work supported by DARPA on Contract Nos. HR0011-12-C-0016 and HR0011-12-C-0014. Any opinions, findings, and conclusions expressed herein do not necessarily reflect the views of DARPA.

References

Omar Alonso and Ricardo Baeza-Yates. 2011. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pp. 153–164. Springer.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *EMNLP-CoNLL*, pp. 858–867.

Tim Buckwalter. 2004. Buckwalter Arabic morphological analyzer version 2.0. *LDC2004L02, ISBN 1-58563-324-0*.

Aitao Chen and Fredric C. Gey. 2003. Combining query translation and document translation in cross-language retrieval. In *CLEF*, volume 3237 of *Lecture Notes in Computer Science*, pp. 108–121. Springer.

Sherri L. Condon, Dan Parvaz, John S. Aberdeen, Christy Doran, Andrew Freeman, and Marwan Awad. 2010. Evaluation of machine translation errors in English and Iraqi Arabic. In *LREC*.

Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. 2010. Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. *Computational Linguistics*, 36(3):505–533.

Jennifer Doyon, Christine Doran, C. Donald Means, and Domenique Parr. 2008. Automated machine translation improvement through post-editing techniques: analyst and translator experiments. In *AMTA*, pp. 346–353.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL*, pp. 363–370.

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. 2001. Improving query translation for cross-language information retrieval using statistical models. In *SIGIR*, pp. 96–104.

Meghan Lammie Glenn, Lauren Friedman, Stephanie M. Strassel, Zhiyi Song, Gary Krug, Kazuaki Maeda, Haejoong Lee, and Christopher Caruso, 2011. *Handbook of Natural Language Processing and Machine Translation*, chapter Human Annotation, pp. 14–64. Springer.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *SIGIR*, pp. 267–274.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *International Conference on Arabic Language Resources and Tools (MEDAR)*, pp. 242–245.

Dilek Hakkani-Tür, Heng Ji, and Ralph Grishman. 2007. Using Information Extraction to Improve Cross-lingual Document Retrieval. In *Proceedings of Multi-source, Multilingual Information Extraction and Summarization, RANLP*.

Benjamin Herbert, Gyorgy Szarvas, and Iryna Gurevych. 2011. Combining query translation techniques to improve cross-language information retrieval. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pp. 712–715.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *ACL-HLT*, pp. 389–397.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *AAAI*, pp. 779–784.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL*, pp. 177–180.

Wei-Yun Ma and Kathleen McKeown. 2009. Where's the verb?: correcting machine translation during question answering. In *ACL-IJCNLP*, pp. 333–336.

Walid Magdy and Gareth J.F. Jones. 2011. An efficient method for using machine translation technologies in cross-language patent search. In *CIKM*, pp. 1925–1928.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondrej Bojar. 2011. Two-step translation with grammatical post-processing. In *WMT*, pp. 426–432.

J. Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *ACL*.

Douglas W. Oard and Julio Gonzalo. 2003. The CLEF 2003 interactive track. In *CLEF*. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318.

Kristen Parton, Kathleen R. McKeown, James Allan, and Enrique Henestroza. 2008. Simultaneous multilingual search for translingual information retrieval. In *CIKM*.

Kristen Parton, Nizar Habash, Kathleen McKeown, Gonzalo Iglesias, and Adrià de Gispert. 2012. Can automatic post-editing make MT more meaningful? In *EAMT*.

Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *SIGIR*, pp. 55–63.

Maja Popović and Hermann Ney. 2007. Word error rates: Decomposition over POS classes and applications for error analysis. In *WMT*, pp. 48–55.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *ACL*, pp. 612–621.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *MT Summit XIII*.

Sara Stymne and Lars Ahrenberg. 2010. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *Conference on Arabic Language Resources and Tools*.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *LREC*, pp. 697–702.