Identifying Effective Translations for Cross-lingual Arabic-to-English User-generated Speech Search

Ahmad Khwileh[1], **Haithem Afli**[2], **Gareth J. F. Jones**[2] and **Andy Way**[2]
ADAPT Centre, School of Computing
Dublin City University
Dublin 9, Ireland
(1) ahmad.khwileh2@mail.dcu.ie
(2) {hafli, gjones, away}@computing.dcu.ie

Abstract

Cross Language Information Retrieval (CLIR) systems are a valuable tool to enable speakers of one language to search for content of interest expressed in a different language. A group for whom this is of particular interest is bilingual Arabic speakers who wish to search for English language content using information needs expressed in Arabic queries. A key challenge in CLIR is crossing the language barrier between the query and the documents. The most common approach to bridging this gap is automated query translation, which can be unreliable for vague or short queries. In this work, we examine the potential for improving CLIR effectiveness by predicting the translation effectiveness using Query Performance Prediction (QPP) techniques. We propose a novel QPP method to estimate the quality of translation for an Arabic-Engish Cross-lingual User-generated Speech Search (CLUGS) task. We present an empirical evaluation that demonstrates the quality of our method on alternative translation outputs extracted from an Arabic-to-English Machine Translation system developed for this task. Finally, we show how this framework can be integrated in CLUGS to find relevant translations for improved retrieval performance.

1 Introduction

The growing archives of online digital content are increasingly diverse in style, media and the language used. Within this content the balance between use of languages is very uneven. An important case of this effect is Arabic multimedia content where the amount of content available is proportionally very small. This results in a significant demand from bilingual Arabic speakers to access information in other languages, most notably English. Cross Language Information Retrieval (CLIR) is an effective tool to bridge the language barrier between user search queries in one language and the target documents in another language (Oard and Diekema, 1998; Khwileh et al., 2016). The simplest and most commonly adopted approach in CLIR is to use machine translation (MT) to translate the user's query. In most cases, MT is used as a black box as an input stage to an otherwise unchanged monolingual search system Many different MT systems have been studied in CLIR research for different tasks, e.g. (Oard and Hackett, 1998; Magdy and Jones, 2014). However, no single MT system has been reported to be effective for all CLIR tasks.

The effectiveness of an MT system for CLIR is primarily evaluated by examining the retrieval quality associated with the translated queries. We follow this practice in this paper, by considering translation quality in terms of measured IR performance on an experimental test collection. We investigation concentrates on a cross-lingual user-generated speech search (CLUGS) task (Khwileh et al., 2015). In this work, we propose a prediction framework that utilises Query Performance Prediction (QPP) methods to estimate expected IR performance for specific query translation based both on the translated query itself and the output of the translation process. As part of our investigation we explore the use of QPP to select from an N-best list of alternative translations for q query generated by an statistical MT systems.

In the next section we give some background and describe the motivation behind our CLUGS task. Section 3 gives an overview of the QPP approaches that we study in this investigation. Section 4 introduces our proposed prediction framework for CLUGS. Section 5 outlines our experimental settings. Section 6 evaluates the proposed framework and section 7 shows this approach can indeed be utilised for finding relevant translations in CLUGS. Section 8 concludes, together with some avenues for future work.

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 100–109,
Valencia, Spain, April 3, 2017.

2 Cross-lingual Arabic-to-English Search for User-generated Speech

The current explosive growth in internet-based social media networks content is creating massive volumes of online multimedia content. This includes User-Generated Speech (UGS) content which is being uploaded to social media sites websites such as YouTube and Facebook. These increasing quantities of UGS data, together with its complex and inconsistent structure, are creating the need for sophisticated Spoken Content Retrieval (SCR) systems to locate relevant material. This presents new challenges and exciting opportunities for IR research. SCR technologies require the combination of speech processing technologies with IR methods. SCR typically utilises Automatic Speech Recognition (ASR) to generate text transcripts of spoken audio. At a simple level, SCR can be considered as the application of IR techniques to ASR transcripts. However, errors in ASR transcripts and the nature of spoken content present significant challenges for SCR (Larson and Jones, 2012).

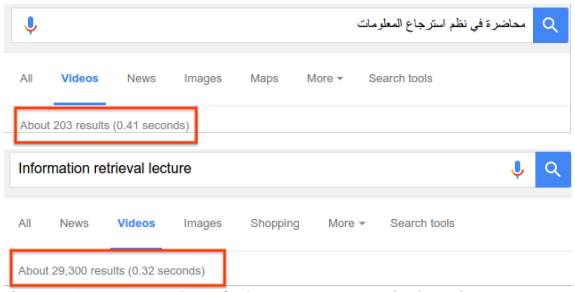

Figure 1: Example of the content variation issue across languages: Video search results for Arabic and English queries.

Beyond these challenges in SCR, further challenges are raised in a multilingual search setting. As noted earlier, one of the scenarios for multilingual search is CLIR where a searcher uses a query in one language to find relevant content in another one, where relevant content in the query language is either sparse or not available.

This is a particularly notable issue for Arabic which is spoken by an estimated 420M speakers in different countries, making it one of the most-spoken languages in the world. Arabic has been the language with the largest growth in Internet users in the last decade with an estimated 2500% growth. In 2016, there were an estimated 168M Internet users with 45% Internet penetration (Internetworldstats.com, 2016). However, the Arabic content available online is still minimal, estimated as being less than 0.1% of the Internet content. The massive gap between available content and speakers of the language means that bilingual Arabic speakers will often seek relevant content in another language. To illustrate this situation consider the example in Figure 1. This shows the search engine with a simple Arabic query محاضرة في نظم استرجاع المعلومات and the equivalent English query *Information retrieval system lecture*

The *Google* video search engine[1] located more than 29,000 matching results in English with all top-ranked results being relevant with high-quality metadata. However, for Arabic, only 203 matching results were located with only one of the top-10 results indicated as relevant.

2.1 Related Work

Addressing CLIR for Arabic speakers provides a real-world use case where research into improved CLIR is important due to its linguistic challenges and political importance (Darwish et al., 2014).

Relevant linguistic features include the complexity of morphological and syntactic structures that require special processing. Indeed, MT for Arabic to English is considered one of the most difficult challenges in CLIR, and effective techniques working with special characteristics of the Arabic language are required (Alqudsi et al., 2014). Previous CLIR work on Arabic has been limited to standard text-based TREC 2001/2002 CLEF[2] tracks (Oard and Gey, 2002; Besançon et al., 2009; Darwish et al., 2014). The data collections used in these tasks were standard Text Arabic news collection collected from AFP. Another larger AFP newswire Arabic collection was released by the INFILE Track in CLEF 2008/2009 (Besançon et al., 2009), but unfortunately received no participation. To date most work on Arabic CLIR has actually focused on the other side of the story, i.e. the retrieval of Arabic documents based on English queries (English-to-Arabic CLIR). Which enabling access to information from Arabic sources, does not address the needs to Arabic speakers. In this work, we study a CLIR task that enables Arabic users to search for the relevant spoken content from the English web. In previous work we investigated the use of Google Translate [3] as a black box off-the-self MT system for this task (2015) We found that the main challenges of this task arise due to noise in the search index for the user-generated data, and how Arabic translation errors can significantly harm retrieval effectiveness.

Despite the problems in translation problems for CLIR encountered when using off-the-shelf

[1] Retrieved from www.google.com/video on 2016-12-01
[2] http://clef2016.clef-initiative.eu/
[3] https://cloud.google.com/translate/

MT systems such as Google and Bing Translate[4], have been observed to outperform open-box MT systems developed by CLIR researchers for many language pairs (Zhou et al., 2012). For instance, during the CLEF 2009 workshop (Leveling et al., 2009a; Leveling et al., 2009b), the best performing non-off-the-shelf MT achieved just 70% of the performance achieved by Google Translate. However, in our earlier work we found that the use of black-box MT for Arabic is still ineffective compared to other languages pairs (such as French-to-English CLIR) (2015).

From examination of the behaviour of MT systems, it is clear that while the "best" translation produced by the MT does not always produce the most effective translation for optimal CLIR performance, better translations are often produced with lower translation confidence by the MT system. In this investigation, we seek to use Query Performance Prediction (QPP) methods developed in the IR community, to improve CLEF effectiveness for Arabic-English search using an open-box MT system. We then study the effectiveness of this approach against standard online black-box MT for a CLUGS search task. In the next sections we describe these QPP techniques and how we utilize them in our CLUGS task.

3 Query Performance Prediction

The motivation behind QPP methods in IR is to estimate the performance of the query at retrieving relevant documents. This inference can be used to tune the retrieval settings to maximize the overall system effectiveness. QPP is divided into pre- and post-retrieval methods. In pre-retrieval QPP, prediction is based on analysing the query difficulty (Hauff et al., 2008; He and Ounis, 2004; He and Ounis, 2006). The estimated query difficulty defines that, given a certain query, whether relevant content is hard (low retrieval performance) or easy (high retrieval performance) to find. Thus, difficulty can be used as an indication of the retrieval performance of the current query. In post-retrieval QPP, the retrieval results of the query are analysed to estimate its performance (Kurland et al., 2011; Shtok et al., 2012). Pre-retrieval methods are more efficient than post-retrieval, causing less overhead to the retrieval system since no retrieval is required for the prediction. In this work we study the application of both pre- and post- methods QPP to predict the translation quality of queries in CLIR. In the next sections we describe the QPP approaches we use in this task.

[4]https://www.microsoft.com/en-us/translator/translatorapi.aspx

3.1 Pre-retrieval QPP

Existing approaches to pre-retrieval QPP are based on measuring the statistics/characteristics of the query terms calculated over the index collection. The most widely used and effective techniques rely on the Inverse Document Frequency (IDF) of query terms, called IDF-based QPP. IDF-based QPP approaches are implemented by taking an aggregation of the IDF values across the query terms such as AvIDF (Average of IDF values), the SUMIDF (the sum of all values) or MAXIDF (the maximum value) (Cronen-Townsend et al., 2002). The IDF value for a term in this work has generally been calculated using the INQUERY formula explained in (Allan et al., 1995) and (He and Ounis, 2006). Another common IDF-based QPP is the Averaged Inverse Collection Term Frequency (avICTF) of the query terms (Plachouras et al., 2004; He and Ounis, 2006). The formula for this predictor is explained in detail in (He and Ounis, 2006). IDF-based predictors have shown positive correlation with query performance over multiple standard IR tasks (Plachouras et al., 2004; He and Ounis, 2004; Hauff et al., 2008; Hauff, 2010).

Other pre-retrieval QPP methods are based on analysing the linguistic features of the query terms such as the the query length (AvQL) which is based on the average number of content words (non stop-words) in a query (He and Ounis, 2004; Mothe and Tanguy, 2005; He and Ounis, 2006), and Query Scope (QS) which makes use of the document frequencies (DF) of the terms (He and Ounis, 2004; He and Ounis, 2006). A higher DF of the query terms indicates that they are very common, and so probably not helpful for finding relevant documents, as they would result in a lower effectiveness of the query.

A more complex technique proposed by Zhao et al. (2008), is the Summed Collection Query similarity (SCQ). SCQ approaches utilise both Term Frequency (TF) and IDF to predict the query performance. Similar to the IDF-based QPP, there are also three different aggregation methods of SQC across the query terms. AvSQC, takes the average across the query terms; SumSQC, which takes the sum of all resultant similarities; and MaxSQC which takes the maximum value among them. SQC is explained in detail in (Zhao et al., 2008; Hauff, 2010). Zhao et al. (2008) also proposed another QPP method that is computationally more expensive called VarTFIDF. which is based on

the distribution of the TF.IDF weights (Zobel and Moffat, 2006) across the query terms. Similar to SQC and IDF QPP approaches, VarTFIDF QPP has a three different versions (SUM,MAX,Avg) based on the used aggregation across the query terms.

In this work, we argue that IDF is not a good predictor for this task. This argument is also supported by our initial investigation of the problem and the following hypothesis. By definition, IDF gives a higher weight to unique terms across the search collection. While this might be useful for a retrieval model to rank documents, using IDF is not reliable for QPP since it also gives high values for translation candidates which are *misleading* terms. We define misleaders as terms that are rare across the collection (hence having high IDF values), but not relevant to the topic of the current query. These misleaders can result in query topic drift (Mitra et al., 1998) and thus negatively impact on retrieval effectiveness. Another source of misleader terms is words which are Out-of-Vocabulary (OOV) with respect to the MT. In this situation the MT system produces incorrect translations of terms which the MT system cannot by definition translate correctly.

To deal with misleaders arising from IDF issues, we propose a new simple prediction technique which we refer to as the Average Term Fluency (AvgFL). Term fluency estimates whether a query contains the same terms that appear in relevant documents. Higher fluency is assumed to lead to better query-document matching, and hence improved QPP effectiveness. We rely on the collection frequency (cf) of each term to indicate its fluency on the given collection D. The cf is normalized by the DF to penalize non-helpful terms which appear in all documents in collection. The proposed AvgFL is calculated as shown in Equation (1); where k is the number of t terms in query Q, cf_t is the cf which is the number of times t appears in the collection D. df_t indicates the DF which is the number of documents contains the term t.

$$AvgFL(Q) = \frac{1}{k}\sum_{t \epsilon Q}^{k}(log(cf_t+1)/(log(df_t+1)+1)) \quad (1)$$

3.2 Post-Retrieval QPP

State-of-the-art post-retrieval QPP techniques use information induced from analyzing the retrieval scores $Score(d)$ of the results set ${D_q}^{[res]}$ produced by retrieval method M, where ${D_q}^{[res]}$ represents the list of document ids retrieved for a query together with their ranks $\mathcal{R}i$ and scores $Score(d)$ sorted according to their relevancy to a query q.

In probabilistic terms, the resultant score $Score(d)$ of a document d represents the estimated relevance probability r of a document d with respect to q $Score(d) \equiv \mathcal{P}(d|q,r)$. These QPP methods are based on analyzing the performance of the top k ranked documents, which includes all documents that have rank $\mathcal{R}i$ that is less than k ($\forall d \epsilon {D_q}^{[res]} d_{\mathcal{R}i}$ where $0 \leqslant ri \leqslant k$) (Zhou and Croft, 2007; Shtok et al., 2012).

WIG is a well-established QPP technique based on the weighted entropy of the top k ranked documents (Zhou and Croft, 2007). This technique works by comparing the scores of the top-k documents $\forall d \epsilon {D_q}^{[k]} Score(d)$ to that obtained by the corpus $Score(D)$. WIG is defined in equation (2).

$$WIG(q,M) = \frac{1}{k}\sum_{d \epsilon D_k} \frac{1}{\sqrt{|q|}}(Score(d) - Score(D)) \quad (2)$$

Another similar post-retrieval QPP technique is the Normalised Query Commitment (NQC) (Shtok et al., 2012). This technique is based on estimating the potential amount of query drift in the list of top k documents by measuring the standard deviation of their obtained retrieval scores. A high standard deviation indicates reduced topic drift and hence probable improved retrieval effectiveness. NQC is defined in equation (3) where $\overline{\mu} = \frac{1}{k}\sum_{d \epsilon {D_q}^{[k]}} Score(d)$.

$$NQC(q,M) = \frac{1}{Score(D)}\sqrt{\sum_{d \epsilon {D_q}^{[k]}} \frac{1}{k}(Score(d) - Score(\overline{\mu})} \quad (3)$$

Both WIG and NQC are tuned to have a strong linear relationship with the performance of the query in which the only variable that needs to be decided is the top-k documents.

For our task, we introduce a modified version of the WIG called Weighted Relevancy Gain (WRG) that focuses on the scores of the top-ranked assumed *relevant* documents vs other top-ranked but assumed *non-relevant* documents. Unlike previous predictors, this approach assumes that the set top-k documents ${D_q}^{[k]}$ for each query is composed of two subsets ${D_q}^{[rel]}$ and ${D_q}^{[nrel]}$ defined as follows: ${D_q}^{[rel]}$ is the set of rel relevant documents that are assumed relevant for query q where $\forall d \epsilon {D_q}^{[res]}$ ($d_{\mathcal{R}i}$ where $0 \leqslant \mathcal{R}i \leqslant rel < k$), and ${D_q}^{[nrel]}$ is the set of documents that are assumed non-relevant and ineffective for Relevancy. These documents are ranked among the top-k documents and right after the rel documents ($rel < nrel < k$) as in $\forall d \epsilon {D_q}^{[res]}$ ($d_{\mathcal{R}i}$ where $rel < \mathcal{R}i \leqslant nrel$).

The WRG predictor aims to analyze the quality of the rel documents by measuring the likelihood that they contain significant variation. This is estimated by measuring the weighted entropy of the assumed rel documents against the

top-ranked yet non-relevant $nrel$ set of documents. Unlike WIG, which uses the $centroid$ of all non-relevant documents ($Score(D)$), WRG uses the $centroid$ of the $nrel$ documents scores : $Cnrel \equiv Cent(Score(D_q{}^{[nrel]})) \equiv \frac{1}{nrel} \sum_{d \epsilon D_q{}^{[nrel]}} Score(d)$ as a reference point for estimating of the effectiveness as shown in equation (4).

$$WRG(q,D_{rel}) = \frac{1}{rel} \sum_{d \epsilon D_{rel}} \frac{1}{\sqrt{|q|}} \left(\frac{Cnrel}{Score(d)}\right) \quad (4)$$

WRG requires 2 parameters: the number of rel documents and the number of $nrel$ documents to perform the actual estimation.

4 Using QPP for CLUGS

We propose to utilize QPP for CLUGS as follows. Assume T_q is an MT translated version of q and $T_q{}^{[n]}$ is the list of n-best translations generated by an MT translation system T. Assuming $\mathcal{Q}$ is the event of being an effective translation of q for getting relevant content in CLUGS, the goal of this prediction task is to estimate $\mathcal{P}(T_q|q,Q)$ (the likelihood of the translation T_q given that a relevance event happens for q), which seeks to answer the following question :

What is the probability $\mathcal{P}(.)$ for each translation candidate T_q from the top n-list generated by translation system T being an effective translation $\mathcal{Q}$ of a query q for CLUGS?

Our proposed framework relies on QPP to rank the best translations $T_q{}^{[n]}$ generated by MT system T based on the probability function $\mathcal{P}(T_q|q,Q)$. We use the previously explained QPP methods in section 3 to predict the retrieval effectiveness of each translation candidate T_q. For example, we assume that AvICTF can be taken as prediction function $\mathcal{F}$ to indicate the effectiveness of translations candidates T_q as $\mathcal{P}(T_q|q,Q) \equiv \mathcal{F}(T_q) \equiv AvICTF(T_q)$.

5 Experimental Setup

In order to evaluate QPP for our CLUGS task we configured three modules as follows. A CLIR system, an MT system to generate the N-best translations, and a QPP system to parse each query candidate of the n-best list and assign a prediction value to it.

The CLUGS task is similar to the one described in (Khwileh et al., 2015). The task is based on the blip1000 collection which contains 14,838 transcripts automatically extracted using an ASR system from videos which were uploaded to a video-sharing website by 2,237 different uploaders, covering a 25 different topics (Schmiedeke et al., 2013).

For the query topic set, we use a modified monolingual *adhoc* version of the 60 different original English topics developed within the MediaEval 2012 Search and Hyperlinking task[5] which was developed by Khwileh et al. (2016).

To setup the CLIR system, similar to the procedure adopted in our earlier investigation (2015), we used two native Arabic (AR) speakers who are also fluent on English (EN) to write their equivalent versions of the queries in Arabic for each of these EN topics.

We configured and trained an AR-to-EN MT system to translate each AR query to EN. Our MT system is a phrase-based (Koehn et al., 2003), that is developed using the Moses Statistical Machine Translation (SMT) toolkit (Koehn et al., 2007). Word alignments in both directions were calculated using a multi-threaded version of the GIZA++ [6] tool (Gao and Vogel, 2008). The parameters of our MT system were tuned on a development corpus using Minimum Error Rate Training (Och, 2003). The AR-to-En MT system was trained using the bilingual training corpora listed in Table 1 from LDC for MSA (Modern Standard Arabic) training. The size of the tuning set is 111.8K and 138.2K of Arabic and English tokens. All AR data are tokenised using MADA-ARZ version 0.4 (Habash et al., 2013).

Corpus	AR genre	AR tokens	EN tokens
bolt	Egyptian	1.70M	2.05M
thy		282k	362k
bbnturk		1.52M	1.58M
bbnegy		514k	588k
gale	MSA	4.28M	5.01 M
fouo		717 k	791k
ummah		3.61M	3.72M
iraqi	Iraqi	1M	1.14M
bbnlev	Levantine	1.59M	1.81M
Total		15.2M	17M

Table 1: The sizes and the genres of bilingual training corpora.

We extracted the top 100 translations list for each query generated by the MT system. The overall number of query candidates generated by was 5,863 with an average of over 90 different translations per query. These queries were used in searching the EN ASR transcripts extracted from the blip1000 collection.

The Terrier retrieval platform[7] was used as the IR component of our experimental setup. Stop

[5] http://www.multimediaeval.org/mediaeval2012/

[6] Available at `http://www.cs.cmu.edu/~qing/`

[7] http://terrier.org/

words were removed based on the standard Terrier list, and stemming performed using the Terrier implementation of Porter stemming. Retrieval was carried out using the PL2 retrieval model using the settings recommended for this CLUGS task in Khwileh et al. (Khwileh et al., 2016), with the empirically-determined hyper-parameters that $c=1$.

5.1 Parameters Tuning for the Post-retrieval QPP

As explained in section 3.2, post-retrieval QPP methods require some parameters to be tuned. For the experiments reported in this work, we used the following approach to tune NQC, WIG and WRG. We used the *optimal paradigm*, proposed in (Shtok et al., 2012), that is based on using values of free parameters that yield optimal prediction performance for each predictor on set of queries. We used the 60 monolingual EN queries as test set to obtain these optimal parameters for each predictor. Parameters k (in WIG and NQC), rel and $nrel$ (in WRG) were tuned through manual data sweeping within the range of [5, 100] with an interval of 5, and through the range of [100,500] with an interval of 50. The optimal k parameters obtained for WIG was 10, while for NQC it was 150, these are indeed similar to those recommended in (Shtok et al., 2012). For the WRG, we found that 30 is the optimal parameter for rel and 60 for $nrel$.

6 Evaluating Prediction Quality

The effectiveness of QPP methods is usually evaluated by measuring correlation between values assigned by the QPP method and the actual performance, in terms of average precision (AP), of each query. The quality of each predictor is evaluated in our CLUGS task by measuring the Pearson linear correlation coefficient ρ between both the AP, which is measured using human relevant assessment for each candidate translation for extracted 100-best and the values assigned to these queries by each prediction method. For each predictor, we follow the implementation reported in the citation shown in the first column of Table 2. For the SQC and VarTFIDF, we report only the best result obtained out of the three aggregations (Max, Avg and Sum) due to space limitations. In addition to Pearson's correlation, we also tested Kendalls tau and Spearman correlations to report the nonlinear relationship between these predictors and the retrieval performance. The prediction quality for each of these predictors on our CLUGS task is shown in Table 2.

6.1 Pre-retrieval Quality

As can be seen from the results shown in Table 2, IDF-based predictors are found to have the least robustness across other predictors. The reliability issue regarding misleading terms (as discussed in section 3.1) significantly impacts the prediction quality of these predictors. To further illustrate this issue, consider the example query
سوفتويرات لتصميم و برمجة مواقع الويبْ.
This query has two candidates EN translations (T1 and T2) as follows:
T1 : " سوفتويرات *for the development and web design"* and
T2 : " سوفتويرات *for the development and design internet"*.

The main difference between these translations is "web" vs "internet". While the word "internet" is more unique term with a higher IDF value, it is considered as a misleader to this query since it shifts the original topic of the query "web design". Thus, this has resulted in a query topic-drift, and hence a false prediction of its performance.

In contrast, prediction quality is improved for all QPP methods which are less focused on the uniqueness of the terms and do not rely *solely* on the IDF in its calculation (i.e. Qs, SQC). AvgFL is shown to have the highest quality over all tested QPP methods, showing a consistent statistically significant prediction across different correlation measures. This arises as result of its robustness in utilising the fluency measure to discriminate between different translations, penalising these which are OOV or very unique words in the collection.

6.2 Post-retrieval Quality

The post-retrieval QPP methods are more robust and perform better than the pre-retrieval methods overall. This is due to the fact that post-retrieval methods are based on the actual scores of the translations in which at least one retrieval run was used for the prediction. Unlike, pre-retrieval methods, post-retrieval requires exhaustive parameter tuning, as explained in section 5.1. Both parameter tuning and the time required to generate the post-retrieval QPP was a major efficiency issue by comparison to pre-retrieval QPR (average time to generate the pre-retrieval QPPs was around 10% to that of the post-retrieval QPPs). WRG has the highest prediction quality across all predictors. The robustness of WRG is due to the fact that it relies on stronger evidence; which is the score of the relevant documents. While NQC/WIG rely only on one parameter, i.e. the top k ranked doc-

	Pearson	Kendall's tau	Spearman's
VarTFIDF (Zhou and Croft, 2007)	-0.20	-0.165	-0.194
SCQ (Zhou and Croft, 2007)	0.248	0.137	0.201
Qs (He and Ounis, 2004)	**-0.319**	-0.221	**-0.29**
AvQL (Mothe and Tanguy, 2005)	-0.193	-0.126	-0.208
SumIDF (Cronen-Townsend et al., 2002)	0.069	0.110	0.163
AvgIDF (Cronen-Townsend et al., 2002)	0.030	0.086	0.128
MaxIDF (Scholer et al., 2004)	-0.044	0.019	0.035
AvICTF (He and Ounis, 2006)	0.118	0.162	0.210
AvgFL (Equation 1)	**0.446**	**0.313**	**0.395**
WRG (Equation 4)	**0.463**	**0.321**	**0.384**
WIG (Equation 2)	**0.405**	0.260	**0.333**
NQC (Equation 3)	**0.385**	0.22	**0.321**

Table 2: Correlation Coefficients vs AP for each query translation from Ar-to-En vs each QPP. Correlation that are significant at the 0.05 confidence level are marked in **bold**.

uments, WRG relies on further tuning of the top k parameter into both the rel and $nrel$ documents to provide better estimation. This, on one hand, helps WRG to identify relevant translations that can in fact distinguish the relevant document from the non relevant ones, but on the other hand, raises efficiency concerns about WRG, since it takes almost twice the time required for WIG/NQC tuning.

7 Finding Relevant Translations in CLUGS

In this section, we investigate the potential for these QPP techniques to be used in an adaptive CLIR algorithm that is able to automatically identify the most relevant translations. The main idea is to use the translation candidate that is predicted to have the highest retrieval effectiveness for each query. Using the same settings explained in section 5, we implement the adaptive CLIR algorithm as follows.

1. For each query, the MT system is used to generate up to the 100-best possible translations which form a selection pool.
2. QPP is used to score each translation candidate from the selection pool based on its estimated retrieval performance.
3. Retrieval is then performed using the translation that is predicted to be most effective.

We investigate using all QPP methods [8] shown in Table 2 to evaluate this adaptive CLIR algorithm.

We compare these adaptive CLIR techniques to three baselines as follows:

Google translate as example of an off-shelf black-box MT tool, similar to our work in (Khwileh et al., 2015); SingleBest, which is the 1-best translation output generated by Moses MT;

100BestAP, which uses the ground-truth data to get the best performing translation in terms of AP from the 100-best translations generated by Moses MT.

[8] We used the same parameters learned for post-retrieval QPP in 5.1

The adaptive and baseline retrieval performance results for the CLIR experiments are shown in Table 3 in terms of the Mean Average Precision (MAP) obtained. For clarity, we also report the percentage of improvement over each of these baselines as an additional columns (i.e. the *over SingleBest* column indicates the improvement in MAP over the SingleBest baseline).

The *Baseline CLIR* results from Table 3 show that black-box Google MT out-performed the SingleBest output from the open-box Moses by 11.8% which confirms the previously reported results in (Leveling et al., 2009b) that using black-box MT can be easier and more effective than just using the Singlebest. On the other hand, the result from the open-box with ideal AP (100BestAP) confirms that the open-box MT can indeed be improved by looking at other translations candidates that are more **relevant** for CLIR.

The *Adaptive CLIR using Pre-Retrieval* block of Table 3 shows how pre-retrieval QPP can be used to find the best translation from the 100-best extracted. This confirms the conclusion obtained from Table 2 where the proposed method is the most effective in getting the single best translation for CLIR with effectiveness comparable to that of the black box MT system and obtained 11% performance improvement over the SingleBest baseline. The *Adaptive CLIR using Post-Retrieval* block of Table 3 shows how the post-retrieval QPP methods are the most effective for finding the most effective translation in CLIR. This confirms previously reported conclusions on comparing pre-retrieval and post-retrieval, i.e. that post-retrieval QPP is always more effective (Hauff, 2010). The WRG predictor is the most effective with significant improvement 28% over the SingleBest and 14% over the black-box. This also confirms that the correlation results reported in Table 2 where WRG has the highest correlation to AP when it comes to predicting the translation quality.

	MAP	over blackbox MT	over SingleBest	Over 100BestAP
Baseline CLIR				
Off-shelf black-box	0.2535	-	11.8%	-28.9% *
100BestAP	0.3566	28.9%*	57.3%*	-
SingleBest	0.2267	-11.8%	-	-57.3%*
Adaptive CLIR using Post-Retrieval				
WRG	**0.2899**	**14.4%**	**27.9%***	**-18.7%**
NQC	0.2379	-6.2%	4.9%	-33.3%*
WIG	0.2423	-4.4%	6.9%	-32.1%*
Adaptive CLIR using Pre-Retrieval				
MAXIDF	0.2082	-17.9%	-8.2%	-41.6%*
QL	0.1827	-27.9%*	-19.4%*	-48.8%*
SumSQC	0.1995	-21.3%*	-12.0%	-44.1%*
AvgFL	**0.2507**	**-1.1%**	**10.6%**	**-29.7%***
avgICTF	0.2219	-12.5%	-2.1%	-37.8%*
SumVarTFIDF	0.1619	-36.1%*	-28.6%*	-54.6%*
Qs	0.2103	-17.0%	-7.2%	-41.0%*

Table 3: Baseline and adaptive CLIR results using both pre-retrieval and post-retrieval QPP. Percentages % with * indicate *statistically significant* different change at 0.05 confidence level

Overall, results from Table 3 indicate that QPP techniques can indeed help re-ranking the translation candidates of open-box MT, and hence improve its translation quality for CLIR purposes. Both AvgFL and WRG predictors, which were designed specifically for this task, served as an adequate reference to find the most effective translations and improve over the SingleBest output that is suggested originally by the MT system. However, none of the reported adaptive CLIR results were able to match or even come close to the ideal performance baseline (100BestAP). This suggests that there is still scope for further improvement. By contrast, these QPP methods are a stand-alone IR metric that is completely unsupervised and works on a query-by-query basis. Training a machine-learning algorithm that combines several QPPs together with other MT-based signals may achieve more robust/accurate prediction for this task. We leave this investigation for future work.

8 Conclusions

This paper has presented a framework for predicting translation quality for a CLUGS task. We proposed a novel unsupervised approach to estimate the effectiveness of a translation when there is no human evaluation of retrieval available. Our experimental investigation reveals that IDF-based prediction is not effective for this task because of the misleading very unique terms which can result in unreliable prediction. We proposed a new Pre-retrieval QPP technique for this task called AvgFL that is designed to detect misleading very unique and OOV words.

For post-retrieval QPP, we also proposed WRG (Weighted Relevancy Gain) that is modified version of the well-established WIG predictor (Zhou and Croft, 2007) and tuned to focus on the information entropy of the relevant documents. Our experimental investigation reports the robustness of these proposed approaches in predicting the translation effectiveness for an Ar-to-En CLUGS task over other state-of-art QPP methods. We found that post-retrieval QPP can be more accurate than pre-retrieval QPP for this task, although it suffers from efficiency issues. Finally, our experiments demonstrated how these predictors could be utilised by a CLIR model that is adaptively able to find the most-relevant translations for IR.

For future work, we plan to experiment with combining different QPP techniques together with other MT-based signals for improved prediction quality. We also plan to use the proposed framework to develop a new CLIR model to estimate the translation quality from different MT systems with different translation outputs.

Acknowledgments

This research was partially supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (`www.adaptcentre.ie`) at Dublin City University.

References

James Allan, Lisa Ballesteros, James P. Callan, W. Bruce Croft, and Zhihong Lu. 1995. Recent experiments with INQUERY. In *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*.

Arwa Alqudsi, Nazlia Omar, and Khalid Shaker. 2014. Arabic machine translation: a survey. *Artificial Intelligence Review*, 42(4):549–572.

Romaric Besançon, Stéphane Chaudiron, Djamel Mostefa, Ismaïl Timimi, Khalid Choukri, and Meriama Laïb. 2009. Overview of CLEF 2009 INFILE track. In *Working Notes for CLEF 2009 Workshop co-located with the 13th European Conference on Digital Libraries (ECDL 2009) , Corfù, Greece, September 30 - October 2, 2009.*

Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 299–306.

Kareem Darwish, Walid Magdy. 2014. Arabic information retrieval. *Foundations and Trends® in Information Retrieval*, 7(4):239–342.

Q. Gao and S. Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, in Columbus, Ohio, USA.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal arabic. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 426–432.

Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 1419–1420.

Claudia Hauff. 2010. *Predicting the effectiveness of queries and retrieval systems*. Ph.D. thesis, Centre for Telematics and Information Technology University of Twente.

Ben He and Iadh Ounis. 2004. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval, 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004, Proceedings*, pages 43–54.

Ben He and Iadh Ounis. 2006. Query performance prediction. *Information Systems*, 31(7):585–594.

Internetworldstats.com. 2016. Internet world users by language top 10 languages. `http://www.internetworldstats.com/stats7.htm`. Retrieved: 2017-01-04.

Ahmad Khwileh and Gareth J. F. Jones. 2016. Investigating segment-based query expansion for user-generated spoken content retrieval. In *14th International Workshop on Content-Based Multimedia Indexing, CBMI 2016, Bucharest, Romania, June 15-17, 2016*, pages 1–6.

Ahmad Khwileh, Debasis Ganguly, and Gareth J. F. Jones. 2015. An investigation of cross-language information retrieval for user-generated internet video. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, pages 117–129. Springer.

Ahmad Khwileh, Debasis Ganguly, and Gareth J. F. Jones. 2016. Utilisation of metadata fields and query expansion in cross-lingual search of user-generated internet video. *Journal of Artificial Intelligence Research*, 55:249–281.

P. Koehn, Franz J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 48–54, Edmonton, Canada.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Oren Kurland, Anna Shtok, David Carmel, and Shay Hummel. 2011. A unified framework for post-retrieval query-performance prediction. In *Conference on the Theory of Information Retrieval*, pages 15–26. Springer.

Martha Larson and Gareth J. F. Jones. 2012. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(45):235–422.

Johannes Leveling, Dong Zhou, Gareth J. F. Jones, and Vincent Wade. 2009a. Tcd-dcu at tel@ clef 2009: Document expansion, query translation and language modeling. In *Working Notes for CLEF 2009 Work-shop co-located with the 13th European Conference on Digital Libraries (ECDL 2009) , Corfù, Greece, September 30 - October 2, 2009.*, volume 30.

Johannes Leveling, Dong Zhou, Gareth J. F. Jones, and Vincent Wade. 2009b. Document expansion, query translation and language modeling for ad-hoc IR. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, pages 58–61.

Walid Magdy and Gareth J. F. Jones. 2014. Studying machine translation technologies for large-data clir tasks: a patent prior-art search case study. *Information Retrieval*, 17(5-6):492–519.

Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving automatic query expansion. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 206–214.

Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop*, pages 7–10.

Douglas W. Oard and Anne R. Diekema. 1998. Cross-language information retrieval. *Annual review of information science and technology*, 33:223–256.

Douglas W. Oard and Fredric C. Gey. 2002. The trec 2002 arabic/english clir track. In *TREC,In The Eleventh Text REtrieval Conference: TREC 2002 (Gaithersburg, MD, Nov. 2002)*, pages 17–26. E.M.Voorhees et al. eds. NIST Special Publication 500-251.

Douglas W. Oard and Paul G. Hackett. 1998. Document translation for cross-language text retrieval at the university of maryland. In *Information Technology: The Sixth Text REtrieval Conference (TREC-6)*, pages 687–696. US Dept. of Commerce, Technology Administration, National Institute of Standards and Technology.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 160–167, Sapporo, Japan.

Vassilis Plachouras, Ben He, and Iadh Ounis. 2004. University of glasgow at TREC 2004: Experiments in web, robust, and terabyte tracks with terrier. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Sebastian Schmiedeke, Peng Xu, Isabelle Ferrané, Maria Eskevich, Christoph Kofler, Martha A Larson, Yannick Estève, Lori Lamel, Gareth J. F. Jones, and Thomas Sikora. 2013. Blip10000: a social video dataset containing spug content for tagging and retrieval. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 96–101. ACM.

Falk Scholer, Hugh E. Williams, and Andrew Turpin. 2004. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology*, 55(7):637–650.

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems (TOIS)*, 30(2):11.

Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, pages 52–64.

Yun Zhou and W. Bruce Croft. 2007. Query performance prediction in web search environments. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 543–550.

Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, 45(1):1.

Justin Zobel and Alistair Moffat. 2006. Inverted files for text search engines. *ACM computing surveys (CSUR)*, 38(2):6.