CLIR for Informal Content in Arabic Forum Posts

Mossaab Bagdouri
Dep. of Computer Science
University of Maryland
College Park, MD, USA
mossaab@umd.edu

Douglas W. Oard
iSchool and UMIACS
University of Maryland
College Park, MD, USA
oard@umd.edu

Vittorio Castelli
T.J. Watson Research Center
IBM
Yorktown Heights, NY, USA
vittorio@us.ibm.com

ABSTRACT

The field of Cross-Language Information Retrieval (CLIR) addresses the problem of finding documents in some language that are relevant to a question posed in a different language. Retrieving answers to questions written using formal vocabulary from collections of informal documents, as with many types of social media, is a largely unexplored subfield of CLIR. Because formal and informal content are often intermingled, CLIR systems that excel at finding formal content may tend to select formal over informal content. To measure this effect, a test collection annotated for both relevance and informality is needed. This paper describes the development of a small test collection for this task, with questions posed in formal English and the documents consisting of intermixed formal and informal Arabic. Experiments with this collection show that dialect classification can help to recognize informal content, thus improving precision. At the same time, the results indicate that neither dialect-tuned morphological analysis nor a lightweight CLIR approach that minimizes propagation of translation errors yet yield a reliable improvement in recall for informal content when compared to a straightforward document translation architecture.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords: CLIR; Informality; Social Media; Evaluation

1. INTRODUCTION

The history of information retrieval research has been strongly dominated by a focus on retrieval of what we might call "formal" content, content written with dissemination in mind. Such content potentially has high value, but constitutes only a tiny fraction of the words produced by our planet's 7 billion people. Recently, activities such as the TREC Blog and Microblog tracks have begun to explore how retrieval systems might be tailored to the unique characteristics of informal content. Perhaps unsurprisingly, it turns out that informal content poses unique challenges for Cross-Language Information Retrieval (CLIR) as well. In this paper, we begin to explore those challenges.

For our experiments, we use what we believe to be the first test collection to focus on CLIR from informal content. We focus on a part of the collection developed initially for the DARPA Broad Operational Language Translation (BOLT) program that includes 11 English questions and 12.6 million Arabic Web forum posts. The questions are well-formed requests, written in formal English. We focus in this paper on retrieval of entire posts using post-scale annotations by independent annotators for relevance and informal language.

The remainder of this paper is organized as follows. Sections 2 and 3 address the consequences of translating documents with a state-of-the-art translation system. We describe a pilot study that shows how this approach results in an adverse selection bias. Section 4 then explores the use of an alternative CLIR architecture based on Probabilistic Structured Queries (PSQ) for recall enhancement, coupled with automatic detection of informal content for precision enhancement. Section 5 summarizes the process of building a collection to test our methods and reports results. We conclude in Section 6 with a discussion of the implications of our results for future work on CLIR from informal content.

2. THE BOLT IR TEST COLLECTION

The collection to be searched consists of Arabic Internet forums, which are Web sites in which users submit *posts* that either originate or extend *threads*; these form tree-like discussions. We used the BOLT Phase 2 IR test collection. This collection was crawled by the Linguistic Data Consortium (LDC) from public Web forums in Egypt with material written between 12 Dec 2001 and 19 May 2012[1] and consists of 12,612,144 posts from 773,861 threads distributed across 272 forums found on 32 Egyptian websites. Our retrieval task is to find relevant *posts* in which useful answers to well formed English questions can be found. For the experiments in this paper, we treat posts as isolated documents, making no use of the thread structure of the forums.

A crucial characteristic of Web forums is that authors will sometimes copy text from various sources (e.g., a news article) into their posts, and that copied text may be written in

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661924.

[1] The collection, LDC2013E08, also contains English and Chinese, but we focus only on Arabic in our experiments. We will make our annotations available to the LDC for inclusion in their public release.

وقال إن ذلك سيؤدي إلي ارتفاع تكلفة الأنتاج للعصائر في مصر بنسبة في حدود20% تقريبا وأوضح
أن هذه الزيادة ستسري علي الإنتاج الجديد الذي سيستخدم المركزات بالأسعار المرتفعة (...)

He said that this would lead to a rise in the cost of production for juices in Egypt at a rate of approximately 20 borders and explained that this increase will apply to the new production, which will be used concentrates on high prices (...)

كلامها صحيح فعلا السياحة باى باى لسه خطفيين اتنين سياح فى سيناء
ذى ما قال الاخ توفيق عكاشة مصر هتشوف ايام ما شفتهاش قبل كده

Its really true words tourism any any Khatfin just two tourists in Sinai like what brother Tawfiq Okasha said Egypt will see days I don't see it before

Figure 1: Two Arabic passages with their machine translations. The translation of the MSA passage on the left is better than that of the Egyptian one on the right. The MT confused the Arabic transliteration of the English word "Bye" (i.e. باى) with the Arabic word بأي meaning "with any". Also, the word خطفيين, meaning kidnappers, was simply transliterated as "Khatfin".

formal language, even when the author otherwise uses informal language. This is different from the behavior observed on short messaging services such as Twitter, where strict limits on the number of characters make the use of Web links to refer to existing content more common. To determine the prevalence of posts containing at least some informal content, we randomly selected 1,000 posts from the collection and the first author of this paper manually annotated them as informal or formal. We found that 819 of the sampled posts contained one or more Egyptian Arabic terms and thus clearly contained some informal content, 50 contained only Modern Standard Arabic (MSA) terms but clearly contained some informal usage, 72 contained only MSA terms but were of debatable formality (i.e., they might be considered formal or informal by different annotators), and 59 contained only MSA terms and were clearly formal. In this paper we consistently treat posts that contain content of debatable formality to be formal. We conclude that approximately 87% (869 of 1000) of the posts in the collection contain some informal content, and that 94% (819/869) of these could be recognized by simply detecting the presence of one or more terms from some Arabic dialect.

The questions for our test collection are drawn from the BOLT Phase 2 IR Evaluation Questions, created also by the LDC (LDC2013E136). In addition to the 17 questions that explicitly targeted Egyptian Arabic, we included 7 questions that do not target any specific language, but that in our judgment are expected to return results from Arabic content. As spelling correction is not a focus of our current work, we removed one question for which a focal term was, in our opinion, incorrectly transliterated. The BOLT program reserved the odd-numbered questions for progress testing, so we used in our study only the 11 even-numbered questions.[2] Of these, we used three for exploratory analysis and eight for the experiments reported in Section 5.

3. EXPLORATORY ANALYSIS

For our exploratory analysis, we selected the three English questions for which the teams were instructed to retrieve answers from the Arabic portion of the test collection alone, and for which both participating systems in the 2013 BOLT IR evaluation returned answer passages that together spanned 25 or more different Arabic posts.[3] The first author of this paper, a native speaker of Arabic who is familiar with Egyptian Arabic, examined every post returned by either system that contained at least one passage marked as relevant by the LDC to identify posts containing informal use of Arabic. For these annotations, and throughout this paper, we defined a post as containing informal use of Arabic if (a) it contains at least one lexical item that is not present in properly written MSA, or (b) if any of the expressions would not be used in a formal MSA document, in the annotator's opinion, even if each individual term is in MSA. We found only about one-third of the posts (32 of 90) in which a relevant passage had been found were annotated as containing any informal use of Arabic. This finding is surprising, as the test collection and the questions had been developed specifically to evaluate CLIR on informal content.

We proffer two possible explanations. Either (1) the small fraction of the collection that contains only formal content is particularly rich in relevant posts, or (2) the participating systems were much better at finding relevant content among the formal posts than among the informal ones. Given the nature of the questions, the first explanation seems unlikely. To understand whether the second one is plausible, we need to look under the hood to see how the two participating systems actually worked. These were complex systems for fully automatic question answering that were still at the time in the midst of development. One notable commonality was that both used a document translation architecture in which statistical Machine Translation (MT) was first used to translate the entire Arabic collection into English, with the question answering process then run on the resulting English translations. Figure 1 shows English translations of two Arabic passages, one of which was written in MSA, the other in Egyptian Arabic. Two effects are evident. First, limitations of the translation model result in transliterations being generated for some words. Second, the MT language model seems to make poor decisions in the vicinity of the transliterated words. The combination of these two effects is substantially more severe when translating Egyptian Arabic than MSA, despite the fact that the translation models were specifically tuned using Egyptian Arabic examples.

4. FINDING RELEVANT INFORMALITY

In this section we describe a three-step process for enhancing our ability to find relevant informal content. First, we clarify our goal. BOLT question answering systems were optimized for finding *relevant* content, with no specific requirement for informality. If it is informal content that we seek, we need evaluation measures that reward success at that task. Second, we need ways of enhancing recall on relevant informal content, even at some cost in precision; we describe two such techniques. Third, we then need to enhance precision on informal content by suppressing retrieval of posts that contain only formal content.

[2] Questions BIR_200052, BIR_200056, BIR_200058, BIR_200060, BIR_200062, BIR_200064, BIR_200066, BIR_200130, BIR_200134, BIR_200138, BIR_200144.

[3] BIR_200056, BIR_200060, BIR_200066.

4.1 Rethinking Evaluation Measures

Our goal is to optimize retrieval of informal content. Two parameters control this measure: Relevance and Informality. For a question $q \in \mathcal{Q}$ and a document $d \in \mathcal{D}$, we want an evaluation function $s : (\mathcal{Q}, \mathcal{D}) \rightarrow [0, 1]$ that, for every pair of documents d_1 and d_2, satisfies:

- If d is not relevant to q, then $s(q, d) = 0$
- If d is formal then $s(q, d) = 0$
- If d_1 and d_2 are of equal relevance, but d_1 is more informal, then $s(q, d_1) \geq s(q, d_2)$
- If d_1 and d_2 are of equal informality, but d_1 is more relevant than d_2, then $s(q, d_1) \geq s(q, d_2)$

When relevance and informality are binary valued functions r: $(\mathcal{Q}, \mathcal{D}) \rightarrow 0, 1$ and i: $\mathcal{D} \rightarrow 0, 1$, this simplifies to:

$$s(q, d) = \begin{cases} 1 & \text{if } d \text{ is relevant to } q \text{ and informal} \\ 0 & \text{otherwise} \end{cases}$$

4.2 Enhancing Recall on Informal Arabic

As the examples above illustrate, the translation model and the language model are both potential sources of error. Translation model errors on informal content are difficult to address. They originate from the greater variability of informal language compared to that of formal language, and from the lack of correspondingly larger training corpora. Presently available sentence-aligned informal-language parallel corpora are comparatively small and thus best used to tune or adapt translation models originally trained on the far larger amount of MSA for which parallel text is available. This approach yields MT results somewhere between that which could be achieved with MSA alone and that which we would expect if large quantities of dialectal Arabic were available as parallel text. One component we can control, however, is the language model. CLIR techniques that lack a word n-gram language model have been shown to yield retrieval results that are about as good as those achieved using an "MT First" document translation architecture. We therefore tried one such CLIR technique, Probabilistic Structured Queries [3], which is known to make good use of translation probabilities. We refer to this approach as "IR First."

Because of lexical and morphological differences between Egyptian Arabic and MSA, we want a form of morphological analysis or stemming that can process either. To this end, we use a combination of the Standard Arabic Morphological Analyzer (SAMA) [7] and the large-scale morphological analyzer for Egyptian Arabic (CALIMA) [4]. Habash et al. reported such a combination results in an analysis coverage of 92.1% [4]. The output for an input word is an unordered list of plausible stems. Inspired by the work of Darwish [2], we disambiguate these candidate stems by returning the one with the highest frequency within the BOLT IR collection.

4.3 Enhancing Precision on Informal Arabic

We are not aware of prior work on classification of informal Arabic, but there has been prior work on the closely related problem of Arabic dialect detection. In particular, Cotterell and Callison-Burch have released a total of 1.25 million words for five Arabic dialects [1]. This training data was originally collected from comments posted on newspaper Web sites and from Arabic Twitter posts. We merged the training examples for all dialects and removed those with fewer than 50 or more than 500 characters, yielding 46,174 positive training examples for the Arabic Dialect condition. As training data for the MSA condition we randomly sampled 59,437 news articles from the Egyptian newspaper Al-Youm Al-Sabe' [10], subject to the same length constraints. We evaluate the accuracy of this classifier on the 1000 random posts from Section 2, of which 869 are informal and the debatable 72 are considered to be formal; and on a balanced set of 118 posts with no debatable content. The accuracy was 88.0% for the former, and 83.1% for the latter. When we apply this classifier to the posts in the the BOLT collection, we find that the prevalence of informal content is estimated to be 93.7%. We also applied the same classifier to individual lines from the posts in this collection, finding that 89.7% of the lines are classified as informal.

5. EXPERIMENTS

We evaluate our methods with independent annotations of pools drawn from specific systems on eight held out topics.

5.1 Systems

In our experiments, we had access to a proprietary system, System **A**, with the MT First architecture. In this system, a question q is analyzed to automatically generate an Indri query, which is issued against the translations of all of the Arabic posts. The retrieved posts are then segmented and each segment is assigned a probability of containing a relevant answer to q by an ensemble of classifiers; the N relevant passages with highest scores are returned. Our system, System **B**, implements the IR First model, and the output is the top N posts. In a third architecture, System **C**, these posts are fed to the segment-scale relevance detection stage of System **A** to return the top N passages. Each of the systems **B** and **C** is controlled by two parameters, each taking two values. The first corresponds to the application of the informality classifier (Section 4.3). We use the subscript i when we prefilter for informal posts and a when we do not. The second parameter is related to the choice made for stemming. c indicates the use of CALIMA (Section 4.2) and l the use of the Lucene's Arabic light stemmer [5].

5.2 Annotations

We hired two annotators: a native speaker of Egyptian Arabic and a native speaker of Arabic who is fluent in Egyptian Arabic. We gave them two independent tasks, each on a 3-point scale. For relevance, a post had to be assessed as relevant, possibly relevant, or not relevant. For informality, a post could be formal MSA, possibly informal, or informal. We trained them independently using one of our exploratory analysis questions (BIR_200060), and we instructed them to discuss the task only with the authors of this paper and not with each other. The annotators then assessed the eight held out topics (four each) for which we report results. They also annotated one training topic (BIR_200056) to measure their agreement. To convert the 3-point scale into a binary judgment, we consider possibly relevant to be not relevant, and possibly informal to be formal. Table 1 shows Cohen's Kappa for this topic. The informality task exhibits a high agreement, with Kappa ranging from 0.794 to 0.867. The relevance task exhibits a high agreement between the first author and Annotator 1 (0.806), but both exhibit more modest agreement (0.502 and 0.459) with Annotator 2.

Table 1: Cohen's Kappa coefficient for topic BIR_200056. Top right triangle: relevance. Bottom left: informality.

Rel Inform	First Author	Annotator 1	Annotator 2
First Author		0.502	0.806
Annotator 1	0.794		0.459
Annotator 2	0.863	0.867	

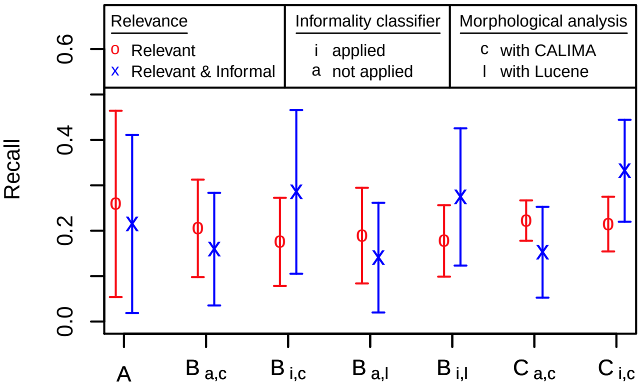

Figure 2: Mean and standard deviation of recall@25 across eight topics assessed by two independent annotators.

5.3 Results

Figure 2 shows the mean and standard deviation for recall at a fixed cutoff of 25 computed over eight topics. We first observe that the traditional recall measure that ignores informality substantially favors systems that are not tuned to retrieve informal content. In fact, the **o** of systems **A**, $\mathbf{B}_{a,c}$, $\mathbf{B}_{a,l}$ and $\mathbf{C}_{a,c}$ are higher than their **x** counterparts, that is, their performance is lower than what a traditional measure indicates. In contrast, the **o** of systems that are tuned to retrieve informal content—namely $\mathbf{B}_{i,c}$, $\mathbf{B}_{i,l}$ and $\mathbf{C}_{i,c}$—are lower than their **x** homologues, that is, their actual performance is higher than what a traditional measure states. Second, when the task is to retrieve relevant posts that are informal, the classifier trained to distinguish between MSA and Arabic Dialects improves precision (and thus recall at a fixed cutoff) substantially. These improvements are: 79% (from 0.1594 to 0.2855) for $\mathbf{B}_{\_,c}$, 95% (from 0.1407 to 0.2744) for $\mathbf{B}_{\_,l}$, and 118% (from 0.1526 to 0.3320) for $\mathbf{C}_{\_,c}$. All of these are statistically significant, at $p<0.05$, using a two-sided paired t-test. Third, no statistically significant difference is seen from CALIMA in the relevant and informal task, with the average recall of 0.2854 for System $\mathbf{B}_{i,c}$ being statistically indistinguishable from the 0.2744 for System $\mathbf{B}_{i,l}$. In contradiction to our expectations, we also find no statistically significant difference between the MT First approach and the IR First approach regardless of the inclusion of the informality condition with relevance; indeed if any undetected recall effect is present between System $\mathbf{C}_{a,c}$ and System **A** it would be a loss, not a gain, in recall.

6. CONCLUSIONS AND FUTURE WORK

We introduced the problem of retrieving informal content from Arabic forums in a CLIR setting. We have shown that traditional evaluation measures like recall that do not consider informality sometimes disadvantage systems that are tuned to retrieve relevant informal content. Our experiments over eleven topics demonstrate that such systems can have their precision enhanced by applying an informality classifier that is actually trained to detect dialectal Arabic. We tested two techniques that might have improved recall at a fixed cutoff on this task, namely Probabilistic Structured Queries, and a morphological analyzer for Egyptian Arabic. Our results do not support that hypothesis.

Other techniques could augment the improvements we have obtained in other ways. Pseudo-Relevance Feedback has recently been shown to enhance the retrieval of informal content [6]. A dialect to MSA MT system such as Elissa [9] could be applied on posts that were identified to be informal, and we might leverage domain adaptation to better tune the morphology of specific dialects [8]. Our annotations should be a resource for exploring such possibilities, although similar annotations for a larger set of questions will ultimately be needed if we are to draw strong conclusions about small differences. Importantly, we have focused only on questions that use formal vocabulary; informal query vocabulary also merits study. We have also compared only to a document translation baseline; query translation and bidirectional translations baselines would also be useful points of comparison. While much remains to be done, we believe our results offer useful insights to help focus this future work.

Acknowledgments

This research was supported in part by DARPA Contract HR0011-12-C-0015. We thank LDC and NIST for making the BOLT IR collection available to us.

7. REFERENCES

[1] R. Cotterell and C. Callison-Burch. A multi-dialect, multi-genre corpus of informal written Arabic. In *LREC*, pages 241–245, 2014.

[2] K. Darwish. Building a shallow Arabic morphological analyzer in one day. In *ACL SEMITIC*, 2002.

[3] K. Darwish and D. Oard. Probabilistic structured query methods. In *SIGIR*, 2003.

[4] N. Habash, R. Eskander, and A. Hawwari. A morphological analyzer for Egyptian Arabic. In *SIGMORPHON*, 2012.

[5] L. Larkey, L. Ballesteros, and M. Connell. Light stemming for Arabic information retrieval. In *Arabic Computational Morphology*. Springer, 2007.

[6] C.-J. Lee and W. B. Croft. Cross-language pseudo-relevance feedback techniques for informal text. In *ECIR*, pages 260–272, 2014.

[7] M. Maamouri, D. Graff, B. Bouziri, S. Krouna, A. Bies, and S. Kulick. Standard Arabic morphological analyzer (SAMA) version 3.1 LDC2010L01. LDC, 2010.

[8] W. Monroe, S. Green, and C. D. Manning. Word segmentation of informal Arabic with domain adaptation. In *ACL*, pages 206–211, 2014.

[9] W. Salloum and N. Habash. Elissa: A dialectal to standard Arabic machine translation system. In *COLING (Demos)*, pages 385–392, 2012.

[10] O. F. Zaidan and C. Callison-Burch. The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In *ACL*, 2011.