Capturing Out-of-Vocabulary Words in Arabic Text

Abdusalam F.A. Nwesri S.M.M. Tahaghoghi Falk Scholer
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia
{nwesri,saied,fscholer}@cs.rmit.edu.au

Abstract

The increasing flow of information between languages has led to a rise in the frequency of non-native or loan words, where terms of one language appear transliterated in another. Dealing with such out of vocabulary words is essential for successful cross-lingual information retrieval. For example, techniques such as stemming should not be applied indiscriminately to all words in a collection, and so before any stemming, foreign words need to be identified. In this paper, we investigate three approaches for the identification of foreign words in Arabic text: lexicons, language patterns, and n-grams and present that results show that lexicon-based approaches outperform the other techniques.

1 Introduction

Arabic words are derived from roots having three, four, or, in rare instances, five characters. The derivation process consistently follows patterns that are based on the three letter verb فَعَلَ (/faʕala/ to do)[1]. Each root word matches a base pattern. Characters are added at the beginning, the middle, or end of the root, but the base characters that match the pattern remain unchanged.

The pronunciation of Arabic characters is associated with short vowels; these are represented by diacritics, and attached to other characters to show how the characters should be pronounced. An Arabic character can be pronounced in several different ways. For example, the letter ب with the diacritic Fatha بَ is pronounced /ba/, with the diacritic Kasra بِ is pronounced /bɪ/, and with having the diacritic Damma بُ is pronounced /bʊ/. Diacritics are not shown in general written Arabic, and the reader must rely on the context to determine the implicit diacritics and therefore the pronunciation of each word. For example, the word ذهب can represent ذَهَبَ (/ðahaba/ = went), ذَهَبْ (/ðahab/ = gold).

Pure Arabic words follow restricted rules in their construction to keep them short and easy to pronounce. Their sounds usually follow the CVCV pattern, where C stands for a consonant and V stands for a Vowel. An Arabic word never has two consecutive consonants nor four consecutive vowels (Al-Shanti, 1996).

Foreign words are words that are borrowed from other languages. Some are remodelled to conform with Arabic word paradigms, and become part of the Arabic lexicon; others are transliterated into Arabic as they are pronounced by different Arabic speakers, with some segmental and vowel changes. The latter are called Out-Of-Vocabulary (OOV) words as they are not found in a standard Arabic lexicon. Such OOV words are increasingly common due to the inflow of information from foreign sources, and include terms that are either new and have yet to be translated into native equivalents, or proper nouns that have had their phonemes replaced by Arabic ones. Examples include words such as مارغرت /margrɪt/ (Margaret) or لينكس /linɪks/ (Linux). This process often results in different Arabic spellings for the same word.

Current Arabic Information Retrieval (AIR) systems do not handle the problem of retrieving the different versions of the same foreign

[1] We represent phonetics using the International Phonetic Alphabet (http://www.arts.gla.ac.uk/IPA/ipachart.html)

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 258–266,
Sydney, July 2006.

word (Abdelali et al., 2004), and instead typically retrieve only the documents containing the same spelling of the word as used in the query.

One solution to this problem has been used in cross-lingual information retrieval, where OOV words in the query are transliterated into their possible equivalents. Transliterating terms in English queries into multiple Arabic equivalents using an English-Arabic dictionary has been shown to have a positive impact on retrieval results (Abduljaleel and Larkey, 2003). However, we are aware of no work on handling OOV terms in Arabic queries.

For this, proper identification of foreign words is essential. Otherwise, query expansion for such words is not likely to be effective: many Arabic words could be wrongly expanded, resulting in long queries with many false transliterations of Arabic words. Furthermore, proper identification of foreign words would be helpful because such words could then be treated differently using techniques such as approximate string matching (Zobel and Dart, 1995).

In this paper, we examine possible techniques to identify foreign words in Arabic text. In the following sections we categorise and define foreign words in Arabic, and follow in section 2 with a discussion of possible different approaches that can identify them in Arabic text. In section 3 we present an initial evaluation of these approaches, and describe improvements in section 4 that we then explore in a second experiment in section 5. We discuss results in section 6 and finally conclude our work in section 7.

1.1 Foreign words in Arabic

Arabic has many foreign words, with varying levels of assimilation into the language. Words borrowed from other languages usually have different style in writing and construction, and Arabic linguists have drawn up rules to identify them. For example, any root Arabic word that has four or more characters should have one or more of the "Dalaga" letters (ف, ر, م, ن, ل, ب). Those that have no such letters are considered foreign (Al-Shanti, 1996). However, while such rules could be useful for linguistic purposes, they have limited application in Information Retrieval (IR); based on rules, many foreign words that have long been absorbed into the language and are spelled consistently would be considered to be OOV. From the IR perspective, foreign words can be split into two general categories: translated and transliterated.

ميلوشفيتش	ميلوسيفيتيش	ميليسوفيتش
ميلوشيفيتش	ميليسيفيتش	مليوسيفيتش
مليشيفيتش	ميلسيفيتش	ميلويسفيتش
ميليشفيتش	ميلوسيوفيتش	ميلسوفيتش
ميليشيفتش	ميلوسيفيتس	ميليوسيفيتش
ميلوزيفيتش	ميلوسوفيتس	مليوسوفيتش
ميلوزفيتش	ميلوشيفيتس	مليوسفيتش
ميلوسيفيتش	ميلوسيفتيش	ميلوسوفيتش
ميلوسيفتش	ميلوسفيتش	ميلوسفيتيش
ميلوسفتيش		

Table 1: Different spelling versions for the name Milosevic

Translated: These are foreign words that are modified or remodelled to conform with Arabic word paradigms; they are well assimilated into Arabic, and are sometimes referred to as Arabicised words (Aljlayl and Frieder, 2002). This process includes changes in the structure of the borrowed word, including segmental and vowel changes, and the addition, deletion, and modification of stress patterns (Al-Qinal, 2002). This category of foreign words usually has a single spelling version that is used consistently. Examples include words such as بستان (/bʊstæn/ = garden), برج (/bʊrʒ/ = tower), راديو (/ræduʊ/ = radio), and قنبلة (/qʊnbula = bomb).

Transliterated: Words in this category are transliterated into Arabic by replacing phonemes with their nearest Arabic equivalents. Although Arabic has a broad sound system that contains most phonemes used in other languages, not all phonemes have Arabic equivalents. In practice, such phonemes may be represented in different ways by different persons, resulting in several spelling versions for the same foreign word. For example, we observed 28 transliterated versions for the name of the former Serbian leader (Milosevic) in the TREC 2002 Arabic collection; these are shown in Table 1.

Transliteration has become more common than translation due to the need for instant access to new foreign terms. It can take considerable time for a new foreign term to be included in reference

dictionaries. However, users often need to immediately use a particular term, and cannot wait until a standard form of the word is created; news agencies form an important category of such users. This transliteration process often results in multiple spellings in common usage.

1.2 Related work

In the context of information retrieval, most work on foreign words in Arabic has been based on transliteration, and carried out under machine translation and cross-lingual information retrieval (CLIR) tasks, where English queries are used to search for Arabic documents, or vice versa. This often involves the use of a bilingual dictionary to translate queries and transliterate OOV words into their equivalents in Arabic.

Expanding a foreign word to its possible variants in a query has been shown to increase the precision of search results (Abduljaleel and Larkey, 2003). However, OOV words in the query are easily recognised based on English rules and an English-Arabic dictionary: capitalised words are marked as nouns, and the remaining words are translated using the dictionary. Words not found in the dictionary are marked as OOV and are transliterated into probable Arabic forms. In contrast, we aim to identify foreign words as a within Arabic text which is made difficult by the absence of such easily perceptible difference.

Stalls and Knight (1998) describe research to determine the original foreign word from its Arabic version; this is known as *back transliteration*. However, rather than using automatic methods to identify foreign words, they used a list of 2 800 names to test the accuracy of the back transliteration algorithm. Of these, only 900 names were successfully transliterated to their source names. While this approach can be used to identify transliterated foreign words, its effectiveness is not known on normal Arabic words as only names were used to test the algorithm.

Jeong et al. (1999) used statistical differences in syllable unigram and bigram patterns between pure Korean words and foreign words to identify foreign words in Korean documents. This approach was later enhanced by Kang and Choi (2002) to incorporate word segmentation.

A related area is language identification, where statistics derived from a language model are used to automatically identify languages (Dunning, 1994). Using N-gram counting produces good accuracy for long strings with 50 or more characters, and moderately well with 10-character-long strings. It is unclear how well this approach would work on individual words with five characters on average.

2 Identifying foreign words

We categorise three general approaches for recognising foreign words in Arabic text:

Arabic lexicon

OOV words can be easily captured by checking whether they exist in an Arabic lexicon. However, the lexicon is unlikely to include all Arabic words, while at the same time it could contain some foreign words. Moreover, this approach will identify misspelled Arabic words as foreign.

Arabic patterns system

Arabic uses a pattern system to derive words from their roots. Roots are three, four or sometimes five letters long. The reference pattern فَعَلَ (/faʕala/ = to do) is often used to represent three-letter root words. For example, the word بَحَثَ (/bʜ θa/ = searched) can be represented by this pattern through mapping بَـ to فَـ, حَـ to عَـ, and ثَـ to لَـ.

Many stems can be generated from this root using standard patterns. For instance, فَاعِلْ (/fæʕɪl/ = doer) , and يَفْعَلُ (/yfʕlʊ/ = is doing) are two different patterns that respectively represent the active participle, and present tense verb from the pattern فَعَلَ. By placing the appropriate core letters and adding additional letters in each pattern, we can generate words such as بَاحِثْ (/bæʜɪθ/ = researcher), يَبْحَثُ (/ybʜθʊ/ = does search) respectively. New words can further accept prefixes and suffixes.

We can recognise whether a word is an Arabic or foreign word by reversing the process and testing the different patterns. If, after all possible affixes have been removed, the remaining stem matches an Arabic pattern, the word is likely to be an Arabic word. For example, to check whether the word وَالْبَاحِثْ (/walbæʜɪθ/ = and the researcher) is a foreign word, we first remove the prefixes و and الـ to get the stem بَاحِثْ; we find that this word matches the pattern فَاعِلْ — it has the same length, and the letter ـا is in the same po-

sition — and conclude that it is therefore an Arabic word. Note that we must perform this determination without relying on diacritics.

This approach is not perfect, as general Arabic text does not include explicit diacritics; if parts of a foreign word match a pattern, it will be marked as being Arabic. Similarly, misspelled words may be classified as foreign words if no matching pattern is found.

N-gram approach

Transliterated foreign words exhibit construction patterns that are often different from Arabic patterns. By counting the N-grams of a sample of foreign words, a profile can be constructed to identify similar words. This approach has been used in language identification, although it is reported to have only moderate effectiveness in identifying short strings (Cavnar and Trenkle, 1994; Dunning, 1994).

2.1 Resources

For the lexicon approach, we used three lexicons: the Khoja root lexicon (Khoja and Garside, 1999), the Buckwalter Lexicon (Buckwalter, 2002), and the Microsoft office 2003 lexicon (Microsoft Corporation, 2002).

The Khoja stemmer has an associated compressed language dictionary that contains well-known roots. The stemmer strips prefixes and suffixes and matches the remaining stem with a list of Arabic patterns. If a match is found, the root is extracted and checked against the dictionary of root words. If no entry is found, the word is considered to be a non-Arabic word. We call this the Khoja Lexicon Approach (KLA).

The Buckwalter morphological analyser is a lexicon that uses three tables and an algorithm to check possible affixes. The algorithm checks a word and analyses its possible prefixes and suffixes to determine possible segmentation for an Arabic word. If the algorithm fails to find any possible segmentation, the word is considered not found in the lexicon. We name this approach the Buckwalter Lexicon Approach (BLA).

The Microsoft office lexicon is the one used in the Microsoft Office 2003 spell-checker. We test whether an Arabic word is found in this lexicon, and classify those that are not in the lexicon to be foreign words. We call this approach the Office Lexicon Approach (OLA).

افعوعل	افعلة	افعلال	افعلاء	افعلل
تفعال	تفاعيل	تستفعل	افعيل	افعوlل
فعالا	فاعول	فاعلة	تفعلل	تفعلة
فعللة	فعلة	فعاليل	فعالى	فعالل
فياعيل	فياعل	فواعيل	فعيلة	فعيلا
مفعلل	مفعلة	مفعلا	مفعالة	مفاعلة
متفعلل	فعولة	فعالة	افعول	تفعل
مفعيلا	مفعيل			

Table 2: Patterns added to the Khoja modified stemmer to implement the KPA approach

To use Arabic patterns, we modified the Khoja stemmer to check whether there is a match between a word and a list of patterns after stemming without further checking against the root dictionary. If there is no match, the word is considered a foreign word. This approach is similar to the approach used by Taghva et al. (2005). We adopted the patterns of the Khoja stemmer and added 37 patterns compiled from Arabic grammar books, these are shown in Table 2. We call these approaches the Khoja Pattern Approach (KPA), and Modified Khoja Pattern Approach (MKP) respectively. A word is also considered to be an Arabic word if the remaining stem has three or fewer letters.

We evaluate the effectiveness of the n-gram method in two ways. First, we extend the n-gram text categorisation method presented by Cavnar and Trenkle (1994). The method uses language profiles where, for each language, all n-grams that occur in a training corpus are sorted in order of decreasing frequency of occurrence, for n ranging from 1 to 5. To classify a text t, we build its n-gram frequency profile, and compute the distance between each n-gram in the text and in each language profile l_j. The total distance is computed by summing up all differences between the position of the n-gram in the text profile and the position of the same n-gram in the language profile:

$$D_j = \sum_{i=1}^{N_i} \mid \frac{rank(t_i, text)}{N_i} - \frac{rank(t_i, l_j)}{N_j} \mid$$

where D_j is the total distance between a text t with N_i n-grams, and a language profile l_j with N_j n-grams; and $rank$ is the position of the n-gram in the frequency-sorted list of all n-grams for either the text or language profile.

In our work, we build two language profiles, one

for native Arabic words and another for foreign words. We compare the n-grams in each word in our list against these two profiles. If the total distance between the word and the foreign words profile is smaller than the total distance between the word and the Arabic words profile, then it is classified as a foreign word. As the two language profiles are not in same size, we compute the relative position of each n-gram by dividing its position in the list by the number of the n-grams in the language profile. We call this approach the n-gram approach (NGR).

We also tried a simpler approach based on the construction of two trigram models: one from Arabic words, and another from foreign words. The probability that a string is a foreign word is determined by comparing the frequency of its trigrams with each language model. A word is considered foreign if the sum of the relative frequency of its trigrams in the foreign words profile is higher than the sum of the relative frequency of its trigrams in the Arabic words profile. We call this approach trigram (TRG).

3 Training Experiments

In this section, we describe how we formed a development data set using Arabic text from the Web, and how we evaluated and improved techniques for identification of foreign words.

3.1 Data

To form our development data set, we crawled the Arabic web sites of the Al-Jazeera news channel[1], the Al-Anwar[2] and El-Akhbar[3] newspapers. A list of 285 482 Arabic words was extracted. After removing Arabic stop words such as pronouns and prepositions, the list had 246 281 Arabic words with 25 492 unique words.

In the absence of diacritics, we decided to remove words with three or fewer characters, as these words could be interpreted as being either Arabic or foreign in different situations. For example, the word بي (/bi/) could be interpreted as the Arabic word meaning "in me", or the English letter B. After this step, 24 218 unique words remained.

We examined these words and categorised each of them either as Arabic word (AW), or a transliterated foreign word (FW). We also had to classify some terms as misspelled Arabic word (MW). We used the Microsoft Office spell-checker as a first-pass filter to identify misspelled words, and then manually inspected each word to identify any that were actually correct; the spell-checker fails to recognise some Arabic words, especially those with some complex affixes. The list also had some local Arabic dialect spellings that we chose to classify as misspelled.

The final list had three categories: 22 295 correct Arabic words, 1 218 foreign words and 705 misspelled words.

To build language models for the trigram approaches (NRG and TRG), we used the TREC 2001 Arabic collection (Gey and Oard, 2001). We manually selected 3 046 foreign words out of the OOV words extracted from the collection using the Microsoft office spell-checker. These foreign words are transliterated foreign words. We built the Arabic language model using 100 000 words extracted from the TREC collection using the same spell-checker. However, we excluded any word that could be a proper noun, to avoid involving foreign words. We used an algorithm to exclude any word that does not accept the suffix haa (ـه), as transliterated proper nouns can not accept Arabic affixes.

3.2 Evaluation

We measure the accuracy of each approach by examining the number of foreign words correctly identified, and the number of incorrect classifications. The precision of each approach is calculated as the ratio of correctly identified foreign words to the total number of words identified as foreign The latter could be correct or misspelled Arabic words identified as foreign plus the actual foreign words identified. The recall is calculated as the ratio of correctly identified foreign words to the number of words marked manually as foreign. Although there is generally a compromise between precision and recall, we consider precision to be more important, since incorrectly classifying Arabic words as foreign would be more likely to harm general retrieval performance. The left-hand side of Table 3 shows the results of our experiments. We have included the MW results to illustrate the effects of misspelled words on each approach

The results show that the n-gram approach (NGR) has the highest precision, while the

[1] http://www.aljazeera.net
[2] http://www.alanwar.com
[3] http://www.elkhabar.com

Appr.	AW #	MW #	FW #	FW R	FW P
OLA	614	698	1 017	0.834	0.437
BLA	384	404	628	0.515	0.443
KLA	1 732	215	745	0.612	0.277
KPA	1 034	135	590	0.480	0.340
MKP	940	126	573	0.470	0.350
NGR	718	95	726	0.596	0.471
TRG	1 591	118	737	0.605	0.301

Appr.	AW #	MW #	FW #	FW R	FW P
OLA	145	248	866	0.711	0.687
BLA	88	149	534	0.438	0.693
KLA	420	83	642	0.527	0.508
KPA	302	52	520	0.430	0.590
MKP	269	51	507	0.416	0.613
NGR	411	69	669	0.549	0.582
TRG	928	85	642	0.527	0.387

Table 3: Identification of foreign words: initial results (left) and results after improvements (right)

lexicon-based OLA approach gives the highest recall. The pattern approaches (KPA) and (MKP) perform well compared to the combination of patterns and the root lexicon (KLA), although the latter produces higher recall. There is a slight improvement in precision when adding more patterns, but recall is sightly reduced. The KLA approach produces the poorest precision, but has better recall rate than the NGR approach.

The results show that many Arabic native words are mistakenly caught in the foreign words net. Our intention is to handle foreign words differently from Arabic native words. Our approach is based on normalising the different forms of the same foreign word to one form at the index level rather than expanding the foreign word to its possible variants at the query level. Retrieval precision will be negatively affected by incorrect classification of native and foreign words. Consequently, we consider that keeping the proportion of false positives — correct Arabic words identified as foreign (precision) — low to be more important than correctly identifying a higher number of foreign words (recall).

Some of the Arabic words categorised as foreign are in fact misspelled; we believe that these have limited effect on retrieval precision, and there is limited value in identifying such words in a query unless the retrieval system incorporates a correction process.

4 Enhanced rules

To reduce the false identification rate of foreign words, we analysed the lists of foreign words, correct Arabic words identified as foreign, and Arabic misspelled words identified as foreign. We noticed that some Arabic characters rarely exist in transliterated foreign words, and used these to separate Arabic words — correctly or incorrectly spelled

Letter	count	letter	count	letter	count
ي	3 839	م	632	ح	2
ا	3 599	د	559	ع	2
و	2 453	ش	514	ص	1
ن	1 660	ج	458	ء	0
س	1 587	ز	334	ؤ	0
ت	1 544	ه	171	أ	0
ر	1 244	خ	84	إ	0
ك	1 070	ث	23	آ	0
ب	900	ق	20	ض	0
ل	863	ط	12	ظ	0
ف	769	ئ	7	ى	0
غ	728	ذ	3	ة	0

Table 4: Frequency of Arabic letters in a sample of 3 046 foreign words

– from true foreign words. Table 4 shows the count of each character in the sample of 3 046 foreign words; foreign words tend to have vowels inserted between consonants to maintain the CVCV paradigm. We also noticed that most of transliterated foreign words do not start with the definite article الـ, or end with the Taa Marbuta ـة. Foreign words also rarely end with two Arabic suffixes.

We also noticed that lexicon based approaches fail to recognise some correct Arabic words for the following reasons:

- Words with the letter ا (Alef) with or without the diacritics Hamza (أ, إ), or the diacritic Madda (آ) are not recognised as correct in many cases. Many words are also categorised incorrectly if the Hamza is wrongly placed above or below the initial Alef or the Madda is absent. In modern Arabic text, the Alef often appears without the Hamza diacritic and

the Madda is sometimes dropped.

- Correct Arabic words are not recognised with particular suffixes. For example, words that have the object suffix, such as the suffix ـها in يعلمونكها (/yuʕalımunakaha/ = they teach it to you).
- Some Arabic words are compound words, written attached to each other most of the time. For example, compound nouns such as عبدالقادر (/ʕbdulqadır/), although composed of two words that are individually identified as being correct, are flagged as incorrect when combined.
- Some common typographical shortcuts result in words being written without white space between them. Where a character that always terminates a word (for example ة) is found in the apparent middle of a word, it is clear that this problem has occurred.

From these observations, we constructed the following rules. Whenever one of the following conditions is met, a word is not classified as foreign:

1. the word contains any of the Arabic characters:
   ئ, ء, ذ, ح, ص, ؤ, أ, إ, آ, ظ, ض, ى, ة;
2. the word starts with the definite article (الـ);
3. the word has more than one Arabic suffix (pronouns attached at the end of the word);
4. the word has no vowels between the second and penultimate character (inclusive); or
5. the word contains one of the strings: ة, ى, ء, اا, يال, رال, زال, دال, ذال, وال, االـ, and when split into two parts at the first character of any sequence, the first part is three characters or longer, and the second part is four characters or longer.

The right-hand side of Table 3 shows the improvements achieved using these rules. It can be seen that they have a large positive impact. Overall, OLA performs the best, with precision at 69% and recall at 71%. Figure 1 shows the precision obtained before and after applying these rules. Improvement is consistent across all approaches, with an increase in precision between 10% and 25%.

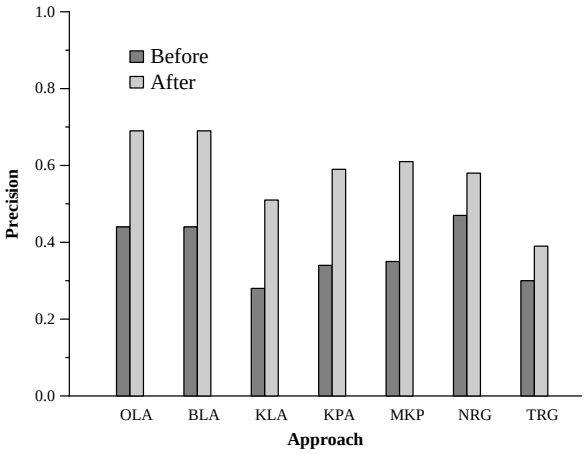

Figure 1: Precision of different approaches before and after Improvements

5 Verification Experiments

To verify our results, we used another data set of similar size to the first to verify our approach. We collected a list of 23 466 unique words from the Dar-al-Hayat newspaper[4]. Words, and classified and marked words in the same way as for the first data set (described in Section 3.1). We determined this new set to comprise 22 800 Arabic words (AW), 536 Foreign words (FW), and 130 Misspelled words (MW). Table 5 shows the initial results and improvements using the enhanced rules obtained by each approach using this data set.

The results on this unseen data are relatively consistent with the previous experiment, but precision in this sample is expectedly lower.

6 Discussion

We have seen that foreign words are not easily recognised in Arabic text, and a large number of Arabic words are affected when we try to exclude foreign words.

We found the lexicon approach to be the best in identifying foreign words. However, current lexicons are relatively small, and the variety of Arabic inflection makes it very difficult to include all correct word forms. Furthermore, current lexicons include many foreign words; for example when using OLA approach, 1 017 foreign words out of 1 218 are OOV, indicating that about 200 foreign words are present in that lexicon. The pattern approach is more efficient but the lack of diacritics in general written Arabic makes it very difficult to precisely match a pattern with a

[4] http://www.daralhayat.com

Appr.	AW #	MW #	FW #	FW R	FW P
OLA	1 189	112	417	0.777	0.242
BLA	780	96	267	0.498	0.234
KLA	1 684	55	312	0.582	0.152
KPA	992	29	238	0.440	0.189
MKP	901	26	231	0.431	0.199
NGR	740	22	286	0.533	0.272
TRG	1 655	19	308	0.575	0.155

Appr.	AW #	MW #	FW #	FW R	FW P
OLA	302	38	307	0.572	0.474
BLA	149	33	184	0.343	0.502
KLA	350	16	216	0.403	0.371
KPA	238	9	166	0.310	0.402
MKP	202	8	162	0.302	0.435
NGR	401	8	245	0.457	0.374
TRG	972	11	235	0.438	0.193

Table 5: Identification of foreign words on the test set: initial results (left) and results after improvements (right)

word, resulting in many foreign words being incorrectly identified as Arabic. Passing the list of all 3 046 manually judged foreign words to the pattern approach, some 2 017 words of this list were correctly judged as foreign, and about one third (1 029) were incorrectly judged to be Arabic. The n-gram method produced reasonable precision compared to the lexicon-based methods. In contrast, TRG had the worst results. This could be due to the limited size of the training corpus. However, we expect that improvements to this approach will remain limited due to the fact that many Arabic and foreign words share the same trigrams. It is clear that all the approaches are improved dramatically when applying the enhancement rules. The improvements of the NGR wasn't as equal as other approaches. This is because some of the rules are implicitly applied within the n-gram approach. The lack of diacritics also makes it very difficult to distinguish between certain foreign and Arabic words. For example, without diacritics, the word كلينتين could be كْلِيْنْتِين (/klɪntʊn/ = Clinton), or كَلِيْنَتَيْن (/kalinatin/ = as two date trees). The pronunciation is different in the two cases, but only context or diacritics can make it clear which word is being used.

7 Conclusion

Identifying foreign words in Arabic text is an important problem for cross-lingual information retrieval, since commonly-used techniques such as stemming should not be applied indiscriminately to all words in a collection.

We have presented three approaches for identifying foreign words in Arabic text: lexicons, patterns, and n-grams. We have presented results that show that the lexicon approach outperforms the other approaches, and have described improvements to minimise the false identification of foreign words. These rules result in improved precision, but have a small negative impact on recall. Overall, the results are relatively low for practical applications, and more work is needed to deal with this problem. As foreign words are characterised by having different versions, an algorithm that collapse those versions to one form could be useful in identifying foreign words. We are presently exploring algorithms to normalise foreign words in Arabic text. This will allow us to identify normalised forms for foreign words and use a single consistent version for indexing and retrieval.

8 Acknowledgements

We thank Microsoft Corporation for providing us with a copy of Microsoft Office Proofing Tools 2003.

References

Ahmed Abdelali, Jim Cowie, and Hamdy S. Soliman. 2004. Arabic information retrieval perspectives. In *Proceedings of the 11th Conference on Natural Language Processing, Journes d'Etude sur la Parole - Traitement Automatique des Langues Naturelles (JEP-TALN)*, Fez, Morocco.

Nasreen Abduljaleel and Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross-language information retrieval. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 139–146. ACM Press.

Jamal B. S. Al-Qinal. 2002. Morphophonemics of loanwords in translation. *Journal of King Saud University*, 13:1–132.

Mohamed Saleh Al-Shanti. 1996. *Al Maharat Allughawia*. Al Andalus for publishing and distribution. 4th edition.

Mohammed Aljlayl and Ophir Frieder. 2002. On Arabic search: improving the retrieval effectiveness via a light stemming approach. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 340–347. ACM Press.

Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0. LDC Catalog No. LDC2002L49.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR-94,*, pages 161–175, Las Vegas, US.

Ted Dunning. 1994. Statistical identification of language. Technical Report MCCS-94-273, Computing Research Lab (CRL), New Mexico State University.

Fredric C. Gey and Douglas W. Oard. 2001. The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries. In *TREC-2001*, volume NIST Special Publication:SP 500-250. National Institute of Standards and Technology.

Kil S. Jeong, Sung Hyon Myaeng, Jae S. Lee, and Key-Sun Choi. 1999. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management*, 35(4):523–540.

Byung-Ju Kang and Key-Sun Choi. 2002. Effective foreign word extraction for Korean information retrieval. *Information Processing and Management*, 38(1):91–109.

Shereen Khoja and Roger Garside. 1999. Stemming Arabic text. Technical report, Computing Department, Lancaster University, Lancaster.

Microsoft Corporation. 2002. Arabic proofing tools in Office 2003.
URL: `http://www.microsoft.com/middleeast/arabicdev/office/office2003/Proofing.asp`.

Bonnie Glover Stalls and Kevin Knight. 1998. Translating names and technical terms in Arabic text. In *COLING/ACL Workshop on Computational Approaches to Semitic Languages*, pages 34–41.

Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *Proceedings of ITCC 2005 Intlernational Conference on Information Technology: Coding and Computing*.

Justin Zobel and Philip Dart. 1995. Finding approximate matches in large lexicons. *Software - Practice and Experience*, 25(3):331–345.