The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database through the Detection and Lemmatization of Unknown Words

Mohammed Attia[1,3] *Younes Samih*[2] *Khaled Shaalan*[1] *Josef van Genabith*[3]
(1) The British University in Dubai, UAE
(2) Heinrich-Heine-Universität, Germany
(3) School of Computing, Dublin City University, Ireland
{mattia,josef}@computing.dcu.ie,
samih@phil.uni-duesseldorf.de,
khaled.shaalan@buid.ac.ae

ABSTRACT

Unknown words, or out of vocabulary words (OOV), cause a significant problem to morphological analysers, syntactic parses, MT systems and other NLP applications. Unknown words make up 29 % of the word types in in a large Arabic corpus used in this study. With today's corpus sizes exceeding 10^9 words, it becomes impossible to manually check corpora for new words to be included in a lexicon. We develop a finite-state morphological guesser and integrate it with a machine-learning-based pre-annotation tool in a pipeline architecture for extracting unknown words, lemmatizing them, and giving them a priority weight for inclusion in a lexical database. The processing is performed on a corpus of contemporary Arabic of 1,089,111,204 words. Our method is tested on a manually-annotated gold standard and yields encouraging results despite the complexity of the task. Our work shows the usability of a highly non-deterministic morphological guesser in a practical and complex application.

TITLE IN ARABIC

القاموس العائم للغة العربية: طريقة آلية لتحديث قاعدة البيانات المعجمية من خلال اكتشاف الكلمات الغير معروفة وردها إلى أصلها

ABSTRACT IN ARABIC

تسبب الكلمات الغير معروفة أو الكلمات الغير مدونة في القواميس مشكلة كبيرة في التحليل الصرفي والإعراب الآلي والترجمة الآلية وغيرها من تطبيقات المعالجة الآلية للغات الطبيعية. فتشكل الكلمات الغير معروفة نسبة 29 % من الكلمات الموجودة في ذخيرة النصوص المستخدمة في هذا البحث. ومع الزيادة الهائلة في حجم ذخائر النصوص التي تتجاوز اليوم مليار كلمة يصبح من المستحيل إجراء أي بحث يدوي عن الكلمات الجديدة لإدراجها في المعاجم الحديثة. ولذلك قمنا بتطوير أداة للتخمين الصرفي قائمة على تقنية آلات الحالة المحدودة وتم دمجها مع أداة قائمة على التعلم الآلي واستخدمناهما معا في عملية تشبه خط الأنابيب تكون مخرجات بعض أجزائه مدخلات لأجزائه الأخرى بحيث نتمكن من استخراج الكلمات الغير معروفة وردها إلى أصلها وإعطائها وزنا يعبر عن الأولوية في الإدراج في قاعدة البيانات المعجمية. ويعتمد هذا البحث على ذخيرة نصوص حجمها 1,089,111,204 كلمة. وقد قمنا باختبار الطريقة التي طورناها باستخدام معيار تم بناؤه يدويا ويقدم نتائج مرضية بالرغم من تعقيد المهمة. وتبين الطريقة التي استخدمناها فائدة أداة التخمين الصرفي الذي يعطي نتائج بها درجة كبيرة من الغموض في تطبيقات عملية ومعقدة.

KEYWORDS : Arabic, unknown words, out of vocabulary words, floating dictionary, lexical enrichment, lexical extension

KEYWORDS IN ARABIC:
اللغة العربية، الكلمات الغير معروفة، الكلمات الغير مدرجة في القواميس، القاموس العائم، الإثراء المعجمي، التوسع المعجمي

Proceedings of COLING 2012: Technical Papers, pages 83–96,
COLING 2012, Mumbai, December 2012.

1 Introduction

Due to the complexity and semi-algorithmic nature of Arabic morphology (that employs numerous rules and constraints on inflection, derivation and cliticization), it has been a challenge for computational processing and analysis (Kiraz, 2001; Beesley 2003). A lexicon is an indispensable part of a morphological analyser (Dichy and Farghaly, 2003; Attia, 2006; Buckwalter, 2004; Beesley, 2001), and the coverage of the lexical database is a key factor in the coverage of the morphological analyser, and limitations in the lexicon will cascade through to higher levels of processing. Moreover, out of vocabulary words (or OOVs) have impact negatively on the performance of parsers (Attia et al., 2010) and MT applications (Huang et al. 2010). This is why an automatic method for updating a lexical database and dealing with unknown words is crucially important.

We present the first attempt, to the best of our knowledge, to address the lemmatization (rather than stemming) of Arabic unknown words. The problem with lemmatizing unknown words is that they cannot be matched against a morphological lexicon. Furthermore, the specific problem with lemmatizing Arabic words is the richness and complexity of Arabic morphological derivational and inflectional processes. For the purposes of this paper, unknown words are words not found by the SAMA morphological analyser (Maamouri et al., 2010) but accepted by the Microsoft Spell Checker. We develop a rule-based finite-state morphological guesser and use a machine learning based disambiguator, MADA (Roth et al., 2008), in a pipeline-based approach to lemmatization.

We test our method against a manually created gold standard of 1,310 types (unique words) and show a significant improvement over the baseline. Furthermore, we devise a novel algorithm for weighting and prioritizing new words for inclusion in a lexicon depending on three factors: number of form variations of the lemmas, cumulative frequency of the forms, and the type of POS (part of speech) tag.

This paper is structured as follows. The remainder of the introduction provides more details on the complexity of the lemmatization process in Arabic, why dealing with unknown words is important, previous work on the topic, and the data used in our experiments. Section 2 presents the methodology we follow in extracting and analysing unknown words. Section 3 provides details on the morphological guesser we develop to help deal with the problem. Section 4 presents and discusses the evaluation results, and Section 5 concludes.

1.1 Complexity of Lemmatization in Arabic

Arabic is an inflectionally rich language with nouns specified for number, gender and case; and verbs specified for tense, number, gender, person, voice and mood. These inflectional processes entail complex alterations on base forms. Arabic is also a clitic language. Clitics are morphemes that have the syntactic characteristics of a word but are morphologically bound to other words (Crystal, 1980). In Arabic, many coordinating conjunctions, the definite article, many prepositions and particles, and a class of pronouns are all clitics that attach themselves either to the start or end of words, and subsequently change the base form according to alteration rules which include assimilation and deletion. These facts complicate the process of lemmatization, or returning the base form given the inflected form.

For English, one can reasonably assume that new words appear very often in their base forms, or the lexical look-up forms. Lindén (2008) indicates that about 86 % of the new words in English appear in their base form. However, in Arabic, which is highly inflectional in nature, only 45 % of new token types in our test set appear in their base form. Moreover, 36 % of the unknown types do not appear in their base form at all in the entire corpus.

1.2 Why Deal with Unknown Words?

Sinclair (1987) introduced the term "Floating Dictionary", a self-updating dictionary that is able to automatically monitor language change. "It would, so to speak, float on top of a corpus, rather like a jelly-fish, its tendrils constantly sensing the state of the language." We think that an electronic 'floating dictionary' should be able to perform at least three major tasks. It should be able to tell which words are not is use anymore, which words have newly appeared in a language, and which word usages or senses have changed based on contemporary data. In this paper we explain our methodology for automatically detecting new words in Arabic, lemmatizing such new words in order to relate multiple surface forms to their base underlying representations, deciding on the word POS tag, collecting statistics on the frequency of use, and modelling human decisions on whether to include the new words in a lexicon or not.

New words are constantly finding their way into any living human language. These new words are either coined or borrowed, or they can be transliterations of proper nouns from other languages. The inclusion of new words in a lexicon is a non-trivial task as it needs to address two important problems. First, there is the problem of detection, or how do we know that a new word has appeared? Second, there is the problem of reaching a decision on the new word, or how do we judge whether the new word is worth adding to the lexicon or not? This is usually done by looking at whether the word is frequent enough, whether it appears in various forms and inflections, and whether it is well-distributed in a corpus. This enables us to determine whether the word constitutes a core lexical item or the usage of the word is just accidental or idiosyncratic.

We address this issue by developing an automatic technique to recognize unknown words and reduce them to their lemmas, predict their POS, and rank them in their order of importance.

1.3 Previous Work

Lemmatization of unknown words has been addressed for Slovene in (Erjavec and Džerosk, 2004), for Hebrew in (Adler at al., 2008) and for English, Finnish, Swedish and Swahili in (Lindén, 2008). Apart from the language involved, our work is different in that we incorporate a finite state guesser in the process. Lemmatization of Arabic words has been addressed in (Roth et al., 2008; Dichy, 2001). The idea of finding and stemming unknown Arabic words has been utilized by Diab et al, (2004). While Diab et al. do not mention unknown words specifically, the fact that they use a character-based classification model and tokenization indicates that they can handle unknown words and perform stemming on them. However, they do not present any evaluation on unknown words specifically. Mohamed and Kübler (2010) handle unknown words explicitly and provide results for known and unknown words in both word segmentation (stemming) and part of speech tagging. They reach a stemming accuracy of 81.39 % on unknown words and over 99 % on known words.

Diab et al.'s and Mohammed and Kübler's work focuses on stemming rather than lemmatization, which are quite distinct albeit frequently confused. The difference between stemming and lemmatization is that stemming strips off prefixes and suffixes and leaves the bare stem, while lemmatization returns the canonical base form. To illustrate this with an example, take the Arabic verb form يقولون 'yqwlwn' "they say". Stemming will remove the present prefix 'y' and the plural suffix 'wn' and leave قول 'qwl' which is a non-word in Arabic. By contrast, full lemmatization will reveal that the word has gone through an alteration process and return the canonical قال 'qAl' "to say" as the base form.

Lemmatization reduces surface forms to their canonical base representations (or dictionary look-up form), i.e, words before undergoing any inflection, which, in Arabic, means verbs in their perfective, indicative, 3rd person, masculine, singular forms, such as شَكَرَ $akara "to thank"; and nominals (the term used for both nouns and adjectives) in their nominative, singular, masculine forms, such as طالِب TAlib "student"; and nominative plural for *pluralia tantum* nouns (or nouns that appear only in the plural form and are not derived from a singular form) , such as ناس nAs "people".

1.4 Data Used

In our work we use a large-scale corpus of 1,089,111,204 words, consisting of the Arabic Gigaword Fourth Edition (Parker et al., 2009) with 925,461,707 words, in addition to 163,649,497 words from news articles crawled from the Al-Jazeera web site. In this corpus, unknown words appear at a rate between 2 % of word tokens (when we ignore possible spelling variants) and 9 % of word tokens (when possible spelling variants are included). In this context spelling variants refer to alternative (sub-standard) spellings recognized by SAMA which are mostly related to the possible overlap between orthographically similar letters, such as the various shapes of *hamzahs* (أ إ ا آ), *taa' marboutah* and *haa'* (ة ه), and *yaa'* and *alif maqsoura* (ي ى).

2 Methodology

To deal with unknown (or out-of-vocabulary) words, we use a pipeline approach which predicts part-of-speech tags and morpho-syntactic features before lemmatization. In the first stage of the pipeline, we use MADA (Roth et al., 2008), an SVM-based tool that relies on the word context to assign POS tags and morpho-syntactic features. MADA internally uses the SAMA morphological analyser (Maamouri et al., 2010), an updated version of Buckalter morphology (Buckwalter, 2004). Second, we develop a finite-state morphological guesser that can provide all the possible interpretations of a given word. The morphological guesser first takes an Arabic surface form as a whole and then strips all possible affixes and clitics off one by one until all possible analyses are exhausted. The morphological guesser is highly non-deterministic as it outputs a large number of solutions. To counteract this non-determinism, all the solutions are matched against the POS and morpho-syntactic tag output for the full surface token by MADA and the analysis with the closest resemblance (i.e. the analysis with the largest number of matching morphological features) is selected.

Beside the complexity of lemmatization described in Section 1.1, the problem is further compounded when dealing with unknown words that cannot be matched by existing lexicons. This requires the development of a finite-state guesser to list all the possible interpretations of an unknown string of letters (explained in detail in Section 3).

To identify, extract and lemmatize unknown Arabic words we use the following sequence of processing steps (Figure 1):

- A corpus of 1,089,111,204 tokens (7,348,173 types) is analysed with MADA.
- The number of types for which MADA could not find an analysis in the Buckwalter morphological analyser is 2,116,180 (about 29 % of the types).

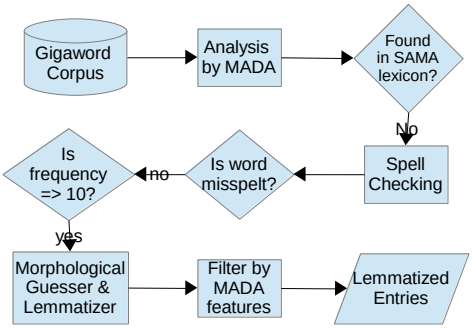

FIGURE 1 – Lemmatization process

- These unknown types were spell checked by the Microsoft Arabic spell checker using MS Office 2010. Among the unknown types of 2,116,180, the number of types accepted as correct is 208,188. The advantage of using spell checking at this stage is that it provides significant filtration of the forms (almost 90 % reduction) and retains a more compact, more manageable, and better quality list of entries to deal with in further processing. The disadvantage is that there is no guarantee that all word forms not accepted by the MS speller are actually spelling mistakes (or that all the ones accepted are correct).
- We select types with frequency of 10 or more of the types accepted by the MS spell checker. This results in a total of 40,277 types.
- We use the full POS tags and morpho-syntactic features produced by MADA.
- We use the finite-state morphological guesser to produce all possible morphological interpretations and relevant lemmatizations.
- We compare the POS tags and morphosyntactic features in MADA output with the output of the morphological guesser and choose the one with the highest matching score.

For testing and evaluation we gold annotate 1,310 words randomly selected from the 40,277 types, providing the gold lemma, the gold POS and lexicographic preference for inclusion in a dictionary. It is to be noted that working with the 2,116,180 types before filtering out possible spelling errors will require annotating a much larger gold standard.

3 Morphological Guesser

Arabic morphotactics allows words to be concatenated with a comparatively large number of clitics (Attia, 2006). Clitics themselves can be concatenated one after the other. Furthermore, clitics undergo assimilation with word stems and with each other, which makes them even harder to handle using surface features only. A verb can comprise up to four tokens (a conjunction, complementizer, verb stem and object pronoun) as illustrated in Table 1. Moreover the verb stem can be prefixed and suffixed with bound morphemes that mark the morpho-syntactic features of tense, number, gender, person, voice and mood. The lemma resides as a nucleus inside layers of proclitics, prefixes, suffixes and enclitics. A verb lemma like شكر '$akara' "to thank" can generate up to 2,552 different valid forms.

Proclitics		**Prefix**	**Lemma**	**Suffix**	**Enclitic**
Conjunction/ question article	*Comp*	*Tense/mood – number/gender*	*Verb*	*Tense/mood – number/gender*	*Object pronoun*
Conjunctions و wa 'and' or ف fa 'then'	ل li 'to'	Imperfective tense (5)	lemma	Imperfective tense (10)	First person (2)
Question word أ > 'is it true that'	س sa 'will'	Perfective tense (1)		Perfective tense (12)	Second person (5)
	ل la 'then'	Imperative (2)		Imperative (5)	Third person (5)

TABLE 1 – Proclitics, enclitics, prefixes and suffixes with Arabic verbs

Proclitics			**lemma**	**Suffix**	**Enclitic**
Conjunction/ question article	*Preposition*	*Definite article*	*Noun*	*Gender/Number*	*Genitive pronoun*
Conjunctions و wa 'and' or ف fa 'then'	ب bi 'with', ك ka 'as' or ل li 'to'	ال Al 'the'	Stem	Masculine Dual (4)	First person (2)
				Feminine Dual (4)	
Question word أ > 'is it true that'				Masculine regular plural (4)	Second person (5)
				Feminine regular plural (1)	Third person (5)
				Feminine Mark (1)	

TABLE 2 – Proclitics, enclitics, prefixes and suffixes with Arabic nouns

Similarly a noun stem can be attached to up to three clitics as shown in Table 2. Although Table 2 shows four clitics, we note that the definite article and the genitive (or possessive) pronoun are mutually exclusive. Nominal stems can also be suffixed with bound morphemes that mark the morpho-syntactic features of number, gender and case. a typical noun like معلم 'muEal~im' 'teacher', generates 519 valid forms.

We develop a finite state (Beesley and Karttunen, 2003; Hulden, 2009) morphological guesser for Arabic that can analyse unknown words with all possible clitics, morpho-syntactic affixes and all relevant alteration operations that include insertion, assimilation, and deletion. Beesley and Karttunen (2003) give some advice on how to create a basic guesser. The core idea of a guesser is to assume that a stem is composed of any arbitrary sequence of non-numeric characters, and this stem can be prefixed and/or suffixed with a predefined set of prefixes, suffixes or clitics. The guesser marks clitic boundaries and tries to return the stem to its default unmarked form, the lemma. Due to the nondeterministic nature of the guesser, there will be a multitude of possible lemmas for each form. The Arabic FST guesser consists of three parts: a lexc file, alteration rules and an XFST compilation file. First, there is the lexc file (Figure 2) with lexicons and continuation classes for the Arabic guesser. The lexc file specifies that there is an optional conjunction, followed by an optional preposition, followed by an optional definite article before the Arabic noun.

```
LEXICON Conjunctions
و-:و-+conj:                    Prepositions;
ف-:ف-+conj:                    Prepositions;
                               Prepositions;

LEXICON Prepositions
ل-:ل-+prep:                    Article;
ك-:ك-+prep:                    Article;
ب-:ب-+prep:                    Article;
                               Article;

LEXICON Article
ال-+defArt                     Nouns;
                               Nouns;

LEXICON Nouns
+noun+fem                      GuessWords;
+noun+masc                     GuessWords;
^ss^خادم^se^+noun+masc         FemMascduFemduMascplFempl;
....
LEXICON GuessWords
^ss^^GUESSNOUNSTEM^^se^        FemMascduFemduMascplFempl;
^ss^^GUESSNOUNSTEM^^se^        FemMascduFemduFempl;
^ss^^GUESSNOUNSTEM^^se^        FemMascduFemdu;
^ss^^GUESSNOUNSTEM^^se^        MascduFempl;
^ss^^GUESSNOUNSTEM^^se^        Mascdu;
^ss^^GUESSNOUNSTEM^^se^        Fempl;
^ss^^GUESSNOUNSTEM^^se^        FemduFempl;
^ss^^GUESSNOUNSTEM^^se^        Femdu;
^ss^^GUESSNOUNSTEM^^se^        NoNumber;
```

FIGURE 2 – Snapshot of the Arabic lexc file

Second, there are the alteration rules which handle the morphological processes of assimilation and deletion. In our system there are about 130 replace rules to handle alterations that affect verbs, nouns, adjectives and function words when they undergo inflections or are attached to affixes and clitics. They take the form of XFST replace rules:

```
A -> w || "+pres" Alphabet _ Alphabet
```

The example rule indicates that ‘A’ changes to ‘w’ under the condition of having the left context ‘+pres’ and a single alphabetical character and the right context of another alphabetical character. Following this rule the verb قال qAl “to say” will change to يقول yaqwl in the present tense form.

Third, there are the XFST compilation rules which bind components together. They replace the multivariable words ‘GUESSNOUNSTEM’ and ‘GUESSVERBSTEM’ with the relevant alphabet using the ‘substitute defined’ command. The XFST commands in our guesser are stated as follows.

```
define Alphabet
define PossNounStem [[Alphabet]^{2,24}] "+Guess":0;
define PossVerbStem [[Alphabet]^{2,6}] "+Guess":0;
substitute defined PossNounStem for "^GUESSNOUNSTEM^"
substitute defined PossVerbStem for "^GUESSVERBSTEM^"
```

This states that a possible noun stem is defined as any sequence of Arabic non-numeric characters of length between 2 and 24 characters. A possible verb stem is between 2 and 6 characters. This word stem is surrounded by prefixes, suffixes, proclitics and enclitics. Clitics are considered as independent tokens and are separated by the ‘@’ sign, while prefixes and suffixes are considered as morpho-syntactic features and are interpreted with tags preceded by the ‘+’ sign. Below we present the analysis of the noun والمُسَوِّقونَ wa-Al-musaw~iquwna “and-the-marketers”, and the verb سَيَأْخُذُنا sa-ya'xu*unA “will-take-us”.

MADA output for wa-Al-musaw~iquwna:

form:wAlmswqwn	num:p	gen:m	per:na	case:n	asp:na	mod:na	vox:na	pos:noun
prc0:Al_det	prc1:0	prc2:wa_conj		prc3:0	enc0:0	stt:d		

Finite-state guesser output for wa-Al-musaw~iquwna:

```
والمسوقون  +adjوالمسوقّ+Guess+masc+pl+nom@
والمسوقون  +adjوالمسوقون+Guess+sg@
والمسوقون  +nounوالمسوقّ+Guess+masc+pl+nom@
والمسوقون  +nounوالمسوقون+Guess+sg@
والمسوقون  +conj@و+defArt@ال+adjمسوقّ+Guess+masc+pl+nom@
والمسوقون  +conj@و+defArt@ال+adjمسوقون+Guess+sg@
والمسوقون  +conj@و+defArt@ال+nounمسوقّ+Guess+masc+pl+nom@      [correct match]
والمسوقون  +conj@و+defArt@ال+nounمسوقون+Guess+sg@
…
```

MADA output for wa-sa-ya'xu*unA:

form:sy>x*nA	num:s	gen:m	per:na	case:na	asp:na	mod:i	vox:a	pos:verb
prc0:0	prc1:0	prc2:0	prc3:0	enc0:1p_poss	stt:na			

Finite-state guesser output for wa-sa-ya'xu*unA:

```
سيأخذنا  +adjسيأخذن+Guess+dual+nom+compound@
سيأخذنا  +adjسيأخذنا+Guess+sg@
سيأخذنا  +nounسيأخذ+Guess+sg@نا+genpron+1pers+@
سيأخذنا  +nounسيأخذنا+Guess+sg@
سيأخذنا  +verb+impسيأخذ+Guess+2pers+masc+sg@نا+objpron+1pers+pl@
سيأخذنا  +verb+impسيأخذن+Guess+2pers+dual@
سيأخذنا  س+fut+art@+verb+pres+pass+3persأخذنا+Guess+masc+sg@
سيأخذنا  س+fut+art@+verb+pres+active+3persأخذ
          +Guess+masc+sg@نا+objpron+1pers+pl@                    [correct match]
سيأخذنا  س+fut+art@+verb+pres+active+3persأخذنا+Guess+masc+sg@
…
```

For a list of 40,277 unknown word types, the morphological guesser produces an average of 12.6 possible interpretations per word. This is highly non-deterministic when compared to the finite state morphological analyser (Attia et al., 2011) which has an average of 2.1 solutions per known word. We also note that 97 % of the gold lemmas in our test set are found among the finite-state guesser's choices, which indicates the high performance of the guesser.

4 Testing and Evaluation

To evaluate our methodology we create a manually annotated gold standard test suite of randomly selected surface form types as mentioned in Section 2. For these surface forms, the gold lemma and part of speech are manually provided. In addition, a human annotator indicates a preference on whether or not to include the entry in a dictionary, that is whether a lemmatized form makes a valid dictionary entry or not. We noticed that most of the forms marked by the annotator as not fitting for inclusion in a dictionary were proper nouns, misspelled words, colloquial words, and words that form a part of a multiword expression. By contrast, nouns, verbs adjectives, and proper nouns with significantly high frequency were marked for inclusion in the lexical database. It is to be mentioned that proper nouns in Arabic are not orthographically distinguished from other words, i.e. there is no capitalization in Arabic as is the case in European languages. This feature of lexicographic preference helps to evaluate our lemma weighting algorithm discussed in Section 4.2. The size of the test suite is 1,310 word form types.

We observe that proper nouns are the most frequent category (45 %) among the unknown words types in the data, and they also cover about 61 % of the unknown token instances in the gold annotated dataset. The POS distribution of the unknown token types of our annotated data is shown in Table 3. As expected, most unknown words are open class words: proper names, nouns, adjectives, and, to a lesser degree, verbs.

Gold POS	**Type Count**	**Ratio**
noun_prop	584	45 %
noun	264	20 %
adj	255	19 %
verb	52	4 %
noun_fem_plural (pluralia tantum)	28	2 %
noun_broken_plural	28	2 %
others: noun_masc_plural (*pluralia tantum*) (4) part (3) pron_dem (1)	8	0.6 %
Excluded		
misspelling	55	4 %
not_known	15	1 %
colloquial	19	1.5 %
Lexicographic relevance		
Include in a dictionary	671	51 %
Don't include in a dictionary	639	49 %

TABLE 3 – Gold tag annotation of the test suite

4.1 Evaluating Lemmatization

In the evaluation experiment we measure accuracy calculated as the number of correct tags divided by the count of all tags. The baseline is given by the assumption that new words appear in their base form, i.e., we do not need to lemmatize them. The baseline accuracy is 45 %. The POS tagging baseline proposes the most frequent tag (proper name) for all unknown words. In our test data accuracy stands also at 45 %. We notice that MADA POS tagging accuracy for unknown words is unexpectedly low (60 %) as shown in Table 4. We use Voted POS Tagging, that is we choose the POS tag assigned most frequently in the data to a lemma. This method has improved the tagging results significantly (Table 4).

As for the lemmatization process, our first experiment in the pipeline-based lemmatization approach obtains a higher score (54 %) than the baseline (45 %) as shown in Table 5.

		Accuracy
	POS tagging	
1	POS Tagging baseline	45 %
2	MADA POS tagging	60 %
3	Voted POS Tagging	69 %

TABLE 4 – Evaluation of POS tagging of unknown words

Examining the data further, we notice that when a proper noun is prefixed with the definite article "Al", the definite article is not stripped off in the gold annotation and is considered as part of the lemma, such as القشيري 'Al-qu$ayriy'. In MADA morpho-syntactic tagging, the definite article is considered as a clitic and not part of the lemma. When this difference is ignored in the second experiment, the lemmatization accuracy increases from 54 % to 63 %. A more detailed error analysis will help devise better heuristics to increase the accuracy of the pipeline-based lemmatization. For example, in the gold annotation some regular feminine and masculine plural forms are considered as *pluralia tantum*, while in the automatic lemmatization they are reduced to their singular forms, such as حجوزات HujuwzAt "bookings".

	Lemmatization	
1	Lemmas found among corpus forms	64 %
2	Lemmas found among fst guesser forms	97 %
3	Lemma selection baseline	45 %
4	Pipeline-based lemmatization (selection decision) with strict definite article matching	54 %
5	Pipeline-based lemmatization (selection decision) ignoring definite article matching	63 %

TABLE 5 – Evaluation of lemmatization of unknown words

The test results indicate significant improvements over the baseline. However, we expect that substantial further improvements can be obtained through further extensive error analysis and developing refined heuristics.

4.2 Evaluating Lemma Weighting

We create a weighting algorithm for ranking and prioritizing unknown words in Arabic so that important words that are valid for inclusion in a lexicon are pushed up the list and less interesting words (from a lexicographic point of view) are pushed down. This is meant to facilitate the effort of manual revision by making sure that the top part of the stack contains the words with highest priority.

In our case we have 40,277 unknown token types. After lemmatization they are reduced to 18,399 types (that is 54 % reduction of the surface forms). This number is still too big for manual validation. In order to address this issue we devise a weighting algorithm for ranking so that the top *n* number of words will include the most lexicographically relevant words. We call surface forms that share the same lemma 'sister forms', and we call the lemma that they share the 'mother lemma'. The weighting algorithm is based on three criteria: number of sister forms, cumulative frequency of the sister forms, and a POS factor. The POS factor gives 50 extra points to verbs, 30 to to nouns and adjectives, and nothing to proper nouns. This is meant to penalize proper nouns due to their high frequency which is disproportionate to other categories. The parameters of the weighting algorithm have been tuned through several rounds of experimentation.

```
Word Weight = ((number of sister forms * 800) +
      cumulative sum of frequencies of sister forms) / 2 +
    POS factor
```

We use the gold annotated data for the evaluation of the lemma weighting criteria, as shown in Table 6. We notice that the combined criteria gives the best balance between increasing the number of lexicographically-relevant words in the top 100 words and reducing the number of lexicographically-relevant words in the bottom 100 words.

Lexicographically-relevant words	In top 100	In bottom 100
relying on Frequency alone (baseline)	63	50
relying on number of sister forms * 800	87	28
relying on POS factor	58	30
using combined criteria	78	15

TABLE 6 – Evaluation of lemma weighting and ranking

Table 7 shows a sample of the entries in the unknown words lexicon. The list includes a spectrum of the different word categories such as proper nouns, adjectives, nouns, broken plural and feminine plural forms, as well as verbs.

#	FST Guessed lemma	Gloss	Weight	Forms
Proper Nouns				
1	أوباما >ubAmA	Obama	40421	لأوباما # أوباما # فأوباما # وأوباما # ولأوباما # بأوباما
2	ساركوزي sArkuwziy	Sarkozy	29361	وساركوزي # فساركوزي # بساركوزي # ساركوزي
3	توتنهام tuwtinhAm	Tottenham	08829	بتوتنهام # وتوتنهام # لتوتنهام # ولتوتنهام # توتنهام
Adjectives				
4	منخرط munxariT	involved	09302	والمنخرطة # ومنخرطة # منخرطات # منخرطة # المنخرطة # المنخرطات
5	متواطئ mutawAti}	conspiring	07016	متواطئان # ومتواطئ # متواطئين # المتواطئ # كمتواطئين # والمتواطئ # والمتواطئون # المتواطئين # ومتواطئين # متواطئون # للمتواطئين # والمتواطئين # متواطئ # ومتواطئون
6	مستتر musotatir	hidden	03329	والمستتر # المستترين # المستتر # مستتر # مستترين # والمستترين # ومستتر
Nouns				
7	اقتياد AqotiyAd	leading	08559	واقتياده # واقتيادهما # اقتياد # لاقتياده # واقتيادهم # لاقتياد # اقتيادها # الاقتياد # باقتياد # باقتياده # وباقتياده # باقتيادهم # اقتيادهم # اقتيادنا # اقتيادي # واقتياد # واقتيادها # اقتيادهما
8	محاصصة muHASaSap	sharing	07056	المحاصصة # ومحاصصة # للمحاصصة # محاصصة # بالمحاصصة # بمحاصصة # والمحاصصة # فالمحاصصة
9	ارتهان ArotihAn	dependence	06616	الارتهان # وارتهان # وارتهانها # ارتهانها # ارتهان # والارتهان # بالارتهان # وارتهانه # ارتهانه # ارتهانهم # لارتهان
Broken Plurals				
10	خصال xiSAl	features	08491	بخصالك # خصال # وخصال # خصاله # الخصال # لخصاله # خصالهم # بخصال # خصالنا # خصالك # بخصاله # بالخصال # خصالها # وخصاله # وخصالها # لخصال # والخصال # وخصالك
11	مكائد makA}id	tricks	05785	للمكائد # مكائدهم # بالمكائد # والمكائد # مكائد # مكائده # المكائد # ومكائده # ومكائدهم # بمكائد # لمكائد # ومكائد # ومكائدها
12	دفوع dufuwE	defences	04418	بالدفوع # الدفوع # دفوعه # دفوعهم # دفوع # والدفوع # ودفوع # دفوعها # بدفوع
Feminine plural forms				
13	صياغة Siyagap	formation	07168	وبصياغات # وصياغات # وصياغاتها # صياغاتهم # بصياغات # صياغتين # والصياغات # لصياغات # صياغاتها # صياغات # الصياغتين # بالصياغات # الصياغات # وصياغاته # صياغاته
14	خصومة xuSuwmap	animosity	06728	والخصومات # خصوماتهم # خصوماته # بخصومات # وخصوماته # خصومات # لخصومات # وخصوماتهم # خصوماتها # خصوماتنا # وخصوماتها # بالخصومات # الخصومات
15	مرارة marArap	bitterness	05339	مراراتها # المرارات # بمراراته # لمرارات # مراراته # والمرارات # بمرارات # ومرارات # مرارات # ومراراتها
Verbs				
16	عسكر Easokara	to militarize	05255	عسكرين # وعسكرت # ليعسكر # يعسكرون # وسيعسكر # العسكرين # سيعسكر # ويعسكر # يعسكر # عسكروا # ستعسكر
17	سيس say~asa	to politicize	04223	سيس # يسيسوا # بسيسي # سيسون # وسيس # تسيس # تسيسوا # وسيسون # يسيسون
18	هندس hanodasa	to design/ engineer	03431	هندس # هندسها # يهندسوا # هندست # يهندسون # هندسوا # يهندس # يهندسها

TABLE 7 – Sample entries selected from the unknown words lexicon

As the corpus is composed mainly of news articles, we assume that the distribution of proper nouns is artificial and arbitrary as it depends, to a large extent, on the specific date and time of an event or series of events that occupies the news for a certain (short-term or long-term) duration. For example, as Table 7 shows, *Obama* and *Sarkozy* ranked top of the list of unknown words, but

now as Sarkozy is no longer the French president and the fate of Obama will be determined in the next presidential election in America, whether these names will continue to maintain the same level of frequency is questionable. This is why verbs, adjectives and nouns constitute the core of the language lexicon, while proper nouns are, to some extent, temporal and transient and the frequency of their use tends to shift from time to time.

Conclusion

We have developed a methodology for automatically updating an Arabic dictionary by extracting unknown words from data and lemmatizing them in order to relate multiple surface forms to their canonical underlying representation using a finite-state guesser and a machine learning tool for disambiguation. We have developed a weighting mechanism for simulating a human decision on whether or not to include new words in a general-domain lexical database. We have shown the feasibility of a highly non-deterministic finite state guesser in an essential application. Out of a word list of 40,255 unknown words we created a lexicon of 18,399 lemmatized, POS-tagged and weighted entries. We have made our unknown word lexicon available as a free open source resource (http://arabic-unknowns.sourceforge.net/).

Acknowledgments

This research is funded by the Irish Research Council for Science Engineering and Technology (IRCSET), the UAE National Research Foundation (NRF) (Grant No. 0514/2011), and the Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

References

Adler, M., Goldberg, Y., Gabay, D. and Elhadad, M. (2008). Unsupervised Lexicon-Based Resolution of Unknown Words for Full Morpholological Analysis. In: Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio.

Attia, M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In: Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK.

Attia, Mohammed, Jennifer Foster, Deirdre Hogan, Joseph Le Roux, Lamia Tounsi and Josef van Genabith. (2010). 'Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French'. First Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010), NAACL HLT. Los Angeles, CA.

Attia, Mohammed, Pavel Pecina, Lamia Tounsi, Antonio Toral, Josef van Genabith. (2011). An Open-Source Finite State Morphological Transducer for Modern Standard Arabic. International Workshop on Finite State Methods and Natural Language Processing (FSMNLP). Blois, France.

Beesley, K. R. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In: The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France.

Beesley, K. R., and Karttunen, L.. (2003). Finite State Morphology: CSLI studies in computational linguistics. Stanford, Calif.: Csli.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer (BAMA) Version 2.0. Linguistic Data Consortium (LDC) catalogue number LDC2004L02, ISBN1-58563-324-0

Crystal, D. (1980). *A First Dictionary of Linguistics and Phonetics*. London: Deutsch.

Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. Proceedings of Human Language Technology-North American Association for Computational Linguistics (HLT-NAACL)

Dichy, J. (2001). On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases. ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects. Toulouse, France.

Dichy, J., and Farghaly, A. (2003). Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built? In: The MT-Summit IX workshop on Machine Translation for Semitic Languages, New Orleans.

Erjavec, T., and Džerosk, S. (2004). Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. *Applied Artificial Intelligence*, 18:17–41.

Huang, Chung-chi, Ho-ching Yen and Jason S. Chang. (2010). Using Sublexical Translations to Handle the OOV Problem in MT. in Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas(AMTA).

Hulden, M. (2009). Foma: a finite-state compiler and library. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09). Stroudsburg, PA, USA.

Kiraz, G. A. (2001). *Computational Nonlinear Morphology: With Emphasis on Semitic Languages*. Cambridge University Press.

Lindén, K. (2008). A Probabilistic Model for Guessing Base Forms of New Words by Analogy. In CICling-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel, pp. 106-116.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., and Kulick, S. (2010). LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.1. LDC Catalog No. LDC2010L01. ISBN: 1-58563-555-3.

Mohamed, Emad; Sandra Kübler (2010). Arabic Part of Speech Tagging. Proceedings of LREC 2010, Valetta, Malta.

Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2009). Arabic Gigaword Fourth Edition. LDC Catalog No. LDC2009T30. ISBN: 1-58563-532-4.

Roth, R., Rambow, O., Habash, N., Diab, M., and Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In: Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio.

Sinclair, J. M. (ed.). (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.