A Jellyfish Dictionary for Arabic

Mohammed Attia, Josef van Genabith

School of Computing, Dublin City University, Ireland
{mattia,josef}@computing.dcu.ie

Abstract

In a festschrift to Martin Gellerstam (Gottlieb and Mogensen, 2007), an article was published by John Sinclair in which he introduced the concept of a *jellyfish dictionary*. It presented the idea of 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 *jellyfish dictionary* should be able to perform three major tasks. It should be able to tell which words have newly appeared in a language, which words are not in use anymore, and which word usages or senses have changed based on contemporary data. In this paper we explain our methodology for realizing a jellyfish dictionary for Arabic by automatically performing the three tasks: detecting new words, flagging obsolete words, and discovering word senses.

Keywords: Arabic; automatic lexical acquisition, detection of new words, obsolete word detection, word senses

1. Introduction

A corpus is the foundation for any lexicographic work, as both a source of lexical knowledge and evidence underpinning theoretical assumptions related to dictionary entries. However, most of the lexicographic work to date has concentrated on the evidence part of the corpus, rather than the knowledge part. Today's dictionaries are inspired and supported by corpora, rather than shaped by them. This is where the need for a *jellyfish dictionary* emerges. The idea of a *jellyfish dictionary* was first introduced in an article published by John Sinclair (Gottlieb and Mogensen, 2007) in which he put forward the concept of 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."

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. The idea of a jellyfish dictionary is to develop intelligent tools to allow the corpus to manage the dictionary from top to bottom. The tendrils of the jellyfish sense changes in the sea of words in the corpus and inform us about new developments.

We uphold that an electronic *jellyfish dictionary* needs to perform three major tasks: detecting new words appearing in a language, flagging obsolete words, and observing word senses by identifying the contexts in which words usually prefer to appear. In

this paper, we present our methodology for performing these three tasks. First, we automatically detect new words in Arabic, lemmatize new words in order to relate multiple surface forms to their base underlying representations, decide on words' part of speech (POS), collect statistics on the frequency of use, and model human decisions on whether to include the new words in a lexicon or not. Second, we signal obsolete words in a traditional dictionary based on statistics from a large corpus and a number of web search sites. Third, we investigate word senses based on their preferred contexts, concentrating on the extraction of subcategorization frames and word trigrams.

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[1]. In this corpus, new words appear at a rate of between 2% of word tokens (when we ignore possible spelling variants) and 9% of word tokens (when possible spelling variants are included). For the purposes of this study, new words are words not recognized by the SAMA morphological analyzer (Maamouri et al., 2010), and 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).

Our techniques and methods in dealing with the extraction and lemmatization of new words are evaluated on a held-out manually-annotated gold standard of 2,103 form types (unique words), improving on previous work by Attia et al. (2012).

This paper is structured as follows. Section 2 presents the methodology we follow in extracting and analysing new words. Section 3 explains how obsolete words are automatically detected. Section 4 provides details on how word senses can be ranked according to their frequency in the corpus in certain contexts (subcategorization frames and trigrams), and Section 5 concludes the paper.

2. Detecting New Words

New words are constantly finding their way into any living human language. These new words are either coined or borrowed and reflect changes in our societies and lives. Words such as تويتر twiytar 'twitter', محاصصة muHASaSap 'allotting shares', عَسْكَرَ Easokara 'to militarize', and سَيَّسَ say~asa 'to politicize' are not included in current Arabic dictionaries. The inclusion of new words in a lexicon needs to address three important problems. First, the detection, or the method by which we know that a new word has appeared. Second, lemmatization, or relating multiple surface forms to their canonical representation. Third, reaching a decision on the new word; that is,

[1] http://aljazeera.net/portal

how we judge whether the new word should be added to the lexicon or not. We address this issue by developing an automatic technique to recognize unknown words in a large corpus of 10^9 words, and reduce them to their lemmas, predict their POS, and rank them in their order of lexicographic importance.

In previous proof-of-concept research, Attia et al. (2012), thereafter referred to as Attia2012, detect a total of 2,116,180 new types. They filter this list using a frequency threshold and a spell checker, creating a subset of 40,277 new types. After lemmatization, the list is reduced to 18,000 possible unique new lemmas. The drawback with filtering in the pre-processing stage through spell checking is that it could be throwing the baby out with the bath water. There is no guarantee that all word forms not accepted by the spell checker used are actually spelling mistakes (or even that all the ones accepted are correct).

In the research presented here we show that filtering in the pre-processing stage actually leads to discarding potentially useful information too early. In our new gold standard of 2,103 types, 1,074 were incorrectly tagged as misspelt by the automatic spell checker, resulting in only 48.93% accuracy for unknown words. Furthermore, of the terms incorrectly tagged as misspellings, 20.58% were nominated to be included in a dictionary (9.59% when excluding proper nouns).

Similar problems arise with the idea of excluding types based on their frequency. Word forms with low frequency may interact with other word forms to support a certain lemma, and throwing them out too early risks losing potentially important information. For example, in our data the word form وديناميانتا wadiynamiy~AtinA 'and-our-dynamics' has a frequency of one, but it interacts with 31 other sister forms (such as والديناميات 'and-dynamics', ديناميانهم 'their-dynamics') with an accumulated frequency of 3,464, to support the lemma دينامية diynamiy~ap 'dynamic'. In our new gold standard test set of 2,103 types, a subset of 701 types is selected from the frequency range of 10 repetitions or less. When analyzed, we found that 306 types of them were valid (43.65%). Of the valid types, 94 (30.72%) participated with other forms to support a certain lemma and all of them were nominated for inclusion in a dictionary.

In the current research we apply our technique to the full list of 2,116,180 unknown types from Attia2012. We test our method against a manually created gold standard of 2,103 types and show a significant improvement over the baseline and Attia2012. Furthermore, we investigate different criteria for weighting and prioritizing new words for inclusion in a lexicon depending on four factors: number of form variations of the lemmas, cumulative frequency of the forms, type of POS tag, and spelling correctness (according to a spell checker).

2.1 Lemmatization

In order to deal with new words we need to address the issue of lemmatization.

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'.

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.

Lemmatization of unknown words has been addressed for Slovene in Erjavec and Džerosk (2004), for Hebrew in Adler at al. (2008), for Spanish in Grefenstette et al. (2002), and for English, Finnish, Swedish and Swahili in Lindén (2008). Lemmatization of Arabic has been addressed in Roth et al. (2008) and Dichy (2001). Mohamed and Kübler (2010) handle Arabic unknown words 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.

Mohammed and Kübler's work, however, focuses on stemming rather than lemmatization, which is 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 words to their canonical base forms. To illustrate this with an example, consider the Arabic verb form يقولون yaquwluwn 'they say'. Stemming will remove the present prefix 'ya' and the plural suffix 'uwn' and leave 'quwl' 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.

We develop a rule-based finite-state (Beesley and Karttunen, 2003; Hulden, 2009) morphological guesser that can deal with morphological concatenations and alterations and integrate it with a machine learning based disambiguator, MADA (Roth et al., 2008), in a pipeline-based approach to lemmatization.

3. Methodology

To deal with unknown (or out-of-vocabulary) words, we use a pipeline approach which predicts POS 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 analyzer (Maamouri et al., 2010), an

updated version of the Buckwalter morphology (Buckwalter, 2004). Second, we use a finite-state morphological guesser that provides all possible interpretations of a given word. The morphological guesser first takes an Arabic surface form as a whole and then strips off all possible affixes and clitics one by one until all possible analyses are exhausted, and it also reverses the effect of morphological alteration rules. The morphological guesser is highly non-deterministic as it outputs a large number of solutions. To counteract this problem, all the solutions are matched against the POS and morpho-syntactic features produced by MADA, and the analysis with the closest resemblance (i.e. the analysis with the largest number of matching morphological features between the FS guesser and MADA) is selected.

For illustration, we present the analysis of the verb form ويتناهشونها wa-yatanAha$uwna-hA 'and-they-snatch-it' by MADA and the different analyses by the finite state guesser sorted according to the number of features that are successfully matched with the MADA analysis of the original surface form.

MADA output for wa-yatanAha$uwna-hA:

form:wytnAh$wnhA num:p gen:m per:3 case:na asp:i mod:i vox:a
pos:verbprc0:0 prc1:0 prc2:wa_conj prc3:0 enc0:3fs_dobj stt:na

Finite-state guesser output for wa-yatanAha$uwna-hA:

9 و+conj@+verb+pres+active+3persتناهش+Guess
+masc+pl+nom@ها+objpron+3pers+sg+fem@
7 و+conj@+verb+pres+active+3persتناهشو+Guess
+fem+pl@ها+objpron+3pers+sg+fem@
-2 و+conj@+adjيتناهشونها+Guess+sg@
-2 و+conj@+nounيتناهشونها+Guess+sg@
-2 +adjويتناهشونها+Guess+sg@
-2 +nounويتناهشونها+Guess+sg@
-3 +adjيتناهشونهو+Guess+dual+nom+compound@
-3 و+conj@+adjيتناهشونه+Guess+dual+nom
+compound@
-3 +nounويتناهشونه+Guess+dual+nom+compound@

The matching uses positive scores for matches and negative scores for features found in the finite state output but not present in the MADA output. The top (highest scoring) analysis is selected as the correct lemma of the word.

Figure 1 shows the steps taken to identify, extract and lemmatize unknown Arabic words, which are summarized as follows:

- A corpus of 1,089,111,204 tokens (7,348,173 types) is analyzed with MADA to produce POS tags and morpho-syntactic features.
- The number of types for which MADA could not find an analysis in the Buckwalter morphological analyzer is 2,116,180 (about 29 % of the types). After removing common spelling variants (as detected by MADA), 1,698,852 types remained.

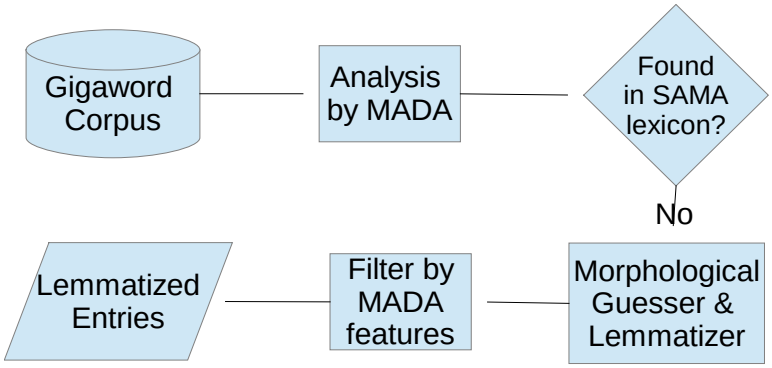

Figure 1: New word extraction and lemmatization process

- Unknown words are analyzed with our finite-state morphological guesser to produce all possible morphological interpretations and relevant possible lemmatizations.
- POS tags and morpho-syntactic features in MADA output are compared with the output of the morphological guesser and the FST guesser analysis with the highest matching score is chosen.

As lemmatization is expected to merge forms having the same lemma together, after lemmatization the list of 1,698,852 types is reduced to 982,886 lemmas, which is too large. We conduct initial filtration by removing word forms that have no supporting morphological variation and which occur only once in the corpus. This basic filtration further reduces the number to 476,349 lemmas.

4. Gold standard Creation

In order to evaluate our methodology we need to create a gold standard from a randomly selected subset of the data. As mentioned earlier, our unknown word list consists of 1,698,852 types. We find that words have varying frequency ranges with a minimum frequency of one, a maximum of 75,885 and a mean of 9.79, as shown in Table 1.

Statistic	Value
Unknown words (after discarding spelling variants)	1,698,852
Minimum frequency	1
Maximum frequency	75,885
Mean	9.79

Table 1: Frequency statistics of the unknown words

When we select a random sample of the data we find that the sample is biased towards low frequency words. Out of 3,000 randomly-selected types, there are 2745

(91.50%) with frequency of 10 or less. This is also true of the entire population where 91.03% of the unknown types have a frequency of 10 or less.

When we investigate the frequency distribution of the unknown words, we see that, as expected, they follow the Zipfian law with a few words having very high frequency and a large number of words having very low frequency (Figure 2).

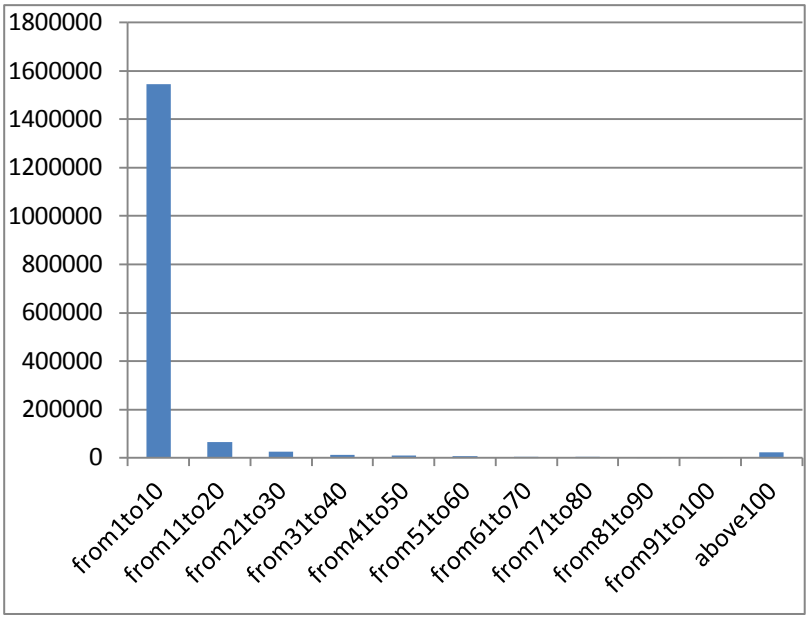

Figure 2: Frequency distribution of the unknown words

In order to avoid the bias towards low frequency words produced by pure randomization, we use a method known in corpus linguistics as 'stratified sampling' or what we may call here 'stratified randomization'. We randomly select 701 words with frequency ≤10, 701 words with frequency >10 and ≤50, and 701 words with frequency >50, so that our test suite becomes representative of three major frequency ranges, as shown in Figure 3.

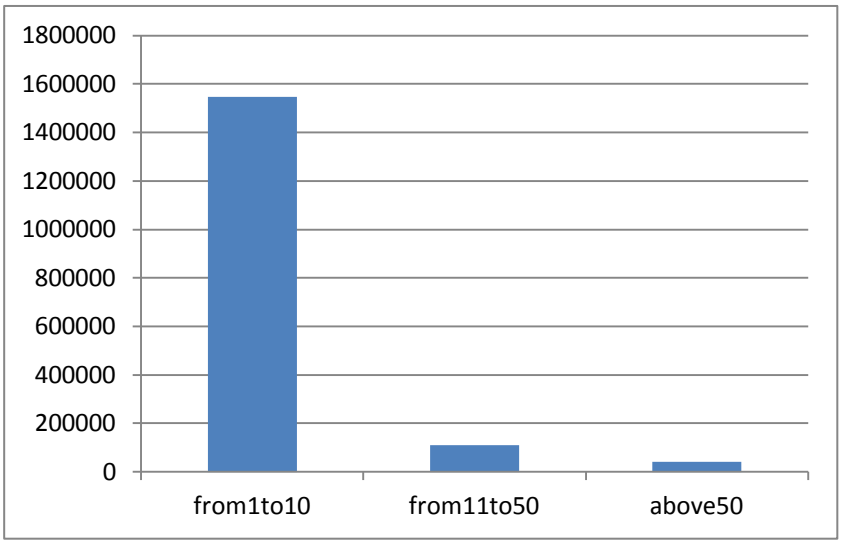

Figure 3: Major frequency ranges of the unknown words

Having created our gold standard of 2,103 unknown types, we ask a human annotator to provide the gold lemma and part of speech for each word form. In addition, the human annotator indicates a preference for 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 the forms marked by the annotator as not fitting for inclusion in a dictionary were mostly misspelled words, colloquial words, and low frequency proper nouns.

Gold Annotation	**Jellyfish2013**				**Attia2012** 1,310 types
	Freq ≤10 701 types	**Freq >10 and ≤50** 701 types	**Freq >50** 701 types	**all** 2,103 types	
Valid Forms: of them	**43.65%**	**75.46%**	**82.31%**	**67.14%**	**93.05%**
noun_prop	70.92%	77.5%	75.74%	75.35%	48.07%
noun	15.03%	10.4%	10.4%	11.4%	21.16%
adj	11.44%	9.26%	9.88%	9.99%	20.75%
verb	2.29%	1.51%	2.25%	1.98%	4.27%
noun_fem_plural (pluralia tantum)	0.33%	0.38%	0.52%	0.42%	2.3%
noun_broken_plural	0.33%	0.38%	1.04%	0.64%	2.3%
Invalid Forms: of them	**56.35%**	**24.54%**	**17.69%**	**32.86%**	**6.95%**
misspelling	60.00%	65.12%	71.77%	63.39%	62.64%
not_resolved	34.68%	19.77%	13.71%	27.21%	16.48%
colloquial	5.06%	15.12%	14.52%	9.26%	20.88%
Lexicographic relevance					
Include in a dictionary	9.84%	13.12%	40.66%	21.21%	51.22%
Include in a dictionary, term not a proper noun (subset of the above)	9.70%	13.12%	16.98%	13.27%	44.35%
Do not include in a dictionary	90.16%	86.88%	59.34%	78.79%	48.78%

Table 2: Gold tag annotation of the test suite

By contrast, nouns, verbs, adjectives, and proper nouns with significantly high frequency were marked for inclusion in the lexical database. This feature of lexicographic preference helps to evaluate our lemma weighting algorithm discussed in the following section.

The POS distribution of the unknown types of our annotated data is shown in Table 2.

Table 2 compares the present gold standard, referred to as Jellyfish2013, to the gold standard presented in Attia et al. (2012), referred to as Attia2012. We observe that proper nouns comprised 48.07% of the valid forms in Attia2012, and 75.35% of the valid forms in Jellyfish2013. We also notice that Attia2012 has fewer invalid forms. Both observations can be explained by the fact that in Attia2012 data passed through filtration by a spell checker which in most cases does not accept infrequent proper nouns. As expected, most unknown words are open class words: proper names, nouns, adjectives, and, to a lesser degree, verbs. It must be noted here that morphological analyzers typically tend to include much more proper nouns than dictionaries. Ordinary dictionaries are usually interested in proper nouns only when they have frequent metonymic use such as *the White House* for 'the US administration' and *Westminster* for 'the UK parliament'.

4.1 Evaluation

We conduct three sets of evaluation experiments to test three aspects of our research on acquiring new words from data: POS tagging, the lemmatization process, and lemma weighting criteria.

4.1.1 POS evaluation

In the first set of experiments we evaluate POS tagging of new words. We measure accuracy calculated as the number of correctly tagged words divided by the number of all valid words. The baseline assigns the most frequent tag (proper name) to all unknown words. In our test data the baseline accuracy stands at 75%. We notice that MADA POS tagging accuracy for unknown words is the same as the baseline, as shown in Table 3. As in Attia2012, we use Voted POS Tagging; that is, we choose the POS tag assigned most frequently by the same tagger (MADA) in the data to a lemma attested more than once. This method has improved the tagging results significantly to 81% which is higher than the baseline. It is also higher than Attia2012, though we use the same method, because of the increased ratio of proper nouns in the gold standard.

		Jellyfish2013 Accuracy	Attia2012 Accuracy
	POS tagging		
1	POS Tagging baseline	75%	45%
2	MADA POS Tagging	75%	60%
3	Voted POS Tagging	81%	69%

Table 3: Evaluation of POS tagging of unknown words

4.1.2 Lemmatization evaluation

In the second set of experiments we test the accuracy of the lemmatization process for new words. 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 65%, as

shown in Table 4. We notice that the baseline in Jellyfish2013 is higher than the baseline in Attia2012 partly due to the increased ratio of proper nouns in the new test suite.

Furthermore, lemmatization has improved significantly because of the revised matching mechanism which penalizes extra features in the guesser that have no matches in the MADA output.

	Lemmatization	Jellyfish2013 Accuracy	Attia2012 Accuracy
1	Lemmas found among corpus forms	81%	64%
3	Lemma selection baseline	65%	45%
5	Pipeline-based lemmatization	84%	63%

Table 4: Evaluation of lemmatization of unknown words

4.1.3 Evaluation of 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 1,698,852 unknown types. After lemmatization and basic filtration, they are reduced to 476,349 (that is a 72% reduction of the surface forms). This number is still too large for manual validation. In order to address this issue we investigate weighting criteria 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 'combined criteria' refers to the weighting algorithm developed in Attia et al. (2012) which 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 nouns and adjectives, and nothing to proper nouns. The reason we give higher frequency for verbs is the fact that verb neologisms are usually less common.

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

We use the gold annotated data for the evaluation of the lemma weighting criteria, as shown in Table 5. In our experiments, relying on the sum of frequency of sister forms obtained the best results, giving an optimal balance between increasing the number of lexicographically-relevant words in the top one tenth of the data and reducing the number of lexicographically-relevant words in the bottom tenth.

Lexicographically-relevant words	In top tenth	In bottom tenth
relying on sum of frequency of sister forms	1032	14
relying on number of sister forms (form variation)	716	55
relying on POS factor	89	178
using combined criteria	770	12

Table 5: Evaluation of lemma weighting and ranking

In Attia2012, the combined criteria gave the best results. We notice our data has a bias towards proper nouns; therefore, it could be the case that the combined criteria will be better able to give appropriate importance to other categories, such as nouns, verbs and adjectives. Below, we list some examples of the new lemmas collected in our research.

Proper nouns: waziyrstAn وزيرستان 'Waziristan'; mAkiyn ماكين 'McCain'; blAkbiyrn بلاكبيرن 'Blackburn'; guwroduwn غوردون 'Gordon'.

Nouns: tasoyiys تسييس 'politicizing'; AHotirAr احترار 'warming'; maAliym معاليم 'landmarks'; tay'iys تيئيس 'putting off'; tawziyr توزير 'appointing as a minister'; muhAtarap مهاترة 'nonsense'; taDomiyd تضميد 'healing'.

Verbs: taEamolaqa تعملق 'to become gigantic'; taqAfaza تقافز 'to jump'; xaSoxaSa خصخص 'to privatize'; AnoHa$ara انحشر 'to squeeze in'; tanAha$a تناهش 'to snatch'; $aroEana شرعن 'to legislate'.

Adjectives: $aEobawiy~ شعبوي 'populist'; baHot بحت 'pure'; muEawolam معولم 'globalized'; munojaz منجز 'accomplished'; manZuwr منظور 'being investigated'; <ixwaniy~ إخواني 'belonging to the Brotherhood'.

5. Flagging Obsolete Words

After a few decades in the life of any dictionary, it becomes burdened with many oddities related particularly to the preservation of obsolete words and senses. This is specifically the case with Arabic dictionaries which suffer from a lack of appropriate systematic maintenance. More than 1,300 years ago, Al-Khalil bin Ahmed Al-Farahidi compiled the first known monolingual Arabic dictionary called *Al-Ain*. Subsequent Arabic dictionaries typically included refinement, expansion, correction, or organisational improvements over previous dictionaries. These dictionaries include *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980), *al-Muheet* by al-Sahib bin 'Abbad (died 995), *Lisan al-'Arab* by ibn Manzour (died 1311), *al-Qamous al-Muheet* by al-Fairouzabadi (died 1414) and *Taj al-Arous* by Muhammad Murtada al-Zabidi (died 1791) (Owens, 1997).

Even relatively modern dictionaries such as *Muheet al-Muheet* (1869) by Butrus al-Bustani and *al-Mu'jam al-Waseet* (1960) by the Academy of the Arabic Language in Cairo were not started from scratch, nor was there an attempt to overhaul the process of dictionary compilation or to make any significant change. The aim was mostly to preserve the language, refine older dictionaries, and accommodate accepted modern terminology. In this way, Arabic dictionaries tend to preserve a fossilized version of the language with each new one reflecting the content of the preceding dictionaries (Ghazali and Braham, 2001).

The Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) includes 40,648 lemmas (consisting of 420 function words and 1,769 proper nouns, and the remaining 38,459 are nouns, verbs and adjectives). BAMA is widely used by the Arabic NLP research community. It is a *de facto* standard tool, and has been described as the "most respected lexical resource of its kind" (Hajič et al., 2005). The latest version of BAMA is renamed SAMA (Standard Arabic Morphological Analyzer) version 3.1 (Maamouri et al., 2010).

Unfortunately, the SAMA lexical database suffers from a legacy of heavy reliance on older Arabic dictionaries, particularly Wehr's Dictionary (Wehr Cowan, 1976), in the compilation of its lexical database. Attia et al. (2011b) estimate that about 25% of the lexical items included in SAMA are outdated. SAMA includes thousands of obsolete words that are no longer used in speaking or writing. For example, BAMA contains six obsolete words for 'desert' (fayfA' فَيْفاء, fadofad فَدْفَد, quwA' قَواء, mawomAp مَوْماة, matolaf مَتْلَف, and sabosab سَبْسَب) which are no longer in current use.

We need to mention that a full study of the diachronic changes in a language (Lass, 1997) will include currency (words becoming obsolete), register (formal or technical words becoming unmarked), region (regional terms becoming global), syntactic behaviour (e.g. a verb acquiring a new subcategorization frame), and meaning (word meaning is changed or extended). Our focus here is only to handle the first type.

Our objective is to automatically detect and extract obsolete words found in SAMA. To do this, we use a data-driven filtering method that combines open web search engines and our pre-annotated corpus. Using frequency statistics[2] on lemmas from three web sites using their own search facilities (Al-Jazeera,[3] Arabic Wikipedia,[4] and the Arabic BBC website[5]), we find that 7,095 lemmas in SAMA have zero hits. On the other hand, frequency statistics from our text corpus described in Section 2.2 above show that 3,604 SAMA lemmas are not used in the corpus at all, and 4,471 lemmas occur less than 10 times. Combining frequency statistics from the web and the corpus,

[2] Statistics were collected in January 2011.

[3] http://aljazeera.net/portal

[4] http://ar.wikipedia.org

[5] http://www.bbc.co.uk/arabic/

we find that there are 29,627 lemmas that returned at least one hit in the web queries and occurred at least 10 times in the corpus. Using a threshold of 10 occurrences here is discretionary, but the aim is to separate the stable core of the language from instances where the use of a word is perhaps accidental or somewhat idiosyncratic. We consider the refined list as representative of the lexicon of MSA as attested by our statistics.

We consider the remaining 8,832 lemmas (38,459 open-class lemmas, not including proper nouns, minus the 29,627 stable lemmas) as obsolete, and we publish them as an open-source resource [6] to allow dictionary compilers to flag these words as outdated in their dictionaries.

6. Detecting Word Senses

The SketchEngine (Kilgarriff and Tugwell, 2002) is a tried-and-tested powerful tool for lexicographic work related to word sense discovery, based on context and significant collocates, and using partial parsing and statistical information. In this work we used a similar approach but with different techniques.

In our research we use a fully-parsed resource, the Penn Arabic Treebank (ATB) (Maamouri and Bies, 2004), to extract subcategorization frames for verbs enriched with probability scores. These subcategorization frames help in showing which word senses are more prominent than others for a given verb. We also show how word senses are tied to word forms captured in terms of co-occurrence frequencies (tri-gram frequencies) extracted from the Arabic Gigaword corpus.

6.1 Encoding of subcategorization frames

The encoding of syntactic subcategorization frames is essential in the construction of computational and paper lexicons alike. Subcategorization frames refer to the predicate argument structure. Traditional dictionaries specify whether verbs are transitive (requiring a subject and an object) or intransitive (requiring no object). Subcategorization frames, as defined by the Lexical Functional Grammar (LFG) theory (Dalrymple, 2001), have a broader coverage as they include all governable grammatical functions. The governable grammatical functions are the arguments required by some predicates in order to produce a well-formed syntactic structure, and they include SUBJ(ect), OBJ(ect), OBJ_{Θ}, $OBL(ique)_{\Theta}$, COMP(lement) and XCOMP. The subcategorization requirements in LFG are expressed in the following format (O'Donovan et al., 2005):

$\pi<gf_1,gf_2,...gf_n>$

[6] http://obsoletearabic.sourceforge.net/

where π is the lemma (predicate or semantic form) and *gf* is a governable grammatical function. The value of the argument list of the semantic form ensures a well-formed sentence.

For example, in the sentence {iEotamada Al-Tifolu EalaY wAlidati-hi اعتمد الطفل على والدته 'The child relied on his mother', the verb {iEotamada 'to rely' has the following argument structure: {iEotamada <(↑SUBJ)(↑OBL$_{>alaY}$)>. By including a subject and an oblique with the preposition >alaY, we ensure that the verb's subcategorization requirements are met and that the sentence is well-formed, or syntactically valid.

Attia et al. (2011a) automatically extract the Arabic subcategorization frames (or predicate-argument structures) from the ATB for a large number of Arabic lemmas, including verbs, nouns and adjectives, as shown in Table 6.

	Verbs	Nouns	Adjectives
lemma-frame pairs in the ATB	6596	855	295

Table 6: Number of subcategorization frames in the ATB

Subcategorization frames are enriched with probability information that provides estimates of the likelihood of occurrence of a certain argument list with a predicate (or lemma). For example, Table 7 show the probability of each subcategorization frame with the verb>abolaga أبلغ 'to inform' which has a frequency of 103 occurrences in the ATB. The subcategorization frames are sorted by probability, ensuring that more frequent subcategorization frames appear on the top.

id	lemma_id	subcats	prob	sense
527	>abolag_1	subj,obj,comp-sbar	0.3398	to inform sb that
525	>abolag_1	subj,comp-sbar	0.165	to announce that
537	>abolag_1	subj,obj	0.1359	let sb be informed
529	>abolag_1	subj,obj,obj2	0.1068	communicate sth to sb
533	>abolag_1	subj,obj,obl-clr@bi	0.068	inform sb of sth

Table 7: Subcategorization frames with probability scores for the lemma '>abolag_1'

6.2 Information on co-occurrence frequencies

In addition to subcategorization frames, the context in which words occur can provide key information on word senses, significant collocates and the various types of idioms, and multiword expressions in which the headword may occur. This is why the recording of co-occurrence frequencies in the corpus is essential.

AraComLex (Attia et al., 2011b), is a useful web application designed specifically for Arabic lexicographic work and provides, among other facilities, the ability to review

word frequencies at various levels: lemma, stem, full form, and contextual examples. Information is sorted by frequency, so that the most prominent senses occupy the top of the lists. Table 8 shows an example of the full forms and stems of the verb >bolaga أبلغ 'to inform'.

id	index_id	full_form	stem	freq
90687	6998	>blg	>abolag	15235
1107949	6998	w>blg	>abolag	9421
31207	6998	>blgt	>abolag	7194
1191154	6998	tblg	bolig	3932
983221	6998	yblg	bolig	3523
838632	6998	wtblg	bolig	3343
492823	6998	wyblg	bolig	3277
114319	6998	>blgh	>abolag	2456

Table 8: Full form variations with frequency for the lemma '>abolag_1'

Furthermore, a lexicographer can go even deeper by reviewing the examples in which the words occurred, sorted according to frequency, as shown in Table 9. For practical reasons and to keep the size of the database within reasonable bound, we only keep records of the word's tri-grams, which in most cases are enough to provide a glimpse of the context and possible collocates.

stem_id	example	freq.	translation
1107949	.#و أبلغ#مصدر	263	a source informed
90687	انه#أبلغ#الى	75	that he communicated to
90687	,#أبلغ#وزير#	70	informed the minister of
90687	.#أبلغ#اداري#	17	an administrative official informed
114319	الذي#أبلغه#أنه	16	who informed him that he

Table 9: tri-gram frequencies for the lemma '>abolag_1'

7. Conclusion

We have developed a set of methods and techniques to equip modern dictionaries with self-updating mechanisms to allow them to discover new words, flush out (or mark) obsolete words and investigate word senses based on co-occurrence information. We automatically extract new words from a large corpus and lemmatize 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. Out of 1,698,852 new words we created a lexicon of 476,349 lemmatized, POS-tagged and weighted entries. We have made our unknown word lexicon available as a free open source resource (http://arabicnewwords.sourceforge.net/).

We deal with the crucial maintenance problem faced by dictionaries in that, over time, they tend to accumulate a large subset of obsolete lexical entries no longer attested in contemporary data. We identify obsolete entries relying on statistics derived from a large pre-annotated corpus and website searches. We also provide essential lexicographic information by automatically building a lexicon of subcategorization frames from the ATB and information on co-occurrence frequencies.

8. Acknowledgements

This research is funded by the Irish Research Council for Science Engineering and Technology (IRCSET), 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.

9. 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, Pavel Pecina, Lamia Tounsi, Antonio Toral, Josef van Genabith. (2011a). Lexical Profiling for Arabic. Electronic Lexicography in the 21st Century. Bled, Slovenia.

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

Attia, Mohammed, Younes Samih, Khaled Shaalan, Josef van Genabith. (2012). The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database through the Detection and Lemmatization of Unknown Words. COLING, Mumbay, India.

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.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Volume 34 of Syntax and Semantics. Academic Press, New York.

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.

Ghazali, S. and Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. Arabic NLP Workshop at ACL/EACL. Toulouse, France

Gottlieb, Henrik and Jens Erik Mogensen (Eds). (2007). *Dictionary Visions, Research and Practice. Selected Papers from the 12th International Symposium on Lexicography*. Copenhagen 2004. Amsterdam/Philadelphia: John Benjamins

Grefenstette, Gregory, Yan Qu, and David A. Evans. (2002). Expanding lexicons by inducing paradigms and validating attested forms. Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Canary Islands, Spain.

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.

Kilgarriff, Adam and David Tugwell. (2002). Sketching words *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Marie-Hélène Corréard (Ed.) EURALEX: 125-137.

Lass, Roger (1997). *Historical linguistics and language change*. 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.

O'Donovan, R., Burke, M., Cahill, A., van Genabith, J. and Way, A. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks, Computational Linguistics, pp. 329-366.

Owens, J.: *The Arabic Grammatical Tradition*. The Semitic Languages. London:Routledge (1997)

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.