Automatic Extraction of Arabic Multiword Expressions

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith
School of Computing, Dublin City University
{mattia,atoral,ltounsi,ppecina,josef}@computing.dcu.ie

Abstract

In this paper we investigate the automatic acquisition of Arabic Multiword Expressions (MWE). We propose three complementary approaches to extract MWEs from available data resources. The first approach relies on the correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages. The second approach collects English MWEs from Princeton WordNet 3.0, translates the collection into Arabic using Google Translate, and utilizes different search engines to validate the output. The third uses lexical association measures to extract MWEs from a large unannotated corpus. We experimentally explore the feasibility of each approach and measure the quality and coverage of the output against gold standards.

1 Introduction

A lexicon of multiword expressions (MWEs) has a significant importance as a linguistic resource because MWEs cannot usually be analyzed literally, or word-for-word. In this paper we apply three approaches to the extraction of Arabic MWEs from multilingual, bilingual, and monolingual data sources. We rely on linguistic information, frequency counts, and statistical measures to create a refined list of candidates. We validate the results with manual and automatic testing.

The paper is organized as follows: in this introduction we describe MWEs and provide a summary of previous related research. Section 2 gives a brief description of the data sources used. Section 3 presents the three approaches used in our experiments, and each approach is tested and evaluated in its relevant sub-section. In Section 4 we discuss the results of the experiments. Finally, we conclude in Section 5.

1.1 What Are Multiword Expressions?

Multiword expressions (MWEs) are defined as idiosyncratic interpretations that cross word boundaries or spaces (Sag et al., 2002). The exact meaning of an MWE is not directly obtained from its component parts. Accommodating MWEs in NLP applications has been reported to improve tasks, such as text mining (SanJuan and Ibekwe-SanJuan, 2006), syntactic parsing (Nivre and Nilsson, 2004; Attia, 2006), and Machine Translation (Deksne, 2008).

There are two basic criteria for identifying MWEs: first, component words exhibit statistically significant co-occurrence, and second, they show a certain level of semantic opaqueness or non-compositionality. Statistically significant co-occurrence can give a good indication of how likely a sequence of words is to form an MWE. This is particularly interesting for statistical techniques which utilize the fact that a large number of MWEs are composed of words that co-occur together more often than can be expected by chance.

The compositionality, or decomposability (Villavicencio et al. 2004), of MWEs is also a core issue that presents a challenge for NLP applications because the meaning of the expression is not directly predicted from the meaning of the component words. In this respect, compositionality varies between phrases that are highly com-

Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 18–26,
Beijing, August 2010

positional, such as, قاعدة عسكريّة *qāʿidatun ʿaskariyyatun*, “military base”, and those that show a degree of idiomaticity, such as, مدينة الملاهى *madiynatu ’l-malāhiy*, “amusement park”, lit. “city of amusements”. In extreme cases the meaning of the expression as a whole is utterly unrelated to the component words, such as, فرس النبىّ *farasu ’l-nabiyyi*, “grasshopper”, lit. “the horse of the Prophet”.

1.2 Related Work

A considerable amount of research has focused on the identification and extraction of MWEs. Given the heterogeneity of MWEs, different approaches were devised. Broadly speaking, work on the extraction of MWEs revolves around four approaches: (a) statistical methods which use association measures to rank MWE candidates (Van de Cruys and Moirón, 2006); (b) symbolic methods which use morpho-syntactic patterns (Vintar and Fišer, 2008); (c) hybrid methods which use both statistical measures and linguistic filters (Boulaknadel et al. 2009; Duan et al., 2009); and (d) word alignment (Moirón and Tiedemann, 2006).

None of the approaches is without limitations. It is difficult to apply symbolic methods to data with no syntactic annotations. Furthermore, due to corpus size, statistical measures have mostly been applied to bigrams and trigrams, and it becomes more problematic to extract MWEs of more than three words. As a consequence, each approach requires specific resources and is suitable for dealing with only one side of a multi-faceted problem.

Pecina (2010) evaluates 82 lexical association measures for the ranking of collocation candidates and concludes that it is not possible to select a single best universal measure, and that different measures give different results for different tasks depending on data, language, and the types of MWE that the task is focused on. Similarly, Ramisch et al. (2008) investigate the hypothesis that MWEs can be detected solely by looking at the distinct statistical properties of their individual words and conclude that the association measures can only detect trends and preferences in the co-occurrences of words.

A lot of effort has concentrated on the task of automatically extracting MWEs for various languages besides English, including Slovene (Vintar and Fišer, 2008), Chinese (Duan et al., 2009), Czech (Pecina, 2010), Dutch (Van de Cruys and Moirón, 2006), Latvian (Deksne, 2008) and German (Zarrieß and Kuhn, 2009).

A few papers, however, focus on Arabic MWEs. Boulaknadel et al. (2009) develop a hybrid multiword term extraction tool for Arabic in the “environment” domain. Attia (2006) reports on the semi-automatic extraction of various types of MWEs in Arabic and how they are used in an LFG-based parser.

In this paper we report on three different methods for the extraction of MWEs for Arabic, a less resourced language. Our approach is linguistically motivated and can be applied to other languages.

2 Data Resources

In this project we use three data resources for extracting MWEs. These resources differ widely in nature, size, structure and the main purpose they are used for. In this section we give a brief introduction to each of these data resources.

Wikipedia (WK) is a freely-available multilingual encyclopedia built by a large number of contributors. Currently WK is published in 271 languages, with each language varying in the number of articles and the average size (number of words) of articles. WK contains additional information that proved to be helpful for linguistic processing such as a category taxonomy and cross-referencing. Each article in WK is assigned a category and may be also linked to equivalent articles in other languages through what are called “interwiki links”. It also contains “disambiguation pages” for resolving the ambiguity related to names that have variant spellings. Arabic Wikipedia (AWK) has about 117,000 articles (as of March 2010[1]) compared to 3.2 million articles in the English Wikipedia. Arabic is ranked 27^{th} according to size (article count) and 17^{th} according to usage (views per hour).

[1] http://stats.wikimedia.org/EN/Sitemap.htm

Princeton WordNet[2] **(PWN)** is an electronic lexical database for English where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms called synsets. In our analysis of PWN 3.0 we find that MWEs are widespread among all the categories, yet with different proportions, as shown by Table 1. Arabic WordNet (AWN) (Elkateb et al., 2006) is constructed according to the methods and techniques used in the development of PWN, but it is limited in size, containing only 11,269 synsets (including 2,348 MWEs).

POS	**Unique Strings**	**MWEs**	**Percentage of MWEs**
Nouns	117,798	60,292	51.18
Verbs	11,529	2,829	24.53
Adjectives	21,479	496	2.31
Adverbs	4,481	714	15.93
Total	155,287	64,331	41.43

Table 1: Size and distribution of MWEs in PWN.

Arabic Gigaword Fourth Edition is an unannotated corpus distributed by the Linguistic Data Consortium (LDC), catalog no. LDC2009T30.[3] It is the largest publicly available corpus of Arabic to date, containing 848 million words. It comprises articles from newspapers from different Arab regions, such as Al-Ahram in Egypt, An Nahar in Lebanon and Assabah in Tunisia, in addition to news agencies, such as Xinhua and Agence France Presse.

3 Methodology

The identification and extraction of MWEs is a problem more complex than can be dealt with by one simple solution. The choice of approach depends on the nature of the task and the type of the resources used. We discuss the experiments we conducted to extract and validate MWEs from three types of data resources each with a different technique and different validation and evaluation methodology. A crucial factor in the selection of the approach is the availability of rich resources that have not been exploited in similar tasks before.

We focus on nominal MWEs because the vast majority of MWEs are nouns, as evidenced by statistics in Table 1 above. We define nominal MWEs as MWEs that act as nouns and have the internal structure of either:

- noun–noun, such as دودة الأرض *duwdatu 'l-ʾarḍ*, “earthworm”;
- noun–adjective, such as إسعافات أوّليّة *ʾisʿāfātun ʾawwaliyyatun*, “first aid”[4];
- noun–preposition–noun, such as, التزلّج على الجليد *al-tazalluǧ ʿalā 'l-ǧaliyd*, “Skiing”, lit. “sliding on ice”; or
- noun–conjunction–noun, such as القانون والنظام *al-qānuwn wa-'l-niẓām*, “law and order”.

We use three approaches to identify and extract MWEs: (a) crosslingual correspondence asymmetries, (b) translation-based extraction, and (c) corpus-based statistics. For each approach we use a number of linguistic and statistical validation techniques and both automatic and manual evaluation.

In the first approach (Section 3.1) we make use of the crosslingual correspondence asymmetry, or many-to-one relations between the titles in the Arabic Wikipedia (AWK) and the corresponding titles in other languages to harvest MWEs. In the second approach (Section 3.2) we assume that automatic translation of MWEs collected from PWN into Arabic are high likelihood MWE candidates that need to be automatically checked and validated. In the third approach (Section 3.3) we try to detect MWEs in a large raw corpus relying on statistical measures and POS-annotation filtering.

3.1 Crosslingual Correspondence Asymmetries

In this approach, our focus is on semantic non-decomposable MWEs and we rely on Crosslingual Correspondence Asymmetries (CCAs) for capturing them. Semantic non-compositionality can be considered as a powerful indication that a phrase is an MWE. Baldwin et al. (2003) classify MWEs, with respect to compositionality, into three categories: (a) non-compositional MWEs, where the expression is semantically impenetrable, such as *hot dog*, (b) idiosyncratically compositional, where the component words are forced to take semantics unavailable outside the MWE, such as *radar footprint*, and (c) simply composi-

[2] http://wordnet.princeton.edu

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T30

[4] In Arabic, the adjective follows the noun.

tional, where the phrase is institutionalized, such as *traffic light*. This, however, can only serve as an approximation, not as a clear-cut division. As Moon (1998) indicates, compositionality can be viewed more as a gradient along a continuum with no clear demarcations, ranging from conventionalized, fully transparent literal expressions to completely opaque idioms.

There are many signs, or indications, of non-compositionality, two well-known among them are "non-substitutability", when a word in the expression cannot be substituted by a semantically equivalent word, and "single-word paraphrasability", when the expression can be paraphrased or translated by a single word. These two indications have been exploited differently by different researchers. Van de Cruys and Moirón (2006) develop an unsupervised method for detecting MWEs using clusters of semantically related words and taking the ratio of the word preference over the cluster preference as an indication of how likely a particular expression is to be an MWE. Melamed (1997) investigates techniques for identifying non-compositional compounds in English-French parallel corpora and emphasises that translation models that take non-compositional compounds into account are more accurate. Moirón and Tiedemann (2006) use word alignment of parallel corpora to locate the translation of an MWE in a target language and decide whether the original expression is idiomatic or literal.

The technique used here is inspired by that of Zarrieß and Kuhn (2009) who rely on the linguistic intuition that if a group of words in one language is translated as a single word in another language, this can be considered as an indication that we have a fixed expression with a non-compositional meaning. They applied their data-driven method to the German-English section of the Europarl corpus after preprocessing with dependency parsing and word alignment, and tested their method on four German verb lemmas.

We also utilize CCAs for the task of MWE extraction. As an approximation we make a binary decision between whether an expression is decomposable or non-decomposable based on the criterion of single word translatability. This technique follows Zarrieß and Kuhn's (2009) assumption that the idiosyncrasy and non-compositionality of MWEs makes it unlikely, to some extent, to have a mirrored representation in the other languages. We consider many-to-one correspondence relationships (an MWE in one language has a single-word translation in another language) as empirical evidence for the detection of MWEs. Here our candidate MWEs are the AWK titles that are made up of more than one word. For each of them we check whether there exists a many-to-one correspondence relation for this title in other languages (the translations are obtained by exploiting the inter-lingual links of AWK). To increase the predictive power of our approach and ensure that the results are more representative we expand the search space into 21 languages[5], rather than only one, as in Zarrieß and Kuhn (2009). This approach helps us with idiomatic MWEs. For non-idiomatic MWEs we rely on the second and third methods discussed in 3.2 and 3.3 respectively.

The steps undertaken in this approach are: (1) Candidate Selection. All AWK multiword titles are taken as candidates. (2) Filtering. We exclude titles of disambiguation and administrative pages. (3) Validation. This includes two steps. First, we check if there is a single-word translation in any of the target languages. Second, we look for the candidate and/or its translations in LRs; the Italian, Spanish, and English translations are looked up in the corresponding WordNets while both the AWK title and its translations are looked up in a multilingual lexicon of Named Entities (NEs), MINELex (Attia et al., 2010). If the candidate title is found in MINELex or if any of its translations is a single word or is found in any WordNet or in MINELex, then the AWK title is classified as a MWE. Otherwise, it is considered a non-MWE.

It is worth-noting that many titles in the AWK are named entities (NEs). We conduct our evaluation on a set of 1100 multiword titles from AWK

[5] These languages are: Dutch, Catalan, Czech, Danish, German, Greek, English, Esperanto, Spanish, French, Hebrew, Indonesian, Italian, Latin, Norwegian, Portuguese, Polish, Romanian, Russian, Swedish and Turkish. The selection is based on three criteria: (a) number of articles, (b) cultural association with Arabic and (c) relevance to scientific terminology.

that have been manually tagged as: non-NE-MWEs (181), NE-MWEs (849) or non-MWEs (70). Given the high percentage of NE-MWEs in the set we derive two gold standards: the first includes NEs as MWEs, and is made up of 1030 MWEs and 70 non-MWEs, and the second drops NEs and hence consists of 251 entries (181 MWEs and 70 non-MWEs). In the experiment these sets are matched with our validation approach. Table 2 compares the results of our experiment (CCA) with a baseline, which considers all multiword titles as MWEs, in terms of precision, recall, $F_{\beta=1}$ and $F_{\beta=0.5}$. We notice that our precision is substantially higher for both sets. Some examples of the CCA method are given in Table 3.

	P	R	F-1	F-0.5
With NEs				
Baseline	93.63	100.00	96.71	94.83
CCA	98.28	44.47	61.23	79.13
Without NEs				
Baseline	72.11	100.00	83.80	76.37
CCA	82.99	21.55	34.21	52.85

Table 2: Evaluation (in percent) of the CCA approach.

Arabic Phrase	Translation	Langs	M-1
فقر دمّ	Anemia	21	100%
التهاب القولون	colitis	12	92%
ورق الحائط	wallpaper	11	82%
قمرة القيادة	cockpit	17	76%
فريق عمل	teamwork	9	67%
فرس النهر	hippopotamus	21	52%
قاعدة بيانات	database	20	45%
فرشاة أسنان	toothbrush	19	37%
فوّهة بركانيّة	volcanic crater	14	21%
فنّ تجريديّ	abstract art	20	15%
دائرة كهربائيّة	electrical network	20	5%
تاريخ الطيران	aviation history	12	0%

Table 3: MWE identification through correspondence asymmetries. The first column shows the Arabic candidate MWE. The second column is the English translation of the expression. The third column is the number of languages that have correspondences for the Arabic expression. The last column is the ratio of many-to-one correspondences where 100% means that all other the languages have the expression as one word, and 0% means that all other languages have a parallel compositional phrase.

3.2 Translation-Based Approach

This approach is bilingual and complements the first approach by focusing on compositional compound nouns which the many-to-one correspondence approach is not likely to identify. We collect English MWEs from PWN, translate them into Arabic, and automatically validate the results. This technique also has an ontological advantage as the translated and validated expressions can be used to extend the Arabic WordNet by linking the expressions to their respective synsets.

This method is partly similar to that of Vintar and Fišer (2008) who automatically extended the Slovene WordNet with nominal multiword expressions by translating the MWEs in PWN using a technique based on word alignment and lexico-syntactic patterns. Here we also use the MWEs in PWN as a starting point to collect MWEs in our target language, Arabic. We depart from Vintar and Fišer (2008) in that instead of using a parallel corpus to find the translation, we use an off-the-shelf SMT system, namely Google Translate. The reason we did not use an alignment-based approach is that word alignment, *per se*, is complex and the quality of the output is dependent on the size and domain of the corpus as well as on the quality of the alignment process itself. Therefore, we use a state-of-art MT system and concentrate on validating the results using frequency statistics.

The rationale behind this technique is that we try to discover MWEs by inducing and analysing a translation model. We assume that an MWE in one language is likely to be translated as an MWE in another language, although we are aware that translations into single words or paraphrases are also possible. First, we extract the list of nominal MWEs from PWN 3.0. This provides us with predefined knowledge of what concepts are likely to be represented as MWEs in the target language. Second, we translate the list into Arabic using Google Translate. Third, we validate the results, by asking a different question: given a list of candidate translations, how likely are they to be correct translations and how well do they correspond to MWEs? We try to answer this question using pure frequency counts from three search engines, namely, Al-Jazeera[6], BBC Arabic[7] and AWK.[8]

We conduct automatic evaluation using as a gold standard the PWN–MWEs that are found in English Wikipedia and have a correspondence in Arabic. The number of gold standard translations is 6322. We test the Google translation without any filtering, and consider this as the baseline, then we filter the output based on the number of combined hits[9] from the search engines. The results are shown in Table 4. The best f-measure achieved is when we accept a candidate translation if it is found only once. The reason for this is that when Google Translate does not know the correct translation of an MWE, it produces an ungrammatical sequence of words that does not return any matches by the search engines. This process gives 13,656 successful MWE candidates from the list of 60,292 translations.

SE Filteration	**Recall**	**Precision**	**F-Measure**
Baseline	100.00	45.84	62.86
1 hit	62.56	73.85	67.74
2 hits	55.58	75.07	63.87
3 hits	50.87	75.29	60.71
4 hits	47.37	74.68	57.97
5 hits	44.51	74.19	55.64
10 hits	36.08	71.99	48.07

Table 4: Automatic evaluation (in percent) of the translation-based approach.

3.3 Corpus-Based Approach

The starting point in this approach is the Arabic Gigaword corpus, which is an unannotated collection of texts that contains 848 million words. In this monolingual setting the only practical solution to extract MWEs is to use lexical association measures based on the frequency distribution of candidate MWEs and to detect any idiosyncratic co-occurrence patterns. Association measures are inexpensive language-independent means for discovering recurrent patterns, or habitual collocates. Association measures are defined by Pecina (2010) as mathematical formulas that determine the strength of the association, or degree of connectedness, between two or more words based on their occurrences and co-occurrences in a text. The higher the connectedness between words, the better the chance they form a collocation.

The corpus is conceived as a randomly generated sequence of words and consecutive bigrams and trigrams in this sequence are observed. Then joint and marginal occurrence frequencies are used to estimate how much the word occurrence is accidental or habitual. For the purpose of this experiment, we use the two following association measures:

Pointwise Mutual Information (PMI) compares the cooccurrence probability of words given their joint distribution and given their individual (marginal) distributions under the assumption of independence. For *two-word* expressions, it is defined as:

$$PMI_2(x, y) = \log_2 \frac{p(x, y)}{p(x, *)p(*, y)}$$

where $p(x, y)$ is the maximum likelihood (ML) estimation of the joint probability (N is the corpus size):

$$p(x, y) = \frac{f(x, y)}{N}$$

and $p(x, *)$, $p(*, y)$ are estimations of marginal probabilities computed in the following manner:

$$p(x, *) = \frac{f(x, *)}{N} = \frac{\sum_y f(x, y)}{N}$$

and analogically for $p(*, y)$. For *three words*, PMI can be extended as follows:

$$PMI_3(x, y, z) = \log_2 \frac{p(x, y, z)}{p(x, *, *)p(*, y, *)p(*, *, z)},$$

Here, the marginal probabilities are estimated as:

$$p(*, y, *) = \frac{f(*, y, *)}{N} = \frac{\sum_{x,z} f(x, y, z)}{N}$$

and analogically for $p(x, *, *)$ and $p(*, *, z)$.

Chi-square compares differences between the observed frequencies $f_{i,j}$ and the expected (under the assumption of independence) frequencies $e_{i,j}$ from the two-way contingency table as follows:

$$\chi_2^2(x, y) = \sum_{i,j \in \{0,1\}} \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}},$$

[6] http://aljazeera.net/portal/search.aspx

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

[8] http://ar.wikipedia.org/wiki/

[9] The hits are combined by taking the aggregate sum of the number of documents returned by the search engines.

where the table cells are referred to by the index pair $i, j \in \{0, 1\}$. The observed frequencies $f_{i,j}$ for a bigram (x, y) are computed in this manner:

$$f_{0,0} = f(x, y), \quad f_{0,1} = f(x, \neg y) = \sum_{v \neq y} f(x, v)$$

and analogically for $f_{1,0}$ and $f_{1,1}$. The expected frequencies $e_{i,j}$ are then estimated using marginal frequencies as in the following equations:

$$e_{0,0} = e(x, y) = \frac{f(x, *)f(*, y)}{N},$$

$$e_{0,1} = e(x, \neg y) = \frac{f(x, *)f(*, \neg y)}{N},$$

and analogically for $e_{1,0}$ and $e_{1,1}$. For three words, the Chi-square formula can be extended and applied to a three-way contingency table as follows:

$$\chi_3^2(x, y, z) = \sum_{i,j,k \in \{0,1\}} \frac{(f_{i,j,k} - e_{i,j,k})^2}{e_{i,j,k}}$$

with the observed $(f_{i,j,k})$ frequencies computed analogically as in this example:

$$f_{0,1,0} = f(x, \neg y, z) = \sum_{v \neq y} f(x, v, z).$$

And similarly for the expected frequencies $(e_{i,j,k})$ with the marginal probabilities as in this example:

$$e_{0,1,0} = f(x, \neg y, z) = \frac{f(x,*,*)f(*,\neg y,*)f(*,*,z)}{N^2}$$

This corpus-based process involves four steps:
(1) We compute the frequency of all the unigrams, bigrams, and trigrams in the corpus.
(2) The association measures are computed for all the bigrams and trigrams with frequency above a threshold which we set to 50. Then the bigrams and trigrams are ranked in descending order.
(3) We conduct lemmatization using MADA (Habash et al., 2009). This step is necessary because Arabic is a clitic language where conjunctions, prepositions and the definite article are attached to nouns which creates data sparsity and obscures the frequency statistics. Using lemmatization helps to collapse all variant forms together, and thus create a more meaningful list of candidates.
(4) Filtering the list using the MADA POS-tagger (Habash et al., 2009) to exclude patterns that generate unlikely collocates and to select those candidates that match the relevant POS patterns. The patterns that we include for bigrams are: `NN NA`, and for trigrams: `NNN NNA NAA`. Table 5 shows the number of phrases extracted for each step.

	$n = 2$	$n = 3$
words	875,920,195	
base form n-grams	134,411,475	
after frequency filtering	1,497,214	560,604
after base-form collapsing	777,830	415,528
after POS filtering	217,630	39,269

Table 5: Bigram and trigram experiment statistics.

The evaluation is based on measuring the quality of ranking the candidates according to their chance to form collocations. To evaluate the results, 3600 expressions were randomly selected and classified into MWE or non-MWE by a human annotator. The performance of the methods is compared by precision scores. The method is focused on two-word (bigram) and three-word (trigram) collocations. The results are reported in Table 6. We notice that the best score for the bigrams is for 10,000 terms using PMI, and for the trigrams 5,000 using χ^2.

	$n = 2$	
# top candidates	PMI_2	χ_2^2
10,000	71	70
25,000	66	69
50,000	57	59
	$n = 3$	
	PMI_3	χ_3^2
2,000	40	46
5,000	56	63
10,000	56	57

Table 6: Bigram and trigram experiment results.

4 Discussion of Experiments and Results

It is an underestimation to view MWEs as a single phenomenon. In fact MWEs encompass a set of diverse and related phenomena that include idioms, proper nouns, compounds, collocations, institutionalised phrases, etc. They can also be of any degree of compositionality, idiosyncrasy and lexical and syntactic flexibility. This complicates the task of MWE identification. Moreover, we have used three data sources with a large degree of discrepancy: (a) titles of articles in the AWK, (b) induced translation of English MWEs collected from PWN, and (b) Arabic Gigaword, which is a collection of free texts.

For each of the data types we apply a different technique that we deem suitable for the task at hand. The results of the experiments have been subjected to testing and evaluation in their respective sections. Table 7 combines and compares the outcomes of the experiments. The column "Intersection" refers to how many MWE candidates are already found through the other methods.

	MWEs	Intersection
Crosslingual	7,792	-
NE–MWEs	38,712	-
Translation-based	13,656	2658
Corpus-based	15,000	697
Union without NEs	33,093	-
Union including NEs	71,805	-

Table 7: Comparison of outcomes from each approach.

We notice that the heterogeneity of the data sources which we used for the task of MWE extraction, helped to enrich our MWE lexicon, as they are complementary to each other. We also notice that the intersection between the corpus-based approach and the other approaches is very low. On examining the results, we assume that the reasons for the low intersection are:

1. A lot of named entities in the news corpus are not famous enough to be included in standard Arabic lexical resources (Wikipedia and WordNet), such as, مناحم مازوز *mināḥim māzuwz*, "Menachem Mazuz".

2. We lemmatize according to clitics and ignore inflection. If we include morphological inflection in the lemmatization this may produce a less marked list and allow better matching, such as, حكمان غيابيّين *ḥukmāni ġiyābiyyayni*, "two sentences in absentia".

3. The set of collocations detected by the association measures may differ from the those which capture the interest of lexicographers and Wikipedians, such as, الخضروات الطازجة *al-ḫuḍrawātu 'l-ṭāziǧatu*, "fresh vegetables".

5 Conclusion

The identification of MWEs is too complex to be dealt with by one simple solution. The choice of approach depends, to a large extent, on the type of data resources used. In this paper, we extract MWEs from heterogeneous data resources using three approaches: (a) crosslingual correspondence asymmetries which relied on the many-to-one relations in interwiki links, (b) translation-based extraction, which employs the automatic translation of PWN–MWEs into Arabic and uses different search engines to filter the translation output, and (c) corpus-based statistics, which applies lexical association measures to detect habitual collocations in a large unannotated corpus. As Arabic has a rich and complex morphology, we lemmatize the text to reduce inflectional forms. These approaches prove to be a fruitful ground for large-scale extraction of Arabic MWEs.

Acknowledgments

This research is funded by Enterprise Ireland (PC/09/037), the Irish Research Council for Science Engineering and Technology (IRCSET), and the EU projects PANACEA (7FP-ITC-248064) and META-NET (FP7-ICT-249119).

References

Attia, Mohammed. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar. In Salakoski, Tapio, Filip Ginter, Sampo Pyysalo, Tapio Pahikkala (Eds.): *Advances in Natural Language Processing*. Vol. 4139, pp. 87–98. Springer-Verlag: Berlin, Heidelberg.

Attia, Mohammed, Antonio Toral, Lamia Tounsi, Monica Monachini and Josef van Genabith. 2010. An automatically built Named Entity lexicon for Arabic. In the 7^{th} International Conference on Language Resources and Evaluation (LREC 2010), pp. 3614–3621. Valletta, Malta.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka and Dominic Widdows. 2003. An Empirical Model of Multiword Expressions Decomposability. In Workshop on Multiword Expressions, the 41^{st} Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 89–96, Sapporo, Japan.

Boulaknadel, Siham, Beatrice Daille, Driss Aboutajdine. 2009. A multi-word term extraction program for Arabic language In the 6^{th} International Conference on Language Resources and Evaluation (LREC 2008), pp. 630–634, Marrakech, Morocco.

Deksne, Daiga, Raivis Skadiņš, Inguna Skadiņa. 2008. Dictionary of Multiword Expressions for Translation into Highly Inflected Languages. In the 6^{th} International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco.

Duan, Jianyong, Mei Zhang, Lijing Tong, and Feng Guo. 2009. A Hybrid Approach to Improve Bilingual Multiword Expression Extraction. In the 13^{th} Pacific-Asia Conference on Knowledge Discovery and Data (PAKDD 2009), pp. 541–547. Bangkok, Thailand.

Elkateb, Sabri, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Christiane Fellbaum. 2006. Building a Wordnet for Arabic. In the 5^{th} International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

Habash, Nizar, Owen Rambow and Ryan Roth. 2009. A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In the 2^{nd} International Conference on Arabic Language Resources and Tools (MEDAR 2009), pp. 102–109. Cairo, Egypt.

Hoang, Huu Hoang, Su Nam Kim and Min-Yen Kan. 2009. A Re-examination of Lexical Association Measures. In the Workshop on Multiword Expressions, the Joint Conference of the 47^{th} Annual Meeting of the Association for Computational Linguistics and the 4^{th} International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009), pp. 31–39, Suntec, Singapore.

Melamed, I. Dan. 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In the 2^{nd} Conference on Empirical Methods in Natural Language Processing (EMNLP 1997), pp. 97–108. Providence, RI.

Moirón, Begoña Villada and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In the Workshop on Multiword Expressions in a Multilingual Context, the 11^{th} Conference of the European Association of Computational Linguistics (EACL 2006), pp. 33–40. Trento, Italy.

Moon, Rosamund 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Clarendom Press, Oxford.

Nivre, Joakim and Jens Nilsson. 2004. Multiword Units in Syntactic Parsing. In Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications, the 4^{th} International Conference on Language Resources and Evaluation (LREC 2004), pp. 39–46. Lisbon, Portugal.

Pecina, Pavel 2010. Lexical association measures and collocation extraction. In *Language Resources and Evaluation* (2010), 44:137-158.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart and Aline Villavicencio. 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In the Workshop on Multiword Expressions, the 6^{th} International Conference on Language Resources and Evaluation (LREC 2008), pp. 50–53. Marrakech, Morocco.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In the 3^{rd} International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), volume 2276 of *Lecture Notes in Computer Science*, pp. 1–15, London, UK. Springer-Verlag.

SanJuan, Eric and Fidelia Ibekwe-SanJuan. 2006. Text mining without document context. In *Information Processing and Management*. Volume 42, Issue 6, pp. 1532–1552.

Van de Cruys, Tim and Begoña Villada Moirón. 2006. Lexico-Semantic Multiword Expression Extraction. In P. Dirix et al. (eds.), *Computational Linguistics in the Netherlands 2006*, pp. 175–190.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron and Fabre Lambeau. 2004. The Lexical Encoding of MWEs. In the Workshop on Multiword Expressions, the 42^{nd} Annual Meeting of the Association for Computational Linguistics (ACL 2004), pp. 80–87. Barcelona, Spain.

Vintar, Špela and Darja Fišer. 2008. Harvesting Multi-Word Expressions from Parallel Corpora. In the 6^{th} International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco.

Zarrieß, Sina and Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In the Workshop on Multiword Expressions, the Joint Conference of the 47^{th} Annual Meeting of the Association for Computational Linguistics and the 4^{th} International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009), pp. 23–30. Suntec, Singapore.