Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

Arab World English Journal (AWEJ) Special Issue on Translation No.5 May, 2016

Pp. 317- 336

Bilingual Lexicon Extraction from Arabic-English Parallel Corpora with a View to Machine Translation

Yasser Muhammad Naguib Sabtan
Department of Languages and Translation,
Dhofar University, Oman
&
Faculty of Languages and Translation,
Al-Azhar University, Egypt

Abstract

Automatic extraction of bilingual lexicons from parallel corpora has been recently exploited to overcome the knowledge acquisition bottleneck in a number of research areas in natural language processing, such as machine translation (MT) and cross-language information retrieval. In this paper the author presents a method for automatic extraction of bilingual lexicons from a parallel Arabic-English corpus annotated with part-of-speech tags for potential use in a full-scale MT system. The extraction method, which does not make use of an initial bilingual dictionary, is based on the statistical technique of co-occurrence frequency in the parallel corpus. In addition, dependency relations for some parallel syntactic constructions are made use of to automatically extract head-dependent translation pairs which are then filtered to obtain one-word translation seeds. These seeds are used as anchor points to resegment the parallel corpus in order to bootstrap the lexicon extraction process. Experimental results show that the accuracy of the extracted lexicons was improved after applying the bootstrapping techniques.

Keywords: Arabic machine translation, bilingual lexicon extraction, computational linguistics, natural language processing, parallel corpora

1. Introduction

The compilation of bilingual lexicons is a major bottleneck in computational linguistics (Fišer & Ljubešić, 2011). That is why a number of research attempts have been carried out recently to automatically extract such lexicons which are required by most cross-lingual natural language processing (NLP) applications. Most of such attempts depend on parallel texts (or corpora) as useful resources for automatically extracting word correspondences between the two languages concerned. Parallel corpora are a key resource as training data for statistical machine translation, and for building or extending bilingual lexicons and terminologies. A parallel corpus is a bilingual corpus consisting of a pair of texts, where one is a translation of the other. In this regard, different researchers have used various techniques, using either purely statistical methods (Brown et al. 1990; Gale & Church, 1991) or a combination of both statistical and linguistic information (Dagan et al. 1991; Kumano & Hirakawa, 1994).

Generally speaking, most approaches to target word selection focus on the word co-occurrence frequencies in the parallel corpus (Gale & Church, 1991; Melamed, 1995; Kaji & Aizono, 1996). Word co-occurrence can be defined in various ways. The most common way is to have an equal number of sentence-aligned segments in the parallel text so that each pair of the source language (SL) and target language (TL) segments are translations of each other. Then, researchers begin to count the number of times that word types in one half of the parallel text co-occur with word types in the other half (Melamed, 2000).

This paper describes the design and development of a method for automatic extraction of bilingual lexicons of open-class words from an Arabic-English parallel corpus. The linguistic resources that are used to annotate the parallel corpus include part-of-speech (POS) tags, using an Arabic tagger (Ramsay & Sabtan, 2009) which was built without using a lexicon and an English tagger (described in Sabtan, 2011) that was also built using the same lexicon-free approach. In addition to POS tags a few number of untyped dependency relations (DRs) are exploited in both languages, using a shallow dependency parser. The author will briefly discuss POS tagging and dependency parsing processes in section five. Furthermore, the main algorithm of lexicon extraction is applied to both word-forms and word stems, using a corpus-based light stemmer which is described in Sabtan (2012). A brief review of the stemming process is presented in the fifth section.

In our approach to automatic extraction of translation equivalents we exploit word co-occurrence frequencies in a parallel corpus. This corpus contains two historically unrelated languages, with the SL (i.e. Arabic) being a morphologically rich language. The used corpus is partially aligned, where a parallel segment is not a sentence but a whole verse that may contain a number of sentences. We will shed more light on the parallel corpus in section four. What is new in our method is that we use (DRs) in both the source and target languages to extract a number of head-dependent translation pairs that are then filtered to obtain one-word translation seeds. These seeds are then used as anchor points to resegment the parallel corpus as a way of bootstrapping the whole process of lexicon building, which improves the accuracy score.

The remainder of this paper is organized as follows: in the following section we give an overview of related work on the exploitation of parallel corpora to extract translation lexicons. In section three we throw light on the linguistic challenges that face Arabic NLP in general, which, in turn, have an impact on the current task. Section four presents the parallel corpus that is used in the experiment. The proposed method for building a bilingual lexicon using a parallel corpus is discussed in section five. In section six we describe the experiments and present the results of the evaluation process. Finally, in section seven we conclude the paper with possible directions for future work.

2. Related Work

Researchers have used various knowledge resources (i.e. linguistic information) along with the statistical technique of co-occurrence for extracting bilingual lexicons. Melamed (1995) shows how to induce a translation lexicon from a bilingual sentence-aligned corpus using both the statistical properties of the corpus and four external knowledge sources that are cast as filters, so that any subset of them can be cascaded in a uniform framework. These filters are (1) POS information (2) machine-readable bilingual dictionaries (3) cognate heuristics (4) word alignment heuristics. Each of these filters can be placed into the cascade independently of the others. He conducts his experiments on the English-French language pair. He concludes that most lexicon entries are improved by only one or two filters, after which more filtering does not result in any significant improvement. Later, Resnik & Melamed (1997) present a word-to-word model of translational equivalence, without using any kind of the above-mentioned linguistic knowledge. They use French-English software manuals of about 400,000 words to test their model. Tiedemann (1998) introduces different methods for the extraction of translation equivalents from parallel corpora for historically related languages. His experiments are conducted on Swedish-English and Swedish-German sentence-aligned parallel corpora. The texts in such corpora are orthographically and structurally similar. Tufiş & Barbu (2002) describe a statistical approach to automatic lexicon extraction from parallel corpora. They implement their approach on six pairs of languages, using a parallel corpus of Orwell's 1984 novel. The TL in these multilingual corpora is English, while the SL is one of the following languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene.

Some other researchers have made use of syntactic contexts to help with the extraction of bilingual equivalents from parallel corpora. For instance, Otero (2005) experiment with learning bilingual equivalents of nouns and adjectives from an English-French parallel corpus that contains over 2 million word tokens, focusing on these contexts that he deems sense-sensitive to link between them in both languages. Such contexts include, for instance, noun-noun, noun-preposition-noun, adjective-noun, and noun-adjective. His approach requires that the parallel texts of both languages should be tokenized, lemmatized, POS tagged and superficially parsed by simple pattern matching to extract sense-sensitive contexts of words. He later extends his approach to learn bilingual equivalents from English and Spanish comparable corpora (2007).

As far as Arabic is concerned, Saleh & Habash (2009) discuss an approach to automatic extraction of a lemma-based Arabic-English dictionary from a sentence-aligned

parallel corpus. They use a morphological disambiguation system to determine the full POS tag, lemma and diacritization (i.e. vocalization).

In another endeavor Morin and Prochasson (2011) develop a model for bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. They make use of a small sized bilingual lexicon induced through parallel sentences included in the comparable corpus. Their experiments were conducted on the French-English language pair.

More recently, Gutierrez-Vasques (2015) present a proposal to perform bilingual lexicon extraction for a distant language pair (Spanish-Nahuatl) using a small parallel corpus. The work was in progress and thus no results were reported.

As far as our approach is concerned, we exploit word co-occurrence frequencies in a parallel corpus, as the case with earlier attempts in the field. What is new in our approach, as mentioned above, is the use of head-dependent relations to extract a seed lexicon that is used as a bootstrapping technique to improve the extraction process. Table 1 shows the accuracy of the above-mentioned approaches to lexicon extraction from parallel corpora, using F-score[1] which comprises both precision and recall.

Table 1. ***F-scores for previous approaches to bilingual lexicon extraction from parallel corpora***

Approach	***Language Pair***	***Data Size***	***Precision***	***Recall***	**F-Score**
Resnik & Melamed (1997)	French-English	400K words	0.94	0.30	0.455
Tiedemann (1998)	Swedish-German	36K short structures	0.967	0.494	0.653
	Swedish-English	36K short structures	0.965	0.283	0.437
Tufiş & Barbu (2002)	Romanian-English	14K words	0.782	0.726	0.753
Otero (2005)	English-French	2M words	0.94	0.74	0.828
Saleh & Habash (2009)	Arabic-English	4M words	0.88	0.59	0.706

3. Arabic Linguistic Challenges

In this section the author throws light on the linguistic challenges that face Arabic NLP, and consequently have an impact on the current task of lexicon extraction. Arabic poses a number of linguistic challenges due to its nature which is massively more ambiguous than English for the following reasons:

- Arabic is normally written without diacritics or short vowels, which results in a great number of ambiguities and consequently represents a challenge for Arabic NLP (Maamouri et al. 2006). The word-form علم*Elm*[2], for instance, is composed of only three letters but has seven different readings, as shown in table 2.

Table 2. ***Ambiguity caused by the lack of diacritics***

Arabic diacritized word	**Meaning**
عِلْمٌ*Eilomu*	knowledge
عَلَمٌ*Ealamu*	flag
عَلِمَ*Ealima*	knew
عُلِمَ*Eulima*	is known
عَلَّمَ*Eal~ama*	taught
عُلِّمَ*Eul~ima*	is taught
عَلِّمْ*Eal~im*	teach!

- Arabic is a highly inflectional language, which makes Arabic morphological analysis a tough process. In Arabic very often a single word will consist of a stem with multiple fused affixes and clitics. Sometimes an Arabic word could stand as a complete sentence, as in فأسقيناكموه *fOsqynAkmwh* "then we gave it to you to drink". This type of complex words is very common in Classical Arabic (CA). This morphological richness is a source of an added increase in ambiguity that is a big challenge to the current task. For instance, the word وجدنا *wjdnA* can be analyzed (among other analyses) as *wa+jad~+u+nA* 'and our grandfather' or as *wajad+nA* 'we found' (Saleh & Habash, 2009).
- Arabic is distinguished by its flexibility of word order, where the orders VSO, VOS, SVO and OVS are all possible orders for the arguments of a transitive verb under appropriate conditions.
- Arabic is a pro-drop language, as the subject may not be explicitly mentioned but implicitly understood as an elliptic personal pronoun. This gives rise to a major syntactic ambiguity, leaving any syntactic parser with the challenge to decide whether or not there is an elliptic pronoun in the subject position. For example, the Arabic sentence أكلت الدجاجة *Aklt AldjAjp* "ate(feminine) the-chicken" has two different interpretations – "The chicken ate" or "(She) ate the chicken" (Attia, 2008: 103).
- Arabic allows for 'equational' or 'verbless' sentences that consist of a noun phrase (NP) and some kind of predication (another NP, a prepositional phrase (PP), an adjectival phrase or an adverbial phrase) (Badawi et al. 2004). These constructions normally contain a zero-copula which is omitted in the present tense indicative, but is present in the negated forms. For instance, the equational sentence الرجل في الدار (*Alrjl fy AldAr* "the man (is) in the house") has a PP predicate في الدار (*fy AldAr* "in the-house").

- Arabic nouns can be used as adjectives or as possessive determiners in a specific type of construction called a construct phrase, with typically little inflectional morphology to mark such uses (Alabbas and Ramsay, 2011). For instance, كتاب الطالب *ktAb AlTAlb* "the student's book" is a construct phrase in which the second noun الطالب *AlTAlb* specifies or limits the particular identity of the first noun كتاب *ktAb*.

4. The Arabic-English Parallel Corpus

In our endeavor to extract translation equivalents we use a parallel Arabic-English corpus. The aim is to test our approach on such a corpus, with a view to be tested in future on any other type of parallel corpora. We use the Qur'anic Arabic text, which is written in CA, and its English translation rendered by Ghali (2005) as our parallel corpus. The parallel corpus has been obtained from a website[3] which contains the Arabic original of the Qur'anic text along with a number of English translations. The Qur'anic text has been chosen because of its availability in both Arabic and English. We have chosen Ghali's translation because it is less explanatory than other translations. When necessary, the explanatory notes are given between parenthetical brackets. This makes them easy to remove by using regular expressions. Furthermore, his translation is source language-oriented, as he emphasizes the "strict adherence to the Arabic text, and the obvious avoidance of irrelevant interpretations and explications" (Ghali, 2005). Accordingly, we can assume that there is a reasonably systematic relationship between lexical items in the Arabic and English versions, so that alignment is not a major issue. The parallel text is a small-sized corpus, containing 77,800 words in the original Arabic text. The Qur'anic text, as the case with many CA texts, is basically diacritized, where diacritics (i.e. short vowels and other orthographic diacritics) are placed above or below letters, as shown in table 2 above. Nonetheless, we use an undiacritized version of the corpus so as to mimic the way Modern Standard Arabic (MSA) is written so that our approach could be extended to an MSA corpus. According to Mubarak et al. (2011), MSA tends to be simpler than CA in grammar usage, syntax structure, and morphological and semantic ambiguity.

The Arabic Qur'anic text is composed of unpunctuated verses which mostly contain long sentences that may exceed 100 words. A Qur'anic verse is one of the numbered subdivisions of a chapter in the Qur'an. A verse may contain one sentence or more, separated by conjunctions rather than punctuation marks. Verse markers are used to denote the end of a Qur'anic verse. But in our discussion we will use the terms verse and sentence interchangeably. The Qur'anic text consists of 114 chapters that contain a total of 6236 verses. The following figure shows a verse in its undiacritized form in parallel with its English translation.

An Arabic Verse	***English Translation***
الحمد لله الذي أنزل على عبده الكتاب ولم يجعل له عوجا *AlHmd llh Al*y Onzl ElY Ebdh AlktAb wlm yjEl lh EwjA*	Praise be to Allah Who has sent down upon His bondman the Book and has not made to it any crookedness.

Figure 1. **A sample of the parallel corpus before adding POS tags**

The 77,800 word tokens in the Arabic corpus contain around 19,000 vocalized word types in the diacritized version, which are reduced to nearly 15,000 unvocalized word types when diacritics are removed. As we can notice, the number of words has collapsed because several words share the same orthographic form but are different with regard to diacritic marks. Thus, many different diacritized word-forms were conflated to fewer forms after removing diacritics. The English translation, on the other hand, contains nearly 162,000 word tokens after normalization (i.e. removing punctuation marks, lowercasing all words and deleting the explanatory words between brackets in the translation). This difference in the number of word tokens between Arabic and English is probably due to the fact that Arabic is known for being morphologically rich, where numerous clitic items (conjunctions, prepositions and pronouns) are attached to words, thus forming complex words that need to be decomposed into a number of words when translating into English.

The Qur'anic text is characterized by unique linguistic or rather rhetorical features, which should pose special interests and challenges for computational linguistics solutions (Sharaf & Atwell, 2009). As pointed out by Dukes and Habash (2010), processing Qur'anic Arabic is a unique challenge from a computational point of view. The linguistic style of the Qur'an makes extensive use of many rhetorical devices such as foregrounding and backgrounding, grammatical shift, metaphors and figurative language, idiomatic expressions, culture-specific items, and lexically compressed items where lengthy details of semantic features are compressed and encapsulated in a single word (Abdul-Raof, 2001). All these features make the current corpus a challenging type of text for the current task. What makes it more challenging is that the Arabic original text is written without punctuation marks, which makes it difficult to know sentence boundaries. We had, thus, to remove punctuation marks from the English translation to resemble the Arabic text in the parallel corpus. All these features refer to the robustness of the adopted approach, since our logical assumption is that experimenting with a less challenging corpus is expected to lead to improvement in accuracy scores.

5. Extracting Translation Equivalents

In this section we present our method for bilingual lexicon extraction (BiLexExtract) from a parallel corpus. The proposed method consists of the following stages:

- Preprocessing Steps
- Bilingual Lexicon Extraction
- Bootstrapping Techniques

5.1 Preprocessing Steps

We have used a number of preprocessing steps for both Arabic and English to annotate the parallel corpus which we use for building the lexicon. The main step is the use of POS tagging to annotate the bi-texts. Having labeled the corpus with POS tags, we wrote a small-sized shallow dependency parser and a stemmer for both languages to be used in our task.

5.1.1 Part-of-speech tagging

We have built a lexicon-free part-of-speech (POS) tagger which is described in detail in Ramsay & Sabtan (2009) to tag the Arabic text in the corpus. This Arabic tagger, which requires very little manual effort, obtains results which are comparable with state-of-the-art taggers for Arabic. As for English, we similarly used a lexicon-free tagger for English which is

based on the BNC basic (C5) tagset (Burnard, 2007), but with some modifications that produce a coarser-grained tagset. The English tagger was also developed with the least manual intervention, as shown in Sabtan (2011). Figure 2 below shows a portion of the parallel corpus after being annotated with POS tags for both Arabic and English.

Arabic POS-tagged Corpus	***English POS-tagged Corpus***
(*AlHmd*,NN)(*llh*,PREP+NN)(*Al*y*,REL PRON)(*Onzl*,VV)(*ElY*,PREP)(*Ebdh*,N N+PRON)(*AlktAb*,NN)(*wlm*,CONJ+ PART)(*yjEl*,VV)(*lh*,PREP+PRON)(*Ew jA*,NN)	(praise,NN)(be,VB)(to,PR)(Allah,NP)(who,PN)(ha s,VH)(sent,NN)(down,AV)(upon,PR)(his,DP)(bon dman,NN)(the,AT)(book,NN)(and,CJ)(has,VH)(no t,XX)(made,VV)(to,TO)(it,PN)(any,DT)(crookedn ess,NN)

Figure 2. **A portion of the parallel corpus after adding POS tags**

The used Arabic and English tagsets to annotate the parallel corpus are described in tables 3 and 4 respectively.

Table 3. ***The used Arabic tagset***

Tag	**Description**
CONJ	Conjunction
DEM	Demonstrative
DET	Determiner
EMPH_PART	Emphatic Particle
INTER_PART	Interrogative Particle
NN	Noun
NUM	Number
PART	Particle
PB_VV	Praise or Blame Verb[4]
PREP	Preposition
PRON	Pronoun
REL_PRON	Relative Pronoun
SP_VV	Special Verb (e.g. *kAna*)[5]
UN_WD	Unknown Word[6]
VV	Verb

Table 4. ***The used English tagset***

Tag	Description	Tag	Description
AJ	Adjective	PN	Pronoun
AT	Article	PO	Possessive Marker*'s*
AV	Adverb	PR	Preposition
CJ	Conjunction	PU	Punctuation Mark
CR	Cardinal Number	TO	Infinitive Marker *to*
DP	Possessive Determiner	VB	Verb 'BE'
DT	Determiner	VD	Verb 'DO'
EX	Existential *there*	VH	Verb 'HAVE'
IT	Interjection	VM	Modal Verbs
NN	Noun	VV	Verb
NP	Proper Noun	XX	Negative Particle *not*
OR	Ordinal Number	ZZ	Alphabetical Symbols

5.1.2 Shallow dependency parsing

We wrote rule-based shallow dependency parsers for both Arabic and English, using dependency grammar framework. A dependency grammar is defined as a set of dependency rules, each of the form 'category X may have category Y as a dependent'. In other words, dependency structure is determined by the relation between a word (a head) and its dependents (Hudson, 1984; Mel'čuk, 1988). For example, the dependency analysis of the sentence "the boy ate an apple" is given the following dependency representation in figure 3 below, with arrows pointing from head to dependent.

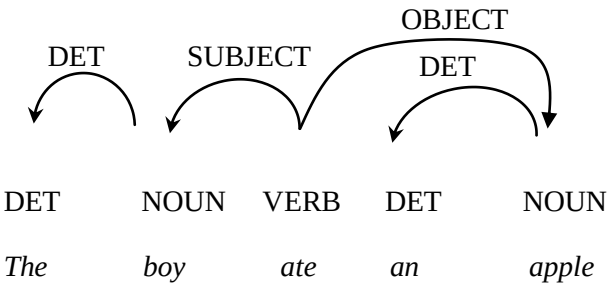

Figure 3. **A dependency graph for an English sentence**

The parser is shallow in the sense that we only use a partial set of dependencies rather than entire dependency trees. The main goal of writing shallow dependency parsers for Arabic and English is to find syntactically related words in the parallel corpus to be used as 'translation seeds' to resegment the corpus and bootstrap the lexicon extraction process. The basic idea behind this activity is that having shorter sentences could improve the accuracy of the extraction process. In our dependency analysis of Arabic we use untyped relations, since it is extremely difficult to label dependents with either 'subject' or 'object' in the absence of a lexicon in which subcategorization frames for verbs are specified. Moreover, Arabic is a relatively free word order language where the subject and object can precede or follow the verb. English word order, in

contrast, is not as free as Arabic. Despite being shallow, both parsers have proven to be useful for our overall task. The details of the bootstrapping techniques will be described in section 5.3 below.

5.1.3 Stemming

We also developed stemmers for Arabic and English. Since it is difficult to obtain the lemmas without using a lexicon, we performed stemming rather than lemmatization. We apply light stemming, using a corpus-based approach (Sabtan, 2012). As for English, we remove inflectional suffixes after grouping word variants based on letter-sequence similarity. This has been done to test the extraction algorithm on both unstemmed and stemmed texts.

5.2 Bilingual Lexicon Extraction

The main task is to automatically build a bilingual lexicon. The lexicon is extracted using statistical methods, based on the following basic principle:

- For each sentence-pair, each word of the target language (TL) sentence is a candidate translation for each word of the aligned source language (SL) sentence.

This principle means that (S, T) is a candidate if T appears in the translation of a sentence containing S. This sentence-pair can be either POS tagged, as made use of in the first stage before bootstrapping, or labeled with dependency relations (DRs), as done later in the bootstrapping techniques. Following the above principle we compute the absolute frequency (the number of occurrences) of each word in the SL and TL sentences. We then compile a bilingual lexicon, giving preference to the target words that have the highest score in the TL sentences that correspond to the SL sentences, providing that they have the same POS tags in both languages. These words are ordered by their frequency of occurrence. This method is the 'baseline' algorithm, which we modify by taking into account the relative distance between SL and TL words in their specific contexts, and then the distance score is squared. This produces our second 'weighted' algorithm. Specifically, this algorithm measures the distance between the positions of SL words in a sentence relative to the positions of corresponding TL words.

As figure 4 shows, we build the bilingual lexicons from texts that are either POS tagged or labeled with DRs. Then we apply the basic principle that we have mentioned above, which we call 'Equivalents Matching Algorithm'. This matching algorithm matches words in the parallel texts based on frequency of occurrence and relative positions as well as the similarity of their POS tags in both languages.

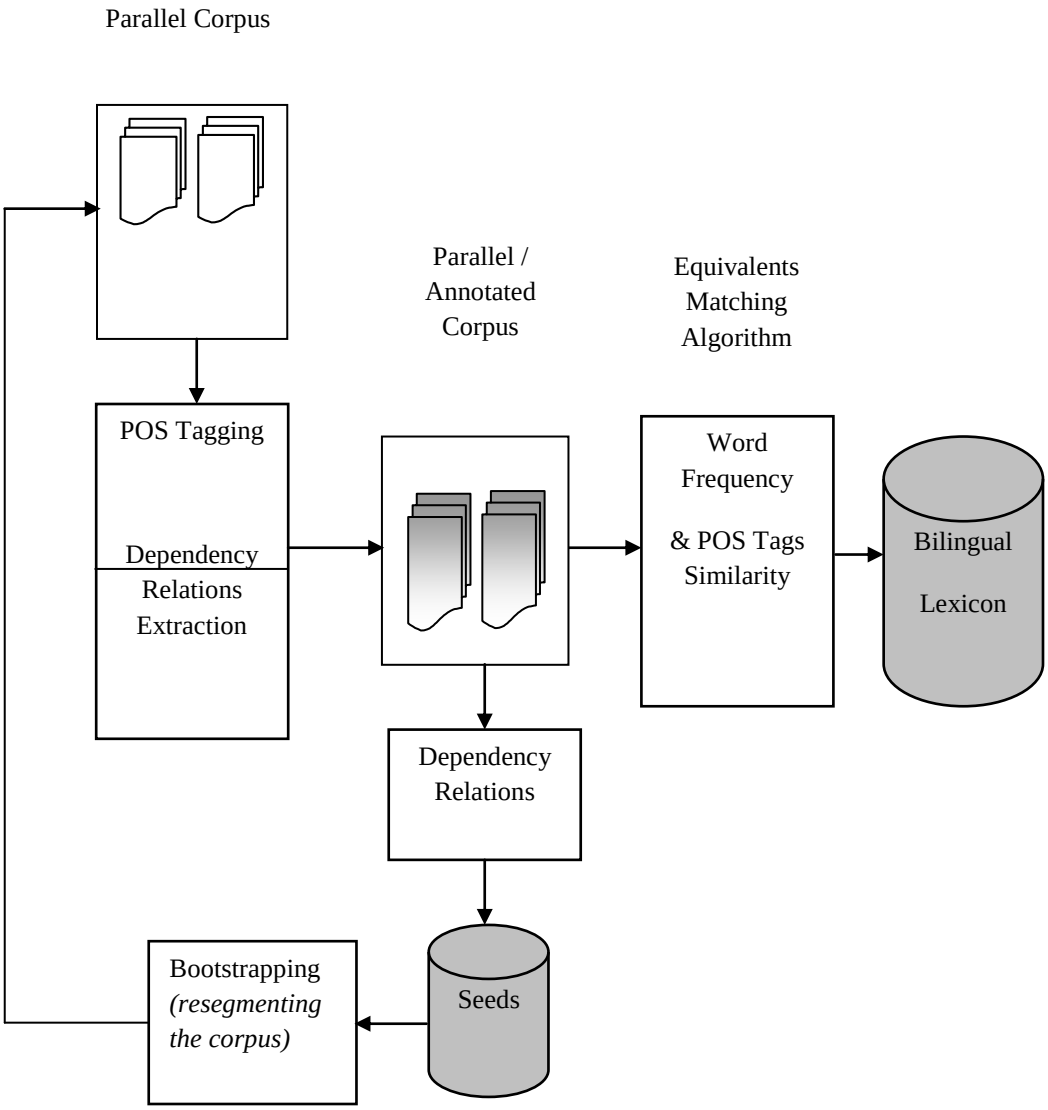

Figure 4. **Automatic lexicon building architecture**

The basic assumption behind matching equivalents based on their POS tags similarity is also emphasized by Melamed (1995, p. 190) when he stated that "word pairs that are good translations of each other are likely to be the same parts of speech in their respective languages". Thus, we apply the following constraint or rather filter.

- A chosen TL candidate for a given SL word must have the same POS tag as that of the SL word.

To clarify what is meant by the similarity of POS tags we will give the following example. If, for instance, a SL word is POS tagged as "verb" and has a number of TL translation candidates that have different POS tags (i.e. "verb", "noun", .etc.), we choose the word that has the same tag as that of the SL word. In this case we select the word that is tagged as "verb".

However, for this approach to be feasible the tagset for Arabic and English should be similar. Since we are mainly interested in open-class words, we have made the tagset for such categories similar in the two languages. Finally, as shown in 5.3, we extract 'head-dependent' pairs for some constructions from the DR-labeled bi-texts. These pairs are then filtered to obtain one-word translation seeds that are used to bootstrap the lexicon extraction process.

The translation candidates in an extracted lexicon are listed in a descending order according to their frequency of occurrence in the context of the SL word in question. As will be shown in the evaluation section, we evaluate the accuracy of our method, using F-measure, on only the first suggested candidate in a translation lexicon. Sometimes the correct equivalent comes in other positions, but we currently score only the top TL candidate.

5.3 Bootstrapping Techniques

To improve the accuracy of an extracted lexicon we have applied a number of bootstrapping techniques, making use of the DRs for some basic constructions in both Arabic and English. We label some constructions in the parallel corpus with DRs to automatically extract a number of translation seeds that we could then use to start our bootstrapping techniques. Specifically, these seeds could be used to resegment the parallel corpus to help improve the matching of equivalents. Broadly speaking, the bootstrapping techniques can be divided into two basic steps: (1) the extraction of seeds and (2) resegmenting the corpus, relying on these seeds.

5.3.1 Extraction of seeds

To automatically extract seeds, we firstly apply the same algorithm described above for extracting the bilingual lexicons but condition that the suggested pair must be labeled with a DR in the parallel corpus. We match those Arabic words that are in a similar DR with their corresponding English words. This matching between Arabic and English pairs is basically between two dependency structures to find corresponding heads and dependents. It should be noted that we map between fragments of sentences in the parallel corpus, as we focus on certain syntactic dependencies in both languages. This can be made clearer through figure 5, which shows a parallel 'verb-object' relation for unstemmed fragments.

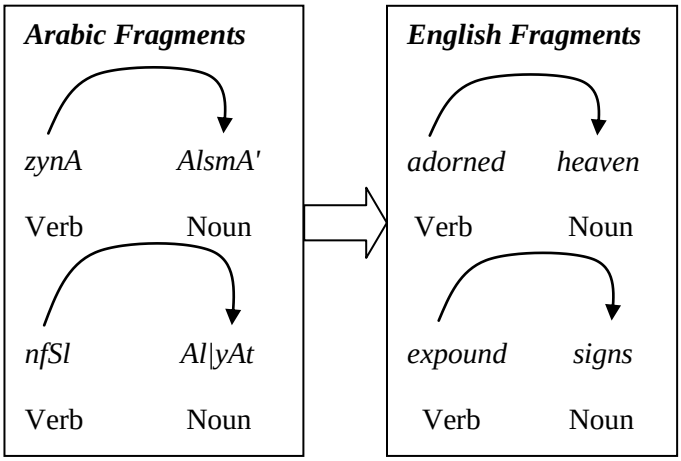

Figure 5. **Mapping between 'head-dependent' pairs in the parallel corpus**

We extract a number of dependency pairs, i.e. 'head-dependent' translation pairs. When examining these pairs we found out that the pairs that occur once in the parallel corpus have low precision. Thus, we automatically extract those pairs that occur at least 2 times in the corpus.

This stage produces a large number of candidates, many of which are wrong. We, therefore, carry out a filtering process to obtain a number of trusted one-word translation seeds. The filter is based on the following constraint:

- A chosen translation candidate for a given element of the dependency pair, the head or dependent, must have an occurrence that is equal to or more than a specific threshold of the total number of occurrence of all other suggested candidates.

We have tried different numbers, setting the threshold to 0.5, 0.3 and 0.2 of the total, but we obtained better scores when we set it to 0.5. For example, the Arabic word أيدي *>ydy* "hands" has a number of translation candidates in a dependency-based extracted lexicon before filtering. These candidates are given along with the number of their occurrence as follows: *>ydy* {'hands': 3, 'legs': 1, 'people': 1, 'angels': 1}. Here the word "hands", which is the right equivalent, will be selected and the three other candidates will be excluded. This is because the word 'hands' occurs 3 times, which is more than 0.5 of the total occurrence of the three other candidates.

Having filtered such candidates, we ended up with a number of one-word translation pairs which we call 'seeds'. Table 5 shows a sample of the extracted seed lexicon which achieved 81% accuracy.

Table 5. ***A sample of the seed lexicon***

Word	POS	Equivalent
|yp آية	noun	sign
$k شك	noun	doubt
sryE سريع	adj.	swift
qAl قال	verb	said
trk ترك	verb	left

These seeds are used as anchor points to resegment the SL sentences and the corresponding TL sentences in the parallel corpus and consequently introduce a new alignment of the sentences in the bi-text.

5.3.2 Resegmenting the corpus

We now use the seeds as anchor points for resegmenting the parallel corpus. Resegmentation is implemented when the SL (Arabic) word in a seed pair is found in a given SL sentence and the TL (English) word in the same seed is found in the corresponding TL sentence. In this case the part of the Arabic sentence that comes before the Arabic seed could align to the part of the English sentence that comes before the English seed and consequently the part of the Arabic sentence that comes after the Arabic seed could align to the part of the English sentence that comes after the English seed.

We carry out three different experiments of resegmentation and test the extraction process after each one of these experiments. These three experiments can be illustrated as follows:

i. Remove seeds from the parallel corpus and start the extraction process on the new bi-texts without the seeds.
ii. Resegment the bi-sentences in the corpus at the places where one of the seeds is found and keep the seeds.
iii. Combine the previous two steps of resegmenting the sentences and removing the seeds.

The third step achieved the best accuracy score for the bootstrapping techniques. The scores obtained before and after bootstrapping will be given below.

6. Evaluation Framework

We have tested our extraction method on a number of extracted lexicons using the 'baseline' and 'weighted' algorithms on POS-tagged texts that are either stemmed or not. We use the standard F-measure, which considers both the precision and the recall of the test to compute the score.

In our framework precision can be simply defined as the number of correct translations proposed by the system divided by the number of all translations which has been suggested in a given test set. Recall, on the other hand, is defined as the number of correct translations proposed by the system divided by the number of all test instances (i.e. the tested samples). We evaluate extracted lexicons with regard to the top translation candidate. Other candidates that occupy any other position in the lexicon are not scored in this framework. It should be noted that our extraction method at present deals only with single words and cannot tackle multi-word expressions (MWEs) which will be a topic for future work.

An extracted bilingual lexicon could contain a number of TL translation candidates for a given SL word. These candidates are listed in order of frequency, and the correct equivalent may occupy any position in the list. However, in some cases no translation candidate is suggested. We use a gold standard which we manually constructed based on the English translation of the corpus we use. Then, we compare the output with the reference translation and compute the score. We have tested three different samples, containing 200 open-class words, from different parts of the corpus. These samples are randomly chosen from the entire corpus. We computed the F-score for each sample then combined the scores for all the three samples to obtain the average score. Table 6 shows the average F-scores for a number of bilingual lexicons before bootstrapping. These bilingual lexicons have been extracted using the Arabic-English bi-text in their stemmed or unstemmed forms.

Table 6. ***F-scores for various types of lexicons before bootstrapping***

Text Type	Algorithm	F-score
Unstemmed Bi-Text	Baseline	0.5362
	Weighted	0.6600
Stemmed Bi-Text	Baseline	0.5708
	Weighted	0.6831

As can be observed in table 6 above, the best F-score was obtained when we used the 'weighted' algorithm on the stemmed Arabic and English bi-text. We will use the 'weighted' algorithm with the stemmed texts, which obtained the best result, to apply the bootstrapping techniques.

Earlier we referred to three different experiments of bootstrapping, (i.e. 1. removing seeds from the corpus, 2. resegmenting the corpus while keeping the seeds, and 3. combining between both steps of resegmenting the corpus and removing seeds). Each experiment resulted in a different F-score. Table 7 summarizes the scores obtained in the three experiments in comparison with the best score obtained before applying the bootstrapping techniques.

Table 7. ***Comparison of F-scores before and after bootstrapping***

Experiments	**Precision**	**Recall**	**F-score**
Before Bootstrapping	0.6849	0.6813	0.6831
1st exp. of bootstrapping	0.7022	0.6985	0.7003
2nd exp. of bootstrapping	0.7350	0.7310	0.7330
3rd exp. of bootstrapping	0.7413	0.7374	0.7393

It is clear in table 7 that the third bootstrapping technique (i.e. resegmenting the corpus and removing seeds) has achieved the best F-score. It should be noted that the removed seeds in the first and third experiments of bootstrapping are computed in the F-score.

Having obtained a new parallel corpus after resegmenting the corpus and removing the seeds, we started to carry out another round of bootstrapping. We hoped that carrying out another round of bootstrapping would improve the situation. However, we did not obtain any extra improvement, and thus did not carry out any further experiments.

Table 8 shows the top 5 translation candidates for an excerpt from the bilingual lexicon that was automatically extracted using stemmed texts in both Arabic and English. We also compare it with the first 5 equivalents in the bilingual lexicon of Google Translate[7]. In the following lines each example in table 8 is discussed so as to throw light on the problems that face the current method for lexicon extraction. The last three words in the table point to three important issues that need to be tackled in future work.

Table 8. ***Comparing top 5 translation candidates in BiLexExtract with Google Translate***

Arabic Word	**POS**	**BiLexExtract**	**Google Translate**
ktAb كتاب	Noun	book, Allah, brought, sent, way	book, volume, publication, work, compilation
jA' جاء	Verb	come, indeed, said,	came, come, arrive, turn up,

		came, say	bring
Elm علم	Noun / Verb	know, knowledge, Allah, say, way	**Noun**: science, flag, knowledge, learning, scholarship **Verb**: know, teach,inform, mark, educate
EZym عظيم	Adj.	magnificent, Allah, tremendous, torment, reward	great, mighty,major, magnificent, fantastic
yb$r يبشر	Verb	good, give, pray, said, esteemed	presage, bode, promise, rasp

The first word, namely كتاب *ktAb* has the correct English equivalent in the first position, i.e. "book". As for the word جاء *jA'*, which is a perfective verb, it has two morphologically related translation candidates in the first and fourth positions. They differ only with respect to their tense. We regard both candidates as correct because we ignore tense differences in our evaluation. It should be noted that generally stemming improved extracted lexicons, as the case with the current word. For example, some of its cliticized (i.e. unstemmed) forms (*jA'h* "came to him" and *jA'k* "came to you") wrongly have the words ("said" and "indeed") as the top translation candidates respectively. However, sometimes the Arabic stemmer produces the illegitimate stem of a word (by keeping clitics) but we score it so long as the TL candidate is the correct equivalent for the SL word's base form.

As regards the word علم *Elm*, it can be grammatically classified as noun or verb, which depends on the context in which the word is used. Table 1 in section 3 above discusses the different meanings for this ambiguous word. This type of ambiguity, which is caused by difference in POS category, is normally called categorical ambiguity and is pervasive in Arabic. This is because, as shown before, Arabic is written without short vowels and other diacritic marks, which leads to many homographs that differ with regard to their POS categories. The first translation candidate "know" is correct for one of the verb senses and the second candidate "knowledge" is correct for the noun sense. As long as the first candidate is the right equivalent for one of the word's senses we consider it right in our current evaluation. In future we plan to handle such homographs to choose the correct equivalent in its sentential context.

Similarly, the word عظيم *EZym* is an ambiguous word that has different senses. In other words, it is a polysemous adjective. The different senses for this word are distinguished according to the linguistic context in which they occur. We use the linguistic context here to mean the syntagmatic relation that deals with co-occurrence patterns. These patterns can be observed on both lexical and structural levels. We are concerned here with the lexical level. One of the relationships that hold between words on the syntagmatically lexical level is 'collocation'. For instance, the word *EZym* may co-occur with the word أجر>*jr* "reward" to mean "magnificent", but it may co-occur with the word عذاب *E*Ab* "torment" to mean "tremendous". The first sense for the word *EZym* occupies the first position in the extracted lexicon, whereas

the second sense occupies the third position in the lexicon. Since the translation candidate in the first position is the correct equivalent for one of the senses of the word in question, we consider it right and score it in our evaluation framework. In future work we aim to handle such ambiguous cases in order to disambiguate them in their context.

Finally, the word يبشر *yb$r* is a lexically compressed verb. This means that the verb contains a number of semantic features that are compressed in a single lexical item. This is a lexical feature of the Qur'anic text that we are using as our corpus, as noted in section 4 above. Such a lexical item needs to be translated into a multi-word expression (MWE) in the target language to convey all of its meanings in the Qur'anic text. Thus, the translation for this verb is normally "give (good) tidings" in most contexts. However, for rhetorical reasons this word is sometimes used instead of its antonym ينذر *yn*r* "warn" to foretell evil things. The extraction method selects one word of the whole MWE, as it does not currently handle MWEs.

7. Conclusion and Future Work

We have presented an approach to bilingual lexicon extraction of open-class words from a parallel corpus for two historically unrelated languages. The extraction method can be applied to any language pair if there is a POS-tagged parallel corpus for the two languages concerned. The used Arabic-English corpus, which is of a religious domain, is composed of mostly long unpunctuated verses, where a verse may contain a number of sentences. The extraction method exploits word co-occurrence frequencies in a POS-tagged parallel corpus. Then, a seed lexicon is extracted, using parallel dependency relations, to bootstrap the entire lexicon extraction process. The best F-score we obtained in the first round of experiments before bootstrapping is 0.6831. Following a number of bootstrapping techniques, the F-score rose to 0.7393. In future we plan to use bigger parallel corpora of other domains to test our method. Moreover, we will investigate how to use extensions of the current algorithm to extract MWEs. Finally, we will handle those lexically ambiguous words that have different senses. Such words include homographs as well as polysemes. We will start with automatic identification of those words with a view to disambiguating them in their contextual sentences.

Notes

1. F-score, or F-measure, is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. In case of MT evaluation, precision can be simply defined as the number of correct translations proposed by the system divided by the total number of all translations which has been suggested in a given test set. Recall, on the other hand, is defined as the number of correct translations proposed by the system divided by the total number of all test instances. The F-score is calculated according to the following equation: F-score = 2 x precision x recall / precision + recall.
2. Throughout this paper, Arabic words are normally presented in the Arabic script followed by Buckwalter transliteration in italic and an English gloss in double quotes. The transliteration scheme is available at: http://www.qamus.org/transliteration.htm
3. http://quran.com
4. This type of verb refers to the Arabic verbs *niEoma* نعم"how good" and بئس *bi}osa* "how bad". They are usually called أفعال المدح والذم "verbs of praise and blame".
5. This refers to كان وأخواتها *kAna waakhwAtuhA*, which place the following subject into the nominative case and the predicate into the accusative case.

6. This tag is used in case the rule-based tagger cannot identify the category of the word under analysis. However, this tag disappears in the output of the final stage of the tagger.
7. Words have been tested by Google Translate in March 2016.

About the Author:

Dr. Yasser Sabtan earned his PhD in Computational Linguistics from the University of Manchester, UK in 2011. He is currently an Assistant Professor at the Department of Languages and Translation, Dhofar University, Oman. Prior to joining Dhofar University in 2015, Dr. Sabtan taught linguistics and translation courses at Al-Azhar University, Egypt. His research interests focus on machine translation, audiovisual translation, Arabic computational linguistics, corpus linguistics and pragmatics.

References

Abdul-Raof, H. (2001). *Qur'an Translation: Discourse, Texture and Exegesis*. London and New York: Routledge.

Alabbas, M. & Ramsay, A. (2011). Evaluation of Dependency Parsers for Long Arabic Sentences. In *Proceedings of the International Conference on Semantic Technology and Information Retrieval (STAIR'11)*, Kuala Lumpur, Malaysia.

Attia, M. A. (2008). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. (Unpublished doctoral dissertation). University of Manchester, UK.

Badawi, E. M., Carter, M. G. & Gully, A. (2004). *Modern Written Arabic: A Comprehensive Grammar*. Routledge.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J.,Jelinek, F., Lafferty, J. D., Mercer, R. L.& Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16* (2), 79–85.

Burnard, L. (2007). *Reference Guide for the British National Corpus (XML Edition)*. The Research Technologies Service at OxfordUniversity Computing Services, Available at http://www.natcorp.ox.ac.uk/docs/URG/

Dagan, I., Itai, A. & Schwall, U. (1991). Two Languages Are More Informative Than One. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California.

Fišer, D. & Ljubešić, N. (2011). Bilingual Lexicon Extraction from Comparable Corpora for Closely Related Languages. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria.

Gale, W. A. & Church, K. W. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, Pacific Grove, California, Morgan Kaufmann Publishers, San Mateo, California.

Ghali, M. M. (2005). *Towards Understanding the Ever-Glorious Qur'an* (5th ed.). Cairo, Egypt: Publishing House for Universities.

Gutierrez-Vasques, X. (2015). Bilingual Lexicon Extraction for a Distant Language Pair Using a Small Parallel Corpus. In *Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW)*, Denver, Colorado.

Hudson, R. (1984). *Word Grammar*. Oxford, England: Basil Blackwell Inc.

Kaji, H. & Aizono, T. (1996). Extracting Word Correspondences from Bilingual Corpora Based

on Word Co-occurrences Information. In *Proceedings of the 16th International Conference on Computational linguistics (COLING'96)*, Copenhagen, Denmark.

Kumano, A. & Hirakawa, H. (1994). Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan.

Maamouri, M., Bies, A. & Kulick, S. (2006). Diacritization: A Challenge to Arabic Treebank Annotation and Parsing. In *Proceedings of the Challenge of Arabic for NLP/MT Conference*. The British Computer Society, London, UK.

Melamed, I. D. (2000). Models of Translational Equivalence among Words. *Computational Linguistics, 26* (2), 221-249.

Melamed, I.D. (1995). Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora (WVLC3)*, Boston, MA, U.S.A.

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University Press of New York.

Morin, E. & Prochasson, E. (2011). Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland, Oregon.

Mubarak, H., Metwally, A. & Ramadan, M. (2011). Analyzing Arabic Diacritization Errors of MADA and Sakhr Diacritizer. In *Proceedings of the 11th Conference on Language Engineering (ESOLEC'2011)*, Cairo, Egypt.

Otero, P. G. (2005). Extraction of Translation Equivalents from Parallel Corpora Using Sense-Sensitive Contexts. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT'05)*, Budapest, Hungary.

Otero, P. G. (2007). Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of the MT Summit XI*, Copenhagen, Denmark.

Ramsay, A. & Sabtan, Y. (2009). Bootstrapping a Lexicon-Free Tagger for Arabic. In *Proceedings of the 9th Conference on Language Engineering (ESOLEC'2009)*, Cairo, Egypt.

Resnik, P. & Melamed, I.D. (1997). Semi-Automatic Acquisition of Domain-Specific Translation Lexicons. In *Proceedings of the 5th ACL Conference on Applied natural Language Processing*, Stroudsburg, PA, USA.

Sabtan, Y. (2011). *Lexical Selection for Machine Translation*. (Unpublished doctoral dissertation). University of Manchester, UK.

Sabtan, Y. (2012). Arabic Stemming: A Corpus-Based Approach. In *Proceedings of the 12th Conference on Language Engineering (ESOLEC'2012)*, Cairo, Egypt.

Saleh, I.M. & Habash, N. (2009). Automatic Extraction of Lemma-based Bilingual Dictionaries for Morphologically Rich Languages. In *Proceedings of the 3rd Workshop on Computational Approaches to Arabic Script-based languages at the MT Summit XII*, Ottawa, Ontario, Canada.

Sharaf, A. & Atwell, E. (2009). A Corpus-based Computational Model for Knowledge Representation of the Quran. In *Proceedings of CL2009 International Conference on Corpus Linguistics*, Liverpool, England.

Tiedemann, J. (1998). Extraction of Translation Equivalents from Parallel Corpora. In *Proceedings of the 11th Conference on Computational Linguistics*, Copenhagen, Denmark.

Tufis, D. & Barbu, A. M. (2002). Revealing Translators' Knowledge: Statistical Methods in Constructing Practical Translation Lexicons for Language and Speech Processing. *The International Journal of Speech Technology, 5*, 199-209.