Automatic Extraction of Lemma-based Bilingual Dictionaries for Morphologically Rich Languages

Ibrahim M. Saleh
Georgetown University
Department of Linguistics
ICC 480, 3700 O Street, NW
Washington, DC 20057-1051
is94@georgetown.edu

Nizar Habash
Columbia University
Center for Computational Learning Systems
475 Riverside Drive
New York, NY 10115
habash@ccls.columbia.edu

Abstract

We present an approach for automatic extraction and filtering of a lemma-based Arabic-English dictionary from parallel corpora. Comparing the results of our system to a manually built dictionary shows a high degree of coverage complementarity. The generated dictionary: (1) has reasonable recall and high precision, (2) is significantly more comprehensive in terms of the covered Arabic-English lemma pairs, and (3) has high potential for future improvement.

1 Introduction

Many research areas in the field of natural language processing (NLP) require the availability of comprehensive up-to-date bilingual machine-readable dictionaries (MRD), e.g. machine translation (MT) and cross-language information retrieval. We can classify MRDs in two dimensions based on their method of creation (manual or automatic) and their form (surface-based or lemma-based).

Manually-built MRDs tend to lack coverage, and are expensive in terms of time and effort, particularly if they are to be kept up-to-date and covering both general and specialized technical terms. Manually-built MRDs are often general-purpose, lemma-based and mostly used in rule-based MT (Habash, 2003; Habash et al., 2006).

In contrast, NLP has benefited greatly from the presence of large amounts of text provided in different languages in the form of parallel corpora (also know as bilingual corpora or bi-texts) and comparable corpora. Such corpora have been used to automatically extract bilingual lexicons for a variety of applications, most prominently, statistical machine translation (Fung 1998; Koehn et al., 2003; Munteanu and Marcu, 2005). Automatically extracted dictionaries (often called translation lexicons or phrase tables) tend to be un-lemmatized surface-based, although often tokenized.

The practical difference between surface-based and lemma-based dictionaries correlates with the degree of morphological richness of the languages under consideration.

In the present study, we examine the challenges of, and present a solution to, extracting lemma-based MRDs from parallel data for a morphologically rich language, Arabic. Our automatically generated MRD is compared and contrasted to a manually built dictionary to highlight the contributions, limitations and potentials of automatic versus manual MRDs. The generated MRD: (1) has reasonable recall and high precision, (2) is significantly more comprehensive in terms of the covered Arabic-English lemma pairs, and (3) has high potential for future improvement.

In the next two sections we present related previous work and Arabic linguistic issues. In sections 4, we present the hand-created dictionary we compare to. In sections 5 and 6, we present and evaluate, respectively, our approach to corpus-based lemma-based dictionary extraction.

2 Previous Work

In spite of the availability of bi-directional and multi-directional electronic dictionaries where Arabic is included (Zughoul and Abu-Alshaar, 2005; Elkateb and Black, 2001; and Black and El-Kateb, 2004), there are not many studies handling the automatic extraction of general-purpose Arabic-English MRDs. The gloss lexicon of the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2004) is a widely used Arabic-English MRD in NLP. We compare to it in the present study. Other online Arabic-English-Arabic dictionaries exist, such as Sakhr's[1] dictionary. However, the lack of published technical documents about them makes them hard to use for research purposes.

We briefly review some relevant research in the area of automatic bilingual lexicon extraction. Melamed (1997) developed a word-to-word translation model that avoids the indirect associations responsible for error in IBM models. Tiedemann (1998) developed strategies for the extraction of translation equivalents from parallel corpora for historically related languages. Vintar (2001) made use of syntactic patterns in developing a system for the extraction of Slovene-English multi-word terms and bilingual conceptual mapping. Similarly, Gamallo (2005) used syntactic contexts that he deemed sense-sensitive to help with Spanish-English bilingual lexicon extraction from parallel corpora. He later extended his approach to Spanish-English bilingual lexicons using comparable corpora (Gamallo, 2007). In the context of MT, Foster et al. (2006) and Martínez and Way (2009) developed state-of-the-art methods for phrase-table filtering, i.e. smoothing and marker based filtering, respectively.

The present study considers the limitations and potentials of automatic versus manual generation of lemma-based MRDs for a morphologically rich language, namely Arabic, from parallel corpora. Our approach combines state-of-the-art Arabic processing with techniques of phrase-table extraction and filtering used in MT.

3 Arabic Linguistic Issues

We present some of the linguistic challenges of working with Arabic and a discussion of the MADA system used to address them.

3.1 Linguistic Challenges

In the context of developing an Arabic-English lexicon using parallel corpora, we distinguish three types of Arabic linguistic challenges: orthographic, morphological, and lexical.

Orthographically, Arabic is written using an alphabet with optional diacritics. Diacritics are used to indicate short vowels, nominal indefiniteness (termed *nunation*), vowel absence and consonantal reduplication (*shadda*). At least one diacritic on a word appears in around 1.5% of all words. Diacritic optionality is a major factor in analytical ambiguity for Arabic. For example, the undiacritized word كاتب *kAtb*[2] can be used to represent: كَاتِبٌ *kaAtibũ* 'a writer [nominative indefinite]' or كَاتَبَ *kaAtaba* 'he corresponded' (among others).

Morphologically, Arabic is a rather rich complex language. One issue is the use of non-concatenative (templatic) processes with roots and patterns to produce both derivational and inflectional forms (Elkateb and Black, 2001; Al-Sughaiyer and Al-Kharashi, 2004). Another issue is Arabic's rich inflectional system which includes gender, number, case, aspect, voice, mood, person, and state. A third issue is the set of attachable clitics representing single-letter conjunctions and particles and object/possessive pronouns. These different morphological issues cause an added increase in ambiguity that is a big challenge to the task at hand. For instance, the word وجدنا *wjdnA* can be analyzed (among other analyses) as *wa+jad~+u+nA* 'and our grand-father [and+grand-father+nominative+our]' (root *jdd*, lemma *jad~*) or as *wajad+nA* 'we found [found+we]' (root *wjd*, lemma *wajad*).

Lexically, we are aware of the challenge of how to define an appropriate lemma in Arabic. Whereas for English, lemmas are easily defined and derived given English's limited morphology (tense and number primarily); in Arabic, the canonical lemma form which abstracts away from all inflected variants is more complex. For nouns, it is the uncliticized singular indefinite (masculine if gender-inflectable) form; and for verbs, it is the uncliticized perfective 3rd person masculine singular form. This is simple enough except for the phenomenon of partial paradigm homonymy (PPH). PPH occurs when two separate words share

[1] http://dictionary.sakhr.com/

[2] Arabic transliteration is in the Habash-Soudi-Buckwalter scheme (Habash et al. 2007).

the same lemma form as defined above. For instance, the lemma بَيْت *bay.t* can represent the word بُيُوت/بَيْت *bay.t/buyuwt* ‘house/houses’ and the word أَبْيَات/بَيْت *bay.t/Âab.yAt* ‘verse/verses’. Although the singular form of the words is homonymic, the plural is not. Habash and Rambow (2006) address this issue by introducing the concept of a Morphological Behavior Class (MBC), which is essential in distinguishing the two lemmas (“lexemes” in Habash and Rambow’s terminology). In this work, we do not address the issue of PPH directly and leave it to future work.

3.2 Morphological Analysis and Disambiguation

On average, an Arabic word form has 1.8 lemmas. If we exclude singleton analyses, the number of lemmas per word rises to 3. We address ambiguity in Arabic by using a morphological disambiguation system, MADA (Habash and Rambow, 2005) as part of preprocessing our resources.

The MADA approach distinguishes between the problems of *morphological analysis* (what are the different readings of a word out-of-context) and *morphological disambiguation* (what is the correct reading in a specific context). Once a word’s morphological analysis is determined in context, we can determine its full POS tag, lemma and diacritization.

Morphological analysis in MADA is done using the Buckwalter Arabic Morphological Analyzer (BAMA). [3] Each BAMA analysis contains a diacritization, a morpheme analysis, a lemma and a gloss. BAMA lexicons have a very good coverage, and most of out-of-lexicon cases tend to be proper nouns. The BAMA glosses are what we compare against in this study (see section 4).

Morphological disambiguation in MADA makes use of 19 orthogonal features to select, for each word, a proper analysis from the list of BAMA potential analyses. The BAMA analysis which matches the most of the predicted features wins. Since MADA selects a complete analysis from BAMA, all decisions regarding morphological ambiguity, lexical ambiguity, tokenization, diacritization and POS tagging are made in one fell swoop (Habash and Rambow, 2005; Roth et al, 2008). The choices not selected are ranked in terms of their likelihood. MADA has over 95% accuracy on basic morphological choice (including tokenization but excluding case/mood/nunation) and, most relevant to this paper, 96% accuracy also on lemma choice (Roth et al, 2008). We do not address MADA errors in this study.

4 Hand-Created Dictionary

The hand-created dictionary discussed in this section is derived from the databases of the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004). BAMA was developed manually over several years and it has been used for many NLP applications, e.g., Penn Arabic Tree Bank (Maamouri et al., 2004). We compare extensively to BAMA in the present study. The BAMA dictionary was created by associating the English glosses of the BAMA *stem* entries with their Arabic lemmas. Since the English gloss is associated with an Arabic stem, the English translation is not strictly an English lemma. For instance, English glosses of irregular (*broken*) plural stems of Arabic nouns are plural (not singular). See Table 1 for examples of these problems.

Lemma	**Stem**	**BAMA English Gloss**
tasAmuH_1	tasAmuH	tolerance
ðAt_1	ðAt	same; self; essence
	ðawAt	selves; beings
ðAt_2	ðAt	having; possessing
	ðawAt	those having; possessing
Ǎin.sAn_1	Ǎin.sAn	human being
	ÂunAs	human beings

Table 1. Examples of some of the BAMA lexicon stem entries and **raw** glosses for three lemmas: تَسامُح tasAmuH, ذَات ðAt and إِنْسَان Ǎin.sAn.

The presence of inflectional information and multi-word translations may be good for a dictionary to have; however, it is problematic when inconsistently used, especially for the goal of evaluating automatically generated lemma-based dictionaries. We address this issue by semi-automatically cleaning the dictionary. We verified the quality of the cleaning step manually for the data we use in development and testing in section 6. Results of cleaning the entries in Table 1 are

[3] Technically, MADA only uses the BAMA lexicons with a different interfacing engine, ALMORGEANA (Habash, 2007).

presented in Table 2. For other problematic BAMA issues, see (Attia, 2006).

Lemma	BAMA English Gloss
tasAmuH	tolerance
ðAt	same/self/essence/ have/posses
Ǎin.sAn	human_being

Table 2. Examples of some of the BAMA lexicon **cleaned** glosses for three lemmas: تَسامُح tasAmuH, ذَات ðAt and إِنْسَان Ǎin.sAn.

The underscores in BAMA lemma forms are indices often used to indicate sense distinctions associated with PPH and MBCs discussed earlier. Unfortunately, this is not done consistently in BAMA, e.g., the lemma *bay.t_3* representing the word *bay.t/buyuwt* has the glosses *house* and *houses*; however, the lemma *bay.t_4* representing *bay.t/Âb.yAt* has the correct plural gloss *verses*, but the inconsistently ambiguous singular gloss *house/verse*. MADA lemmatization produces indexed lemmas by default. In the present study, we use these indexed lemmas in building our corpus-based dictionary (next section); however we ignore them in evaluation.

Overall, the BAMA dictionary has over 36K lemmas (38.5K indexed lemmas) appearing in over 74K pairings with English lemmas.

5 Corpus-based Dictionary

The process of creating our corpus-based dictionary is divided into four stages: preprocessing, alignment, extraction and filtering. We use 4 million words of an Arabic-English sentence aligned corpus.[4]

5.1 Preprocessing

Arabic preprocessing includes separating punctuation from words and lemmatization using MADA (Habash and Rambow, 2005; Roth et al, 2008). We currently take the top-1 choice in MADA ranking but plan to consider the top n ranked lemmas in the future. English preprocessing includes separating punctuation from words, splitting off *'s*, lemmatization and down-casing. Although punctuation and function words are not a target of lexicon building, we keep them because we think they may help the alignment process. In the future, we will consider variants of this pipeline that exclude them for comparison purposes.

5.2 Word Alignment

The lemmatized parallel corpus is word-aligned using GIZA++ (Och and Ney, 2003).

5.3 Translation Extraction

Translation extraction is done using the Pharaoh system tool for phrase-table extraction (Koehn, 2004). We use all default settings except for *MaxPhraseLength*, which is set to 1 to disallow multi-word phrases. The resulting phrase table is basically our unfiltered bilingual lemma-based dictionary. In this phrase table, each Arabic lemma is aligned to one or more English translations. For example the lemma شَرَقِيّ *šar.qiy~*, glossed in BAMA as 'eastern/east/oriental,' aligns to 11 English lemmas: some good (e.g., *east, eastern, oriental, sharkeya, sharkia* and *sharqia*), some bad (*northeastern*) and some ugly (*belly, of, province* and the punctuation mark ","). Associated with each Arabic-English lemma pair are four probability scores: phrase translation probability φ(Arabic|English), lexical weighting lex(Arabic|English), phrase translation probability φ(English|Arabic), and lexical weighting lex(English|Arabic) (Koehn, 2004).

5.4 Lexicon Filtering

The alignment-based bilingual lexicon generated in the previous step contains many noisy pairings. We use Ripper (Cohen, 1996), a rule-based machine learning classifier, to learn noise-filtering rules. Although Ripper may not be as competitive as other machine learning systems, it is quite fast and produces human-readable rules that allow better understanding of the decisions made (Elming and Habash, 2007).

The features we use to train with Ripper are the four probabilities from the extraction step, the product of the translation probabilities (1st and 3rd

[4] The parallel text includes Arabic News, eTIRR, English translation of Arabic Treebank, and Ummah. All resources are distributed by the Linguistic Data Consortium (http://www.ldc.upenn.edu).

values) and whether the English translation is a stop word or not. Stop words are non-content words such as closed-class function words and punctuation. We do not include any other lexical features nor any word/lemma forms, as we want to learn generalizations.

We train Ripper using a held-out development set randomly extracted from our bilingual lexicon. Two sets of gold classification targets are created: a manual set (supervised) and an automatic set produced by matching against an existing dictionary (semi-supervised). Further details are in section 6. We experimented with various loss ratio settings for Ripper and determined empirically that 0.5 was optimal for maximal precision and recall. We discuss our results in the next section.

6 Evaluation

In this section we present our evaluation. The following three subsections describe the various data sets we use, our results and an error analysis of our best performing system.

6.1 Data description

Our test suite focuses on a randomly selected a set of 201 Arabic lemmas and their English lemma glosses. The Arabic lemmas are not BAMA indexed. They correspond to 262 indexed lemmas. We identify three types of data sets.

A. Basic Sets

PHR is an automatically generated dictionary through word alignment as described in section 5. It is essentially the phrase table for the 201 Arabic lemmas. PHR contains 1541 Arabic-English pairs (an average of 7.7 translations per lemma).

BAM is a dictionary derived from the manually created BAMA lexicons as described in section 4. Out of the 201 Arabic lemmas, 12 cases (~6%) are not analyzable by BAMA and as such receive no translations. All of these BAMA out-of-vocabulary cases are proper nouns such as *kytAjymA* 'Kitajima', *nywbwrt* 'Newport', and *swkArnwbwtry* 'Sukarnoputri'. This is a typical weakness in manually created dictionaries and is a strong advantage to automatically generated dictionaries. Out of the basic 515 Arabic-English pairs in BAM, 37 (~7%) are multi-word translations such as فَتْوَى fat.waý 'legal opinion.' Since we explicitly exclude multi-word expressions, we split all of these cases into multiple entries. The final BAM dictionary we use contains 549 pairs for 189 Arabic lemmas (an average of 2.9 translations per lemma).

B. Human-Filtered Sets

PHR#HUM is a human-filtered subset of PHR containing 676 pairs (~44% of PHR) and covering 170 lemmas at an average of 4 translations per lemma. In PHR, there are 31 lemmas that do not receive a single correct translation through alignment.

GLB is a *super-gold* dictionary constructed by taking the union of BAM and PHR#HUM. GLB contains 994 pairs (an average of 4.9 translations per lemma). Around 32% of GLB entries are exclusively from BAM; 45% are exclusively from PHR#HUM; and the rest (23%) is shared by both BAM and PHR#HUM.

C. Automatically Filtered Sets

We present three filtered sets. First is a simple baseline filter, **PHR#STP**, which exploits the observation that many erroneous English translations are stop-word-like, i.e., closed classes such as prepositions and determiners; and punctuation. PHR#STP contains only pairs whose English lemmas are not stop-words. PHR#STP has 1196 pairs covering 194 Arabic lemmas (an average of 6.2 translations per lemma).

We also use two Ripper filters trained on a held-out development set as described in section 5.4. The development data (**DEV**) contains a set of 201 Arabic lemmas different from PHR but is also extracted from the automatic word alignments. Two versions of DEV are used: **DEV#HUM**, which is filtered by a human; and **DEV#BAM**, which is filtered by BAM. The Ripper filters produce two automatically filtered dictionaries from PHR:

- **PHR#RIP$_{\text{HUM}}$** is generated with a Ripper filter trained on DEV#HUM. It contains 458 pairs covering 151 Arabic lemmas at 3 translations per lemma.

- **PHR#RIP$_{BAM}$** is generated with a Ripper filter trained on DEV#BAM. It contains 322 pairs covering 168 Arabic lemmas at 1.9 translations per lemma.

6.2 Results

We present two sets of results evaluating against PHR#HUM and against GLB. In all cases we present our results in terms of precision, recall and F-score (harmonic mean of precision and recall).

Using PHR#HUM as Gold

	Recall	**P**recision	**F**-score
PHR	100%	44%	61%
PHR#STP	99%	56%	71%
PHR#RIP$_{HUM}$	59%	88%	71%
PHR#RIP$_{BAM}$	42%	89%	57%

Table 3. Summary of Recall, Precision, and F-score against PHR#HUM.

Table 3 contains the results of evaluating against PHR#HUM. In the first two data rows, we have two baselines: a no-filter baseline (PHR) and a simple filter baseline (PHR#STP). The next two data rows present our Ripper-based filters PHR#RIP$_{HUM}$ and PHR#RIP$_{BAM}$. The PHR baseline degenerately gets perfect recall of course, while showing how hard the problem of filtering is through its low precision. The stop-word filter maintains a high recall while increasing precision by an absolute 12%, a nice result using a simple technique. Our two Ripper-based filters both improve precision (almost doubling it) at a tradeoff with recall. PHR#RIP$_{HUM}$ outperforms PHR#RIP$_{BAM}$ in recall but is close in terms of precision.

Using GLB as Gold

For the sake of completeness, we compare our filtered sets against GLB, as a more realistic general human-checked dictionary than either BAM or PHR#HUM alone. Table 4 presents the results of this comparison including other relative data sets to help contextualize the value of the new dictionaries.

	Recall	**P**recision	**F**-score
BAM	55%	100%	71%
PHR#HUM	68%	100%	81%
PHR	68%	44%	53%
PHR#STP	67%	56%	61%
PHR#RIP$_{HUM}$	40%	88%	55%
PHR#RIP$_{BAM}$	29%	89%	43%
BAM ∪ PHR	100%	53%	70%
BAM ∪ PHR#STP	99%	65%	79%
BAM ∪ PHR#RIP$_{HUM}$	80%	93%	86%
BAM ∪ PHR#RIP$_{BAM}$	69%	95%	80%

Table 4. Summary of Recall, Precision, and F-score against GLB.

Table 4 can be divided into three areas:

- The first pair of rows show the contribution of BAM and PHR#HUM to GLB for reference.
- The second and third pairs of rows are comparable to Table 3 in showing scores against two baselines and two Ripper-filtered sets. The results are also comparable with the main difference being the drop in recall resulting from BAM being part of GLB on top of PHR#HUM.
- The fourth and fifth pairs of rows are comparable to Table 3 also. They are different from the second and third pairs of rows in that they compare the sets unioned with BAM. The point of doing this is to provide a fair estimate of the quality of extending BAM with the automatically generated resources. Overall, adding the human-tuned Ripper-filtered set to BAM has the best combination of precision and recall (highest F-score). Interestingly, adding unfiltered PHR to BAM produces the inverse recall-precision ratios (same F-score) as using BAM as is.

6.3 Error analysis

We present an analysis of the errors in PHR#RIP$_{HUM}$ against PHR#HUM (presented in Table 3).

The 12% error in precision (false positives) comprises 56 pairs involving 39 lemmas. Of these lemmas, 13 (33%) do not have a single possible good translation in PHR. The erroneous pairs fall into four classes. First, over two-thirds (67%) of the errors involve incorrect lemma form that can be attributed to an error in MADA lemmatization. For example, the lemma ظَلَّام *Ďal~Am* (*tyrant*) is incorrectly paired with *darkness*, whose correct Arabic lemma is ظَلَام *ĎalAm*. Second, less than a quarter of the errors (23%) involve a word misalignment. For example, the lemma سَاعَة *sAςaħ* (*O'clock, hour*) is incorrectly paired with fifteen. Third, around 6% of the cases involve multi-word expressions that we do not currently handle. For example, the lemma شَرَقِيّ *šar.qiy~* (*eastern*) is incorrectly paired with *belly*. This is a result of mishandling the phrase رَقْص شَرَقِيّ *raq.S šar.qiy~* 'belly dancing (lit. eastern dancing)'. Finally, 4% of errors are caused by incorrect translations and typos in the parallel text. For example, the lemma مُسْتَوْرِد *mus.taw·rid* (*importer*) is incorrectly translated as *importuners*.

The 41% recall error (false negatives) comprises 325 pairs (involving 122 Arabic lemmas). Of these lemmas, 32 (26%) receive not a correct translation in PHR#RIP$_{HUM}$ although at least one is present in PHR. A sample analysis of these pairs shows that most of the errors can be attributed to low probability scores resulting from bad alignments.

7 Conclusion and Future Plans

In this paper, we present and evaluate a technique for the automatic creation of a lemma-based MRD for Arabic, a morphologically rich language. Our evaluation and error analysis of the precision and recall of our generated dictionary show that it has a high degree of complementarity to a widely-used hand-created dictionary. The precision of our dictionary is high, although its recall can be improved. The technique we describe is very fast and can be easily scaled up to more parallel data.

In the future, we plan to address measures to improve recall and precision, e.g. using improved word alignments and alternative machine learning approaches to filtering. We plan to investigate extensions to multi-word expressions and address lemma failures in MADA by using top-n lemmas. We also plan to use more parallel data and study better models of lemmas that address partial paradigm homonymy.

Acknowledgments

The first author was supported by a Fulbright scholarship through the Council for International Exchange of Scholars. The second author was partially supported by Defense Advanced Research Projects Agency Contract No. HR0011-08-C-0110. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. We would like to thank George Wilson, Mona Diab and Owen Rambow for helpful conversations.

References

Al-Sughaiyer, Imad A., and Ibrahim A. Al-Kharashi. 2004. Arabic Morphological Techniques: A Comprehensive Survey. *Journal of the American Society for Information Science and Technology*, 55(3):189-213.

Attia, Mohammed. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks. In *Proceedings of The Challenge of Arabic for NLP/MT Conference. The British Computer Society*, London, UK.

Black, William J., and Sabri El-Kateb. 2004. A Prototype English-Arabic Dictionary based on WordNet. In *Proceedings of 2nd Global WordNet Conference*, GWC2004, Czech Republic.

Buckwalter, Tim. 2004. Buckwalter Arabic Morphological Analyzer (version 2.0).

Cohen, William. 1996. Learning Trees and Rules with Set-valued Features. In *14th Conference of the American Association of Artificial Intelligence (AAAI)*.

Dichy, Joseph, and Ali Farghaly. 2007. Grammar-Lexis Relations in the Computational Morphology of Arabic. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Soudi, Abdelhadi, van den Bosch, Antal, and Neumann, Günter (Eds.).

Elkateb, Sabri, and William Black. 2001. Towards the Design of English-Arabic Terminological Knowledge Base. In *Proceedings of the Workshop on Arabic natural Language Processing, ACL* 2001, Toulouse, France.

Elming, Jakob, and Nizar Habash. 2007. Combination of Statistical Word Alignments Based on Multiple Preprocessing Schemes, In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL)*, Rochester, New York.

Foster, George, Roland Kuhn, and Howard Johnson. 2006. Phrase Table Smoothing for Statistical Machine Translation. In *Proceedings of the Conference on*

Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia.

Fung, Pascale. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In *Proceedings of the Association for Machine Translation in the Americas* (AMTA), Langhorne, PA.

Gamallo Otero, Pablo. 2005. Extraction of Translation Equivalents from Parallel Corpora Using Sense-Sensitive Contexts. In *Proceedings of the European Association for Machine Translation conference* (EAMT), Budapest.

Gamallo Otero, Pablo. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of MT Summit XI*, Copenhagen, Denmark.

Habash, Nizar, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods. Soudi, Abdelhadi; van den Bosch, Antal; Neumann, Günter (Eds.)*, Springer.

Habash, Nizar. 2007. Arabic Morphological Representations for Machine Translation. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods. Soudi, Abdelhadi, van den Bosch, Antal, and Neumann, Günter (Eds.)*, Springer.

Habash, Nizar, and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of Conference of the Association for Computational Linguistics* (ACL), Sydney, Australia.

Habash, Nizar, Bonnie Dorr, and Christof Monz. 2006. Challenges in Building an Arabic-English GHMT System with SMT Components. In *Proceedings of* AMTA, Cambridge, MA.

Habash, Nizar, and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging, and Morphological Disambiguation in One Fell Swoop. In *Proceedings of ACL*, Ann Arbor, MI.

Habash, Nizar. 2003. Matador: A Large Scale Spanish-English GHMT System. In *Proceedings of MT Summit IX*, New Orleans, LA.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of HLT-NAACL 2003*, Edmonton, Canada.

Koehn, Philipp. 2004. Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In *Proceedings of AMTA*, Washington, DC.

Maamouri, Mohamed, Ann Bies, and Tim Buckwalter. 2004. The Penn Arabic Treebank: Building a Large-scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Martínez, S. Felipe and Andy Way. 2009. Marker-based Filtering of Bilingual Phrase Pairs for SMT. In *Proceedings of EAMT*, Barcelona, Spain.

Melamed, Dan. 1997. A Word-to-word Model of Translational Equivalence. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics* (EACL), Madrid, Spain.

Munteanu, Dragos Stefan & and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-parallel Corpora. *Computational Linguistics* 31 (4), 477-504.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

Roth, Ryan, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of ACL*, Columbus, OH.

Tiedemann, Jorg. 1998. Extraction of Translation Equivalents from Parallel Corpora. In *the 11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark.

Vintar, Špela. 2001. Using Parallel Corpora for Translation-oriented Term Extraction. *Babel Journal*, 47(2), 121-132.

Zughoul, Muhammad R. and Awatef Miz'il Abu-Alshaar. 2005. English/Arabic/English Machine Translation: A Historical Perspective. *META*, 50(3), 1022–1041.