Design, Construction and Validation of an Arabic-English Conceptual Interlingua for Cross-lingual Information Retrieval

Nizar Habash[1], Clinton Mah[2], Sabiha Imran[2], Randy Calistri-Yeh[2] and Páraic Sheridan[2]

[1] Columbia University, Center for Computational Learning Systems, 475 Riverside Drive, New York, NY 10115
[2] TextWise LLC, An MNIS Company, 401 South Salina Street, Syracuse, NY 13202
habash@cs.columbia.edu, paraic@textwise.com

Abstract

This paper describes the issues involved in extending a trans-lingual lexicon, the TextWise Conceptual Interlingua (CI), with Arabic terms. The Conceptual Interlingua is based on the Princeton English WordNet (Fellbaum, 1998). It is a central component in the cross-lingual information retrieval (CLIR) system CINDOR (Conceptual INterlingua for DOcument Retrieval). Arabic has a rich morphological system combining templatic and affixational paradigms for both inflectional and derivational morphology. This rich morphology poses a major challenge to the design and building of the Arabic CI and also its validation. This is because the available resources for Arabic, whether manually constructed bilingual lexicons or lexicons automatically derived from bilingual parallel corpora, exist at different levels of morphological representation. We describe here the issues and decisions made in the design and construction of the Arabic-English CI using different types of manual and automatic resources. We also present the results of an extensive validation of the Arabic CI and briefly discuss the evaluation of its use for CLIR on the TREC Arabic Benchmark collection.

1. Introduction

Over several years of research and development effort, TextWise has designed and constructed an innovative technology for cross-lingual information retrieval across a variety of languages, called CINDOR. This product is built around an innovative multilingual framework known as the Conceptual Interlingua that supports a broad range of multilingual and cross-lingual information analysis applications.

In recent years, there has been increasing interest in the application of cross-lingual technologies to the Arabic language. During two years of evaluating Arabic cross-lingual search systems at the annual Text Retrieval Conference (TREC) hosted at the U.S. National Institute for Standards and Technology (NIST), a dozen different groups spanning both academia and industry evaluated and presented a wide range of different systems for cross-lingual English-Arabic search (Gey & Oard 2001; Oard & Gey 2002).

In contrast to many of these research systems however, TextWise has developed CINDOR to offer a scalable cross-language capability that works with several commercial search platforms commonly used in enterprise search environments.

This paper describes the issues involved in extending the CINDOR Conceptual Interlingua (CI) for cross-lingual Arabic search. We focus particularly on the issues presented by the rich morphology of the Arabic language, the new automated techniques we used to map terms to concepts within the CI, and issues around validation of Arabic terms in the final CI.

Section 2 presents a background description of the CI and its previous use. Section 3 gives a quick overview of Arabic morphology focusing on issues relevant to this work. Section 4 and 5 describe the design and construction of the Arabic CI respectively. Finally, the validation and evaluation of the Arabic CI are presented in Section 6.

2. The Conceptual Interlingua

The use of a Conceptual Interlingua (CI) for CLIR search has been thoroughly investigated and benchmarked for English, French, Spanish, Japanese and Chinese (Ruiz et al., 1999; Ruiz et al., 2000). Within the current CINDOR approach, the user's English query is first translated into its relevant CI concepts (effectively, WordNet synsets), which are then disambiguated automatically or with the help of the user who issued the original search query. The disambiguation happens in English without requiring foreign language knowledge from the user. After disambiguation, the query is translated through the chosen concepts and monolingual document retrieval then takes place in the target language using an underlying search platform. The CINDOR system is designed to work with several independent third-party search platforms commonly used in enterprise search environments.

The CI has been developed based on the Princeton English WordNet (Fellbaum, 1998). The use of this WordNet framework provides the additional advantage of potentially tapping into a broad network of groups who are independently compiling language resources around the common paradigm of the Princeton WordNet.

Other groups that have focused on multilingual versions of WordNet include the EuroWordNet

consortium[1], which has compiled versions in Dutch, Italian, Spanish, French, German, Czech and Estonian. These resources are fully compatible with the Conceptual Interlingua framework and provide a starting point for introducing CINDOR capabilities in these languages. The Global WordNet Association[2] also lists ongoing development of WordNet resources in Bulgarian, Danish, Greek, Hebrew, Hindi, Icelandic, Kannada, Latvian, Moldavian, Norwegian, Romanian, Russian, Slovenian, Swedish, Tamil, Thai, Turkish, and Yugoslavian. The CINDOR approach is always to leverage existing WordNet resources in different languages as far as possible to provide coverage quickly across many languages. In several cases however, including our current work in Arabic, TextWise has been asked to create Conceptual Interlingua resources for cross-lingual search in languages for which no other WordNet resources existed. While there is now a new Princeton effort to build an Arabic WordNet (Black et al., 2006), the work reported here pre-dates that initiative.

3. Arabic Morphology

A key challenge in the design and construction of an Arabic Conceptual Interlingua was the rich morphology of the Arabic language.

In discussing Arabic morphology, it is important to separate two different aspects: structure/form versus semantics/meaning. In terms of the form of its morphology, Arabic has both templatic and affixational morphemes. Templatic morphemes (roots and patterns) interdigitate to form a word stem. For example the word كاتب kAtib *'writer'* is constructed from the root كتب ktb *'writing-related'* and the pattern 1A2i3 *'active participant/doer'*.[3] Affixational morphemes (prefixes, suffixes and circumfixes) are concatenated to the word stem. For example the word والكاتبان wAlkAtibAn *'and the two writers'* is constructed from the prefix conjunction +و w+ *'and'*, the prefix definite article +ال Al+ *'the'* and the suffix dual marker +ان +An .

The root in Arabic morphology is a central concept which is usually thought of as a shared semantic component among the set of words derived from it. For example, the words كتب katab 'write', كاتب kAtab 'correspond', كتّب kat~ab 'dictate', كتاب kitAb 'book' and كاتب kAtib 'writer' all share the same root كتب ktb 'writing-related'. However, there are examples where the relationship to the root is rather *idiosyncratic*. For example, the word كتيبة katiybap 'brigade' is derived from the same root كتب ktb but seems less related to other derivations. Moreover, some roots have homonyms. For example, the root لحم lHm has at least two major distinct senses as 'flesh-related' and 'welding-related'. As a result, the word لحام laH~Am means both '*butcher'* and '*welder'*. These examples suggest that the root is too abstract and the process of derivation too idiosyncratic to be trusted as a representation in the Arabic CI.

The second aspect of Arabic morphology has to do with the distinction between derivational and inflectional morphology, which is similar to that in other languages. Derivational morphology is concerned with creating words from other words/stems/roots where the core meaning is modified. For example, the Arabic كاتب kAtib '*writer*' can be seen as derived from the root كتب ktb the same way the English *writer* can be seen as a derivation from *write*. The exactness of where a word is derived from and how its meaning came to be can be as elusive in Arabic as it is in any other language. More importantly in the context of this work, derivational morphology is at best idiosyncratic as was shown earlier. In inflectional morphology, the core meaning of the word is still intact and the extensions are predictable and compositional. For example, the semantic relationship between كاتب kAtib '*writer*' and كتاب kut~Ab *'writers'* maintains the sense of the kind of person described, but only varies the number.

The relationship between inflectional/derivational morphology and templatic/affixational morphemes is orthogonal (Habash, 2006). Some inflectional features are realized using pattern changes (templatic morphology). For example, the word كتاب kut~Ab *'writers'* is created using a pattern change from the singular form كاتب kAtib '*writer*'. This phenomenon of *broken plurals* is quite common in Arabic. Similarly, some affixational morphemes can be derivational. For example, the adjective كتبي kutubiy~ 'book-related' is derived from the stem كتب kutub 'books' (a broken plural of كتاب kitAb 'book') and the derivational suffix +ي +iy.

Given the variability in the relationship between morpheme types and inflectional/derivational morphology, it is necessary to define the *lexeme* as a set of forms differing only in inflectional morphology. The traditional citation form of a lexeme used in dictionaries is the perfective 3rd person masculine singular for verbs and the singular masculine form for nouns and adjectives. If there is no masculine form, the feminine singular is used. In all cases, any affixational inflections are removed. As such, the lexeme [كتاب] [kitAb] *'book'* normalizes over all the different inflectional forms of كتاب kitAb such as كتبنا kutubnA *'our books'* and الكتابان AlkitAbAn *'the two books'*.

4. Design of the Arabic CI

The structure of the Arabic CI follows that of the Princeton English WordNet. The design questions of the Arabic CI are focused on three issues. The first two relate to Arabic morphology: (1) defining an appropriate level of representation for Arabic entries that balances Arabic's rich morphology while maintaining the semantic wellformedness of WordNet entries; and (2) defining a proper interface between the entries in the CI and the generated forms in a query translation. The third is more general to building a trans-lingual resource: (3) addressing missing concepts in WordNet. Additional practical considerations are presented in the next section.

Level of Representation: On the first issue, we agree with the assessment of (Diab 2004) that lexemes are the proper level of representation. The lexeme is defined as a set of word forms sharing a common meaning and differing only in inflectional morphology. The number of inflectional features for an Arabic word can be rather large (Habash & Rambow 2005). The

[1] http://www.illc.uva.nl/EuroWordNet/

[2] http://www.globalwordnet.org/

[3] The numbered positions indicate the root radical slot.

lexeme citation form is typically the least inflected member of the set, e.g., for verbs, the perfective 3rd person masculine singular is used. The lexeme reflects the right level of representation in terms of avoiding derivational idiosyncrasies and abstracting over inflectional variations. The other possible candidate representations are the root and the stem (Dichy & Farghaly, 2003). The root is a very high level abstraction that is semantically rather coarse in granularity and full of idiosyncrasies. The examples in the previous section serve to highlight this well. The stem on the other hand is too shallow a level of representation that will miss a lot of inflectionally related terms. Additionally, there are many ways to cut up Arabic words using a stemmer all of which trade off longer stems that add unwanted distinctions and shorter stems that conflate unrelated words. Finally, the lexeme level of representation happens to be what is traditionally used in Arabic dictionaries to specify the term whose meaning is discussed. Although Arabic dictionaries do section terms based on common roots, the actual entries are not the roots themselves but rather the lexemes.

It is important to point out that even with the use of lexemes, homonyms will still be a problem as in any other language. For example, the lexeme [قاعدة] qAEidap (noun, proper noun) has at least three meanings: (a.) principle/rule [language, mathematics] (b.) military base and (c.) AlQaida organization. Some machine-readable dictionaries, including one we use in the next section, do use markers to distinguish these different lexemes.

Search System Interface: Second, the output of an English-to-Arabic query translation using the CI is a list of Arabic lexemes that are sent to an Arabic search system. Since CINDOR must work with a variety of search platforms and since different search systems might take different approaches to Arabic word stemming for indexing/retrieval purposes, the query translation process should be independent of any specific system. This requires the Arabic output of the query translation be in a natural inflected form so that the algorithm used by a specific Arabic search system to index the document terms can be used on the query terms to guarantee consistency. To that end, the translation process could be augmented with a generation step to produce inflected forms from the lexemes (Habash 2004). The generation process could be controlled to determine the degree of this query expansion. Though we currently do not use morphological generation as part of the complete CINDOR system, this is an avenue for potential future work.

Missing Synsets: The problem of concept mismatch between different pairs of languages is important to the design of a bilingual WordNet (Vossen *et al* 1997). In practice, we find there are many concepts that would occur in English search queries that are not found in WordNet (e.g. in WordNet v2.1, the sense of 'oil' as 'petroleum' is absent). These missing synsets have not been specifically addressed in a comprehensive fashion in our work, though the CINDOR search system does have a complementary 'Linguistic Toolkit' that provides for ongoing expansion of concepts and terms in the Conceptual Interlingua.

For a subset of the problem of missing synsets however, we do allow placing terms in a synset of a different part of speech if it captures some of the underlying semantics and we have evidence to support it. An Arabic-English example of cross POS translation is the word عمامة EamAmap *turban* which has a verbal form in Arabic تعمم taEam~am *'to wear a turban'*. In English there is no single verb to capture this meaning of the Arabic verb. An expansion of the English query '*turban*' into both nominal and verbal forms in Arabic might be needed. So we allow the Arabic verb to be part of the noun synset.

5. Construction of the Arabic CI

The construction of the Arabic CI can be broken into two tasks: first, the collection of Arabic-English translation pairs; and second, merging those pairs and using them to selectively populate the CI concepts (WordNet synsets) with Arabic terms. In this paper we discuss in detail the first task only. The second task is accomplished using a variety of language-independent techniques and additional proprietary techniques similar to previously published work (Diab 2004; Ruiz et al. 2000).

The first task of collecting translation pairs is complicated by the need to combine entries from a variety of resources with their own types of representation into a CI at the lexeme level of representation. The use of different resources is intended to increase coverage and robustness of the overall system. We make a distinction between manually created machine-readable dictionaries and automatic dictionaries constructed from parallel corpus data. Since existing dictionaries were created to serve different purposes (e.g. as part of a morphological analyzer), they may have made different decisions on what constitutes a "stem".

Parallel data is preprocessed and tokenized before being aligned automatically. The translation pairs created from parallel text are further processed (both sides, English and Arabic) to infer the underlying lexemes. Translation pairs from parallel data are noisier than those from dictionaries, but they have the potential to provide domain-specific vocabulary that is not available in existing dictionaries. In the rest of this section we describe the specific resources we used for the construction of the Arabic CI and present the statistics on the final Arabic CI constructed.

5.1. Raw Arabic Resources

The Arabic CI was constructed from six different resources, each with its unique challenges for incorporating into a unified framework. The contributions of each of these resources are detailed in Section 5.2. The following are the different resources:

1. **The Buckwalter Lexicon (BUCK):** This is the lexicon for the Buckwalter Morphological Analyzer (Buckwalter, 2002). Though not intended for use as a stand-alone dictionary, it does include English glosses and provides Arabic lexemes (lemmas). It was the most valuable single resource for construction of the Arabic CI.
2. **The NMSU Arabic-English Lexicon (NMSU):** This resource was created as part of an Arabic

morphological analyzer (http://crl.nmsu.edu/). Its Arabic entries are not lexemes as found in regular dictionaries, but rather stems that we filtered and in some cases modified (by morphological category) to reconstruct a lexeme form. Its overall coverage of Arabic is weak, but it did make some useful contributions to the Arabic CI.

3. **Tufts Dictionary (TUFTS):** This is an XML encoding of the 19th Century Salmoné Arabic Learner's Dictionary. The vocabulary is somewhat outdated, but can still be helpful for cultural or religious terms.
4. **The UN Parallel Corpus (UN):** This consists of official UN documents in Arabic and English with alignment of sentences done by the Linguistic Data Consortium (LDC). We used 250,000 sentence pairs. The Arabic side was tokenized and part-of-speech (POS) tagged using ASVMT (Diab et al. 2004). The English side was POS tagged and lemmatized using the Connexor Machinese Phrase Tagging software (www.connexor.com). We used the automatic word alignment system GIZA++ (Och & Ney 2003) to derive a translation lexicon from alignments using POS tags and translation probabilities to constrain the choices. This provided an additional 8,000 unique Arabic terms to be assigned to the CI.
5. **The USBGN Database of Official Geographic Names (USBGN):** The online GNS system (http://geonames.usgs.gov/) is a potentially rich resource for proper nouns in an Arabic CI. The current Arabic CI includes only the place names for Iraq as a demonstration of how such terms should be handled.
6. **Manual translations (NEW):** We identified 2,721 English terms that occurred frequently in a domain-relevant corpus that was being used to guide coverage objectives of the Arabic CI. These English terms were present in WordNet 2.0, but had no Arabic translations from the five resources listed above. We therefore manually translated these English terms, producing a total of 4,160 new Arabic terms. This manual translation further served as a benchmark for comparison of time/cost of this approach versus the automated processing used on the other resources.

5.2. Resulting Arabic CI

The final overall statistics for the Arabic CI derived from the 6 resources outlined above are as follows:

	Distinct Terms	Concepts	Term-Concept Assignments
Arabic	67,499	41,328	126,584
English	63,575	41,328	86,628

The Arabic terms in the CI broken out by POS are:

POS	N	PN	V	AJ	AV
	34947	13192	12400	12166	908
	51.8%	19.6%	18.4%	18.0%	1.3%

The percentages across a row will not add up to 100% because an Arabic term can be more than one part of speech.

The 41,328 concepts in the final CI by POS are:

POS	N	PN	V	AJ	AV
	17797	13006	5059	4947	519

The following chart shows the number of concepts, Arabic terms, and English terms that each resource contributes, and also counts how many are uniquely contributed by that resource. Some Arabic terms are replicated as they may appear in different orthographic variants. This makes the Arabic unique term count appear larger than it should be. Nevertheless, the resulting numbers provide a useful insight about the relative contributions of linguistic resources for Arabic.

	Concepts		Arabic Terms		English Terms	
	Resource	Unique	Resource	Unique	Resource	Unique
BUCK	23683	5519	37386	5114	37022	6273
NMSU	17870	1183	32879	9751	29646	1183
TUFTS	14681	4214	22525	6047	24680	4399
UN	6566	1736	7781	1309	12214	1900
USBGN	8288	8185	8352	7816	16456	16174

6. Evaluation of the Arabic CI

Independent assessment of the Arabic CI was done external to the project team using two professional Arabic-English translators who are native Arabic speakers (specifically they were Egyptian). A detailed instruction manual in addition to a web-based interface for validation was provided to the two judges.

We asked the judges to mark each Arabic term assigned to a synset as one of four classes: 'accepted', (accepted-but-)'inflected', 'rejected' or 'unsure'. The instructions to the assessors varied from semantic-level instructions that had to do with understanding the correct meaning of synsets and spotting homonyms, to how to construct Arabic validation examples from English WordNet examples. We also provided guidelines on identifying inflectional variants that, although semantically correct, only managed to get in the CI as a result of using heterogeneous resources. The judges were instructed not to count crossing POS as rejections if such cross-POS assigned terms were semantically compatible with the synset they appear in. Finally, due to time limitations, the judges were asked to only add missing terms if they felt strongly about a term that did not appear in a synset but should have.

The external assessment of the Arabic CI was designed to be a strict assessment of *translation* quality independent of the application of the Arabic CI for cross-lingual *retrieval*. While our interest in building an Arabic CI is primarily for its use in our cross-lingual search system, we wanted to first assess the CI in the strictest manner in order to highlight systematic issues that could be addressed in our process or concept-mapping algorithms.

The validation of the CI entries generated from Buckwalter, NMSU, Tufts and UN (collectively, DICTS) resources was done together using one sample with a small overlap to compute inter-judge agreement. The CI entries from USBGN and NEW resources were judged separately. Different sample sizes were used for these latter assessments. As a result we calculate CI correctness separately for these resources. The sample sizes were as follows:

Resource	Sample	Overlap
DICTS	10,281	Yes
USBGN	669	No
NEW	4,160	No

6.1. Inter-Judge Agreement

Exact inter-judge agreement (the ratio of terms receiving the same judgment from both judges) was 84.82%. We also computed a soft inter-judge agreement that conflated inflected and accepted judgments and allowed unsure judgments to match anything. The soft inter-judge agreement was a high 97.20%. We additionally performed an analysis of a sample of the judgments with the help of an independent Arabic linguist. The linguist examined the judgments on 5% of the DICTS synsets, and on 4% of the NEW manual translation synsets. The linguist agreed with the external judges 91.1% of the time on DICTS terms, and 91.2% of the time on NEW terms. Disagreement was mostly over inflected/accepted choices for multi-word terms, incorrect rejection of valid English transliterations into Arabic, and whether an "unsure" judgment was warranted. We did not conduct a formal analysis of USBGN judgments, but an informal inspection of 100 terms from the USBGN dataset suggested 99% agreement between the Arabic linguist and the external judges. The higher agreement is to be expected since validating the USBGN terms was less complex than the other resources.

6.2. CI Correctness for CLIR

In measuring the "correctness" or "quality" of the sampled CI, we will use the Accept/Reject Ratio (ARR). This is computed as:

ARR = (Accepted + Inflected) / (Accepted + Inflected + Rejected)

The ARR combines accepted and inflected judgments since both are allowed in the CI for cross-lingual search purposes. It ignores unsure judgments. It also ignores manually added terms.

Our initial strict assessment of CI quality by external judges indicated a high level of rejection of CI entries generated automatically from dictionary resources (DICTS) at 44%. To better understand the terms that the judges rejected, we performed an in-depth analysis of 130 rejected nouns and adverbs from the DICTS dataset. We found that 54% of the rejected terms were properly rejected. Some of these rejections are due to incorrect transliteration or difficulties parsing dictionary definitions. But many are just intrinsic to building a large-coverage CI.

One third of the initial rejection cases (or 18% of the sample) were traced to mapping problems in the automated algorithms that map translation pairs into concepts (synsets). These problems, once identified, were then fixed. One example of such a problem is the class of bad assignments due to noisy word alignments from the UN Parallel Corpus. For instance, the term أوزبكستان >wzbkstAn 'Uzbekistan' was incorrectly aligned to the word '*consecutively'* in one parallel text, and as a result became a member of the synset *(in a consecutive manner; "he was consecutively ill, then well, then ill again")*. This problem was fixed by raising the threshold on the translation probability for terms pairs to be used.

In other cases, our second review indicated there was strong evidence that the term should have been accepted. For example, the less commonly used spelling of the Arabic term تبت tibit was rejected as a synonym for 'Tibet' even though both Buckwalter and the UN Parallel Corpus show this usage. Overall, 24% of the terms rejected by the judges were found to be properly rejected in the strict definition of our external assessment but are still considered acceptable synset assignments for purposes of cross-lingual retrieval. For example, the Arabic term هَسيس hasiys *(Buckwalter: whisper)* was rejected from the synset '*Speaking' (the utterance of intelligible speech)*, but would still be useful for retrieving relevant documents.

Since the goal of our analysis was to properly assess the quality of the Arabic CI for its use in the CINDOR cross-lingual search system, we present our results in terms of 'Retrieval ARR' which accounts for the corrections to mappings after our initial external validation and accepts term mappings which are judged acceptable for a search application though may not be correct for a general translation situation.

We assess the overall quality of the Arabic CI by taking the weighted average of the ARR values from each of the three classes of input resources, in proportion to the relative size of each resource in the full CI. The results of our evaluation are presented below, with an overall correctness assessment for retrieval of 82%.

Resource	Retrieval ARR
DICTS	79%
NEW	83%
USBGN	99%
Total	**82%**

6.3. Cross-Lingual Retrieval Effectiveness

As outlined in Sections 4 and 5 above, the Arabic CI was designed and constructed with the application of Cross-Lingual Retrieval within the CINDOR system clearly in mind. In evaluating retrieval effectiveness we have compared the relative performance of our English-Arabic cross-language CINDOR search system to an equivalent monolingual search using Arabic queries[4]. This allows us to compare the effectiveness of an English-speaking user running cross-lingual searches to

[4] The terms governing our use of commercial search systems limit our ability to publish evaluation results.

that of a native Arabic speaker running Arabic queries directly against the underlying search engine.

We conducted two formal tests using 25 test queries and 383,872 Arabic documents from the TREC-11 test set (Oard & Gey 2002). In the first test, we established a monolingual (Arabic-to-Arabic) baseline using the Arabic versions of the TREC queries provided by NIST. These queries were created for the TREC conference by native Arabic speakers, and the results represent how well the underlying search engine handles Arabic monolingual retrieval. We then benchmarked our CINDOR Arabic cross-lingual system against the same collection of 383,872 Arabic documents, but instead used the English versions of the 25 TREC-11 queries. Native English speakers, who knew no Arabic at all, used the CINDOR system to retrieve Arabic documents, including an interactive step to select specific concepts (meanings) for each query term.

The search application for which the Arabic CI is being designed emphasizes recall (finding all relevant documents) over precision (finding only relevant documents). The results of our evaluation showed CINDOR achieved cross-lingual precision in the top-20 documents at 66% of that achieved in the monolingual Arabic case, while the cross-lingual searches through CINDOR returned 18% more relevant document (increased recall) over the monolingual system. This is consistent with the concept-driven CLIR approach used in CINDOR, which expands queries with synonyms from the CI.

7. Conclusions

The work presented here is the first attempt to building a large-scale Arabic-English resource modeled around WordNet for CLIR using a variety of heterogeneous resources. Many of the design, construction and validation issues are relevant to other languages, such as Semitic languages and languages with rich morphologies and to the general process of building or extending WordNets in foreign languages. In particular, much of the work of extracting translation correspondences from heterogeneous resources and mapping them into the WordNet framework has been automated through a comprehensive set of processes and tools resulting from this work.

8. Acknowledgments

We thank Owen Rambow and Mona Diab for many helpful discussions, and we would like to thank the anonymous reviewers for their helpful comments.

This material is based upon work funded in whole or in part by the U.S. Government and any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

9. References

Aljlayl, M. and O. Frieder. (2002). On Arabic search: Improving the retrieval effectiveness via a light stemming approach. In *Proceedings of ACM Conference on Information and Knowledge Management*.

Al-Sughaiyer, I. and I. Al-Kharashi. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Black, W., S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. (2006). Introducing the Arabic WordNet Project. In *Proceedings of the third International WordNet Conference (GWC-06)*.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. *Linguistic Data Consortium*, University of Pennsylvania. Catalogue No.: LDC2002L49.

Darwish, K. (2002). Building a Shallow Morphological Analyzer in One Day. In *Proceedings of the Association for Computational Linguistics (ACL) workshop on Computational Approaches to Semitic Languages*.

Diab, Mona. (2004) The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. In *Proceedings of the Arabic Language Technologies and Resources Conference,* NEMLAR.

Diab, M., K. Hacioglu, and D. Jurafsky. (2004). Automatic tagging of arabic Arabic text: From raw text to base phrase chunks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Dichy, J. and A. Farghaly. (2003). Roots & Patterns vs. Stems plus Grammar-Lexis Specifications. In *Proceedings of MT Summit IX Workshop on MT for Semitic Languages*.

Fellbaum, Christiane. (1998) WordNet: An Electronic Lexical Database. MIT Press.

Gey, F. and D. Oard, D.., (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries., In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*.

Habash, N. (2004). Large Scale Lexeme Based Arabic Morphological Generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*.

Habash, N. (2006). "Issues in Arabic Morphological Analysis for Machine Translation." Book Chapter. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Editors Antal van den Bosch and Abdelhadi Soudi.

Habash, N. and O. Rambow (2005). Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In *Proceedings of ACL'05*.

Oard, D. and Gey, F. (2002). The TREC 2002 Arabic/English CLIR Track. In *Proceedings of TREC-11*.

Och, Franz Josef. J. Ochand, Hermann H. Ney. (2003). "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistics*, volume 29, number 1, pp. 19-51.

Ruiz, M., A. Diekema, and P. Sheridan. (1999). CINDOR conceptual interlingua document retrieval: TREC-8 Evaluation. In *Proceedings of TREC-8*.

Ruiz, M., S. Rowe, M. Forrester and P. Sheridan. (2000). CINDOR TREC-9 English-Chinese Evaluation. In *Proceedings of TREC-9*.

Soudi, A., V. Cavalli-Sforza, and A. Jamari. (2001). A Computational Lexeme-Based Treatment of Arabic Morphology. In *Proceedings of the Arabic Natural Language Processing Workshop, (ACL'01)*.

Vossen, P., Diez-Orzas, P., and Peters, W. (1997)., The Multilingual Design of EuroWordNet. In *Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, *(ACl'97)*.