Preliminary Lexical Framework for English-Arabic Semantic Resource Construction

Anne R. Diekema
Center for Natural Language Processing
4-206 Center for Science & Technology
Syracuse, NY, 13210 USA
diekemar@syr.edu

Abstract

This paper describes preliminary work concerning the creation of a Framework to aid in lexical semantic resource construction. The Framework consists of 9 stages during which various lexical resources are collected, studied, and combined into a single combinatory lexical resource. To evaluate the general Framework it was applied to a small set of English and Arabic resources, automatically combining them into a single lexical knowledge base that can be used for query translation and disambiguation in Cross-Language Information Retrieval.

1 Introduction

Cross-Language Information Retrieval (CLIR) systems facilitate matching between queries and documents that do not necessarily share the same language. To accomplish this matching between distinct vocabularies, a translation step is required. The preferred method is to translate the query language into the document language by using machine translation, or lexicon lookup. While machine translation may work reasonably well on full sentences, queries tend to be short lists of keywords, and are often more suited for lexical lookup (Oard and Diekema, 1998).

This paper describes a preliminary framework for the creation of a lexical resource through the combination of other lexical resources. The preliminary Framework will be applied to create a translation lexicon for use in an English-Arabic CLIR system. The resulting lexicon will be used to translate English queries into (unvocalized) Arabic. It will also provide the user of the system with lexical semantic information about each of the possible translations to aid with disambiguation of the Arabic query. While the combination of lexical resources is nothing new, establishing a sound methodology for resource combination, as presented in this paper on English-Arabic semantic resource construction, is an important contribution. Once the Framework has been evaluated for English-Arabic resource construction, it can be extended to additional languages and resource types.

2 Related Work

2.1 Arabic-English dictionary combination

As pointed out previously, translation plays an important role in CLIR. Most of the CLIR systems participating in the (Arabic) Cross-Language Information Retrieval track[1] at the Text REtrieval Conference (TREC)[2] used a query translation dictionary-based approach where each source query term was looked up in the translation resource and replaced by all or a subset of the available translations to create the target query (Larkey, Ballesteros, and Connell, 2002), (Gey and Oard, 2001), (Oard and Gey, 2002). The four main sources of translation knowledge that have been applied to CLIR are ontologies, bilingual dictionaries, machine translation lexicons, and corpora.

Research shows that combining translation resources increases CLIR performance (Larkey et al., 2002) Not only does this combination increase translation coverage, it also refines translation probability calculations. Chen and Gey used a combination of dictionaries for query translation and compared retrieval performance of this dictionary combination with machine translation (Chen and Gey, 2001). The dictionaries outperformed MT. Small bilingual dictionaries were created by Larkey and Connell (2001) for place names and also inverted an Arabic-English dictionary to English-Arabic. They found that using dictionaries that have multiple senses,

[1] There have been two large scale Arabic information retrieval evaluations as part of TREC. These Arabic tracks took place in 2001, and 2002 and had approximately 10 participating teams each.

[2] http://trec.nist.gov

though not always correct, outperform bilingual term lists with only one translation alternative. Combining dictionaries is especially important when working with ambiguous languages such as Arabic.

Many TREC teams used translation probabilities to deal with translation ambiguity and term weighting issues, especially since a translation lexicon with probabilities was provided as a standard resource. However, most teams combined translation probabilities from different sources and achieved better retrieval results that way (Xu, Fraser, and Weischedel, 2002), (Chowdhury et al., 2002), (Darwish and Oard, 2002). Darwish and Oard (2002) posit that since there is no such thing as a complete translation resource one should always use a combination of resources and that translation probabilities will be more accurate if one uses more resources.

2.2 Resource combination methodologies

Ruiz (2000) uses the term *lexical triangulation* to describe the process of mapping a bilingual English-Chinese lexicon into an existing WordNet-based Conceptual Interlingua by using translation evidence from multiple sources. Recall that WordNet synsets are formed by groups of terms with similar meaning (Miller, 1990). By translating each of the synonyms into Chinese, Ruiz created a frequency-ranked list of translations, and assumed that the most frequent translations were most likely to be correct. By establishing certain translation evidence thresholds, mappings of varying reliability were created. This method was later augmented with additional translation evidence from a Chinese-English parallel corpus.

A methodology to improve query translation is described by Chen (2003). The methodology is intended to improve translation through the use of NLP techniques and the combining of the document collection, available translation resources, and transliteration techniques. A basic mapping was created between the Chinese terms from the collection and the English terms in WordNet by using a simple Chinese-English lexicon. Missing terms such as Named Entities were added through the process of transliteration. By customizing the translation resources to the document collection Chen showed an improvement in retrieval performance.

3 Establishing a Preliminary Framework

The preliminary Framework provides a methodology for the automatic combination of various lexical semantic resources such as machine readable dictionaries, ontologies, encyclopedias, and machine translation lexicons. While these individual resources are all valuable individually, automatic intelligent lexical combination into one single lexical knowledge base will provide an enhancement that is larger than the sum of its parts. The resulting resource will provide better coverage, more reliable translation probability information, and additional information leveraged through the process of lexical triangulation. In an initial evaluation of the preliminary Framework, it was applied to the combination of English and Arabic lexical resources as described in section 4.

The preliminary Framework consists of 9 stages:

1) establish goals
2) collect resources
3) create resource feature matrix
4) develop evidence combination strategies and thresholds
5) construct combinatory lexical resource
6) manage problems that arise during creation
7) evaluate combinatory lexical resource
8) implement possible improvements
9) create final version of combinatory lexical resource.

Stage 1: The first stage of the Framework is intended to establish the possible usage of the combinatory lexical resource (resulting form the combination of multiple resources). The requirements of this resource will drive the second stage: resource collection.

Stage 2: Two types of resources should be collected: language processing resources such as stemmers and tokenizers; and lexical semantic resources such as dictionaries and lexicons. While not every resource may seem particularly useful at first, different resources can aid in mapping other resources together. During the second stage, conversion into a single encoding (such as UTF-8) will also take place.

Stage 3: Once a set of resources has been collected, the resource feature matrix can be created. This matrix provides an overview of the types of information found in the collected resources and of certain resource characteristics. For example, it is important to note what base form the dictionary entries have. Some dictionaries use the singular form (for nouns) or indefinite form (for verbs), some use roots, others use stems, and free resources from the web often use a combination of all of the above. By studying the feature matrix the evidence combination strategies for stage four can be developed.

	Arabic	English	word	stem	root	vocalized	unvocalized	pos	English definition	Arabic definition	synonyms	sense information
Arabeyes	x	x	x				x					
Ajeeb	x	x	x			x		x		x		x
Buckwalter	x	x		x		x	x	x	x			x
Gigaword	x		x				x					
WordNet 2.0		x						x	x		x	x

Table 1: Resource feature matrix

Stage 4: An intelligent resource combination strategy should be informed by the features of the different resources. It may be, for example, that one resource uses vocalized Arabic only and that another resource uses both vocalized and unvocalized Arabic. This fact should be taken into account by the combination strategy since the second resource can serve as an intermediary to map the first resource. Thresholding decisions are also part of stage four because the certainty of some combinations will be higher than others.

Stage 5: Stage five involves writing programs based on the findings in stage four that will automatically create the combinatory lexical resource. The combination programs should provide output concerning problematic instances that occur during the creation i.e. words that only occur in a single resource, so that these problems may be handled by alternative strategies in stage six.

Stage 6: Most of the problems in stage six are likely to be uncommon words, such as named entities or transliteration. A transliteration step, where for example English letters, i.e. *r*, are mapped to the closest Arabic sounding letters, i.e. ر , may be applied for languages that do not share the same orthographies.

Stage 7: After the initial combinatory lexical resource has been created it needs to be evaluated. First the accuracy (quality) of the combination mappings of the various resources needs to be assessed in an intrinsic evaluation. After it has been established that the combination has been successful, an extrinsic evaluation can be carried out. In this evaluation the combinatory lexical resource is tested as part of the actual application the source was intended for, i.e. CLIR. (For a more detailed description of evaluation see Section 5 below.)

Stage 8: These two evaluations will inform stage eight where possible improvements are added to the combination process.

Stage 9: The final version of the combinatory lexical resource can be created in stage nine.

4 Application of the Framework to English-Arabic

The preliminary Framework as described in section 3 was applied to five English and Arabic language resources as a kind of feasibility test. Following the Framework, we first established the goals of the combinatory lexical resource. It was determined that the resource would be used as a translation resource for CLIR that would aid query translation as well as manual translation disambiguation by the user. This meant that the combinatory lexical resource would need translation probabilities as well as English definitions for Arabic translations to enable an English language user to select the correct Arabic translation. We collected five different resources: WordNet 2.0[3], the lexicon included with the Buckwalter Stemmer[4], translations mined from Ajeeb[5], the wordlist from the Arabeyes project[6], and the LDC Arabic Gigaword corpus[7]. After the resources were collected the feature matrix was developed (see Table 1).

[3] http://www.cogsci.princeton.edu/~wn

[4] http://www.qamus.org

[5] http://english.ajeeb.com

[6] http://www.arabeyes.org

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T12

The established combinatory lexical resource goals and resource feature matrix were used to determine the combination strategy. Since the resource should provide the user with definitions of Arabic words and WordNet is most comprehensive in this regard, it was selected as our base resource. The AFP newswire collection from the Gigaword corpus was used to mine Ajeeb. As is evident in the matrix, all resources contain English terms as a common denominator. The information used for evidence combination was as follows. Evidence used for mapping the Ajeeb and Buckwalter lexicons is part-of-speech information. Additionally, these two resources also provide vocalized Arabic terms/stems that can be used for a more reliable (less ambiguous) match. The Arabeyes lexicon is not terribly rich but was used as additional evidence for a certain translation through frequency weighting. The combinatory lexical resource was constructed by mapping the three lexical resources into WordNet using the evidence as discussed above (see Table 2).

world, human race, humanity, humankind, human beings, humans, mankind, man, all of the inhabitants of the earth
all of the inhabitants of the earth
عالم بشر ناس كون أرض دنيا ورى أناس آنام أنام أُناس انام برية آناس ملكوت اناس رهط المعمورة اِمْرَأ بَشَرِيّ الكائنات أنام أوادِم اِمْرُؤ بشري البشر أَنَس إِنْسانِيّ إِنْس اِمْرِئ انس مَرْء راجِل مرؤ مَرْأ مَرْؤ عالَم الإنسانية مَرْئ انساني راجل إنساني امرئ مرأ الرجل مرء عِباد وَرا البشرية إمرؤ امرؤ العالم مرئ امرأ ورا رَجُل اوادم إمرئ رجل ناسُوت أنس عباد أوادم بَشَر إنس ناسوت إمرأ دُنْيا وَرَى

Table 2: Combinatory lexical resource entry example resulting from Step 5

After examining the combinatory lexical resource we found that the Arabeyes Arabic terms could not be compared directly to the Arabic terms in the other lexical resources since the determiner prefixes are still attached to the terms (as in العالم for example). More problematic were the translations mined from Ajeeb since the part-of-speech information of the Arabic term did not necessarily match the part-of-speech of the translations:

```
   حرس#VB#2.1.2#خَفَر#حَرَسَ
#do_sentry_duty,keep_watch_over,
guard,watchdog,oversee,sentinel,
shield,watch,ward
```

The first problem is easily fixed by applying a light stemmer to the dictionary. At this point it is not clear however, how to fix the second problem. It was also decided that the translation reliability weighting by frequency is too limited to be useful. A back-translation lookup needs to determine how many other terms can result in a certain translation. This data can then update the reliability score.

5 Comprehensive Evaluation

While we only have carried out a preliminary evaluation, we envision a comprehensive evaluation in the near future. As part of this evaluation three different types of evaluation can be carried out:

1) evaluate the process of applying the Framework;
2) evaluate the combinatory lexical resource itself; and
3) evaluate the contribution of the combinatory lexical resource to the application the resource was created for.

Evaluation of the process of applying the Framework will provide evidence as to the advantages and disadvantages of our Framework, and where it may have to be adjusted.

The construction of a Combinatory Lexical Resource by applying the Framework is the first step toward an effective evaluation of the full Framework. The construction process detailed in Section 3 should be carefully documented. The evaluation will focus on the time and effort spent on the process, difficulties or ease with resources that are acquired, managed and processed, as well as problems or issues that arise during the process.

The intrinsic evaluation of the combinatory lexical resource indicates the quality of the newly created combinatory lexical resource. For this evaluation a large random number of entries will need to be evaluated for correctness. The evaluation will provide accuracy and coverage measures for the resource. Also, descriptive statistics will be generated to provide general understanding of the lexical resource that has been produced.

The extrinsic evaluation of the combinatory lexical resource is intended to measure the contribution of the resource to an application (i.e. CLIR, Information Extraction). The application of choice should be run with the combinatory lexical resource, and without. Performance metrics appropriate for the type of application can be collected for both experiments and then compared.

6 Conclusion and future research

A general Framework for lexical resource construction was presented in the context of English-Arabic semantic resource combination. The initial evaluation of the Framework looks promising in that it was successfully applied to combine five English-Arabic resources. The stages of the Framework provided a useful guideline for lexical resource combination and can be applied to resources in any language. We plan to extend the evaluation of the Framework to a more in depth intrinsic evaluation where the quality of the mappings is tested. An extrinsic evaluation should also take place to evaluate the combinatory lexical resource as part of the CLIR system. As for future research we hope to extend the evidence combination algorithms to include more sophisticated information using back translation and transliteration.

7 Acknowledgements

This work is supported by the U.S. Department of Justice.

References

A. Chen, and F. Gey. 2001. *Translation Term Weighting and Combining Translation Resources in Cross-Language Retrieval*. In “Proceedings of the Tenth Text REtrieval Conference (TREC-10)”, E.M. Voorhees and D.K. Harman ed., pages 529-533, NIST, Gaithersburg, MD.

J. Chen. 2003. The Construction, Use, and Evaluation of a Lexical Knowledge Base for English-Chinese Cross-Language Information Retrieval. Dissertation. School of Information Studies, Syracuse University.

A. Chowdhury, M. Aljalayl, E. Jensen, S. Beitzel, D. Grossman, O. Frieder. 2002. *IIT at TREC-2002: Linear Combinations Based on Document Structure and Varied Stemming for Arabic Retrieval*. In “Proceedings of the Eleventh Text REtrieval Conference (TREC-11)”, E.M. Voorhees and C.P. Buckland ed., pages 299-310, NIST, Gaithersburg, MD.

K. Darwish and D.W. Oard. 2002. *CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval*. In “Proceedings of the Eleventh Text REtrieval Conference (TREC-11)”, E.M. Voorhees and C.P. Buckland ed., pages 703-710, NIST, Gaithersburg, MD.

F.C. Gey, and Oard, D.W. 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)”, E.M. Voorhees and D.K. Harman ed., pages 16-25, NIST, Gaithersburg, MD.

L.S. Larkey, J. Allan, M.E. Connell, A. Bolivar, and C. Wade. 2002. *UMass at TREC 2002: Cross Language and Novelty Tracks*. In “Proceedings of the Eleventh Text REtrieval Conference (TREC-11)”, E.M. Voorhees and C.P. Buckland ed., pages 721-732, NIST, Gaithersburg, MD.

L.S. Larkey, L. Ballesteros, M. Connell. 2002. *Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis*. In “Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval”, M. Beaulieu et al. ed., pages 275-282, ACM, NY, NY.

L.S. Larkey, and M. E. Connell. 2001. *Arabic Information Retrieval at UMass in TREC-10*. In “Proceedings of the Tenth Text REtrieval Conference (TREC-10)”, E.M. Voorhees and D.K. Harman ed., pages 562-570, NIST, Gaithersburg, MD.

G. Miller. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), Special Issue.

D. Oard and A. Diekema. 1998. Cross-Language Information Retrieval. *Annual Review of Information Science*, 33: 223-256.

D.W. Oard, and Gey, F.C.2002. *The TREC-2002 Arabic/English CLIR Track*. In “Proceedings of the Eleventh Text REtrieval Conference (TREC-11)”, E.M. Voorhees and C.P. Buckland ed., pages 17-26, NIST, Gaithersburg, MD.

M.E. Ruiz, et al. 2001. *CINDOR TREC-9 English-Chinese Evaluation*. In “Proceedings of the 9th Text REtrieval Conference (TREC-9)”, E.M. Voorhees and D.K. Harman ed., pages 379-388, NIST, Gaithersburg, MD.

J. Xu, A. Fraser, R. Weischedel. 2002. *Empirical Studies in Strategies for Arabic Retrieval.* In “Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval”, M. Beaulieu et al. ed., pages 269-274, ACM, NY, NY.