

Arabic-English NLP at CRL

Rémi Zajac, Ahmed Malki, Ahmed Abdelali

Computing Research Laboratory, New Mexico State University
zajac, amalki, ahmed@crl.nmsu.edu

Abstract

This paper gives a brief overview of Arabic-English Cross Language Information Retrieval and Machine Translation efforts at CRL. We also describes Arabic resources available from CRL.

1 Introduction

CRL started developing Arabic natural language processing systems within the Temple project (Vanni & Zajac 97). In this project, we developed a morphological analyzer, an Arabic-English dictionary, a number and date tagger and a small Arabic phrasal dictionary, all of which were used to build a simple glossary-based Arabic-English machine translation system. The tools and resources were further developed in subsequent projects. The URSA Unicode-based information retrieval engine (Davis 99) was adapted to Arabic and was the basis of a cross-language Arabic-English information retrieval system. This paper gives a brief overview of the tools and resources developed at CRL. We are currently packaging these tools and resources to make them available to the academic community.

2 Arabic-English MT and CLIR

Arabic-English MT. The machine translation strategy is fairly simple. The first phase of the machine translation process is a morphological analysis of the text, followed by a bilingual dictionary look-up. The next phase is the glossary-based machine translation component which uses both a bilingual glossary and the result of the dictionary look-up to produce a phrase-by-phrase translation, and in the case when no phrase can be recognized in a sentence, a word-by-word translation. The result is passed on to the English morphological generator.

Thus, the linguistic components of the system are few and simple: a morphological analyzer, a bilingual dictionary, a bilingual glossary and a morphological generator. Since the translation component does not perform syntactic analysis, it is clear that the complexity of the system lies in the glossary-based machine translation engine and that the quality of the translation depends heavily on the quality

and the size of the glossary. The structure of a glossary entry is simple: a source phrase pattern plus a list of corresponding target patterns. A glossary entry can easily be added to the system by a user who has minimal linguistic training.

This system, originally developed in the Temple project (Vanni & Zajac 97), has been reimplemented using the modular Corelli MT architecture (Amtrup & Zajac 00).

Arabic-English CLIR. The Arabic Information Retrieval system offers the possibilities of querying an Arabic resource and retrieving Arabic documents using the URSA Unicode-based retrieval engine developed at CRL.¹ The system has been demonstrated using an archive of 5 months of the Al-Rayah daily newspaper from Qatar. The system offers the option to use a full word mode or using a morphological analyzer. In the latter option, documents are indexed using stems, and the query is also processed using the morphological analyzer. See a web demo at <http://crl.nmsu.edu/~ahmed/>.

The Arabic-English CLIR system offers the possibility of querying the Arabic text using an English query. The English query is processed interactively and the user can improve the translated query. The translated (Arabic) query will be sent to the Arabic IR system to retrieve relevant Arabic documents. The user has a full set of tools to display and browse the results, and translate back into English the retrieved documents. The functional complete system (Malki 01) is shown in Figure 4. See a web demo at <http://crl.nmsu.edu/~ahmed/test/news>.

3 Arabic Tools

Morphological Analyzer. The Arabic word *wasayaktubunahaa* can be analyzed as follows: the root morpheme *ktb* (“to write”) combines with the verb pattern morpheme CCuC (present/future tense) to form the stem *ktub*, to which are attached the prefixes *wa* (“and”), *sa* (future tense), *ya* (3rd person), and the suffixes *wuna* (masc.

¹ See <http://crl.nmsu.edu/Research/Projects/tipster/ursa/> for papers, technical documentation and download of the URSA engine.

pl.) and *haa* (“it”): *wasayaktubuunahaa=wa-sa-ya-ktub-uuna-haa*. This word can be glossed as: “and they (masc.pl.) will write it.”

For the purpose of computer analysis, Arabic words are treated as having three elements: prefix, stem, and suffix (Buckwalter 98). Using this approach, the word *wsyktbuunhaa* is segmented as *wsy-ktb-uunhaa*. All valid prefix and suffix concatenations are stored in respective lexicons (78 prefixes with 17 prefix categories, and 318 suffixes with 51 suffix categories). Likewise, all valid stems (i.e. combinations of root and pattern morphemes) are stored in a lexicon of stems (about 60,000 stems with 129 stems categories). Co-occurrence constraints are simply expressed as boolean combinations of prefix-stem categories, prefix-suffix combinations, and stem-suffix combinations.

This design is very efficient as a word is decomposed into prefix-stem-suffix by looking up the sub-lexicons (implemented as tries) and checking boolean constraints. When a stem is not found in the stem dictionary, the analyzer produces all possible stems by checking appropriate combination of prefix and suffix. The morphological analyzer is parameterized in order to correctly process regional variations in spelling (e.g., the Egyptian orthography), as well as for handling various codesets, including Unicode.

Number and date tagger. The tagger works on cp-1256 Arabic texts and produces a tagged text with tags `date`, `date expressions`, `number (ordinal or digit)`. Each tag has a value attribute which contains the decimal value of the Arabic number or date.

```
<date value=130193>13 ynAyr 1993</date>
<number value=2>2</number>
<ordinal value=1>A1>w1</ordinal>
<number value=1993>1993</number>
<date value=110193>A1>vnyn</date>
```

The tagger can be run on the command line using the script `numberDateTagger` as:

```
numberDateTagger <inputfile> <outfile>
```

4 Arabic Resources

Corpora. The Arabic corpora collected at CRL include a collection of more than 60MB of news from the AFP news agency. All documents in the collection are listed by date and tagged using SGML. Tags include format the number of the story, date, headlines and a footer. An example is showed in Figure 1.

We also have a several collections of on-line newspapers articles from eleven countries (Algeria, Iraq, Jordan, Kuwait, Lebanon, Mauritania,

```
<DOC>
<DOCNO>yyyyymmdd_AFP_ARB.dddd
</DOCNO>
<HEADER>Arabic Text</HEADER>
<BODY>
<HEADLINE>Arabic Text</HEADLINE>
<TEXT>
<P>One or More Paragraphs of
Arabic Text</P>
</TEXT>
<FOOTER>Arabic Text</FOOTER>
</BODY>
<TRAILER>Arabic Text</TRAILER>
</DOC>
```

Figure 1: SGML template

Morocco, Oman, Qatar, Saudi Arabia and Syria). These news articles were collected between February and November 2000. This collection has been tagged and indexed and deployed in some information retrieval experiments. The collection can be queried using a Unicode-based IR system developed at the CRL (Davis 99). The indexes are built in two different ways: based on the form of the word, and using the morphological analyzer. The system also allow the user to choose stem the query when searching the collection either by single country `-newspaper-` or the whole collection. The collection is accessible from our web site at <http://crl.nmsu.edu/~ahmed/test/news/>.

General Arabic-English Dictionary. The Arabic-English dictionary is essentially a morphological dictionary with English translations. It does not contain usual part-of-speech information nor proper citation forms. Instead, an entry key (field `$H` below) is a morphological stem, typically a sub-string of an inflected word. All stem variants for the same word are listed. Each entry contains a morphological category (number of the inflectional paradigm for that stem, field `$C`). English translations are listed in field `$I`. The dictionary contains approximately 72,000 pairs of (stems, category) and about 43,000 unique stems. An entry looks like:

```
$H dxAry
$C R001
$I savings;storage;
$$
```

Proper Names. The Arabic proper names dictionary was developed to serve as a main resource for the proper nouns, capture some specific characteristics and categorization (First name, Last Name, Place,...) of the proper nouns, the dictionary contains 1694 entries which are mainly Arabic proper

\$H ا. او سميت	\$H زيوزينسكايا
\$C R058	\$C R058
\$M a . O . Smith	\$M Zyuzinskaya
\$pn name	\$pn name
\$I a . O . Smith	\$I Zyuzinskaya
\$\$	\$\$
\$H العقيب	\$H الايبيل
\$C R058	\$C R058
\$M Al a Aqib	\$M Al Abyar
\$pn place	\$pn place
\$I Al a Aqib	\$I Al Abyar
\$\$	\$\$

Figure 2: Onomasticon entries

names and places and organizations. The dictionary also include the English translation, or transliteration in some cases.

English-Arabic Onomasticon: is a large collection (204,606 entries) of names of persons and companies, and also includes geographical places, countries and cities around the world. Extracted mainly from English newspapers and journals. The Arabic side of the Onomasticon is an Arabic transliteration of the western names (Figure 2).

The format of entry is essentially the same as the general Arabic-English dictionary mentioned above.

English-Arabic Dictionary. English-Arabic Dictionary: the English Arabic dictionary contains 113,208 entries. Each entry contains the part of speech (POS) of the English word and one or more translation based on the meaning and the context of usage of the English word (Figure 3).

Glossary. The machine translation system uses an Arabic-English phrasal dictionary (“glossary”) containing approximately 12,000 phrases. This glossary was built by automatically extracting phrasal patterns from an Arabic corpus of news articles and technical documentation. Translations were added manually.

References

Amtrup, Jan W. and Rémi Zajac. “A Freely Available Toolkit for Machine Translation”. *COLING-2000*, July 31-August 4 2000, Saarbrücken, Germany.

\$H adjutant
\$C R059
\$M مساعد
\$I army officer responsible for administrative work in a battalion
\$\$

Figure 3: English-Arabic dictionary entry

- Buckwalter, Tim. 1998. “Technical Report on the Arabic Morphological Analyzer”. NMSU-CRL Technical Report, 1998. (see also <http://www.itsnet.com/~qamus/>)
- Davis, Mark. 1999. “URSA: The Unicode Retrieval System Architecture”. *14th International Unicode Conference*, Boston, MA, March 1999.
- Malki, Ahmed. 2001. “Arabic Interactive Cross-Language Information Retrieval Via Natural Language Processing”. PhD Dissertation, New Mexico State University. CRL Technical Report MCCS-01-326, 2001.
- Vanni, Michelle & Rémi Zajac. 1997. “Glossary-Based MT Engines in a Multilingual Analyst’s Workstation Architecture”. *Machine Translation 12*, Special Issue on New Tools for Human Translators. pp131-157.

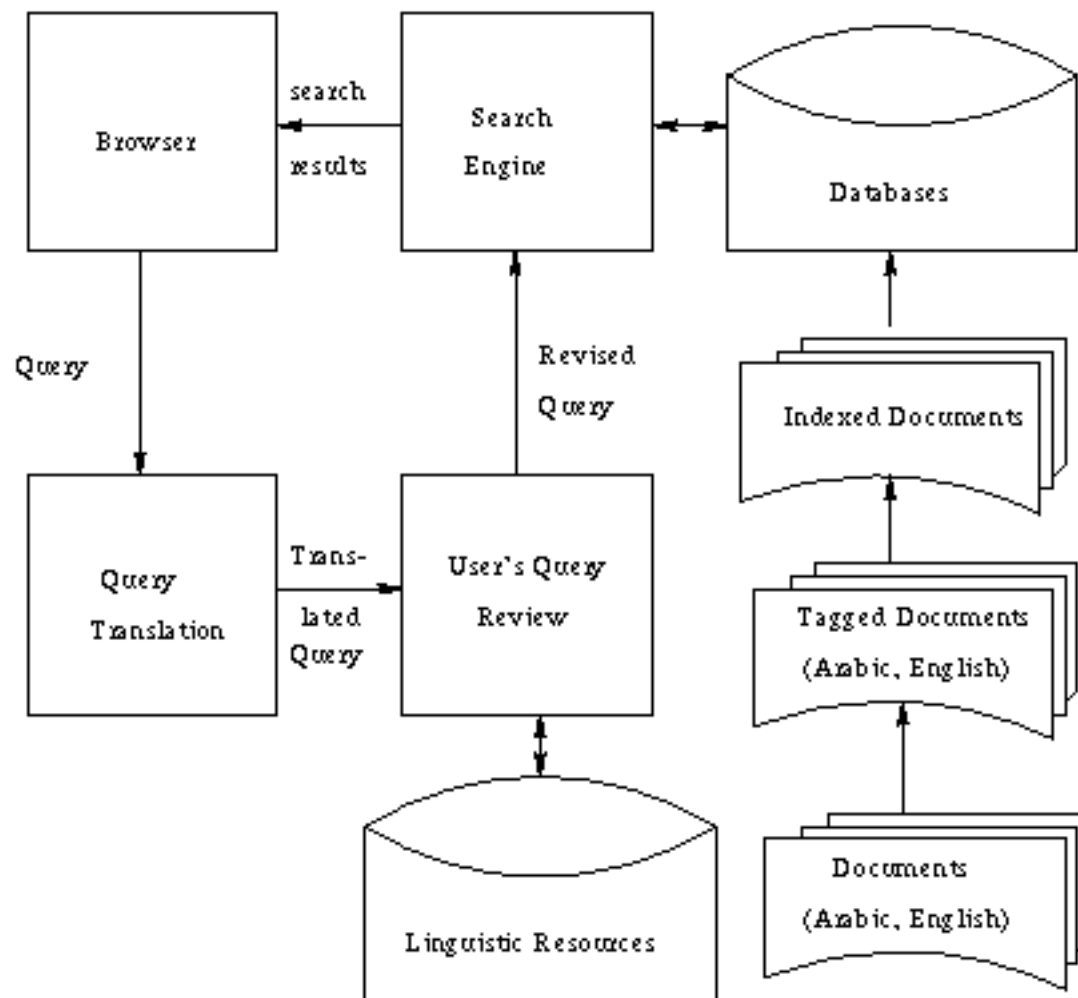


Figure 4: Functional View of the Interactive Cross-Language System