Arabic Named Entity Recognition from Diverse Text Types

Khaled Shaalan and Hafsa Raza

Faculty of Informatics, The British University in Dubai, P.O Box 502216, Dubai, UAE
Khaled.shaalan@buid.ac.ae, hafsa.raza@gmail.com

Abstract. Name identification has been worked on quite intensively for the past few years, and has been incorporated into several products. Many researchers have attacked this problem in a variety of languages but only a few limited researches have focused on Named Entity Recognition (NER) for Arabic text due to the lack of resources for Arabic named entities and the limited amount of progress made in Arabic natural language processing in general. In this paper, we present the results of our attempt at the recognition and extraction of 10 most important named entities in Arabic script; the person name, location, company, date, time, price, measurement, phone number, ISBN and file name. We developed the system, Name Entity Recognition for Arabic (NERA), using a rule-based approach. The system consists of a whitelist representing a dictionary of names, and a grammar, in the form of regular expressions, which are responsible for recognizing the named entities. NERA is evaluated using our own corpora that are tagged in a semi-automated way, and the performance results achieved were satisfactory in terms of precision, recall, and f-measure.

Keywords: Information extraction; Named entity recognition; Arabic natural language processing.

1 Introduction

NER system is a significant tool in NLP research since it allows identification of proper nouns in open-domain texts. Larkey have conducted a study that showed the importance of the proper names component in language tasks involving searching, tracking, retrieving, or extracting information [9]. Another study by Crestan & de Loupy showed that named entity extraction helps users to browse large document collections more quickly and efficiently [2]. This seems plausible as, according to Gey 30% of the content-bearing words in news are proper names [5]. Abuleil [12] and Chinchor [11] stated that the valuable information in text is usually located around proper names, to collect this information it should be found first.

We have adopted the rule-based approach using linguistic grammar-based techniques to develop NERA. The approach is motivated by the characteristics and peculiarities of Arabic language. The recognition process takes two cycles, using the whitelist component and then applying the grammar rules. This open architecture approach provides flexibility and adaptability features in our system and it can be

A. Ranta, B. Nordström (Eds.): GoTAL 2008, LNAI 5221, pp. 440–451, 2008.

easily configured to work with different languages, NLP applications, and domains. We present the results of our attempt at the recognition and extraction of 10 most important named entities in Arabic script that is, the person name, location, company, date, time, price, measurement, phone number, ISBN and file name. The NERA system is evaluated using a reference corpus that is tagged with names in a semi-automated way. The achieved system performance results were satisfactory when evaluated against the standard measures; precision, recall, and f-measure.

The rest of this paper is structured as follows. Section 2 presents previous related work in Arabic NER. Section 3 describes the data collection methods used. Section 4 explains in detail our approach to NER in terms of system architecture. Section 5 is dedicated to show the reference corpora we built to carry out our experimental work. In Section 6 we present the results of our experiments, whereas in the Section 7 we highlight how our system NERA, provides solutions to challenges posed by Arabic language. Finally, in Section 8, we draw some conclusions and discuss future works.

2 Related Work

Name identification has been worked on quite intensively for the past few years, and has been incorporated into several products. Many researchers have attacked this problem in a variety of languages but only a few limited researches have focused on NER for Arabic text. This is due to the lack of resources for Arabic NE and the limited amount of progress made in Arabic NLP in general.

Maloney and Niv developed TAGARAB an Arabic name recognizer that uses a pattern-recognition engine integrated with morphological analysis. The role of the morphological analyzer is to decide where a name ends and the non-name context begins. The decision depends on the part-of-speech of the Arabic word and/or its inflections. The performance achieved for the Person NE recognition was 86.2%, 76.2% and 80.9% whereas for the Location NE it was 94.5%, 85.3% and 89.7% for precision, recall and f-measure respectively [7].

Abuleil presented a technique to extract proper names from text to build a database of names along with their classification that can be used in question-answering systems. This work was done in three main stages: 1) marking the phrases that might include names, 2) building up graphs to represent the words in these phrases and the relationships between them, and 3) applying rules to generate the names, classify each of them, and saves them in a database. The NE recognition accuracy was estimated in terms of precision by the author; People (90%), Location (93%) and Organization (92%) [12].

Samy has used parallel corpora in Spanish, and Arabic and an NE tagger in Spanish to tag the names in the Arabic corpus. For each sentence pair aligned together, they use a simple mapping scheme to transliterate all the words in the Arabic sentence and return those matching with NEs in the Spanish sentence as the NEs in Arabic. While they report high precision (84%) and recall (97.5%), it should be noted that their approach is applicable only when a parallel corpus is available [3].

Zitouni has adopted a statistical approach for the entity detection and recognition (EDR). In this work, a mention can be either named (e.g. John Mayor), nominal (the president) or pronominal (she, it). All are referring to one conceptual entity. The

performance of this mention detection system is given by the author in terms of precision (64.4%), recall (55.7%) and f-measure (59.7%) [6].

3 Data Collection

For training and testing purposes, we have compiled corpora containing texts which are diverse in terms of domain, format, style and genre. This aims to ensure that the system can cope adequately with any kind of text, and that its future use is not limited to any particular text type. Techniques used for acquiring such data include:

- **Automatic collection of named entities instances and indicators from annotated corpora:** The Automatic Content Extraction (ACE[1]) and Arabic Treebank (ATB[2]) are some great resources that facilitate corpus based studies of many interesting linguistic phenomena in Modern Standard Arabic (MSA). These corpora were exploited for the data collection task. These corpora, which are tagged with great linguistic details, were first analyzed and the commonly occurring patterns were studied. These identified patterns were then used to extract useful data.
- **Name Database provided by government organization:** The person and company name dictionaries were also build from names collected from some organizations including Immigration Departments, Educational bodies, and Brokerage companies.
- **Internet Resources**[3]: Names were retrieved further from various websites[4] containing lists of Arabic names, company names and locations. Some of these names are Romanized (written using the Latin alphabet) and had to be transliterated from English to Arabic.

The NEs compiled by processing corpora, internet resources and various organizations, had to be further processed to ensure that the compiled data is clean. The raw data received had to be further processed to make it suitable for incorporation into the system.

4 The Architecture for NERA System

The NERA system was implemented through incorporation into the FAST ESP framework, [5]. Figure 1 shows the abstract architecture of the NERA system. The system requires two main processing resources: a *whitelist* (gazetteer) and a finite state transduction *grammar*. A *filtration mechanism* is also employed that enables revision capabilities in the system.

[1] ACE reference: http://projects.ldc.upenn.edu/ace/
[2] Treebank Corpus reference: http://www.ircs.upenn.edu/arabic/
[3] Web sites include: http://en.wikipedia.org/wiki/List_of_Arabic_names, http://www.islam4you.info/contents/names/fa.php, and http://www.mybabynamessite.com/list.php?letter=a
[4] Web sites include: http://en.wikipedia.org/wiki/List_of_Arabic_names, http://www.islam4you.info/contents/names/fa.php, and http://www.mybabynamessite.com/list.php?letter=a

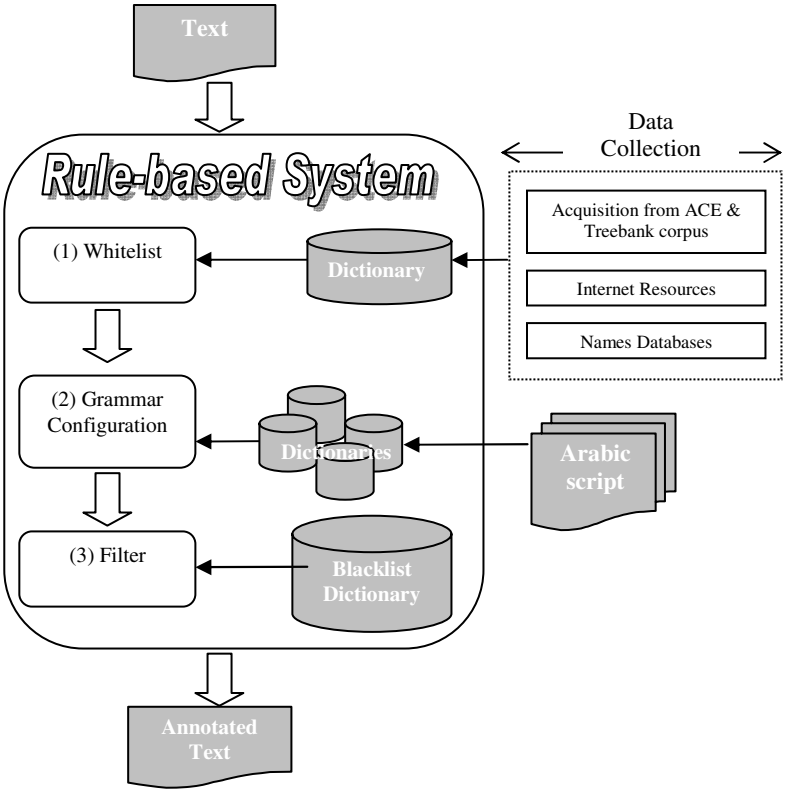

Fig. 1. Architecture of the System

4.1 Whitelist

The whitelist plays the role of fixed static dictionaries of various named entities. It is a mechanism that accepts matches which are reported as a result of an intersection between the dictionary and the input text. A Whitelist is a list of strings that must be recognized independent of the rules. It contains entries in the format:

عبدالرحمن قاسم الشيراوى|Abdulrahman Qasim Mohammed Alshirawi

The English transliterations of the Arabic names are included in the dictionary as meta-data in order to allow for incorporation with various applications.

4.2 Grammar

The grammar performs recognition and extraction of Arabic named entities from the input text based on derived rules. It describes patterns to match NEs, thereby annotations being created as a result. Due to the peculiarities and complexities in the Arabic language, grammar rules are a vital processing resource for the recognition system. For instance the lack of capitalization for proper nouns can be very well compensated

by using NE indicators to formulate recognition rules. These NE indicators were obtained as a result of the deep contextual analysis of various Arabic scripts that were performed during the data collection phase. The indicators are referred to as trigger words within our system, forming a window around a named entity, which helps in identifying a NE within text but does not get recognized itself.

- Person Title: السيدة (Mrs.), السيد (Mrs.)
- Job title: الدكتورة (the doctor), أستاذ العلوم (the sciences professor)
- Company indicator: ذات مسئوليه محدودة (LLC)
- Country Post-indicators: الاتحادية (the federal), الديمقراطية (the democracy)
- City Post-indicators: عاصمة المالية (the finance capital)
- Measurement: ملليجرامات (miligrams), كيلوا مترات (kilometers)
- Price: جنيه مصري (Egyptian Pound), درهم إماراتي (dirham Emirati)

Moreover inflections within Arabic language can be well dealt with using handcrafted rules, which enables stripping off of the prefixes and suffixes from the stem word, prior recognition. Thus ensuring the recognition of the actual NE instance alone. For each type of named entity several rules were built and each one was applied in a particular order to ensure that the most comprehensive recognition result was achieved.

Example rule for *Person name* recognition

```
((honorfic+ws(location(ية|ي)+ws)?)+firsts_v
(ws+lasts_v)?ws+(number)?)
```

The above rule recognizes a person name composed of a first name followed by optional last name based on a preceding person indicator pattern, or the trigger words. The following name would be recognized by this rule:

الملك عبد الله	[The king Abdullah]
الملك الأردني عبد الله	[The Jordanian king Abdullah]
الملك الأردني عبد الله الثاني	[The Jordanian king Abdullah II]
الملكة الأردنية رانيا	[The Jordanian queen Rania]

Apart from contextual cues, the typical Arabic naming elements were used to formulate rules such as nasab, kunya, etc. Thereby the rules resulted in a good control over critical instances by recognizing complex entities.

Example rule for *Location* recognition

```
((مدينة | Administrative division) + ws)? + city name
+ws + direction
```

The rule above recognizes a city name (existing in the dictionary of city names). The following name would be recognized by this rule:

مدينة اغادير جنوب... [Agadir City south of …]

4.3 Filter

A *filtration* mechanism is used in the form of a *Blacklist* (rejecter) within the grammar configuration to filter matches, returned by rules, which appear after named entity indicators but are invalid entities. Consider the following example:

'وزير الخارجية العراقي الامين العام' [The Iraqi Foreign Minister the Secretary-General]

In this example, the words following the person indicator ('وزير الخارجية العراقي' [The Iraqi Foreign Minister]) that is, 'الامين العام' (the Secretary-General) is not a valid person name. The role of the blacklist, another set of rules, is rejecting such incorrect matches.

Apart from the *Blacklist* component certain heuristic *Filter rules* are used for post-processing the system's extraction results in order to disambiguate extracted named entities. When applying a set of single-slot extraction rules to the input text i.e. sets of rules which extract particular types of named entity one after the other, one cannot exclude the possibility of identical or overlapping textual matches within the document, among different rules for different named entities. For instance, different sets of rules for extracting instances of both the named entities *person* and *location names* may overlap or exactly match in certain text fragments, resulting in ambiguous named entities. Among these named entities, the correct choice must be made. The *filter rule* is an intelligent way of specifying how to get the correct choice, with respect to the context in which the ambiguous situation may arise.

The following example illustrates an ambiguous situation in Arabic script:

احمد اباد لديه اهتمام بالغ بالفلسفة
(**Ahmed Abad** has a keen interest in philosophy)

In this example the bold text fragment represents both a person name and a location. Hence when NERA is applied here, both the Person and Location Extractors will return matches as 'احمد اباد' (Ahmed Abad). The developer can tune the system to resolve some kinds of ambiguous situations by the virtue of filter rules. One solution to disambiguate this situation is to use the following filter rule:

```
If a possible match M1 for a location entity intersects
with a match M2 that was previously reported by the
person extractor, then the match as a location name
will be discarded.
```

Thus in case of an intersection, the match for person names is preferred over location names. The filter rules defined within the system play a significant role to handle such situations and resolve ambiguity. However, it should be built upon careful analysis of the ambiguous situations in order to get accurate results.

5 Resources Build for Arabic NER within NERA

To develop the Arabic NER, we had to build our own corpora due to the unavailability of free Arabic corpora for research purposes. Moreover, the commercially available Arabic corpora are oriented towards newswire which we found lacks the coverage of the 10 named entities involved in our research. Further, we have also built the whitelist (gazetteer) component, which is a vital processing resource for many NLP tasks. Following, we present the main characteristics of the developed resources for Arabic.

5.1 Corpora for Person, Location, Date, Time, Price and Measurement NE

ACE (Automatic Content extraction, version 5.3.3 2005.05.31) and ATB (Arabic Treebank, version 2.0, LDC catalog number LDC2003T06) corpora by LDC are some great Arabic NLP resources. These corpora contain text taken from newswire documents and broadcast news which was used to create the entity tagged reference corpora for evaluating Person, Location, Date, Time, Price, and Measurement extractors within NERA.

For efficiency purpose the reference corpus build was divided into sets of test corpora, each being approximately 100KB in size. The total number of test sets for these named entities is 34, with 24 created from ACE corpus and 10 created from ATB corpus. The total size of the reference corpus is around 4MB composed of 300000 words. The size and content of the corpus is such that it contains a representative amount of occurrences of the following NE: Person name includes 500+ entities, location includes 500+ entities, date includes 394 entities, time includes 110 entities, price includes 400 entities, and measurement includes 386 entities.

5.2 Corpus for Company Named Entity

The ACE and ATB corpora do not include representative number of entities for company names. We sought another corpus, that is, Corpus of Contemporary Arabic (CCA[5]) [8]**.** We used CCA to create of the reference corpus for evaluating the *company extractor*. For building up the company test corpus we created two reference corpus set (each 100 KB in size) from randomly selected text from the CCA corpus. Both the two sets were hand tagged to mark company names within it. A total of *226 company name* instances have been hand tagged.

5.3 Corpus for Phone Number, ISBN and File Name Named Entities

Arabic available corpus resources are quite limited and restrained to coverage of the most important NEs such as person, location etc. Hence various Arabic websites (e.g. Real Estate, Newspaper etc) were analyzed to collect *Phone number, ISBN* and *file name* entities. The corpus build was hand tagged with 191 Phone number entities, 100 entities for ISBN, and 139 entities for File name.

5.4 Whitelists

NERA gathers three different manually built gazetteers or whitelist:

1. *Person Whitelist*: This contains a list of 263,598 complete names of people collected from DNRD (Dubai Naturalization & Residency Department), Brokerage companies, and existing Arabic corpora and internet resources. Further the names were split into dictionaries of first names with 175,502 names and last names with 33,517 names;
2. *Location Whitelist*: This consists of 4,900 names of continents, countries, cities, states, political regions, towns and villages found in the Arabic version of Wikipedia and other websites;

[5] CCA is freely downloaded online http://www.comp.leeds.ac.uk/eric/latifa/research.htm

3. *Organizations Whitelist*: This consists of a list of 273,491 names of companies including areas such as media and newspaper, construction, banks & insurance, airlines, telecommunications and many more.

6 Experiment

The evaluation of the NERA extractors was performed using our own reference corpora which highlight the Arabic resources built during this project work. Since the corpora were tagged in a semi-automated way, certain named entities were left untagged. In the recognition results these NEs were recognized correctly by the system, but since they were not tagged in the test corpora the evaluation tool marked these as false positives when in reality they were true positives. To overcome this issue, the entities marked as false positives by evaluation tool were extracted and retagged in the reference corpora. This iterative tagging of the corpus ensured quality. Moreover this tool can perform evaluation on a corpus with size limited to 100 KB. Hence the 5MB of evaluation corpora composed of 397,069 words was divided into 46 sets of corpus files.

6.1 Evaluation Method

We have adopted the standard evaluation measures in the IE community [1] (i.e. precision, recall and F-measures), to evaluate and compare the results. It was introduced to provide a single figure to compare different systems' performances.

6.2 Results

Table 1 summarizes the accumulative recognition accuracy, in terms of precision & recall, achieved by all the 10 extractors within NERA, against the reference corpora.

With respect to the extractors' person, location and company some of the entries within the whitelist component built were extracted from the same corpus used also for creating the reference corpora for evaluation. However, the evaluation results achieved are accurate since they indicated recognition of named entities not included in the *whitelist* but being recognized by the grammar rules within the pattern matching component.

Table 1. Accumulated accuracy of the 10 named entities

No	NE	Precision	Recall	f-measure
1	Person	86.3%	89.2%	87.7%
2	Location	77.4%	96.8%	85.9%
3	Company	81.45%	84.95%	83.15%
4	Date	91.2%	92.3%	91.6%
5	Time	97.25%	94.5%	95.4%
6	Price	100%	99.45%	98.6%
7	Measurement	97.8%	97.3%	97.2%
8	Phone Number	94.9%	87.9%	91.3%
9	ISBN	94.8%	95.8%	95.3%
10	File name	95.7%	97.1%	96.4%

One important factor that has greatly influenced the above achieved results is the non-standardization of written Arabic text. Majority of them are unstructured loaded with inconsistencies due to the lack of control over written forms of Arabic script. Standard practices in publishing written Arabic resources can help achieve far better accuracy results

7 Solutions to Challenges in NERA

7.1 Inflections

Arabic is a highly inflected language. So, within the handcrafted rules, we added the possibilities of breaking down the inflected form into a stem (or numeric figure) and affixes in order to recognize the stem as a name entity. Table 2 shows some inflected named entity examples which have been dealt with in the grammar file for the respective entity type.

7.2 Non-casing Language

Due to the lack of capital letters in Arabic script, we used keywords or indicator words to guide us to the place where one could find them in the text. The method adopted is to derive a set of heuristic rules that parse the phrases to extract the name entities. Some examples of keywords used for identifying the names are:

- Personal names (title): **Mr.** John Adams → **السيّد** جون آدامز
- Personal names (job title): **President** John Adams → **الرئيس** جون آدامز

Table 2. Examples of inflections in Arabic text

Arabic Ex.	English Trans.	Entity Type	Affix (clitics)
بـ ٢٠,٢٦٦ دولارا	For $20,266	Price	‘ب’ (baa)
الـ ٢٩٢٥ متر	The 2925 meter	Measurement	‘الـ’(al)
بالولايات المتحدة	For the United States	Location	‘بال’ (baa, alif, laam)
ومصر	And Egypt	Location	‘و’ (Waw)
لهيئة الاذاعة البريطانية "بي بي سي”	for the British Broadcasting Corporation "BBC"	Company	‘ل’ (laam)

7.3 Spelling Variants

Spelling of translated and transliterated proper names in general tends to be inconsistent in Arabic text. Table 3 shows some examples of the inconsistency, although some can be considered as typos.

The extractor can handle, to some extent the above mentioned spelling variants. Such issues were dealt with within the context sensitive rules and dictionary build within the NERA system.

Table 3. Examples of variations in Arabic text

Arabic Ex.	English Trans.	Entity Type
أندونيسية / أندونيسيا	Indonesia	Location
جلدر/ جيلد /غلدر/غيلدر	Guilder	Price (currency)
لوس انجيليس /لوس انجيلس/لوس انجلوس/لوس انجليس	Los Angeles	Location
رقم الموبيل: ٥٧٥٦٤٥٣ / الجوال: ٥٧٥٦٤٥٣	Mobile no: 3546575	Phone number
جوهانسورغ /جوهانسوبورغ/جوهانسبرغ/ جوهانسبورغ	Johannesburg	Location

7.4 Typographic Variants

The extractor is capable of recognizing variations in written Arabic text for the various named entities being recognized. Table 4 contains some example NE indicating typographic variations.

7.5 Ambiguity

This commonly found problem in Arabic script is encountered within NERA when ambiguous matches are returned by different extractors. Table 5 shows some of the ambiguous situations that the system can handle. These situations can be handled by specifying a filter rule that gives preference on one extractor over the other.

Table 4. Examples of typographic variations in Arabic text

Arabic Ex.	English Trans.	Entity Type	Typographic variation
أستراليا/استراليا	Australia	Location	drop hamza (initially, medially, or finally)
السعوديه/السعودية	Saudi Arabia	Location	two dots removed from taa marbouta
ليرة/ليره	Lira	Price	Two dots inserted on final haa
آسيا/اسيا	Asia	Location	Drop of the madda from aleph
إلاربع/ألاربع	4th	Date (day)	Hamza (below or above aleph)

Table 5. Ambiguous examples

Ambiguous Ex.	English Trans.	Incorrect	Correct
فرنك سويسري 1.6985	1.6985 Swiss Franc	Person	Price
١٥ رمضان الكريم ٢٠٠٥	15th of Ramadan Al karim 2005	Person	Date
جاسم المتحدة للعقارات والصيانة العامه	Jussim united for real estate and general maintenance	Person	Company
١,٥ بليون دولار سنغافورة	1.5 billion Singapore dollar	Location	Price
شركة أرامكو السعودية	Saudi Aramco	Location	Company
في المساء اليزابيث الثانية	In the evening Elizabeth II	Time	Person
نقطة تحول في سبتمبر سنة … ١٩٥٤… قدم مارتن	…a turning point in September 1954 Martin presented…	Measurement	Date

8 Conclusion

The work done in this project is an attempt to broaden the coverage for entity extraction by incorporating the Arabic language, thereby paving the path towards enabling search solutions to the Arabian market. Various data collection techniques were used for acquiring dictionary name lists. The rule-based approach employed with great linguistic expertise provided a successful implementation of the NERA system by accomplishing challenges posed by Arabic language. Rules are capable of recognizing inflected forms by breaking them down into stems and affixes. A filtration mechanism is employed in the form of a rejecter within the grammar configuration that helps in deciding where a name ends and the non-name context begins. Further the intelligent use of filter rules helps in dealing with ambiguity between named entities. We have evaluated our system performance using a reference corpus that is tagged in a semi-automated way. The average Precision and Recall achieved by NERA extractors for each named entity type, against the reference corpora were satisfactory.

Acknowledgement

This work is funded by the "Named Entity Recognition for Arabic" joint project between The British Univ. in Duabi, Dubai, UAE and FAST search & Transfer Inc., Oslo, Norway. We thank the FAST team. In particular, we would like to thank Dr. Petra Maier and Dr. Jürgen Oesterle for their technical support. Any opinions, findings and conclusions or recommendations expressed in this material are the authors, and do not necessarily reflect those of the sponsor.

References

1. Sitter, A.D., Calders, T., Daelemans, W.: A Formal Framework for Evaluation of Information Extraction, University of Antwerp, Dept. of Mathematics and Computer Science, Technical Report, TR 2004-0. (2004), http://www.cnts.ua.ac.be/Publications/2004/DCD04
2. Eric, C., de Loupy, C.: Browsing Help for a Faster Retrieval. In: Coling2004 proceedings, Geneva, August 2004, pp. 576–582 (2004)
3. Samy, D., Moreno, A., Guirao, J.M.: A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus. In: International Conference RANLP, Borovets, Bulgaria, pp. 459–465.
4. FAST ESP, http://www.fastsearch.com/thesolution.aspx?m=376
5. Frederic, G.: Research to Improve Cross-Language Retrieval – Position Paper for CLE. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, pp. 83–88. Springer, Heidelberg (2001)
6. Zitouni, I., Sorensen, J., Luo, X., Florian, R.: The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution. In: Proceedings of the ACL workshop on Computational Approaches to Semitic Languages, 43rd Annual Meeting of the Association of Computational Linguistics (ACL2005), Ann Arbor, Michigan, USA, pp. 63–70 (2005)

7. Maloney, J., Niv, M.: TAGARAB: A Fast, Accurate Arabic Name Recogniser Using High Precision Morphological Analysis. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, August, pp. 8–15 (1998)
8. Al-Sulaiti, L., Atwell, E.: Extending the Corpus of Contemporary Arabic. In: Proceedings of Corpus Linguistics conference 2005. University of Birmingham, UK (2005)
9. Larkey, L.S., Jaleel, N.A., Connell, M.: What's in a Name?: Proper Names in Arabic Cross Language Information Retrieval CIIR Technical Report IR-278 (2003)
10. Maamouri, M.: Language education and human development: Arabic diglossia and its impact on the quality of education in the Arab region. In: The Mediterranean Development Forum. The World Bank, Washington (1998)
11. Chinchor, N.: Overview of MUC-7. In: Proceedings of the Seventh Message Understanding Conference (MUC-7) (1998)
12. Abuleil, S.: Extracting Names from Arabic Text for Question-Answering Systems. In: Proceedings of Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 2004), Avignon, France, pp. 638–647 (2004)