Cue-based bootstrapping of Arabic semantic features

Khaled Elghamry [1,a], Rania Al-Sabbagh [a], Nagwa El-Zeiny [b]

[a] Faculty of Al-Alsun (Languages), Ain Shams University, Cairo, Egypt

[b] Faculty of Arts, Helwan University, Cairo, Egypt

Abstract

Motivated by the fact that semantic features are understudied in Arabic Natural Language Processing (ANLP) in spite of being essential for some Natural Language Processing (NLP) tasks such as Anaphora Resolution (AR), Word Sense Disambiguation (WSD) and Prepositional Phrase (PP) attachment, this paper presents a cue-based algorithm to build an Arabic lexicon that tackles such semantic features. The lexicon, whose entries are extracted from the World Wide Web (WWW) using bilingual and monolingual cues, achieves a performance rate of 89.7% measured according to a gold standard set of 3000 entries. Moreover, using such a lexicon raises the performance of an AR algorithm for Arabic generic corpora from 74.4% to 87.4% which is a state-of-the-art performance rate. To the best of the authors' knowledge, this paper presents the first attempt to deal with Arabic semantic features beyond the features of gender and number.

Keywords: Arabic semantic features, cue-based bootstrapping, web as corpus.

1. Introduction

Semantic features, according to Silzer (2005), are the constituents of the meaning of the word expressed by plus (+) and minus (–) signs. They include a set of abstract concepts such as gender, number, rationality (being able to think or unable to), animacy etc. For example, the semantic features of the noun *woman* are +HUMAN, +ADULT, +ANIMATE, +RATIONAL, –PLURAL and –MALE.

In Natural Language Processing (NLP), semantic features are used for a variety of tasks such as Anaphora Resolution (AR) (Lappin and Leass 1994, Al-Sabbagh 2007), Word Sense Disambiguation (WSD) (Turney 2004) and Prepositional Phrase (PP) attachment (Hartrumpf et al. 2006). For most cases, these semantic features are used to *filter* a set of possible candidates from the candidates whose semantic features do not match the target linguistic unit; that is, the linguistic unit to be disambiguated like the pronoun in the case of AR, the ambiguous word(s) in WSD and the verb in PP attachment.

For instance, Al-Sabbagh (2007) used semantic features as filters for an AR algorithm for Arabic generic corpora so that only the candidates that agree with the semantic features of the pronoun are used as input for the AR algorithm. In sentence (1) below, there are two possible candidate antecedents for the pronoun هم /*hm*/[2] (their) whose distinctive semantic feature is +PLURAL. The two candidates are الحوار /*AlHwAr*/ (the conversation) which is –PLURAL and

[1] Revision made on May 29th, 2008, concerning the mention of the first author (Khaled Elghamry).

[2] Buckwalter's Transliteration Scheme (Buckwalter 2002). URL: www.qamus.org/transliteration.htm

المثقفين /*Almvqfyn*/ (the cultured) which is +PLURAL. Using semantic features lead to excluding the former and correctly choosing the latter as the correct antecedent.

(1) الحوار مفتوح للمثقفين بمختلف مشاربهم

Transliteration:

/*AlHwAr mftwH llmvqfyn bmxtlf m$Arbhm*/

Translation:

The conversation is open for all the cultured with their different interests[3]

In spite of being essential for many tasks, semantic features are usually understudied, especially for such languages as Arabic. To the best of the authors' knowledge, there are only two NLP systems that deal with Arabic semantic features: AraMorph (Buckwalter 2002) and MADA (Habash and Rambow 2005). Moreover, they are not included in current Arabic ontologies such as Arabic WordNet (Elkateb et al. 2006).

As a result, this paper presents a cue-based algorithm that uses both bilingual and monolingual cues to build a lexicon whose entries are enriched with semantic features. As a proof-of-concept, the paper focuses on Arabic nouns and some of their semantic features such as gender, number and rationality. The rest of the paper falls in four parts: the first outlines related work to Arabic semantic features and cue-based bootstrapping, the second discusses the cue-based algorithm, the third outlines the evaluation methodologies and the last highlights future work.

2. Related Work

2.1. Arabic Natural Language Processing Systems and Arabic Semantic Features

To the best of the authors' knowledge, there are two Arabic Natural Language Processing (ANLP) systems that deal with Arabic semantic features. These systems are AraMorph (Buckwalter 2002) and MADA (Habash and Rambow 2005) which are briefly discussed in the following subsections.

2.1.1. AraMorph (Buckwalter 2002)

Buckwalter's AraMorph (2002) deals with the semantic features of gender and number only. It marks them only when they are morphologically marked; that is, when they are indicated by a gender and/or number suffix.

Arabic has the set of four gender-marking suffixes and a set of five number-marking suffixes which are outlined in table (1) below.

[3] Translation is the authors'.

Gender-Marking Suffixes		
The Suffix	**The Semantic Feature indicated**	**Example**
ة /*p*/	–MALE	طالبة /*TAlbp*/ (a female student)
ون /*wn*/	+MALE	محامون /*mHAmwn*/ (male lawyers; in the nominative case)
ين /*yn*/	+MALE	محامين /*mHAmyn*/ (male lawyers, in the genitive case)
ات /*At*/	–MALE	طالبات /*TAlbAt*/ (female students)
Number-Marking Suffixes		
ة /*p*/	–PLURAL	طبيبة /*Tbybp*/ (a doctor)
ون /*wn*/	+PLURAL	صحفيون /*SHfywn*/ (journalists; in the nominative case)
ين /*yn*/	+PLURAL	صحفيين /*SHfyyn*/ (journalists; in the genitive case)
ات /*At*/	+PLURAL	طالبات /*TAlbAt*/ (female students)
ان /*An*/	+DUAL	طالبان /*TAlbAn*/ (two students; in the nominative case)
ين /*yn*/	+DUAL	طالبين /*TAlbyn*/ (two students; in the genitive case)

Table (1): Gender and Number Suffixes in the Arabic Language

Since Buckwalter's AraMorph (2002) tags the gender and number features of the words based on their suffixes, it manages to tag only 13% of the nouns in a 3000-word corpus and 35.5% of a 20-million-word corpus.

2.1.2. MADA (Habash and Rambow 2005)

Like AraMorph (Buckwalter 2002), the Morphological Analysis and Disambiguation (MADA) tool of Habash and Rambow (2005) deals only with the semantic features of gender and number which are used among other morphosyntactic features to disambiguate morphologically ambiguous words. The semantic features of gender and number are extracted from the output of Aragen (Habash 2004) which tags gender and number features only in the case that they are morphologically marked. The two semantic features of gender and number achieve an accuracy rate of 98.8% in the output of MADA (Habash and Rambow 2004). However, to the best of the authors' knowledge, there is no clear information concerning their recall rate.

2.2. Cue-Based Bootstrapping

Bootstrapping is "the process of attaining new knowledge on the basis of already existing knowledge" (Elghamry 2004: 31). It typically relies on *cues* which represent the initial knowledge that starts the knowledge acquisition process. Cue-based bootstrapping is used to classify rhetorical relation in English texts (Sporleder and Lascarides 2005), to acquire English verb subcategorization frames (Elghamry 2004) among other functions.

In ANLP, cue-based bootstrapping is used both monolingually and bilingually (Darwish and Oard 2002, Diab et al. 2004). Bilingual bootstrapping refers to acquiring knowledge using the cues of a second language (here English). Monolingual cue-based bootstrapping relies directly on cues extracted from the target language itself (here Arabic). Diab (2004) uses cues from parallel corpora and the English WordNet (Miller 2005) to bootstrap and Arabic WordNet. She finds that 52.3% of the Arabic nouns, verbs and adjectives correspond to the definitions of the English WordNet. Similarly, Darwish and Oard (2002) use cues from parallel corpora and translation lists to build translation probability tables for Arabic-in-English translation and vice versa.

3. The Cue-Based Algorithm

The algorithm uses both bilingual and monolingual cues to bootstrap a semantic-features lexicon, whose entries are extracted from the web documents. The algorithm informally works as follows:

1. Using bilingual cues[4] (here English cues) to bootstrap English words with the relevant semantic features from the web documents.
2. Translating the English words into Arabic using Machine Translation (MT) systems.
3. Validating the translated Arabic words using an Arabic corpus and a set of Arabic cues. Meanwhile, using the Arabic cues to enlarge the lexicon.
4. Only the words that are validated are added to the lexicon.

The following subsections discuss in detail each step and highlight its relevant results.

3.1. Bilingual Cues

Bilingual cues are divided into two categories: syntactic and lexical cues. Syntactic cues are based on English function words that are indicative of some semantic features such as *number* and *rationality*. These words are summarized in table (2).

[4] All monolingual and bilingual used are scholarly fed by the authors.

English Cues	Their Semantic Features	Example[5]
An/A This/That Every/Each/No	Followed by –PLURAL nouns	How can a *girl* make her voice sound like a *boy*'s? ... *girl* and *boy* are –PLURAL
... which is/was ... ... who is/was .... is/was	Preceded by –PLURAL nouns	You are on heavy *ground* which is saturated with water. …. *ground* is –PLURAL
... which are/were ... ... who are/were ... ... are/were	Preceded by +PLURAL nouns	What are some natural *resources* which are now being non-renewable? … *resources* is +PLURAL
These/Those Many/Few Numbers	Followed by +PLURAL nouns	Please follow these *directions* to submit a … … *directions* are +PLURAL
... which is/was/are/were ...	Preceded by –RATIONAL	American fighters established their own *rules* which were few … *rules* is –RATIONAL
... who is/was/are/were ...	Preceded by +RATIONAL	Visas are offered to *people* who are going on business or social visits. … *people* is +RATIONAL

Table (2): English Function Words Used as Bilingual Cues for Semantic Features Acquisition

In order for these cues to have a good recall rate, the authors used the web as corpus being a free, instantly available source of immense amounts of documents, representing almost all possible languages and genres (Kilgarriff and Grefenstette 2003). Two search engines are used to search the web documents; these engines are discussed in table (3).

[5] All examples in table (2) are extracted from www.answers.com

The Search Engine	Description
www.answers.com	It aggregates dictionary and encyclopedia content from more than 100 sources in all fields such as Wikipedia and Computer Desktop Encyclopedia[6].
www.search.com	It searches Google, Ask.com, LookSmart and dozens of other leading search engines[7].

Table (3): Search Engines Used to Extract the Lexicon Entries from the Web Documents

The phase of bilingual cues results in the following lists of English words:

The Semantic Feature	Its Variations	Total Number of Words
Number	**Singular**	8,628
	Plural	4,132
Rationality	**Rational**	613
	Irrational	1000

Table (4): Output Lists of Bilingual Cues

3.2. Translating the Extracted Words into Arabic

The output English lists that resulted from bilingual cues are submitted to English-Arabic MT systems. Two publicly available MT systems are used to avoid bias to the most common sense of the word. Table (5) briefly reviews each MT system.

The MT System	Description
Google Translation Tool	A Statistical MT system based on the state-of-the-art technology and is publicly available through: www.google.com
Golden Al-Wafi Translator	A dictionary-based MT system that makes use of Arabic–English general and specialized dictionaries

Table (5): The MT Systems Used to Translate the Cue-Based Extracted English Words

The two MT systems translate ~ 80% of the English lists whose details are shown in table (6).

[6] Source: Online Document. Accessed 9 Oct. 2007. URL: www.pcmag.com.

[7] Source: homepage of www.search.com. Accessed: 9 Oct. 2007.

The Semantic Feature	Its Variations	Total Number of Words after Translation
Number	**Singular**	6,902
	Plural	3,298
Rationality	**Rational**	510
	Irrational	800

Table (6): The Translated Lists

3.3. Validating and Expanding Translated Words

English and Arabic are typologically different languages. The semantic features of a word in one language may be different from the semantic features of the same word in the other language. For example, *information* is an uncountable noun in English, but it is countable in Arabic with its singular form being معلومة /*mElwmp*/ (a piece of information) and its plural form being معلومات /*mElwmAt*/ (pieces of information). Therefore, Arabic translated words are to be validated against an Arabic corpus using a set of Arabic cues. Not only are Arabic cues used for validation, but also they are used to expand the semantic features lists and to add a new semantic feature to the entries of the lexicon, namely, *gender*.

Arabic cues used are both syntactic and lexical. Syntactic cues – outlined in table (7) – are based on Arabic relative pronouns, demonstratives and coordination tools.

Arabic Cue	Cue Type	Semantic Features	Example[8]
هذا /*h*A*/ (this) ذلك /**lk*/ (that)	Demonstrative	–PLURAL +MALE	وقال ان *هذا الفتى* يسرق ... */wqAl An h*A AlftY ysrq/* (and he said that *this boy* steals) ... الفتى /*AlftY*/ (the boy) is –PLURAL and +MALE
هذه /*h*h*/ (this) تلك /*tlk*/ (that)	Demonstrative	–MALE	ماذا فعلت *تلك الفتاة* فى المطار؟ */mA*A fElt tlk AlftAp?/* (What did *that girl* do?) ... الفتاة /*AlftAp*/ (the girl) is –MALE
هذان /*h*An*/ (these) هذين /*h*yn*/ (these)	Demonstrative	+DUAL +MALE	*هذان النظامان* الشريران. */h*An AlnZAmAn Al$ryrAn/* (These *two evil systems*) ... النظامان /*AlnZAmAn*/ (the two systems) is +DUAL and +MALE
هاتان /*hAtAn*/ (these) هاتين /*hAtyn*/ (these)	Demonstrative	+DUAL –MALE	*هاتين العائلتين* المتنافستين */hAtyn AlEA}ltyn AlmtnAfstyn/* (These *two competing families*) … العائلتين /*AlEA}ltyn*/ (the two families) is +DUAL and –MALE

[8] All examples in table (2) are extracted from www.answers.com.

هؤلاء /*h&lA'*/ (these)	Demonstrative	+PLURAL	*هؤلاء القوم* ... /*h&lA' Alqwm*/ (*these people*) ... القوم /*Alqwm*/ (the people) is +PLURAL
أولئك /*>wl}k*/ (those)	Demonstrative	+PLURAL +MALE	*أولئك الأطفال* الذين ... /*>wl}k Al>TfAl Al*yn*/ (*Those children* who ...) ... الأطفال /*Al>TfAl*/ (children) is +PLURAL and +MALE
الذي /*Al*y*/ (who/which)	Relative Pronoun	–PLURAL +MALE	*الشخص الذي* يستخدم السحر ... /*Al$xS Al*y ystxdm AlsHr*/ (The *person who* uses magic) ... الشخص /*Al$xS*/ (the person) is –PLURAL and +MALE
التي /*Alty*/ (who/which)	Relative Pronoun	–MALE	تابع الكثيرون *الحملة التي* بدأها ... /*TAbE Alkvyrwn AlHmlp Alty bd>hA*/ (Many have followed up *the campaign which* was launched by …) … الحملة /*AlHmlp*/ (the campaign) is –MALE
اللذان /*All*An*/ (who/which) اللذين /*All*yn*/ (who/which)	Relative Pronoun	+DUAL +MALE	*الجنديان اللذان* خطفهما ... /*AljndyAn All*An xTfhmA*/ (The *two soliders who* were kidnapped) ... الجنديان /*AljndyAn*/ (the two soliders) is +DUAL and +MALE
اللتان /*AlltAn*/ (who/which) اللتين /*Alltyn*/ (who/which)	Relative Pronoun	+DUAL –MALE	وصول *الطائرتين اللتين* تقلان ... /*wSwl AlTA}rtyn Alltyn tqlAn ...*/ (The arrival of *the two airplanes which* carry ...) ... الطئرتين /*AlTA}rtyn*/ (the two airplanes) is +DUAL and –MALE
الذين /*Al*yn*/ (who/which)	Relative Pronoun	+PLURAL +MALE +RATIONAL	*أسطورة الرجال* الذين ... /*>sTwrp AlrjAl Al*yn ...*/ (The legend of *the men who* ...) ... الرجال /*AlrjAl*/ (men) is +PLURAL, +MALE and +RATIONAL

Table (7): Arabic Cues Used for Gender and Number Semantic Features

Lexical cues include a set of Arabic verbs which are typically used followed by a +RATIONAL. These verbs are as follows:

The Verb	Meaning
ذكر /*kr/	Mention
صرح /SrH/	Declare
أعلن />Eln/	Announce
قال /qAl/	Say
زعم /zEm/	Claim
ناقش /nAq$/	Discuss
قدم /qdm/	Present
أوضح />wDH/	Clarify
عرف /Erf/	Know
وصف /wSf/	Describe
عرض /ErD/	Show
اعتبر /AEtbr/	Consider

Table (8): Indicating Arabic Verbs for the Rationality Semantic Feature

The validation and expansion phase results in the following final lists:

The Semantic Feature	Its Variations	Total Number of Words
Gender	**Feminine**	16,370
	Masculine	18,289
Number	**Singular**	26,401
	Plural	7,935
Rationality	**Rational**	40,21
	Irrational	20,355

Table (9): Final Lists of Semantic Features

What follows is a complete example for the cue-based algorithm:

- Searching the web using the aforementioned English cues results in '*a boy*' that is tagged as –PLURAL since it follows the article '*a*'.
- The output word '*boy*' is submitted to Google MT systems which translates it as فتى /*ftY*/ (boy) and to Golden Al-Wafi which translates is as ولد /*wld*/ (boy).
- Both فتي /*ftY*/ and ولد /*wld*/ are considered as potential –PLURAL Arabic nouns.
- The two nouns are validated using the aforementioned Arabic cues. The search engine www.answers.com yields 25,800 hits for هذا الفتى /*h*A AlftY*/ (this boy) and 28,000 hits for هذا الولد /*h*A Alwld*/ (this boy). The other search engine – www.search.com – gives 10,420 hits for هذا الفتى /*h*A AlftY*/ (this boy) and 12,520 hits for هذا الولد /*h*A Alwld*/ (this boy).
- Therefore, both الفتى /*AlftY*/ and الولد /*Alwld*/ are added to the lexicon and are tagged as –PLURAL Arabic nouns.

4. Evaluation

The semantic features lexicon is meant as a lexical resource for ANLP applications. Consequently, two evaluation methodologies are used: the first is based on a gold standard set to evaluate the lexicon on its own, whereas the second evaluated the lexicon against an ANLP task, namely AR.

4.1. Gold Standard Evaluation

A 3000-word gold standard set is built by the authors in order to evaluate the lexicon as a lexical resource on its own. According to the gold standard evaluation, the lexicon achieves a recall rate of 85% and a precision rate of 95% and thus an F-measured performance rate of ~ 89.7%.

4.2. Task-Based Evaluation

Since semantic features are used for many NLP tasks, the lexicon is integrated with an AR statistical algorithm (Al-Sabbagh 2007) and manages to improve the performance rate by 13% and increases it from 74.4% to 87.4%.

5. Conclusion and Future Work

This paper presented a cue-based algorithm for Arabic semantic features acquisition with a performance rate of 87.7%. The resulting lexicon improves performance rate for some ANLP tasks such as AR by 13%. The contributions of this paper are:

- Dealing with a new Arabic semantic feature that has not been dealt with before; that is, rationality
- Highlighting the possibility of bilingual bootstrapping of Arabic semantic features
- Using the web as corpus to provide immense corpora for cue-based bootstrapping

For future work, the authors are adding more features such as animacy and abstraction. Moreover, they are expanding the gold standard set and are using new search engines which are mainly designed for Arabic such as www.ayn.com.

References

Al-Sabbagh R. (2007). *Pronominal Anaphora Resolution in Arabic English Machine Translation Systems*. Unpublished MA Thesis: Forth coming. Ain Shams University, Egypt.

Buckwalter T. (2002). *Buckwalter Arabic Morphological Analyzer. Version 1.0*. LDC Catalog No. LDC2002L49, ISBN 1-58563-257-0.

Darwish K. and Oard D. (2002). CLIR Experiments at Maryland for TREC 2002: Evidence Combination for Arabic-English Retrieval. *Proceedings of CLIR.*

Diab M., Hacioglu K. and Jurafsky D. (2004). Automatic Tagging of Arabic Text: from Raw Text to Base Phrase Chunks. In Dumas, S., Marcus, D. and Roukos, S. (Eds.). *HLT-NAACL 2004: Short Papers* (pp.140-152). Boston: Association for Computational Linguistics.

Elghamry K. (2004). *A Generalized Cue Based Approach to the Automatic Acquisition of Subcategorization Frames*. PhD Thesis. Department of Linguistics, Indiana University.

Elkateb S., Black W., Rodriguez H., Al-Khalifa M., Vossen P., Pease A. and Fellbaum C. (2006). Building a WordNet for Arabic. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

Habash N. and Rambow O. (2005). Arabic Tokenization, Morphological Analysis and Part-of-Speech Tagging in One Fell Swoop. *Proceeding of the Conference of American Association for Computational Linguistics (ACL'05)*, 573-580.

Habash N. (2004). Large Scale Lexeme Based Arabic Morphological Generation. *Proceedings of JEP-TALN 2004, Session Traitement Automatique de l'Arabe*.

Hartrumpf S., Helbig H. and Osswald R. (2006). Semantic Interpretation of Prepositions for NLP Applications*. Proceedings of the 3rd ACM-SIGSEM Workshop on Prepositions*, Trento, Italy, 29-37.

Kilgarriff and Grefenstette. (2003). Web as Corpus. *Computational Linguistics.* 29: 3. 333-347.

Lappin S. and Leass H. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics,* No.20, 535-561.

Miller G. (2005). WordNet: A Lexical Database of the English Language. Online URL: http://wordnet.princeton.edu/. Accessed: 24 October 2007.

Silzer P. (2005). *Working with Language: An Interactive Guide to Understanding Language and Linguistics*. Supplementary Course Material for the Department of TESOL and Applied Linguistics, Biola University, California, USA.

Sporleder C. and Lascarides A. (2005). Using Automatically Labeled Examples to Classify Rhetorical Relations: An Assessment. *Natural Language Engineering.* Vol. 1.

Turney P. (2004). Word Sense Disambiguation by Web Mining for Word Co-occurrence Probabilities. *Proceedings of the 3rd International Workshop on the Evaluation of the Semantic Analysis of Text (SENSEVAL-3)*, Barcelona, Spain, 239-242.