Revisiting Arabic Semantic Role Labeling using SVM Kernel Methods

Laurel Hart, Hassan Alam, Aman Kumar
BCL Technologies, San Jose, California, USA
{lhart, hassana, amank}@bcltechnologies.com

ABSTRACT

As a critical language, there is huge potential for the usefulness of an Arabic Semantic Role Labeling (SRL) system. This task involves two subtasks: predicate argument boundary detection and argument classification. Based on the innovations of Diab, Moschitti, and Pighin (2007) in the field of Arabic Natural Language Processing (NLP), SRL in particular, we are currently developing a system for automatic SRL in Arabic.

KEYWORDS: Arabic, semantic role labeling, SRL, predicate argument, boundary detection, argument classification.

Proceedings of COLING 2012: Demonstration Papers, pages 207–214,
COLING 2012, Mumbai, December 2012.

1 Introduction

The automatic detection and identification of semantic roles in a sentence—a process known as Semantic Role Labeling (SRL)—has many potential applications within computational linguistics. Imagine the uses for improving machine translation, information extraction, and document analysis, among other innovations. As the computational linguistics field has expanded, so has the amount of research into SRL. However, as with much language technology research, the main focus has been on English. Because of this, Arabic-language [1] technologies and methods are often adapted from tools that have succeeded for English, rather than developed on their own. Recent years have produced powerful development resources, such as an Arabic Treebank and Propbank, in both pilot and revised forms. The number of resources available for automatic parsing, POS -tagging, chunking, of Arabic still lags behind that of English, but has grown noticeably. As a critical language, there is huge potential for an Arabic SRL system to revolutionize Arabic-language tools.

Based on the innovations of Diab, Moschitti, and Pighin (2007) in Arabic SRL, we are currently developing a system for automatic SRL in Arabic. We are looking for feedback from the conference before fully implementing and reporting the performance of this system.

2 Arabic NLP

Arabic has a number of challenges which aren't present for other languages. It is unlike English in many ways, which suggests that directly applying English- language technology many not be the absolute optimal approach for an effective system. On the other hand, there is no reason not to utilize the work that has been done on SRL if it can be used in a cross- linguistic way. The best approach will be to build upon previous work and customize it to Arabic linguistic features, thus using the differences between English and Arabic to advantage.

One such feature is Arabic's rich morphology. In Arabic, nouns and adjectives encode information about number, gender, case, and definiteness. Verbs are even richer, encoding tense, voice, mood, gender, number, and person. These features are often expressed via diacritics, short vowels marked above or below letters [2].

A word in Arabic is typically formed by selecting one of approximately 5,000, 3-to-5-consonant roots, and adding affixes. [3] Diacritics are sometimes the only thing that specifies semantic differences between words, and for certain genres, especially online communication, they are often left out. In this case, it is a special challenge for an automated system to determine the difference.

Another feature is word order. While not completely free, Arabic allows for subject-verb- object (SVO, as in English), verb- subject-object (VSO), and more occasionally OSV and OVS. In the Arabic Treebank, SVO and VSO each equally account for 35% of the sentences [4].

[1] In this paper, the term “Arabic” is used to refer to Modern Standard Arabic. Other dialects will be specifically noted as such.
[2] Diab et al., 2007.
[3] Abbasi and Chen, 2005.
[4] Palmer et al., 2008.

Arabic allows for noun phrases which are more complex than those in English, particularly the possessive construction called *idafa*, which relies on the definiteness of the nouns to convey precise meaning.

Another feature worth mentioning, but not currently handled by this system, is pro-drop. Because nouns, adjectives, and verbs encode so much information, it is fairly common to completely drop the subject of a sentence because it is implied. An example of this given in (Palmer et al., 2008) is: *Akl AlbrtqAl* 'ate-[he] the-oranges.' In this context, the verb *Akl,* 'ate,' expresses the subject 'he,' which is not directly said. (Palmer et al., 2008) notes that 30% of the sentences in the Arabic Treebank are pro-dropped.

3 Related Work

One of the things that challenges Arabic NLP is that much progress in computational linguistics is focused on English. As a result, many Arabic language technologies are based on research done on English, and then revised to better fit the characteristics of Arabic. This is not necessarily detrimental, but it does affect the development process. This is the route taken in (Diab et al., 2007) : adapting techniques which have proven successful for English SRL for use in an Arabic system. Specifically, Moschitti's SVM-light-TK is trained and tested on Arabic data, using features which have proven effective for English and some other languages, which are referred to as the "standard set". The features were grouped into:

a) Phrase Type, Predicate Word, Head Word, Position and Voice, based on (Gildea and Jurafsky 2002);
b) Partial Path, No Direction Path, Head Word POS, First and Last Word/POS in Constituent and SubCategorization based on (Pradhan et al., 2003);
c) Syntactic Frame, based on (Xue and Palmer, 2004)

(Diab et al., 2008) extends the first, rudimentary system by tailoring it to Arabic- specific features. This tailoring manifests in the form of feature selection for the SVM. The new, Arabic-specific features consist of inflectional morphology, including number, gender, definiteness, mood, case, person; derivational morphology including lemma form of the words with explicitly-marked diacritics; the English gloss; vocalized form with full diacritics (much like lemma but including inflections); and unvowelized word as it appears in the sentence. Tree kernels were specifically chosen for the initial set of experiments in order to be able to deal with the immense set of possible features. Adding the Arabic-specific features was shown to significantly improve performance.

4 Experiment Design

4.1 Machine Learning Algorithm

As in (Diab et al., 2007) and (Diab et al., 2008), this system will be using SVM-light-TK. The SVM algorithm has been shown in numerous studies to handle noisy data and large feature sets well. Using Moschitti's Tree Kernel SVM will allow for an extensible system, better suited to the addition of Arabic features. For the present, however, mostly the standard SVM-light capability will be utilized with a polynomial kernel.

4.2 Data

The SVM will be trained on annotated data collected from news-oriented, Arabic-language blogs. For initial testing, the corpus is relatively small, with just over 100 sentences. These were run through Benajiba and Diab's AMIRA 2.0 POS tagger and BP chunk parser. They were then annotated with ARG0 and ARG1 in PropBank-like style, with adverbial and prepositional (ARGM) phrases annotated for later use. For this implementation, only ARG0 and ARG1 will be labeled. All of the sentences in this corpus included explicit subjects rather than pro-drop. Additionally, most of the sentences were of SVO form, with very little variation.

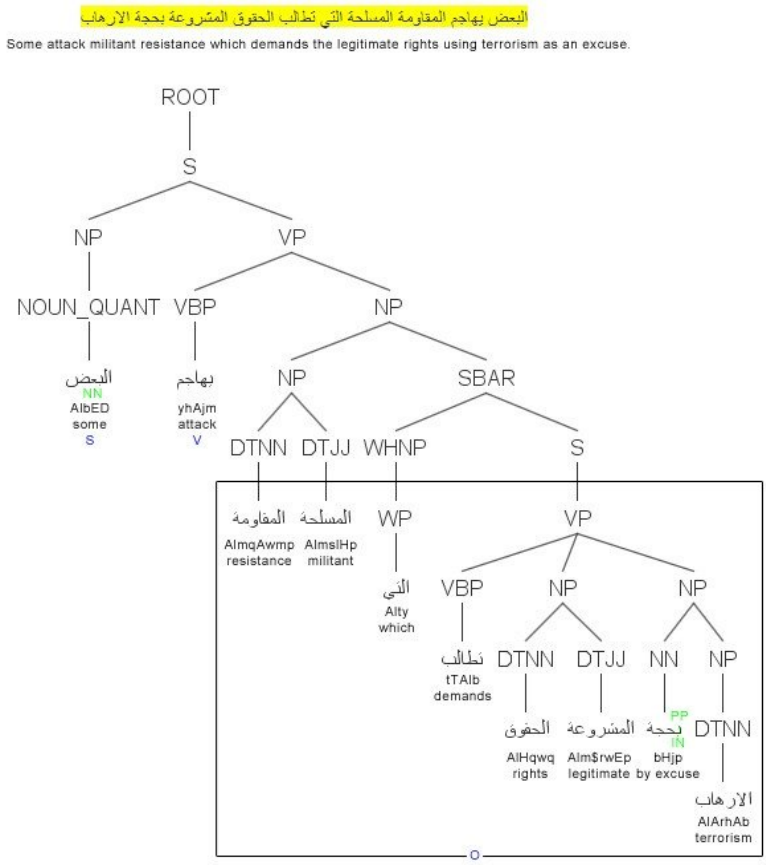

FIGURE 1 – A syntactic parse tree created using the Stanford Parser (factored Arabic grammar), then marked up to show semantic roles. “S” denotes the subject (ARG0), “V” denotes the predicate, “O” denotes the object (ARG1).

4.3 Predicate Argument Extraction and Argument Classification

Any SRL system involves extracting and labeling predicate structures. At the sentence level, this is constituted of two tasks: detecting the word span of arguments within the sentence, and classifying the arguments found by type (ARG0, ARG1, ARGM). (Diab, 2007 et al.) describes the general algorithm for doing so by the steps:

1. given a sentence from the *training-set*, generate a full syntactic parse-tree;
2. let *P* and *A* be the set of predicates and the set of parse-tree nodes (i.e. the potential arguments), respectively;
3. for each pair <*p, a*> ∈ *P* × *A*:

- Extract the feature representation set,;
- If the subtree rooted in *a* covers exactly the words of one argument of *p*, put $F_{p,a}$ in T^+ (positive examples), otherwise put it in T^- (negative examples).

The T^+ and T^- sets then serve to train the boundary classifier. With some restructuring, T^+ and T^-can also be used to train the argument classifier.

4.4 Features

The aspect that most distinguishes a system built for Arabic from one built for English is the selection of features. As mentioned above, (Diab et al., 2007)'s SRL system initially used the set of standard features before later adding Arabic-specific ones. Similarly, this system will first be developed using the standard set before experimenting with Arabic features. Following are brief descriptions of the feature types to be included.

- *Phrase Type*: The syntactic category of the phrase expressing the semantic roles. (Gildea and Jurafsky 2002)
- *Predicate Word*: Lemma form of the predicate word.
- *Head Word*: Syntactic head of phrase. (Pradhan et al., 2003)
- *Position*: Position of constituent relative to predicate (before or after).
- *Voice* : Active or passive classification of a sentence. This is included due to correspondence between active voice and subject, passive voice and object.
- *Path*: Syntactic path linking the argument and its predicate. For example, the path of ARG0 in Figure 1 is NQ↑NP↑S↓VP↓VBP.
- *Partial Path*: Section of the path that connects argument to the common parent of the predicate.
- *No Direction Path*: Path without directions.
- *Head Word POS*: Syntactic part of speech of the head word.
- *First and Last Word/POS in Constituent*: First and last words and part of speech of phrase.
- *SubCategorization*: Production rule expanding the predicate parent node. (Diab et al., 2008)
- *Syntactic Frame*: Noun phrase positions relative to the predicate.

5 Expected Results

Using the official CoNLL evaluator, (Diab et al., 2007)'s initial system was able to achieve overall F1 scores of 77.85 and 81.43 on classifying the arguments of the development and testing sets, respectively. Boundary detection results were also quite impressive, with F1 scores of 93.68 and 94.06. These results were yielded by use of only the standard features listed above. By adding in Arabic-specific features, utilizing tree kernels, and testing across a variety of models, (Diab,

2008 et al.) were able to increase the F1 score for automated boundary detection and argument classification to 82.17.

By drawing on previous work such as that of (Diab et al., 2007; Diab et al., 2008), we hope to achieve similar measures, possibly even improving upon them by applying research performed after the publication of (Diab et al., 2008).

Conclusion and future work

Many of the next steps for expanding the system are quite clear, as the system has not been fully implemented yet.

In the future, more types of arguments will be labeled. This will be done by comparing the results of multiple 1-vs-ALL passes through the SVM trained for different argument types and selecting the highest score.

The system will also be tested on a larger, more representative corpus that possesses sentences exhibiting pro-dropping and more word-order variation. For our purposes, the SRL system should be able to detect these patterns.

During the development of the Arabic SRL system, we will continue to tailor it specifically to Arabic, and make more use of its unique linguistic features.

Constructive feedback on the design of this SRL system is welcomed.

References

Abbasi, Ahmed, and Hsinchun Chen. "Applying Authorship Analysis to Extremist Group Web Forum Messages." *IEEE Intelligent Systems, Special Issue on Artificial Intelligence for National and Homeland Security* Sept. (2005): 67-75.

Benajiba, Yassine, and Mona Diab. *AMIRA 2.0*. Columbia University. Web. 2012. <http://nlp.ldeo.columbia.edu/amira/>.

Diab, Mona, Alessandro Moschitti, and Daniele Pighin. "CUNIT: A Semantic Role Labeling System for Modern Standard Arabic." *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)* June (2007): 133-36. Web.

Diab, Mona, Alessandro Moschitti, and Daniele Pighin. "Semantic Role Labeling Systems for Arabic using Kernel Methods."*Proceedings of ACL-08: HLT* June (2008): 798-806.

Gildea, Daniel, and Daniel Jurafsky. "Automatic Labeling of Semantic Roles." *Computational Linguistics* 28.3 (2002).

Green, Spence, and Christopher D. Manning. "Better Arabic Parsing: Baselines, Evaluations, and Analysis." *COLING 2010* (2010).

Joachims, Thorsten. "SVM^{light}." Cornell University, 14 Aug. 2008. Web. 2012.

Moschitti, Alessandro. "Tree Kernels in SVM-Light." University of Trento, Italy. Web. 2012. <http://disi.unitn.it/moschitti/Tree-Kernel.htm>.

Moschitti, Alessandro. "Making Tree Kernels practical for Natural Language Learning." *Proceedings of the Eleventh International Conference on European Association for Computational Linguistics* (2006). Print.

Palmer, Martha, Olga Babko-Malaya, Ann Bies, Mona Diab, Mohammed Maamouri, Aous Mansouri, and Wajdi Zaghouani. "A Pilot Arabic Propbank." (2008).

Pradhan, Sameer, Kadri Hacioglu, Wayne Ward, James H. Martin, and Daniel Jurafsky. "Semantic Role Parsing: Adding Semantic Structure to Unstructured Text." *Proceedings of ICDM-2003* (2003).

Xue, Nianwen, and Martha Palmer. "Calibrating Features for Semantic Role Labeling." (2004).

W. Zaghouani , M. Diab , A. Mansouri , S. Pradhan , M. Palmer, “The Revised Arabic PropBank,” in: *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*, Uppsala, Sweden, 15-16 July 2010, pp. 222–226.

Zaghouani, Wajdi, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. "The Revised Arabic PropBank." *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010* 15 July (2012): 222-26.