The Revised Arabic PropBank

Wajdi Zaghouani[♣] , Mona Diab[♠] , Aous Mansouri[‡], Sameer Pradhan[◊] and Martha Palmer[‡]

[♣]Linguistic Data Consortium, [♠]Columbia University,
[‡]University of Colorado, [◊]BBN Technologies

wajdiz@ldc.upenn.edu, mdiab@ccls.columbia.edu, aous.mansouri@colorado.edu,
pradhan@bbn.com, martha.palmer@colorado.edu

Abstract

The revised Arabic PropBank (APB) reflects a number of changes to the data and the process of PropBanking. Several changes stem from Treebank revisions. An automatic process was put in place to map existing annotation to the new trees. We have revised the original 493 Frame Files from the Pilot APB and added 1462 new files for a total of 1955 Frame Files with 2446 framesets. In addition to a heightened attention to sense distinctions this cycle includes a greater attempt to address complicated predicates such as light verb constructions and multi-word expressions. New tools facilitate the data tagging and also simplify frame creation.

1 Introduction

Recent years have witnessed a surge in available automated resources for the Arabic language.[1] These resources can now be exploited by the computational linguistics community with the aim of improving the automatic processing of Arabic. This paper discusses semantic labeling.

Shallow approaches to semantic processing are making large advances in the direction of efficiently and effectively deriving application relevant explicit semantic information from text (Pradhan et al., 2003; Gildea and Palmer, 2002; Pradhan et al., 2004; Gildea and Jurafsky, 2002; Xue and Palmer, 2004; Chen and Rambow, 2003; Carreras and Marquez, 2005; Moschitti, 2004; Moschitti et al., 2005; Diab et al., 2008). Indeed, the existence of semantically annotated resources in English such as FrameNet (Baker et al., 1998) and PropBank (Kingsbury and Palmer, 2003; Palmer et al., 2005) corpora have marked a surge in efficient approaches to automatic semantic labeling of the English language. For example, in the English sentence, 'John enjoys movies', the predicate is 'enjoys' and the first argument, the subject, is 'John', and the second argument, the object, is 'movies'. 'John' would be labeled as the *agent/experiencer* and 'movies' would be the *theme/content.* According to PropBank, 'John' is labeled Arg0 (or enjoyer) and 'movies' is labeled Arg1 (or thing enjoyed). Crucially, that independent of the labeling formalism adopted, the labels do not vary in different syntactic constructions, which is why proposition annotation is different from syntactic Treebank annotation. For instance, if the example above was in the passive voice, 'Movies are enjoyed by John', 'movies' is still the *Theme/Content* (Arg1) and (thing enjoyed), while 'John' remains the *Agent/Experiencer* (Arg0) and (enjoyer). Likewise for the example 'John opened the door' vs. 'The door opened', in both of these examples 'the door' is the *Theme* (Arg1). In addition to English, there are PropBank efforts in Chinese (Xue et al., 2009), Korean (Palmer et al. 2006) and Hindi (Palmer et al., 2009), as well as FrameNet annotations in Chinese, German, Japanese, Spanish and other languages (Hans 2009). Being able to automatically apply this level of analysis to Arabic is clearly a desirable goal, and indeed, we began a pilot Arabic PropBank effort several years ago (Palmer et al., 2008).

In this paper, we present recent work on adapting the original pilot Arabic Proposition Bank (APB) annotation to the recent changes that have been made to the Arabic Treebank (Maamouri et al., 2008). These changes have presented both linguistic and engineering challenges as described in the following sections. In Section 2 we discuss major linguistics changes in the Arabic Treebank annotation, and any impact they might have for the APB effort. In Section 3 we discuss the engineering ramifications of adding and deleting nodes from parse trees, which necessitates mov-

[1] In this paper, we use Arabic to refer to Modern Standard Arabic (MSA).

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 222–226,
Uppsala, Sweden, 15-16 July 2010.

ing all of the APB label pointers to new tree locations. Finally, in Section 4 we discuss the current APB annotation pipeline, which takes into account all of these changes. We conclude with a statement of our current goals for the project.

2 Arabic Treebank Revision and APB

The Arabic syntactic Treebank Part 3 v3.1 was revised according to the new Arabic Treebank Annotation Guidelines. Major changes have affected the NP structure and the classification of verbs with clausal arguments, as well as improvements to the annotation in general.[2]

The Arabic Treebank (ATB) is at the core of the APB annotations. The current revisions have resulted in a more consistent treebank that is closer in its analyses to traditional Arabic grammar. The ATB was revised for two levels of linguistic representation, namely morphological information and syntactic structure. Both of these changes have implications for APB annotations.

The new ATB introduced more consistency in the application of morphological features to POS tags, hence almost all relevant words in the ATB have full morphological features of number, gender, case, mood, and definiteness associated with them. This more comprehensive application has implications on agreement markers between nouns and their modifiers and predicative verbs and their arguments, allowing for more consistent semantic analysis in the APB.

In particular, the new ATB explicitly marks the gerunds in Arabic known as maSAdir (singular maSdar.) MaSAdirs, now annotated as VN, are typically predicative nouns that take arguments that should receive semantic roles. The nouns marked as VN are embedded in a new kind of syntactic S structure headed by a VN and having subject and object arguments similar to verbal arguments. This syntactic structure, namely S-NOM, was present in previous editions/versions of the ATB but it was headed by a regular noun, hence it was difficult to find. This explicit VN annotation allows the APB effort to take these new categories into account as predicates. For instance [تكبد]**VN** [-هم]**ARG0** [خسائر كبيرة]**ARG1**, transliterated as takab~udi-, meaning 'suffered' is an example of predicative nominal together with its semantically annotated arguments ARG0 transliterated as -him, meaning 'they' and ARG1 transliterated as xasA}ira kabiyrap, meaning 'heavy losses'.

Other changes in the ATB include *idafa* constructions (a means of expressing possession) and the addition of a pseudo-verb POS tag for a particular group of particles traditionally known as "the sisters of إنّ <in~a 'indeed' ". These have very little impact on the APB annotation.

3 Revised Treebank processing

One of the challenges that we faced during the process of revising the APB was the transfer of the already existing annotation to the newly revised trees -- especially since APB data encoding is tightly coupled with the explicit tree structure. Some of the ATB changes that affected APB projection from the old pilot effort to the new trees are listed as follows:

i. Changes to the tree structure
ii. Changes to the number of tokens -- both modification (insertion and deletion) of traces and modification to some tokenization
iii. Changes in parts of speech
iv. Changes to sentence breaks

The APB modifications are performed within the OntoNotes project (Hovy et al. 2006), we have direct access to the OntoNotes DB Tool, which we extended to facilitate a smooth transition. The tool is modified to perform a three-step mapping process:

a) De-reference the existing (tree) node-level annotations to the respective token spans;

b) Align the original token spans to the best possible token spans in the revised trees. This was usually straight forward, but sometimes the tokenization affected the boundaries of a span in which case careful heuristics had to be employed to find the correct mapping. We incorporated the standard "diff" utility into the API. A simple space separated token-based diff would not completely align cases where the tokenization had been changed in the new tree. For these cases we had to back-off to a character based alignment to recover the alignments. This two-pass strategy works better than using character-based align-

[2] For a complete description of the new Treebank annotation guidelines, see (Arabic Treebank Morphological and Syntactic Annotation Guidelines 2008) at http://projects.ldc.upenn.edu/ArabicTreebank/.

ment as a default since the diff tool does not have any specific domain-level constraints and gets spurious alignments;

c) Create the PropBank (tree) node-pointers for the revised spans.

As expected, this process is not completely automatic. There are cases where we can deterministically transfer the annotations to the new trees, and other cases (especially ones that involve decision making based on newly added traces) where we cannot. We automatically transferred all the annotation that could be done deterministically, and flagged all the others for human review. These cases were grouped into multiple categories for the convenience of the annotators. Some of the part of speech changes invalidated some existing annotations, and created new predicates to annotate. In the first case, we simply dropped the existing annotations on the affected nodes, and in the latter we just created new pointers to be annotated. We could automatically map roughly 50% of the annotations. The rest are being manually reviewed.

4 Annotation Tools and Pipeline

4.1 Annotation process

APB consists of two major portions: the lexicon resource of Frame Files and the annotated corpus. Hence, the process is divided into framing and annotation (Palmer et al., 2005).

Currently, we have four linguists (framers) creating predicate Frame Files. Using the frame creation tool Cornerstone, a Frame File is created for a specific lemma found in the Arabic Treebank. The information in the Frame File must include the lemma and at least one frameset.

Previously, senses were lumped together into a single frame if they shared the same argument structure. In this effort, however, we are attempting to be more sensitive to the different senses and consequently each unique sense has its own frameset. A frameset contains an English definition, the argument structure for the frameset, a set of (parsed) Arabic examples as an illustration, and it may include Arabic synonyms to further help the annotators with sense disambiguation.

Figure 1 illustrates the Frameset for the verb **استمع isotamaE}** 'to listen'

Predicate: **{isotamaE استمع**
Roleset id: f1, to listen
Arg0: entity listening
Arg1: thing listened

Figure 1. The frameset of the verb {isotamaE

Rel: {isotamaE, استمع
Arg0: -NONE- *
Gloss: He
Arg1: الى مطالبهم
Gloss: to their demands
Example: استمع الى مطالبهم

Figure 2. An example annotation for a sentence containing the verb {isotamaE

In addition to the framers, we also have five native Arabic speakers as annotators on the team, using the annotation tool Jubilee (described below). Treebanked sentences from the ATB are clearly displayed in Jubilee, as well as the raw text for that sentence at the bottom of the screen. The verb that needs to be tagged is clearly marked on the tree for the annotators. A drop-down menu is available for the annotators to use so that they may choose a particular frameset for that specific instance. Once a frameset is chosen the argument structure will be displayed for them to see. As a visual aid, the annotators may also click on the "example" button in order to see the examples for that particular frameset. Finally, the complements of the predicate are tagged directly on the tree, and the annotators may move on to the next sentence. Figure 2 illustrates a sample annotation.

Once the data has been double-blind annotated, the adjudication process begins. An adjudicator, a member of the framing team, provides the Gold Standard annotation by going over the tagged instances to settle any differences in the choices. Occasionally a verb will be mis-lemmatized (e.g. the instance may actually be **سَهَّل sah~al** 'to cause to become easy' but it is lemmatized under سَهُل sahul-u 'to be easy' which looks identical without vocalization.) At this point the lemmas are corrected and sent back to the annotators to tag before the adjudicators can complete their work.

The framers and annotators meet regularly at least every fortnight. These meetings are important for the framers since they may need to convey to the annotators any changes or issues with the frames, syntactic matters, or anything else that may require extra training or preparation for the annotators. It is important to note that while the framers are linguists, the annotators are not. This means that the annotators must be instructed on a number of things including, but not limited to, how to read trees, and what forms a constituent, as well as how to get familiar with the tools in order to start annotating the data. Therefore, little touches, such as the addition of Arabic synonyms to the framesets (especially since not all of the annotators have the same level of fluency in English), or confronting specific linguistic phenomena via multiple modalities are a necessary part of the process. To these meetings, the annotators mostly bring their questions and concerns about the data they are working on. We rely heavily on the annotator's language skills. They take note of whether a frame appears to be incorrect, is missing an argument, or is missing a sense. And since they go through every instance in the data, annotators are instrumental for pointing out any errors the ATB. Since everything is discussed together as a group people frequently benefit from the conversations and issues that are raised. These bi-monthly meetings not only help maintain a certain level of quality control but establish a feeling of cohesion in the group.

The APB has decided to thoroughly tackle light verb constructions and multi-word expressions as part of an effort to facilitate mapping between the different languages that are being PropBanked. In the process of setting this up a number of challenges have surfaced which include: how can we cross-linguistically approach these phenomena in a (semi) integrated manner, how to identify one construction from the other, figuring out a language specific reliable diagnostic test, and whether we deal with these constructions as a whole unit or as separate parts; and how? (Hwang, et al., 2010)

4.2 Tools

Frameset files are created in an XML format. During the Pilot Propbank project these files were created manually by editing the XML file related to a particular predicate. This proved to be time consuming and prone to many formatting errors**.** The Frame File creation for the revised APB is now performed with the recently developed Cornerstone tool (Choi et al., 2010a), which is a PropBank frameset editor that allows the creation and editing of Propbank framesets without requiring any prior knowledge of XML.

Moreover, the annotation is now performed by Jubilee, a new annotation tool, which has improved the annotation process by displaying several types of relevant syntactic and semantic information at the same time. Having everything displayed helps the annotator quickly absorb and apply the necessary syntactic and semantic information pertinent to each predicate for consistent and efficient annotation (Choi et al., 20010b). Both tools are available as Open Source tools on Google code.[3]

4.3 Current Annotation Status and Goals

We have currently created 1955 verb predicate Frame Files which correspond to 2446 framesets, since one verb predicate Frame File can contain one or more framesets. We will reconcile the previous Arabic PropBank with the new Treebank and create an additional 3000 Frame files to cover the rest of the ATB3 verb types.

5 Conclusion

This paper describes the recently revived and revised APB. The changes in the ATB have affected the APB in two fundamentally different ways. More fine-grained POS tags facilitate the tasks of labeling predicate argument structures. However, all of the tokenization changes have rendered the old pointers obsolete, and new pointers to the new constituent boundaries have to be supplied. This task is underway, as well as the task of creating several thousand additional Frame Files to complete predicate coverage of ATB3.

Acknowledgments

We gratefully acknowledge a grant from the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, subcontract from BBN, Inc. We also thank Abdel-Aati Hawwary and Maha Saliba Foster and our annotators for their invaluable contributions to this project.

References

Boas, Hans C. 2009. Multilingual FrameNets. In Computational Lexicography: Methods and Applications. Berlin: Mouton de Gruyter. pp. x+352

Carreras, Xavier & Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*, Ann Arbor, MI, USA.

[3] http://code.google.com/p/propbank/

Chen, John & Owen Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan.

Choi, Jinho D., Claire Bonial, & Martha Palmer. 2010a. Propbank Instance Annotation Guidelines Using a Dedicated Editor,Cornerstone. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*,Valletta, Malta.

Choi, Jinho D., Claire Bonial, & Martha Palmer. 2010b. Propbank Instance Annotation Guidelines Using a Dedicated Editor,Jubilee. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10),*Valletta, Malta.

Diab, Mona, Alessandro Moschitti, & Daniele Pighin. 2008. Semantic Role Labeling Systems for Arabic using Kernel Methods. In *Proceedings of ACL*. Association for Computational Linguistics, Columbus, OH, USA.

Gildea, Daniel & Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Gildea, Daniel & Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA, USA.

Gusfield, Dan. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Habash, Nizar & Owen Rambow. 2007. Arabic diacritization through full morphological tagging. In *HLT-NAACL* 2007; Companion Volume, Short Papers, Association for Computational Linguistics, pages 53–56, Rochester, NY, USA.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of HLT-NAACL* 2006, New York, USA.

Hwang, Jena D., Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue & Martha Palmer. 2010. PropBank Annotation of Multilingual Light Verb Constructions. In *Proceedings of the LAW-ACL 2010*. Uppsala, Sweden.

Maamouri, Mohamed, Ann Bies, Seth Kulick. 2008. Enhanced Annotation and Parsing of the Arabic Treebank. In *Proceedings of* INFOS 2008, Cairo, Egypt.

Márquez, Lluís. 2009. Semantic Role Labeling. Past, Present and Future . TALP Research Center. Technical University of Catalonia. Tutorial at ACL-IJCNLP 2009.

Moschitti, Alessandro. 2004. A study on convolution kernels for shallow semantic parsing. In *proceedings of the 42th Conference on Association for Computational Linguistic (ACL-2004)*, Barcelona, Spain.

Moschitti, Alessandro, Ana-Maria Giuglea, Bonaventura Coppola, & Roberto Basili. 2005. Hierarchical semantic role labeling. In *Proceedings of CoNLL-2005*, Ann Arbor, MI, USA.

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona Diab, Mohamed Maamouri, Aous Mansouri, Wajdi Zaghouani. 2008. A Pilot Arabic Propbank. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Palmer, Martha, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, & Fei Xia. 2009. Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure. In *The 7th International Conference on Natural Language Processing* (ICON-2009), Hyderabad, India.

Palmer, Martha, Daniel Gildea, & Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31, 1 (Mar. 2005), 71-106.

Palmer, Martha, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, & Yeongmi Jeon. 2006. LDC Catalog LDC2006T03.

Pradhan, Sameer, Kadri Hacioglu, Wayne Ward, James H. Martin, & Daniel Jurafsky. 2003. Semantic role parsing: Adding semantic structure to unstructured text. In *Proceedings of ICDM-2003*, Melbourne, USA.

Pradhan, Sameer S., Wayne H Ward, Kadri Hacioglu, James H Martin, & Dan Jurafsky. 2004. Shallow semantic parsing using support vector machines. In Susan Dumais, Daniel Marcu, & Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 233–240, Boston, MA, USA.

Xue, Nianwen & Martha Palmer. 2004. Calibrating features for semantic role labeling. In Dekang Lin & Dekai Wu, editors, *Proceedings of ACL-EMNLP 2004*, pages 88–94, Barcelona, Spain.

Xue, Nianwen & Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15 Jan. 2009, 143-172.