Universal Dependencies for Arabic Tweets

Fahad Albogamy
School of Computer Science,
University of Manchester,
Manchester, M13 9PL, UK
albogamf@cs.man.ac.uk

Allan Ramsay
School of Computer Science,
University of Manchester,
Manchester, M13 9PL, UK
allan.ramsay@cs.man.ac.uk

Abstract

To facilitate cross-lingual studies, there is an increasing interest in identifying linguistic universals. Recently, a new universal scheme was designed as a part of universal dependency project. In this paper, we map the Arabic tweets dependency treebank (ATDT) to the Universal Dependency (UD) scheme to compare it to other language resources and for the purpose of cross-lingual studies.

1 Introduction

Universal Dependency (UD) is a common scheme proposed by (Agic et al. 2015) to support cross-lingual studies and to compare the scheme in question to other language resources. Dependency treebanks have been developed for many languages such as Arabic, Czech and Turkish. However, each treebank has different labelling annotation and was built according to a specific linguistic theory. Due to these variations, a treebank for one language cannot be easily compared to a treebank for another language, so using a universal framework is an appealing method of overcoming these variations. From a parsing point of view, it is difficult to compare parser output in one language to that in another if their training data is based on different labelling schemes, because parsing results can be influenced by the number of annotation labels and different linguistic analyses across languages, as demonstrated by (McDonald et al. 2011) through a cross-lingual parsing study.

To facilitate cross-lingual studies, there is an increasing interest in identifying linguistic universals. In POS tagging, the Google universal POS tagset was developed by (Petrov et al. 2011), which contains 12 main POS tags aiming to cover the common categories that exist in any language. In parsing, a set of 41 dependency labels and a universal annotation scheme was developed by (McDonald et al. 2013) and used to convert ten language treebanks for the purpose of multilingual parsing. Recently, a new universal scheme was designed as a part of universal dependency project (Agic et al. 2015; Nivre et al. 2016). This scheme is based on the universal Stanford dependency (De Marneffe et al. 2006) and the Google universal POS tagset. We have mapped mapped the Arabic tweets dependency treebank (ATDT) (Albogamy et al. 2017) to the Universal Dependency (UD) scheme to compare it to other language resources and for the purpose of cross-lingual studies. The ATDT is a corpus of Arabic tweets that have been annotated with information on deep syntactic structure. This paper summaries the conversation and mapping of the ATDT to the UD scheme and outlines some structural changes and specific dependency labels introduced during the mapping process.

2 Mapping the Arabic tweets POS tagset to the universal POS tagset

To facilitate cross-lingual POS tagging and parsing studies, the (Google) universal POS tagset was designed by (Petrov et al. 2011). Its aim is to simplify POS tagsets and unify them across languages. The ATDT used a tagset obtained by unifying the tagsets from AMIRA, MADA and Stanford. The resulting tagset consists of the main POS tags both coarse- and fine-grained in addition to Twitter-specifics tags (Albogamy and Ramsay 2016). Table 1 shows the mapping of Arabic tweets POS tagset to the universal POS tagset. Most of the POS mappings made from the Arabic Tweets POS tagset to the universal POS tagset are intuitive. There are two types of mapping: one-to-one (e.g. JJ → ADJ), many-to-one; fine-

Proceedings of Recent Advances in Natural Language Processing, pages 46–51,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_007

Universal POS Mappings		
UD Tagset	**Arabic Tweets Tagset**	**Gloss**
ADJ	JJ	Adjective
ADP	IN	Preposition
ADV	RB	Adverb
	WRB	Wh-adverb
AUX	AUX	auxiliary
CCONJ	CC	Coordinating conjunction
DET	DET	Definite,article
INTJ	UH	Interjection
NOUN	NN	Common noun
NUM	CD	Cardinal number
PART	RP	Particle
PRON	DT	Demonstrative pronoun
	PRP	Subject pronoun
	SPRP	Clitic personal pronoun
	WP	Relative pronoun
PROPN	NNP	Proper noun
PUNCT	PUNC	Punctuation
SCONJ	CO	Subordinating conjunction
SYM	-	symbol
VERB	VB	Verb
X	-	other
-	AC	Accusative mark
-	AGR	Nouns agreement
-	FUT	Future mark
-	PERS	Person mark for verbs
-	TNS	Verb tense
-	EMOJ	Emoji
-	EMOT	Emoticons
-	LINK	Url or link
-	MEN	MEN
-	REP	Reply
-	RET	Retweet
-	USERN	Username

Table 1: Mapping of Arabic tweets POS to Universal POS tagset. '-' marks unused POS tags.

grained tags mapped to a coarse-grained tag (e.g. DT, PRP, SPRP and WP → PRON). However, there are Arabic tweets POS tags that cannot be mapped to the universal POS tagset. Some of them are related to Twitter phenomena (i.e. REP, MEN, LINK, USERN, RET, EMOT and EMOJ), whereas the others resulting from splitting clitics (i.e. AC, AGR, FUT, PERS and TNS). Therefore, they should be taken into consideration when mapping to the universal tagset. It should be noted that information is lost when fine-grained tags are mapped to coarse-grained tags. Consequently, the relationship and meaning between words and structure is often lost. For example, if we map PRP and WP to PRON then the parser cannot distinguish between ordinary pronouns and relative pronouns. As a result, we will not see that relative clauses have different structures from other clauses. '@' token which is attached with username in tweets has multiple tags. So, assigning tags to tweet items is not an entirely trivial activity. The tagger has also to learn when '@' is a reply and when it is a retweet or a mention (Albogamy and Ramsay 2016).

It is worth mentioning that taggers are generally more accurate on coarse-grained tagsets than fine-grained ones. (Marton et al. 2013) showed that using a fine-grained tagset by a tagger (e.g. MADA) can decrease the accuracy of the parser. This is because the tagger is likely to make more mistakes when using a fine-grained tagset, and these mistakes can have substantial knock-on effects on the performance of the parser. Therefore, we use the coarse-grained tagset in our tagger.

3 Mapping the Arabic tweets dependency scheme to the Universal Dependency scheme

The Universal Dependency scheme (UD15) consists of 41 dependency labels. Table 2 shows the mapping of the Arabic tweets dependency treebank (ATDT) scheme to the UD15 scheme. Some labels in the UD15 annotation scheme do not apply to the Arabic tweets language, marked '-' in Table 2. The mapping process involves some structural changes and it introduces more specific dependency labels as described below.

3.1 Structural changes

coordination The Arabic tweets treebank treats a coordinating conjunction (e.g. و 'and') as the head

and the coordinates as its daughter. On the other hand, the universal dependency annotation scheme treats the first coordinate as the head of the coordination, and the rest of the phrase as its daughter to the right.

cop In Arabic language, a copula is used in past tense forms and negated sentences. The copula is treated as a verb in the Arabic tweets treebank. So, it can function as the root of a dependency tree. However, the UD15 scheme treats the predicate as the head of the sentence, and the copula as its daughter.

case vs prepcomp The UD15 scheme attaches the head of a preposition phrase to the verb, and makes the preposition a daughter of the object, saying that it is a case-marker on the noun. On the other hand, the Arabic tweets treebank makes the preposition the head of a prepositional phrase, with the noun labelled as prepcomp (i.e. as the complement of the preposition)

aux vs auxcomp In the Arabic tweets treebank, the non-main (auxiliary) verb in a sentence functions as the head, and this relation is labelled as auxcomp. In the UD15, the auxiliary is usually taken to be a daughter of a verb.

zero-copula Arabic has zero-copula feature. A zero-copula sentence consists of an NP and predications (another NP, an adjective, a PP). In the Arabic tweets treebank, the subject is taken to be the head of a zero-copula sentence whereas the UD15 assumes that the predication is the head of the sentence.

The choice of whether to make a preposition the head or a daughter of the following noun phrase (NP), and of whether to make an auxiliary the head or a daughter of the following verb, depends on the underlying linguistic theory. We are using a version of the grammar of Arabic described by (Alabbas and Ramsay 2012) in which a preposition is taken to be the head of a preposition phrase (PP) and an auxiliary is taken to be the head of a sentence. They undertook experiments that showed that Arabic dependency parsing is more accurate when using the above structures.

3.2 Twitter-specific relations

To use a parser to extract Arabic tweets syntactic structure, we should be familiar with the grammatical structure of Arabic tweets and train the parser on it. Tweets have many phenomena such as mentions, replies, retweets, hashtags, links and etc. (Albogamy and Ramsay 2015). These elements become parts of tweets text and they will play grammatical roles in this context. In this section, we will discuss the grammatical structure of Arabic tweets.

UD Dependency Label Mappings		
Universal Label	**Arabic Tweets label**	**Gloss**
root	root	root
acl	Whmod	clausal modifier of noun (adjectival clause)
advcl	advcl	adverbial clause modifier
advmod	Advmod	adverbial modifier
amod	-	adjectival modifier
appos	-	appositional modifier
aux	auxcomp	auxiliary
case	predcomp	case marking
cc	cc	coordinating conjunction
ccomp	xcomp	clausal complement
clf	-	classifier
compound	-	compound
cop	-	copula
csubj	Subj	clausal subject
dep	-	unspecified dependency
det	det	determiner
discourse	-	discourse element
dislocated	-	dislocated elements
expl	-	expletive
fixed	-	fixed multiword expression
flat	-	flat multiword expression
goeswith	-	goes with
iobj	-	indirect object
list	-	list
mark	-	marker
nmod	nmod	nominal modifier
nsubj	subj	nominal subject
nummod	-	numeric modifier
obj	obj	object
obl	-	oblique nominal
orphan	-	orphan
parataxis	-	parataxis
punct	-	punctuation
reparandum	-	overridden disfluency
vocative	vocative	vocative
xcomp	xcomp	open clausal complement
-	Link	verbal argument or modifier
-	Men	verbal argument
-	Reply	usually the root
-	Retweet	usually the root
-	Usern	daughter of retweet, or mention or reply
-	Emot	modifier
-	Emoj	modifier

Table 2: Mapping of Arabic tweets treebank scheme to UD15 scheme. '-' marks unused dependency labels.

Arabic tweets have new elements which are not part of the normal Arabic language. These elements have syntactic function in tweets. For

example, both the hashtags and the mention in (1) are parts of the tweet syntactic structure.

(1) عندك سؤال عن القبول في #الجامعة ؟
#محادثة_الجامعةاسال عن طريق برنامج
لتحصل على اجابة من@UniAdmission

These new elements cannot be assigned the traditional POS tags. This means that they cannot easily be dealt with using traditional grammar, and hence it is important to discover their grammatical roles. As seen above, these elements are part of tweets so they must have grammatical relations with the rest of tweets, but because they are new parts and do not exist in the MSA grammar we need to know what kinds of relations they have. There are some efforts in the literature aiming at parsing English tweets. In (Kong et al. 2014) they developed a dependency parser for English tweets, but they omitted most of the tweets elements from the material to be parsed which leads to losing parts of the content of the tweet. In contrast, in this research we will try to discover the structures of Arabic tweets by taking into account all tweet constructions and we argue that all the new tweet elements play grammatical roles. One way to discover their grammatical roles is to rewrite tweets in ordinary more formal Arabic and try to preserve the meaning as far as possible, which will help us to understand the relationships between these elements and to analyse tweets structures. We will show a few examples of grammatical functions which can be played by these elements and present their dependency trees.

1. The link in tweet (2) is a Twitter-specific element and functions as the subject of a zero-copula senetence:
(2) Original tweet صورة لاحد عشاق جيرارد http://t.co/vY0feFK3F2
Transliteration Gerrard's fans one-of picture http://t.co/vY0feFK3F2
Paraphrase to MSA http://t.co/vY0feFK3F2 هذه صورة لاحد عشاق جيرارد
Paraphrase to English http://t.co/vY0feFK3F2 (is) a picture of one of Gerrard's fans

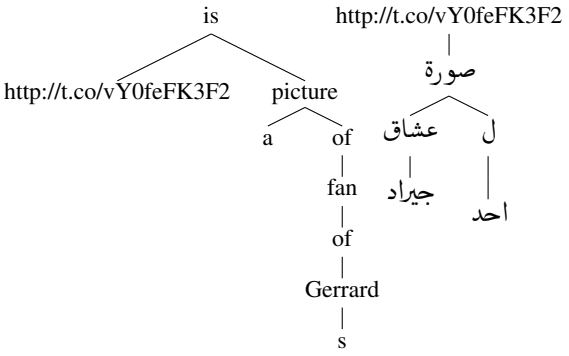

Figure 1: Dependency tree for tweet (2). Left: (English) Right: (Arabic).

2. The link in tweet (3) is a Twitter-specific element and works as a modifier:
(3) Original tweet http://t.co/mT2feDS5n5:انها تمطر
Transliteration http://t.co/mT2feDS5n5 : raining it's
Paraphrase to MSA http://t.co/mT2feDS5n5 انها تمطر:انظر الصورة
Paraphrase to English It is raining : see the picture http://t.co/mT2feDS5n5

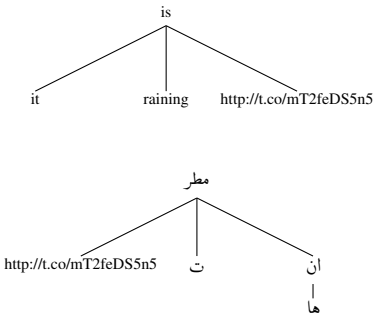

Figure 2: Dependency tree for tweet (3). Top: (English) Bottom: (Arabic).

3. The reply mark is a Twitter-specific element and it is equivalent to "reply to someone" phrase. It works as a verb in tweet (4):
(4) Original tweet انا سويت هذا @AhamedMoh
Transliteration this did I @AhamedMoh
Paraphrase to MSA رد على @AhamedMoh: انا سويت هذا
Paraphrase to English Reply to @AhamedMoh: I did this

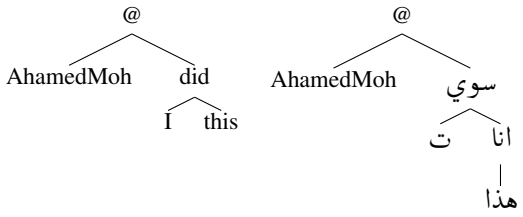

Figure 3: Dependency tree for tweet (4). Left: (English) Right: (Arabic).

4. The mention mark is a Twitter-specific element and works as an object in tweet (5):
(5) Original tweet @FahadTiger انا قابلت اليوم
Transliteration @FahadTiger today met I
Paraphrase to MSA @FahadTiger انا قابلت اليوم
Paraphrase to English I met @FahadTiger today

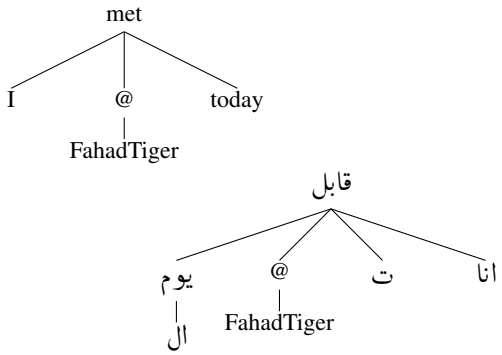

Figure 4: Dependency tree for tweet (5). Left: (English) Right: (Arabic).

5. The re-tweet is a discourse marker and it is equivalent to "someone says" phrase. It works as a verb in tweet (6) :
(6)Original tweet الدراسة صعبة مررررة @MomiAH:
Transliteration very difficult study @MomiAH:
Paraphrase to MSA @MomiAH يقول أن الدراسة صعبة جدا
Paraphrase to English @MomiAH says that the study is very difficult

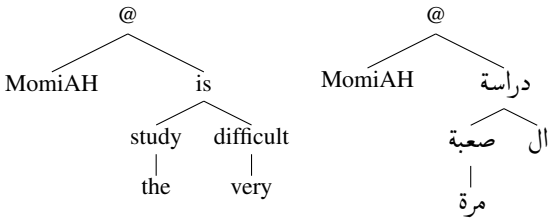

Figure 5: Dependency tree for tweet (6). Left: (English) Right: (Arabic).

4 Conclusion

In this paper, we have explained the importance of a universal annotation scheme in cross-lingual studies (e.g. cross-evaluate parsers). We have described the mapping of the Arabic POS tagset to the universal POS tagset and Arabic tweets dependency treebank to the Universal Dependency scheme. We have explained the mapping and conversion process in detail including structural changes. We have also discussed linguistic analyses of Arabic tweets and motivation for introducing specific dependency labels during the mapping process.

Acknowledgments

The authors would like to thank the anonymous reviewers for their encouraging feedback and insights. Fahad would also like to thank King Saud University for their financial support. Allan Ramsay's contribution to this work was partially supported by Qatar National Research Foundation (grant NPRP-7-1334-6 -039).

References

Agic, Ž., M. J. Aranzabe, A. Atutxa, C. Bosco, J. Choi, M.-C. de Marneffe, T. Dozat, R. Farkas, J. Foster, F. Ginter, et al. (2015). Universal dependencies 1.1. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague 3*.

Alabbas, M. and A. Ramsay (2012). Arabic treebank: from phrase-structure trees to dependency trees. In *Proceedings of the META-RESEARCH Workshop on Advanced Treebanking at the 8th International Conference on Language Resources and Evaluation (LREC)*, pp. 61–68.

Albogamy, F. and A. Ramsay (2015). POS tagging for Arabic tweets. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 1. Citeseer.

Albogamy, F. and A. Ramsay (2016). Fast and robust POS tagger for Arabic tweets using agreement-based bootstrapping. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Albogamy, F., A. Ramsay, and H. Ahmed (2017). Arabic tweets treebanking and parsing: A bootstrapping approach. In *Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP)-EACL*, pp. 94.

De Marneffe, M.-C., B. MacCartney, Manning, and C. D (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Volume 6, pp. 449–454. Genoa.

Kong, L., N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. Smith (2014). A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar*.

Marton, Y., N. Habash, and O. Rambow (2013). Dependency parsing of modern standard Arabic with lexical and inflectional features. *Computational Linguistics 39*(1), 161–194.

McDonald, R., S. Petrov, and K. Hall (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 62–72. Association for Computational Linguistics.

McDonald, R. T., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. B. Hall, S. Petrov, H. Zhang, O. Täckström, et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the Association for Computational Linguistics (ACL)(2)*, pp. 92–97.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1659–1666.

Petrov, S., D. Das, and R. McDonald (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.