I.J. Information Technology and Computer Science, **2021, 1, 54-61**
Published Online February 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2021.01.04

ADPBC: Arabic Dependency Parsing Based Corpora for Information Extraction

Sally Mohamed
Computer Science Department, Faculty of Computer and Information, Menoufia University, Egypt
Higher Institute of Engineering and Technology, Tanta, Egypt
E-mail: smbm222@yahoo.com

Mahmoud Hussien
Computer Science Department, Faculty of Computer and Information, Menoufia University
E-mail: fci_3mh@yahoo.com

Hamdy M. Mousa
Computer Science Department, Faculty of Computer and Information, Menoufia University
E-mail: hamdimmm@hotmail.com

Received: 04 June 2020; Revised: 13 August 2020; Accepted: 06 September 2020; Published: 08 February 2021

Abstract: There is a massive amount of different information and data in the World Wide Web, and the number of Arabic users and contents is widely increasing. Information extraction is an essential issue to access and sort the data on the web. In this regard, information extraction becomes a challenge, especially for languages, which have a complex morphology like Arabic. Consequently, the trend today is to build a new corpus that makes the information extraction easier and more precise. This paper presents Arabic linguistically analyzed corpus, including dependency relation. The collected data includes five fields; they are a sport, religious, weather, news and biomedical. The output is CoNLL universal lattice file format (CoNLL-UL). The corpus contains an index for the sentences and their linguistic meta-data to enable quick mining and search across the corpus. This corpus has seventeenth morphological annotations and eight features based on the identification of the textual structures help to recognize and understand the grammatical characteristics of the text and perform the dependency relation. The parsing and dependency process conducted by the universal dependency model and corrected manually. The results illustrated the enhancement in the dependency relation corpus. The designed Arabic corpus helps to quickly get linguistic annotations for a text and make the information Extraction techniques easy and clear to learn. The gotten results illustrated the average enhancement in the dependency relation corpus.

Index Terms: Dependency parsing, Arabic corpora, information extraction.

1. Introduction

Resources, such as corpora, are essential for researchers working on Arabic Natural Language Processing (NLP)[1,2]. One of the most critical problems in the corpora is a general weakness and low accuracy that refer to the complexity of the Arabic langue that has more morphologies than other languages [3]. There are some limitations in the number of available Arabic linguistic resources for research purposes; all of them are not including dependency relation and the scale of the datasets not sufficiently large [4]. Which, in turn, led to obtaining the degree of efficiency required in most of the research related to the Arabic NLP, especially the researches related to the information extraction process. The Information extraction is considered as a base of researches interested in the classifyingpolarity sentiments, summarizing the reviews, and classifying the emotions or opinion mining depending on the dependency parsing corpus. The most recent research in NLP in universal dependency have achieved promising results in different applications and languages[5]. Wherefore, this research aims at taking advantage of the universal dependency annotation in Arabic by making an available, revised and reliable corpus to use in the future.

In recent years, there are several Arabic treebank and corpora that have been introduced. However, they did not include dependency except the treebank in The Linguistic Data Consortium (LDC) which available at [6] and not free available there has another corpus. However, it does not include dependency and is used for a social subject like sentiment and classification [7]. Wherefore, this paper presents a new corpus includes mainly on the dependency parsing. Arabic text essays from the web have been collected in five different fields they are a sport, religion, weather, news, and biomedical. After that, for producing CoNLL-U formatted files, UDPipe tool is used for completing the

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

following tasks: 1) tokeniezation, 2) morphological analysis, 3) part-of-speech tagging, 4) lemmatization and 5) dependency parsing [8][9]. Our corpus was initially designed for training new techniques in information extraction. Multiple natural language models have been combined including Stanford, Farasa and Universal dependencies (UDpipe) to adds a different kind of annotation to the corpora. This corpus is motivated to quickly and painlessly get linguistic annotations for a text and make the information extraction according to be useful for other ANLP application. This article is structured as follows; section 2 is introducing some of the previous Arabic corpora with a specific field in Arabic natural language Section 3 includes dependency annotation schema. Section 4 building the corpus. Section 5 is Evaluation finally section 6 the conclusion and future work.

2. Related Work

Recently, number of Arabic corpora types is presented such as Raw Text Corpora, Annotated Corpora, Lexical Databases Speech Corpora and handwriting. Most of the corpora are designed to be used in specific Arabic natural language tasks such as optical character recognition for both handwriting and printing text [10], and speech recognition [11]. Also, ARALEX online is a dataset interested in the morphology and steam of the word which is used in translation [12]. Furthermore, QUWI aims to develop the database in size to make it more advanced in the handwriting [10]. Saad et al. presented a method based on transfering annotation from language to another with subjective/objective labels. They collected dataset from Wikipedia and euro-news website in English and Arabic. The experiments show that there are distinction between subjective/objective texts [13].

Alsolamy et al. built manually corpus-based sentiment lexicon for Arabic opinion Mining [7]. Furthermore, Gamal et al. constructed a benchmark dataset of Arabic Dialect Tweets [14]. Zaghouani et al. built Annotation Procedure that consists annotation management and MT post-editing annotation for Modern Standard Arabic corpora Machine Translation. [15]. Also, there is Arabic Scripts Dataset for writer identification [16]. It is also, there is annotated corpora that useful for corpus-based building automatic spelling correction tools [17]. Although these corpora are useful for different Arabic NLP tasks, there is not any available revised dataset for the Arabic dependency parsing that could be used in Arabic information extraction [16].

3. Dependency Annotation Scheme

The Arabic language accommodates more than one of noun compounds named the nested noun compound, which makes the technique of extraction extra difficult [18]. Many parsers for Arabic langue have been implemented such as; In MaltParser, Each token has five characteristics in MaltParser as follows: 1) word's ordinal position in the sentence, 2) part-of-speech (POS) tag, 3) form of word, 4) parent word identifer, and 5) the current word and its parent dependency relation (deprel) [19].

Stanford Arabic parser uses Buckwalter Arabic morphological analyzer (Bama) to complete segmentation. The strategy of Bama system is concatenative lexicon-driven. This system contains the lexicon, the compatibility tables and the analysis engine [20]. In bama, the pronouns, prepositions and conjunctions clitics are almost separated off as separate words. However, Inflectional morphology, derivational morphology and the clitic determiner "Al" (ال) is not separated off. So, Stanford Arabic Parser augments "augmented Bies" tag set for representing words that starts by the determiner "Al" (ال) and uses extra tags. these extra tags appear for all parts of speech (NN, CD, etc.) preceded by determiner (DT). To the best of our knowledge, there is no typed dependencies (grammatical relations) analysis available for Arabic from an architectural perspective; Stanford CoreNLP does not attempt to do everything. It is nothing more than a straightforward pipeline architecture [21][22].

Farasa parser uses a Support vector machine (SVM) for ranking [23]. In [23], The authors used Farasa Base and Farasa Lookup. The concatenated stop-words list is used in Farasa Base and all words are directly segmented by the classifier. In Farasa Lookup FarasaLookup, those words that were previously segmented or appeared more than four times during training are stored, but the unseen words are classified.

UDPipe used the Columbia Arabic Conversion Tool. It creates cross-linguistically consistent annotation guidelines that facilitate the creation of treebanks and the structural basis and the same label sets. The trainable pipeline (UDPipe) splits the sentence into segments and tokens, removes affixes and produces lemma (dictionary form) of all tokens, defines POS tagging and generates dependency parsing for sentence [24]. Currently, UDPipe solves the problem by inserting a module consisting of a trivial lookup in the dictionary of multiword tokens which is generated from training data [25].

4. Building Corpora

The building of this corpus built is a preliminary work for intended future research in Arabic NLP. The ADPBC consists of comma-separated values (CSV) file in the CoNLL universal lattices (CoNLL-UL) format that is created using text documents collected from the web, after that some processes have been done as illustrated in Fig.1 for building the corpus. The conducted corpora are available at "https://github.com/salsama/Arabic-Information-Extraction-Corpus".

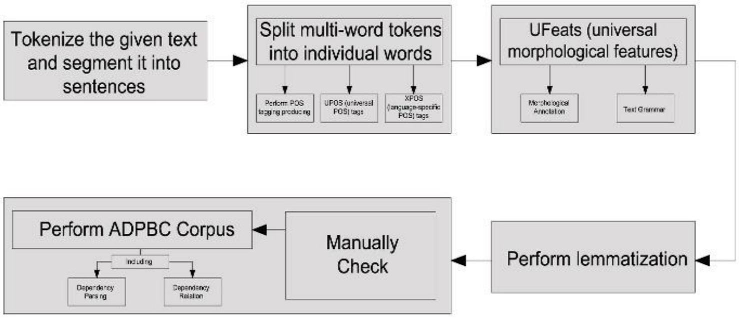
Tokenize the given text and segment it into sentences
Split multi-word tokens into individual words
Perform POS tagging producing
UPOS (universal POS) tags
XPOS (language-specific POS) tags
UFeats (universal morphological features)
Morphological Annotation
Text Grammar
Perform lemmatization
Manually Check
Perform ADPBC Corpus
Including
Dependency Parsing
Dependency Relation

Fig.1. The steps of corpus building

This corpus contains the words and their dependency relation produced by performing some steps. Firstly, segment the text sentence to individual words to produces the word-token, the second stage is performing a number of linguistic processes, which are POS Tagging and Lemmatization. POS tagging is the process of label every word in the text by its grammatical category and is automatically performed using the POS tagger technique [26], [27]. Lemmatization is a process for recovering the lemma from the word-surface form[28]. Secondly, performing the dependency parsing step which facilate the parsing of large scale text which is important to be accurate and fast.

In this stage, we used the Arabic PADT and Farasa model to performed POS Tagging and Lemmatization. The Arabic PADT (Prague Arabic Dependency Treebank) contains a multi-level description comprising functional morphology, analytical dependency syntax, and text grammatical representation of linguistic meaning. The PADT morphology succeeds to determine all of the contextual and lexical parameters [29]. Farasa has a new Arabic segmentation package that is more efficiency and faster than Stanford's Arabic segmenter and is used in this step. Finally, the output has been converted into CoNLL format as illustrated in Fig. 2 [30].

In the next step, manual checks have been performed for the corpus including the POS tagging and morphological annotation. As an example, the corresponding dependency-parsing tree for the sentence "The Russian authorities recommended that they leave the village of Nyonoxa days after an explosion of a nuclear nature" is illustrated in Fig. 2 and Table 1 illustrates the CONLL-U file format for the sentence.

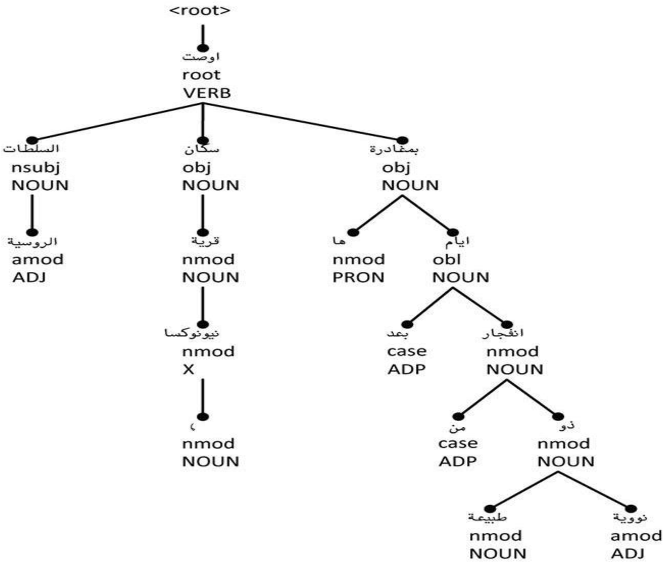

Fig.2. Dependency parsing tree

Dependency trees (DT) are real word order for representing the relations between head and dependent in the sentence. DT outputs common tag set that used for various languages, It starts from the verb that is the root of the sentence that connected with three nouns and these nouns connected by other adjectives and adverbial. The good dependency tree is simply a tree with the appropriate nodes, their nodes map, one to one, to the tokens as a result of the morphological analysis and tokenization, and their roots collect the nodes according to the division into sentences or paragraphs. The edges comply with the labels and valency limitations determined by the lexical entries. The task of conversion for these trees was easy as the linguistic representation was already what we want [31].

As shown in Table 1, the output contains sent_id: the sentence identifier, the text of the sentence (text) for which this token is part of and eight features; These features are: 1) word index (Id), integer starting at 1 for each new sentence; maybe a range for multiword, 2) Word form or punctuation symbol (Form), 3) stem or lemma of word form (Lemma), 4) Universal part-of-speech tag (UPos Tag), 5) particular language-part-of-speech tag (XPos Tag); which is underscore if not longer available, 6) morphological features list of the universal feature inventory (Feats), language-specific extension; which is underscore if not available, 7) the current word identification (Head) that takes zero valus or a value of word index (Id) and 8) Dependency Relation to the Head (DepRel), which is root if Head = 0 or defined depending on the treebank [32]. Universal Dependency Relation has several types, as shown in Table 2:

Table 1. The output in CONLL format

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel
# newdoc							
# newpar							
# sent_id = 1							
# text = أوصت السلطات الروسية سكان قرية نيونوكسا ،بمغادرتها بعد أيام من انفجار ذو طبيعة نووية							
1	أوصت	أَوصَى	VERB	VP-A-3FS--	Aspect=Perf|Gender=Fem|Number=Sing|	0	Root
2	السلطات	سُلطَة	NOUN	N------P1D	Case=Nom|Definite=Def|Number=Plur	1	Nsubj
3	الروسية	رُوسِيّ	ADJ	A-----FS1D	Case=Nom|Definite=Def|Gender=Fem|N	2	Amod
4	سكان	سُكَّان	NOUN	N------S4R	Case=Acc|Definite=Cons|Number=Sing	1	Obj
5	قرية	قَريَة	NOUN	N------S2R	Case=Gen|Definite=Cons|Number=Sing	4	Nmod
6	نيونوكسا	نيونوكسا	X	X---------	Foreign=Yes	5	Nmod
7	،	،	PUNCT	G---------	_	6	Punct
8-9	بمغادرتها	_	_	_	_	_	_
8	بمغادرة	بمغادرة	NOUN	N------S4R	Case=Acc|Definite=Cons|Number=Sing	1	Obj
9	ها	هُوَ	PRON	SP---3FS2-	Case=Gen|Gender=Fem|Number=Sing|Pe	8	Nmod
10	بعد	بَعدَ	ADP	PI------4-	AdpType=Prep|Case=Acc	11	Case
11	أيام	يَوم	NOUN	N------P2I	Case=Gen|Definite=Ind|Number=Plur	8	Obl
12	من	مِن	ADP	P---------	AdpType=Prep	13	Case
13	انفجار	اِنفِجَار	NOUN	N------S2R	Case=Gen|Definite=Cons|Number=Sing	11	Nmod
14	ذو	ذُو	NOUN	N------S2R	Case=Gen|Definite=Cons|Number=Sing	13	Nmod
15	طبيعة	طَبِيعَة	NOUN	N------S2I	Case=Gen|Definite=Ind|Number=Sing	14	Nmod
16	نووية	نَوَوِيّ	ADJ	A-----FS2I	Case=Gen|Definite=Ind|Gender=Fem|Nu	14	Amod

Table 2. Dependency Parsing Relation

Rel	Definition
Root	points to the root of the sentence.
Nsubj	A nominal subject.
Amod	Adjectival modifier.
Obj	Direct object.
Fixed	certain fixed grammaticized.
Cc	coordinating conjunction.
Parataxis	Used to connect to sentences together.
Obl	adverbial attaching to a verb, adjective.
Case	providing a more uniform analysis of nominal elements.
Nmod	nominal modifier.
Det	Determiner.
Acl	an adverbial clause.
Advmod	adverbial modifier.
Aux	Auxiliary.
Conj	Conjunct.
Xcomp	open clausal complement.
Comp	comparison constructions.

The Universal Dependencies framework has been chosen because it provides a comparable format and help in matching the different types of dependency relations in various languages. There are seventeen types of dependency relation have been provided by parsers trained on Arabic-padt-ud-2.4-data. For example; the subject and object, adnominal clauses, conjunction, relative, auxiliary, adverbial, and parataxis [26].

5. Confusion Matrix Evaluation

This section presents an evaluation of conducted corpus. UDPipe can produces errors like any parser, and it was essential to assess the output to configure and estimate these mistakes [26]. The evaluation is performed manually, and a confusion matrix has been conducted for the 17 types of dependency parser annotation. The confusion matrix compares between the data was yielded from the UDPipe model and manually corrected data. The results show the mismatching accrue between the root and Nmod, nsubj and OBJ and Amod and case. The results in Table 3 illustrates the enhancement in the corrected data.

Table 3. confusion matrix

Gold Auto	ROOT	NSUBJ	AMOD	OBJ	FIXED	CC	PARATIX	OBL	NMOD	DET	Acl	Case	comp	Advmod	Aux	Conj	Xcomp
Root	80%	5%	0%	3%	0%	0%	0%	0%	10%	0%	1%	0%	0%	0%	0%	1%	0%
Nsubj	0%	80%	0%	6%	0%	0%	4%	3%	5%	0%	0%	0%	0%	0%	0%	2%	0%
Amod	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	0%	0%
Obj	0%	0%	0%	72%	0%	0%	0%	.02%	9%	0%	0%	0%	0%	0%	0%	20%	0%
Fixed	0%	0%	0%	0%	80%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%
Cc	0%	0%	0%	0%	0%	99%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%
Parataxis	18%	0%	0%	0%	0%	0%	36%	9%	0%	0%	9%	9%	0%	0%	0%	18%	0%
Obl	8%	8%	8%	0%	0%	0%	0%	33%	41%	0%	0%	0%	0%	0%	0%	0%	0%
Nmod	1%	4%	0%	4%	1%	1%	0%	5%	91%	0%	1%	0%	0%	0%	0%	2%	1%
Det	0%	0%	0%	0%	0%	0%	0%	25%	25%	50%	0%	0%	0%	0%	0%	0%	0%
Acl	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	95%	0%	5%	0%	0%	0%	0%
Case	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	97%	0%	0%	0%	0%	0%
Comp	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	99%	0%	0%	0%	0%
Advmod	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	99%	0%	0%	0%
Aux	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	40%	40%	0%	0%
Conj	0%	1%	3%	0%	0%	0%	13%	3%	0%	0%	0%	0%	0%	0%	0%	75%	0%
Xcomp	0%	0%	0%	0%	0%	0%	0%	42%	0%	0%	0%	0%	0%	0%	0%	14%	28%

This corpus contains seventeenth morphological annotations, as illustrates in Fig. 3 where identify the textual forms of a discourse lexically and recognize their grammatical properties.

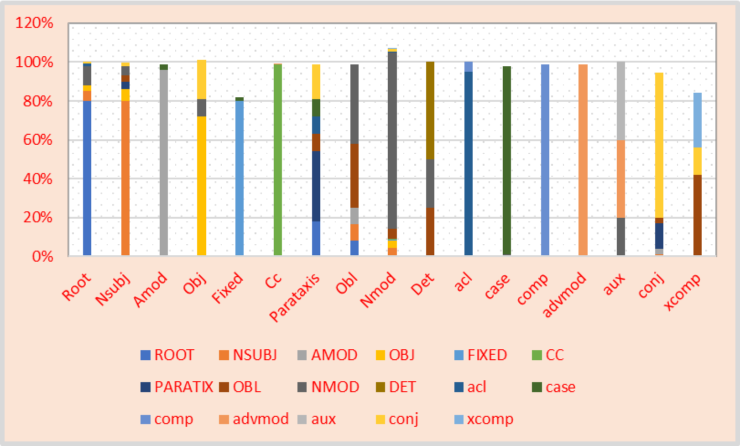

Fig.3. Percentage of miss matching relation

The analytical syntactical process describes the superficial dependency structures within the discourse, whereas the text grammatical illustration and detects the underlying dependency structures and restores the meaning of linguistic. Annotations on the analytical level are described and represented by dependency trees. The most frequent relation errors are mismatched between different dependency relation annotations, for example, the root and Nmod, root and Nsbj, root and obj, Parataxis and root, Parataxis and Adv, Obj and Nmod.

The ADPBC dataset is viable to start simple large-scale experiments with syntax-aware models while it is not required for long and resource-intensive preprocessing. This data set is uploaded online and its public to modified and improvement. The confusion matrix result indicates that the dataset is suitable for the future researcher in Arabic, especially in information extraction. Various domains and other opinion mining topics are covered into our corpus. The extra meta-data information of a document can be used in various Arabic natural language processes such as machine translation. The system tested on five categories of documents that performed from them, and the average number of sentences per document is 25 sentences. The average number of words per sentence is 15 words.

6. Conclusion and Future Work

This paper presented ADPBC corpus, which is automatically annotated using UDPipe model then manually corrected. The text has been extracted from the web. After Spell checking is performed for the text, the parsing process is applyed. The non syntax errors will not affect the syntactic complexity evaluation. The paper contains features used in the evaluation of automatic dependency parsing. The corpus will be utilized in numerous contexts starting from information extraction as well as the core of all Arabic natural language process to unsupervised induction of word senses and frame structures.

Additionally, it has several characteristics as well as morphological annotation depends on. That main valuable property is dependency parsing that makes availability to improving the Late in performance in overall ANLP. Finally, the results indicated that the annotation and the parsing processes will be improving. In the future work and the conducted corpora will use mainly in intended future research interested on the information extraction.

References

[1] Noureddine Doumi, Ahmed Lehireche, Denis Maurel, Ahmed Abdelali,"A Semi-Automatic and Low Cost Approach to Build Scalable Lemma-based Lexical Resources for Arabic Verbs", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.2, pp.1-13, 2016. DOI: 10.5815/ijitcs.2016.02.01

[2] M. El-haj and R. Koulali, “KALIMAT a Multipurpose Arabic Corpus,” Second Work. Arab. Corpus Linguist., pp. 22–25, 2013.

[3] S. Ali, H. Mousa, and M. Hussien, “A Review of Open Information Extraction Techniques,” IJCI. Int. J. Comput. Inf., vol. 6, no. 1, pp. 20–28, 2019.

[4] H. Mahmoud, S. S. Kareem, and T. El-Shishtawy, “A Semantic Retrieval System for Extracting Relationships from Biological Corpus,” Int. J. Comput. Sci. Inf. Technol., vol. 10, no. 1, pp. 43–53, 2018.

[5] M. Straka, J. Hajiˇ, and J. Strakov, “UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization , Morphological Analysis , POS Tagging and Parsing.”

[6] “https://catalog.ldc.upenn.edu/LDC2010T08.” p. https://catalog.ldc.upenn.edu/LDC2010T08.

[7] Afnan Atiah Alsolamy, Muazzam Ahmed Siddiqui, Imtiaz Hussain Khan, " A Corpus Based Approach to Build Arabic Sentiment Lexicon", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.11, No.6, pp. 16-23, 2019. DOI: 10.5815/ijieeb.2019.06.03

[8] M. Straka, J. Hajič, and J. Straková, “UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing,” Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016, pp. 4290–4297, 2016.

[9] D. Taji, N. Habash, and D. Zeman, “Universal Dependencies for Arabic,” pp. 166–176, 2017.

[10] S. Al Maadeed, W. Ayouby, A. Hassaïne, and J. M. Aljaam, “QUWI: An Arabic and English handwriting dataset for offline writer identification,” Proc. - Int. Work. Front. Handwrit. Recognition, IWFHR, no. September, pp. 746–751, 2012.

[11] Moner N. M. Arafa, Reda Elbarougy, A. A. Ewees, G. M. Behery," A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.10, No.4, pp. 31-38, 2018.DOI: 10.5815/ijigsp.2018.04.04

[12] W. Zaghouani, “Critical Survey of the Freely Available Arabic Corpora,” Proc. Work. Free. Arab. Corpora Corpora Process. Tools Work. Program., pp. 1–8, 2017.

[13] M. Saad, D. Langlois, and K. Smäili, “Building and modelling multilingual subjective corpora,” Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014, no. May, pp. 3086–3091, 2014.

[14] Donia Gamal, Marco Alfonse, El-Sayed M.El-Horbaty, Abdel-Badeeh M.Salem, "Twitter Benchmark Dataset for Arabic Sentiment Analysis", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.1, pp. 33-38, 2019.DOI: 10.5815/ijmecs.2019.01.04

[15] W. Zaghouani, N. Habash, O. Obeid, B. Mohit, H. Bouamor, and K. Oflazer, “Building an Arabic machine translation post-edited corpus: Guidelines and annotation,” Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016, pp. 1869–1876, 2016.

[16] T. Arts, Y. Belinkov, N. Habash, A. Kilgarriff, and V. Suchomel, “ArTenTen: Arabic Corpus and Word Sketches,” J. King Saud Univ. - Comput. Inf. Sci., vol. 26, no. 4, pp. 357–371, 2014.

[17] A. O. Al-Thubaity, “A 700M+ Arabic corpus: KACST Arabic corpus design and construction,” Lang. Resour. Eval., vol. 49, no. 3, pp. 721–751, 2015.

[18] N. Omar and Q. Al-Tashi, “Arabic Nested Noun Compound Extraction Based on Linguistic Features and Statistical Measures,” GEMA Online® J. Lang. Stud., vol. 18, no. 2, pp. 93–107, 2018.

[19] Y. Marton, N. Habash, and O. Rambow, “Dependency parsing of modern standard arabic with lexical and inflectional features,” Comput. Linguist., vol. 39, no. 1, pp. 161–194, 2013.

[20] Nizar Y. Habash, Introduction to Arabic Natural Language Processing. 2010.

[21] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr., pp. 55–60, 2014.

[22] N. Bhutani, Y. Suhara, W.-C. Tan, A. Halevy, and H. V. Jagadish, “Open Information Extraction from Question-Answer Pairs,” pp. 2294–2305, 2019.

[23] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, “Farasa: A Fast and Furious Segmenter for Arabic,” vol.”, Proceedings of the 2016, Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2016) (Demonstrations), San Diego, California, J pp. 11–16, June 12-17, 2016.

[24] M. Straka, “UDPipe 2 . 0 Prototype at CoNLL 2018 UD Shared Task,” *Proc. CoNLL 2018 Shar. Task Multiling. Parsing from Raw Text to Univers. Depend.*, pp. 197–207, 2018.

[25] F. Albogamy and A. Ramsay, “Universal dependencies for Arabic tweets,” *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, vol. 2017-Septe, pp. 46–51, 2017.

[26] O. Lyashevkaya and I. Panteleeva, “Automatic Dependency Parsing of a Learner English Corpus Realec,” *High. Sch. Econ. Res. Pap. No. WP BRP.*, 2018.

[27] A. Panchenko, E. Ruppert, S. Faralli, S. P. Ponzetto, and C. Biemann, “Building a web-scale dependency-parsed corpus from common crawl,” *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, pp. 1816–1823, 2019.

[28] M. Serrano Morales Antoni Badia Cardús, “Treball de fi de màster What is Modern Standard Arabic NLP? Definition and Tools (or How to understand Arabic even if you do not know a word),” 2015.

[29] H. K. El-Najjar and R. S. Baraka, “Improving Dependency Parsing of Verbal Arabic Sentences Using Semantic Features,” *Int. J. Speech Technol.*, pp. 86–91, 2018.

[30] D. Zeman *et al.*, “CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies,” pp. 1–21, 2018.

[31] S. Buchholz and E. Marsi, “CoNLL-X shared task on multilingual dependency parsing,” *Proc. Tenth Conf. Comput. Nat. Lang. Learn.*, no. June, pp. 149–164, 2006.

[32] A. More *et al.*, “Conll-UL: Universal morphological lattices for universal dependency parsing,” *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, pp. 3847–3853, 2019.

Authos' Profiles

Sally Mohamed: received BSc and MSc in computer and automatic control engineering from Tanta 2004 and 2014 respectively. Her main research interest includes Data Mining, Machine Learning and Embedding Network

Mahmoud Hussein: received his BSc. and MSc. in Computer Science from Menoufia University, Faculty of Computers and Information in 2006 and 2009 respectively and received his PhD in Software Engineering from Swinburne University of Technology, Faculty of Information and Communications Technology in 2013. His research interest includes Software Engineering, Data Mining, Machine Learning, Data Privacy, and Security

Hamdy M. Mousa: received the B.S. and M.S. in Electronic Engineering and Automatic control and measurements from Menoufia University, Faculty of Electronic Engineering in 1991 and 2002, respectively and received his Ph.D. in Automatic control and measurements Engineering (Artificial intelligent) from Menoufia University, Faculty of Electronic Engineering in 2007. His research interest includes intelligent systems, Natural Language Processing, privacy, security, embedded systems, GSP applications, intelligent agent, Bioinformatics, Robotics.

How to cite this paper: Sally Mohamed, Mahmoud Hussien, Hamdy M. Mousa, "ADPBC: Arabic Dependency Parsing Based Corpora for Information Extraction", International Journal of Information Technology and Computer Science(IJITCS), Vol.13, No.1, pp.54-61, 2021. DOI: 10.5815/ijitcs.2021.01.04