Parsing Arabic using deep learning technology

Rahma Maalej [1], Nabil Khoufi [2] and Chafik Aloulou [3]

[1,3] *University of Sfax, ANLP Research Group, MIRACL Lab, Sfax, Tunisia*
[2] *ANLP Research Group, MIRACL Lab, Sfax, Tunisia*

Abstract
Syntactic Parsing present a fundamental step in the process of automatic analysis of the language since it is the crucial task of determining the syntactic structures sentences. In this paper, we propose to syntactically analyze sentences for the Arabic language using deep learning techniques. We present our methodology and expose evaluation results using several deep learning architectures.

Keywords
Natural language processing, NLP, Syntactic Parsing, standard Arabic, deep learning, Machine Learning, LSTM, BILSTM, RNN.

1. Introduction

In the domain of Automatic Natural Languages Processing (NLP), syntactic parsing represents a crucial step in NLP applications such as machine translation, spelling correction, etc. It is from this stage that we will be able to generate the syntactic structure of a text which makes it possible to clarify the relations between the different linguistic units to construct a semantic representation. It is a complicated step because of the complexity and the richness of Arabic language. Moreover, a bad decomposition of the sentence or a bad choice of the grammatical category of the grammatical part will influence the semantic interpretation. Therefore, many works have been done using different approaches: the linguistic approach, the statistical approach and the hybrid method [1]. The linguistic approach is based on a lexical knowledge and on a precise linguistic rule to eliminate the ambiguities of the words of the sentence and to adapt the structure of the sentence [20]. The statistical approach works with statistical models. The hybrid method combines both of linguistic and statistical methods [12][9].

Due to the complexity of the Arabic language, we still need to improve the results of Arabic parsing task.

The objective of this work is to realize a statistical syntactic parser for Modern Standard Arabic (MSA) based on deep learning techniques.

2. Related works

Several works have been done using a linguistic approach for MSA syntactic parsing.

[17] applied an idea which consists in integrating a grammar based on the use of concepts instead of words. This method will achieve a greater rate of coverage of the characteristics of the Arabic language [16]. In fact, they built a probabilistic out-of-context grammar that relies on schemes from 100,000 Arabic words. They obtained an accuracy of 63.35%.

[11] proposed to build an Arabic parser. This analyzer is based on a PCFG grammar (Probabilistic context free grammar). Their method was divided into two phases. The first phase is the induction of grammar. The objective of this phase is to automatically deduce a PCFG from the annotated corpus ATB. This process consists of two steps: The first step is to derive CFG rules from the ATB. The second step consists in assigning a probability to each deduced rule[10]. The second phase consists in

Tunisian Algerian Conference on Applied Computing (TACC 2021), December 18–20, 2021, Tabarka, Tunisia
EMAIL: rahmamaalej1234@gmail.com (R. Maalej) ; nabil.khoufi@fsegs.rnu.tn (N. Khoufi); chafik.aloulou@fsegs.tnu.tn (C. Aloulou)

implementing the deduced grammar (results of the first phase) to perform the syntactic analysis. They tested this analyzer with a set of sentences taken from the ATB Treebank (1650 sentences). The authors found an accuracy of 83.59%, a recall equal to 82.98% and an F-measure equal to 83.23%.

[2] proposed to use an Arabic property grammar to evaluate and enrich the Stanford Parser. In fact, this parser is based on the verification of the satisfaction of the syntactic constraints, also called properties, based on the analysis results of the corpus. Moreover, they enriched the simple representation of the result of the analysis with syntactic properties. This makes it possible to clarify several implicit information which present the relations between syntactic units.[2] obtained an F-score of 77.62%. with a recall value of 70.2% and a precision of 86.81%.

Recently,[8] proposed to construct a formal grammar for use in automatic processing applications of the Arabic language. Their thesis work aimed to set up a grammar in order to describe the syntax and semantics of Modern Standard Arabic. They presented a complete meta-grammatical description that allows to achieve a syntactically and semantically rich representation of this language. For the evaluation of the syntactic analysis, they constructed manually a test corpus by extracting 1000 syntactically correct sentences from the Tunisian school book (8th grade level). They obtained as results, an accuracy of 82.33%, a Recall of 88.10%and an F1-score equal to 85.11%.

[4] proposed two approaches; the first consists in detecting and correcting syntactic errors based on the automatic generation of correct sentences [18]. The second is aimed at the automatic correction of case termination errors based on parsing. He built a corpus that consists of 360sentences. The system achieved an accuracy rate equals to 92.01%, a recall rate which is equal to 84.83% and an F-score of 88.27%.

The statistical approach is based on statistical models. We have not found recent works using deep learning for syntactic parsing of MSA. On the other hand, we found works dealing with French and English. We can cite the work of [6] worked on lexicalized parsing for de-constituting grammars.
[6] had a main objective which consisted in adapting an underlying statistical model to these new representations. They proposed a study of three neuronal architectures of increasing complexity and show that the use of a non-linear hidden layer makes it possible to take advantage of the information given by the embedding. They found as results an F-measure of 80.7% for the given tags and 78.3% for the predicted tags.

[4] performed the analysis of advantages of a pre-trained language model such as BERT (bidirectional encoder representations from transformers) to the syntactic analysis in discontinues constituents in English (PTB, Penn Treebank). They found as result F-score equal to 95%.

[9] have implemented a Hybrid Standard Arabic parser that combines two parsing approaches: statistical parsing with linguistic parsing. The objective of this analyzer is to reduce the analysis time which is exponential with the size of the sentence.

[19] present the results of an evaluation on the integration of data from a syntactic lexicon, the Grammar lexicon, in a parser. They have carried out the evaluation according to the cross-validation method on 10% of the French Treebank (FTB). They have obtained an F-score of 85.32%.

Table1

Related works study recap

Authors	Testing data	Results
[17]	100000 Arabic sentence	Accuracy= 63.35%
[10]	ATB	-
[11]	ATB Treebank (1650 sentences)	Accuracy = 83.59%, Recall = 82.98% and F-measure = 83.23%.
[2]	ATB, PTB, 100 Arabic phrases obtained from stories for Arab children	F-score=77,62%. Recall= 70,2%, Precision=86,81%.
[8]	corpus1: 1000 grammatically correct sentences and corpus2: consisting of 250 ungrammatical sentences	Precision=82,33%, Recall=88,10% and F1-score=85,11%.

[4]	360 Arabic sentences	= 92%, $Recall$ = 84%$andF - score$ = 88.27%
[6]	French Corpus SPMRL	F-score = 80.7 % for the data tags and 78.3 % for the predicted tags
[4]	English corpus PTB	F-score=95%
[9]	ATB	F-score=83,24%
[19]	French Treebank	F-score=85.32%

3. Proposed Approach

This section presents the general architecture of our proposed approach.

Our statistical method for parsing Arabic based on deep learning techniques is divided into three steps: the preprocessing step, the training step and the validation and testing step. The first step presents the data preprocessing by extracting the syntactic levels for each sentence of the data. The second, does the training of our models and finally, the third step is the step of validation and testing. The steps of our proposed method are illustrated in the following figure:

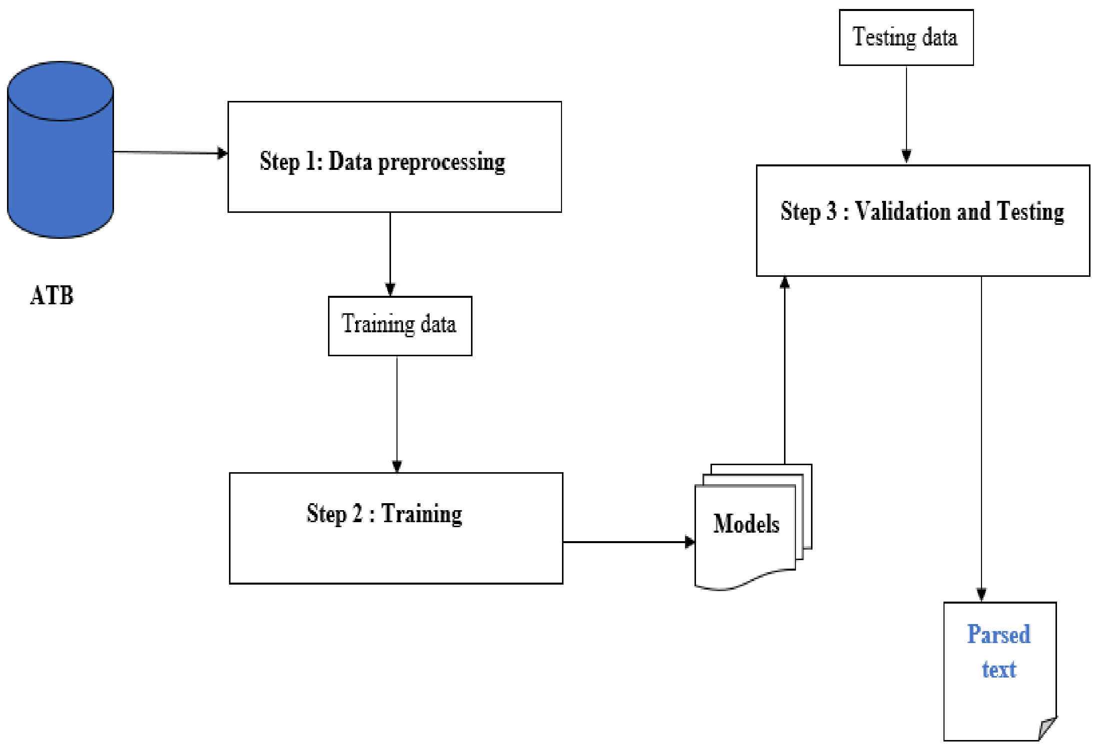

Figure 1: Architecture of the proposed method

3.1. The preprocessing step

The preprocessing step consists of preparing data for the training step.

3.1.1. Corpus ATB

In this work, we used the annotated Penn Arabic Treebank (ATB) corpus [13].

The ATB corpus comprises data extracted from linguistic sources written with the standard and modern Arabic language. It behaves 599 texts taken from the Lebanese newspaper ≪ An Nahar≫. The texts are non-vowel or partially vowel and segment. Each word is annotated with several information such as the morphological trait, the part of speech and the English translation. Also, it contains the syntax trees for each sentence [14].

In addition, The ATB corpus is very rich in information, such as gender, number and rationality. It deals with large and important annotations. On the one hand, the corpus encompasses a set of 498 annotations to describe all the morphological functionalities, it also uses 22 annotations to describe grammatical classes. There are 20 other notations, which describe the semantic relationships between words. On the other hand, stop words are also marked (stop words exist with a large number in Standard Arabic) with specific annotations. In our work, we use version 3.2 which contains 402,291 words. This version has been created with several different formats to help the user for his research: SGM, POS, XML, penntree, Integreted.

For the preprocessing step of our ATB corpus, we mainly used the two formats ≪ XML ≫ and ≪ Tree ≫ because they have a hierarchical and readable representation and in addition, they contain the essential information for our work objective and for experimentation.

We have presented the eight syntactic levels with a set of morpho-syntactic labels for each sentence extracted from the corpus ATB [3].

Features. These features indicate the information used from the annotated corpus during the

training step, which is the morphological annotations in the syntactic annotations. In fact, we extracted the morpho-syntactic labeling from ATB. And, we have reduced the number of labels in order to avoid complexity and to be prepared to learn and predict.

For example: NP+ ADJP: NP.

3.2. The training step

Many machine learning algorithms need a vector representation as an input. In this work, we have used embedding resources such as input from our neural systems. we have represented our data resource with word2vec.

3.2.1. Continuous Representations of Words

Word2vec is a popular method of constructing word embedding. It was introduced by [15] Two variations of word2vec have been proposed for learning word embedding: Skip-gram and BOW (Bag of Words). The BOW (Bag Of Words) is an architecture that predicts the current target word (the central word) based on the source context words (surrounding words) [15].

The skip-gram aims to predict the words of the context given an input word [15].

In this work, we applied the bag of words (BOW) representation because it is the most popular method for NLP applications.

After obtaining the word embedding with the word2vec method, we start the training by applying the deep learning techniques: LSTM, GRU, BI-LSTM. These models will present results after training and we will compare them to choose the model having a better result to apply it to perform parsing.

Today, neural networks present the systems most used in different machine learning tasks. They are widely used, for example, in the fields of computer vision (image classification, detection of an object, segmentation, etc.) and automatic language processing (automatic translation, voice recognition, language models, etc.).

In this work, we are interested in the creation of a model for each syntactic level. This model has an important objective, is to determinate the different constituents of a sentence and the different relations between them.

- Long Short-Term Memory Network (LSTM) presents a building unit for the layers of a recurrent neural network (RNN). An RNN made up of LSTM units is often referred to as an LSTM network. A common LSTM unit is presented with a cell, an input gate, an output gate and a forget gate. The cell is responsible for” storing” values. Each of the three gates can be considered as a” conventional” artificial neuron, as in a multilayer neural network (or feedforward): that is to say, they calculate an activation (using an activation function) of a weighted sum. Indeed, they can be considered as regulators of the flow of values. There are connections between these doors and the cell.

- Gated Recurrent Unit (GRU) is a gated recurrent network similar to the LSTM network, but it receives fewer parameters than it.
- BILSTM is a Bidirectional LSTM, present a sequence processing model that include two LSTMs: one is receives the input in a forward direction, and the other in a backwards direction. BILSTMs increase the amount of information available on the network and improve the context available for the algorithm.

4. Results

The evaluation of our method is realized in a validation and test step. To achieve this, we have divided the ATB corpus in two parts, one for learning (70%) and one for evaluation (30%) between validation (0.15%) and testing (0.15%).

We used deep learning techniques for train and test our models for our eight levels. we applied the RNN models: LSTM model, GRU model and BILSTM model.

The results are illustrated in table 2. The table shows analysis performance by levels.

Levels	**Accuracy**		
	LSTM-model	**GRU-model**	**BILSTM-model**
Level0	99,00%	99,08%	99,60%
Level1	99.18%	99.31%	99.53%
Level2	98,84%	98,91%	99,17%
Level3	99,09%	99.15%	99.36%
Level4	99,36%	99.39%	99.60%
Level5	99,21%	99.28%	99.49%
Level6	99.25%	99.32%	99.57%
Level7	99.26%	99.35%	99.60%

Table 2
Evaluation results

We have obtained good results which are encouraging to realize the syntactic parsing for the standard Arabic language with a deep learning model. We can compare the results obtained for the models: LSTM, GRU, BILSTM. We deduced that that the models BILSTM have best results.

5. Conclusion and Perspectives

In this paper, we presented our numerical method to perform parsing for Standard Arabic using deep learning techniques. We have built a model and we obtained encouraging results. As a perspective, we think that we can develop the learning stage by adding other features besides the POS features. we believe that the enrichment of the features can give better results. We plan to evaluate our method with another external Arabic corpus.

6. References

[1] C. Aloulou.Une approche multi-agent pour l’analyse de l’arabe: Modélisation de la syntaxe.PhD thesis, Thèse de doctorat en informarique, Ecole Nationale des Sciences de l ..., 2005.

[2] R. B. Bahloul, N. Kadri, K. Haddar, and P. Blache. Evaluation and enrichment of stanfordparser using an arabic property grammar. In A. F. Gelbukh, editor,Computational Linguisticsand Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest,Hungary, April 17-23, 2017,

Revised Selected Papers, Part I, volume 10761 ofLecture Notes inComputer Science, pages 170–182. Springer, 2017.

[3] H. B. Barhoumi, Aloulou and Z. (2015). Analyse syntaxique statistique de la langue arabe.2015.

[4] M. Chouaib.Contributions à la correction automatique des erreurs syntaxiques dans la langueArabe. PhD thesis, Faculté des Sciences Ben M'sik Université Hassan II Casablanca, 2020.

[5] M. Coavoux. Qu'apporte bert à l'analyse syntaxique en constituants discontinus? une suitede tests pour évaluer les prédictions de structures syntaxiques discontinues en anglais(what does bert contribute to discontinuous constituency parsing? a test suite to evaluate discontinuous constituency structure predictions in english). InActes de la 6e conférenceconjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique desLangues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatiquepour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: TraitementAutomatique des Langues Naturelles, pages 189–196, 2020.

[6] M. Coavoux and B. Crabbé. Comparaison d'architectures neuronales pour l'analysesyntaxique en constituants. InActes de la 22e conférence sur le Traitement Automatique desLangues Naturelles. Articles longs, pages 291–302, 2015.

[7] M. Coavoux and B. Crabbé. Prédiction structurée pour l'analyse syntaxique en constituantspar transitions: modèles denses et modèles creux.Traitement Automatique des Langues,57(1), 2016.

[8] C. B. Khelil.Construction semi-automatique d'une grammaire d'arbres adjoints pour l'analysesyntaxico-sémantique de l'arabe. PhD thesis, Université d'Orléans; Université de la Manouba(Tunisie), 2019.

[9] N. Khoufi. Une approche hybride pour l'analyse syntaxique de la langue arabe. PhD thesis,Université de Sfax (Tunisie), 2017.

[10] N. Khoufi, C. Aloulou, and L. H. Belguith. Arabic probabilistic context free grammarinduction from a treebank.Res. Comput. Sci., 90:77–86, 2015.

[11] N. Khoufi, C. Aloulou, and L. H. Belguith. A framework for language resource constructionand syntactic analysis: Case of arabic. InInternational Conference on Intelligent TextProcessing and Computational Linguistics, pages 356–365. Springer, 2016.

[12] N. Khoufi, C. Aloulou, and L. H. Belguith. Toward hybrid method for parsing modernstandard arabic. In2016 17th IEEE/ACIS International Conference on Software Engineering,Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pages 451–456.IEEE, 2016.

[13] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. The penn arabic treebank: Building alarge-scale annotated arabic corpus. InNEMLAR conference on Arabic language resourcesand tools, volume 27, pages 466–467. Cairo, 2004.

[14] M. Maamouri, A. Bies, and S. Kulick. Enhancing the arabic treebank: a collaborative efforttoward new annotation guideline. InLREC, pages 3–192. Citeseer, 2008.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representationsin vector space.arXiv preprint arXiv:1301.3781, 2013.

[16] M. A. B. Mohamed, S. Mallat, M. A. Nahdi, and M. Zrigui. Exploring the potential ofschemes in building nlp tools for arabic language.International Arab Journal of InformationTechnology (IAJIT), 12(6), 2015.

[17] M. A. B. Mohamed, S. Zrigui, A. Zouaghi, and M. Zrigui. N-scheme model: An approachtowards reducing arabic language sparseness. In2015 5th International Conference onInformation & Communication Technology and Accessibility (ICTA), pages 1–5. IEEE, 2015.

[18] C. Moukrim, A. Tragha, T. Almalki, et al. An innovative approach to autocorrecting grammatical errors in arabic texts.Journal of King Saud University-Computer and InformationSciences, 2019.

[19] A. Sigogne, M. Constant, and E. Laporte. Int\’egration des donn\’ees d’un lexique syntax-ique dans un analyseur syntaxique probabiliste.arXiv preprint arXiv:1404.1872, 2014.

[20] E. Wehrli and L. Nerima. The fips multilingual parser. InLanguage Production, Cognition,and the Lexicon, pages 473–490. Springer, 2015.