Automatic discourse parsing of Arabic text: Building discourse structure of text

Hela Bettaieb[1], Ines Boujelben[2] and Iskandar Keskes[3]

[1] FSEGS, faculty of economics and management Sfax
Miracl lab-Sfax
[2] Gabes University, Tunisia
[3]Gafsa University, Tunisia
ANLP Research group, Miracl lab-Sfax
Helabettaieb1991@gmail.com
boujelben_ines@yahoo.fr
IskandarKeskes@gmail.com

Abstract. Discourse parsing of Arabic texts presents an important task in natural language processing (NLP) and it plays a critical role in discourse analysis. It is considered as a modern scientific concern. Its importance lies in its ability to determine the semantic and rhetorical meaning between discourse units through a coherent structure. Discourse structure analysis can benefit a variety of NLP applications such as question answering, machine translation, text categorization, etc. The rhetorical analysis is based on three pillars. The first pillar consists to segment the text into discourse units. The second pillar is to look for structural links (attachments) between different discourse units. The third pillar connects these units to each other via discourse relations. In this context, our task of automatic discourse parsing of Arabic text falls within the second pillar of rhetorical analysis. This perception of rhetorical analysis is based on the Segmented Discourse Representation Theory (SDRT) as framework of our proposed method and on the Random Forest classifier. Our method achieves encouraging results in terms of the F-measure (73%) when applied to the reference corpus Discourse Arabic TreeBank (D-ATB).

Keywords: Attachments, Discourse structure, Discourse Relations, Discourse analysis, Arabic text.

1 Introduction

According to the 2017 edition of Ethnologue [1], the Arabic language has around 290 million native speakers and a total of around 422 million total speakers. There are three types of the Arabic language: classical Arabic, modern standard Arabic (MSA) and dialectical Arabic. In addition, it is a Semitic language which differs from foreign languages by its syntactical complexity and its morphological richness because it is derivational and inflectional [2] and semantically wildness. It has no limits to its semantics meanings and its grammatical and morphological bases [3]. So to solve this complexity and to deal with the challenges of Arabic language, we should simplify the rhetorical structure of the Arabic texts [4]. It's noteworthy that the research realized in this framework of discourse parsing for Arabic language is very limited. Among the applications, we can mention Arabic social Media Analysis and translation [5], transform based Arabic sign language recognition [6] and Irony detection system for Arabic in social media [7]. Compared to other languages including for example the English language transition-based Neural RST parsing with implicit syntax feature [8], for the Chinese language we have joint modeling of structure identification and nuclearity recognition macro Chinese Discourse Treebank [9] and unified RvNN framework for end to end Chinese discourse parsing [10], and so forth.

The structural and hierarchical structure of the text is based on the principle of the rhetorical analysis which relies on three important phases. The first phase is to segment the text into text units (arguments). The second phase attempts to find structural links between the various units of text which is our focus. The last phase is to identify the different discourse relations that bind the text units.

To illustrate these three phases, we take the example (1) with its rhetorical structure in Figure1.

Example (1):

[Several national federations have initiated some activities [led by trade and transport sectors] _2 in pressuring the government]_1 [to impose increases on various materials, they considered it inevitable] _3 [to ensure a comfortable profit margin, based on increases] _4 [carried by the text of the finance Bill of 2018] _5 [and its direct impact on the prices of various raw materials] _6 [the president of national federation of private carriers declared, in statements to “Chourouk” yesterday] _7

[شرعت عدة اتحاديات و فدراليات وطنية ممثلة لبعض الانشطة، [تتقدمها قطاعات التجارة و النقل،]_2 في الضغط على
الحكومة]_1 [لفرض زيادات على مختلف المواد و الخدمات،اعتبروها حتمية]_3 [لضمان هامش ربح مريح، بناء على
الزيادات]_4 [التى حملها نص مشروع قانون المالية لسنة 2018]_5 [و انعكاساته المباشرة على اسعار مختلف المواد
الأولية.]_6 [وذكر رئيس الاتحادية الوطنية للناقلين الخواص، في تصريحات ل"الشروق"امس،]_7

The example (1) illustrates the three stages of the discourse analysis. Starting by the segmentation of text into elementary discourse units and according to the segmented discourse representation theory attributed to [11]. We note that each unit is numbered in order within the text. This example contains seven text units. The segmentation into discourse units aids the transition to the second phase, which attempts to search for the structural links that combine the various text units. In example (1), the first unit connects structurally with the complex unit that collects the third and the four unit and the simple unit: one and seven. This link between the first unit and the other units has a semantic meaning and a variety of discourse relations including for example “Temporal Ordering”, “Goal”, “Explanation” and “Description”.

We conclude that these structural links reflect the coherent picture of the rhetorical structure of the text. This structural connection is based entirely on the semantic and rhetorical meaning between discourse units. This presents the role of the third phase of discourse analysis by defining the discourse relations between discourse units. For example the third unit of the text is structurally related to the fourth unit through the discourse relation “explanation”.

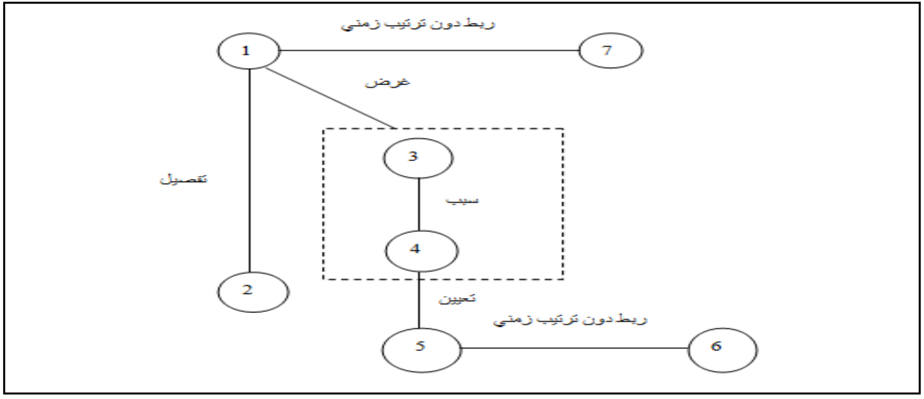

Figure 1. The discourse structure of the example (1)

Our method is based on the Segmented Discourse Representation Theory (SDRT) to search for structural links between discourse units. It is based on semi-supervised learning through the application of the Random Forest classifier [12]. In this context, we used the corpus (D-ATB) [13] which consists of a set of texts accompanied by its rhetorical structures and annotated by linguistic expert. The importance of such annotation makes the task of empirical and exploratory studies possible to realize on a range of texts. In addition, it aims to simplify the linguistic content and the rhetorical context. Our proposed method is implemented through the creation of new automated system “Discursive analysis of Arabic text” to build automatically the discourse structure of Arabic texts.

The rest of this article is organized as follows. Section 2 defines our corpus. Section 3 explains our proposed method of discourse parsing. Section 3 describes the features used in our way. The final section presents the different experiments from which we discuss the reported results.

2 The Discourse Arabic TreeBank corpus (D-ATB)

Our corpus consists of 50 annotated texts with morphological and grammatical information gathered from the Arabic TreeBank corpus [14]. Our corpus is composed of 1272 sentences, 2788 elementary discourse units and 372 (7.49%) embedded units. These texts are associated with their rhetorical structures according to the segmented discourse representation theory. Furthermore, it is annotated by a linguistic expert, providing a discourse structure to the reader.
The annotation path begins with the segmentation of the text into units based on a combination of morphological, lexical, grammatical and punctuation features [15]. The example (2) is segmented into two units.

Example (2):

[The news paper “Babylon”]_1 [supervised by Saddam Hussein, “officials in Turkey”] _2

[وذكّرت صحيفة "بابليون"] _1 [التي يشرف عليها صدام حسين "المسئولين في تركيا"] _2

In order to activate this segmentation, we used two analyzers of the context: Alkhalil [16] [17]. It provides a range of characteristics related to the word's morphological possibilities, weight, root, form and references. As for the morphological analyzer SAMA [18], it gives a set of characteristics related to the word represented in all its morphological possibilities with its root, its trunk, its pronunciation, its precedents and its suffixes.

After completing the segmentation stage, we move to the stage of determining discourse relations between different discourse units [19]. This stage relied on a set of lexical, semantic and morphological features and on a dictionary containing 174 connectors.

The relation between segments can be explicit, introduced directly through connectors from a context, or implicitly inferred through the semantic content between discourse units. If we retake the discourse relation between the first unit and the complex unit that collects the third and the fourth units of example (1), we can notice the explicit relation "Goal" expressed through the connector "to".

The discourse structure of the text is characterized by a coherent structural connection that links simple and complex units. As the case between the first unit and the complex unit composed of the third and the four units of the example (1) where we find a goal discourse relation. For example, the first unit related to the seventieth unit through a non-adjacent discourse relation that expresses a "Temporal Ordering".

To construct our corpus, we applied a set of processes. These processes are divided into two stages: the cleaning stage and the enrichment stage. The first stage is characterized by its ability to configure our data to extract the rules and the discourse relations between them. We note that before this stage, texts are presented as a succession and juxtaposition of discourse units which making difficult the task of discourse structure analysis. So, the structure of text has to be reformulated by segmenting it into paragraphs. Each paragraph has its own structure. So, the graph of text is composed of a set of sub-graphs. In our corpus, we have 130 paragraphs.

3 Proposed method

3.1 Segmented Discourse Representation Theory (SDRT)

The Segmented Discourse Representation Theory (SDRT) is a special framework that combines representative, dynamic and pragmatic aspects. In order to provide a coherent discourse structure called SDRS that embodies the hierarchical structure of discourse. Besides, it aims at identifying two types of discourse relations (horizontal and vertical) Figure 2. The example (3) shows that the discourse relation between the second and the third unit "Slow Ordering" is a horizontal relation. In contrast, the discourse relation "Background" between the first unit and the complex unit (second unit and third) is a vertical relation. Furthermore, this theory is based on two basic concepts. The first concept is DRS for the simple units and SDRS for the complex units. The second concept is the discourse relations which relate the discourse units. The complex structure of SDRS is a recursive and coherent structure. The addition of any new unit must be connected to the previous unit by discourse relations. This condition generates a restriction [20] that highlights the connection process on the right frontier. This

constraint requires each new unit to be connected to the last unit analyzed or to the units that dominates it hierarchically.
This theory is characterized by its ability to adopt the multiple attachments phenomenon. Indeed, two discourse units can be connected to several discourse relations.

Example (3):

[Zimbabwe’s government may face harsh sanctions after a « flawed » election renewed for Mugabe]_1 [After the re-election of President Robert Mugabe in an election that has been condemned by the international community] _2 [The Zimbabwe’s government is expected to face tougher international sanctions] _3

[حكومة زيمبابوي قد تواجه عقوبات قاسية بعد انتخابات "معيبة" جددت لموغابي]_1[بعد التجديد للرئيس
روبرت موغابي في انتخابات ندد المجتمع الدولي بما شابها من عيوب،]_2[يتوقع ان تواجه حكومة زيمبابوي
عقوبات دولية اشد قسوة،]_3

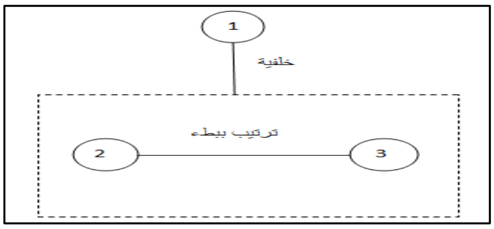

Figure 2. The discourse structure of the example (3)

3.2 Method description

Our proposed method of automatic discourse parsing applying to Arabic texts is based on the SDRT. Besides, it is principally based on a variety of features at different levels morphological, lexical, semantic and others. We applied a semi-supervised learning machine using the Random forest classifier. This classifier contributes to the treatment of unbalanced data during the classification phase. Figure 3 shows our method’s working model divided into three steps of discourse analysis in order to obtain a coherent structure based on the SDRS structure and its constraints. These constraints state the principle of the available units and their position on the right frontier. The type of the discourse relations contributes to guiding the discourse structure.

3.3 Proposed features

Our method is based on a variety of features, which are the basis of building the discourse structure. We compiled various types of features to describe our dataset:

- Punctuation marks

It has the ability to refer the binding sites and to detect some of the discourse relations between text units such as description, attribution. We have five punctuation marks that indicate the possible linkage situations, which are (? /. /, /: /!).

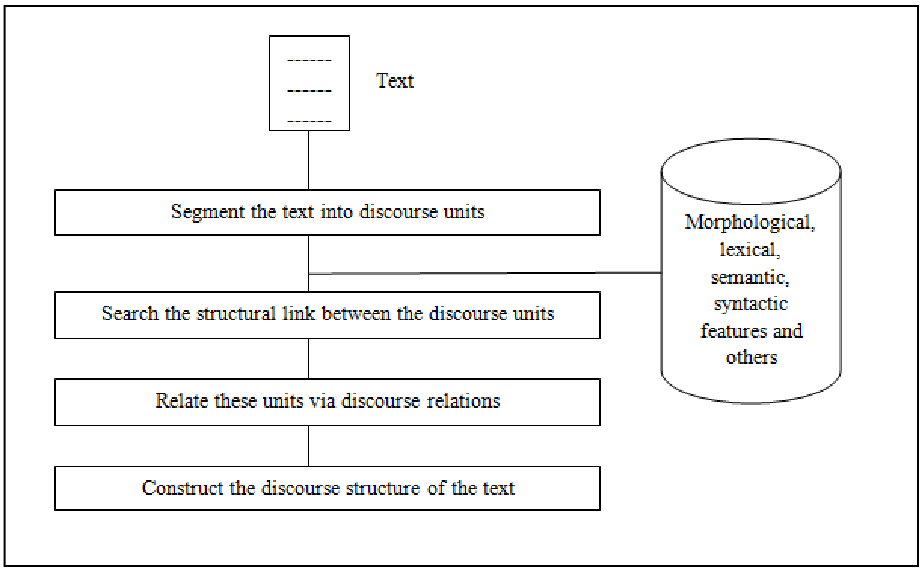
Text
Segment the text into discourse units
Search the structural link between the discourse units
Relate these units via discourse relations
Construct the discourse structure of the text
Morphological, lexical, semantic, syntactic features and others

Figure 3. The steps of the proposed method

Example (4):
[Fisher said:]_1 [“as long as the evidence not efficient”,] _2

[وقال فيشر:]_ 1["ما دام أن لادليل دامغاً،]_2

In this example, we find that the punctuation mark “:” links the arguments (1) and (2) with the discourse relation of attribution.

- The modals

This feature is based on a manually built lexicon consisting of 50 modal verbs such as (announce, declare, confirm, etc).

Example (5):

[The independent Russian agency “Interfax” announces]_1 [that a Russian military delegation went yesterday to the United States] _2

[أفادت وكالة "انترفاكس" الروسية المستقلة]_ 1 [أن وفداً عسكرياً روسياً توجه أمس إلى الولايات المتحدة]_2

The previous arguments (1) and (2) are linked by the modal verb “confirmed” which expresses an attribution relation.

- Polarity

It is mainly based on a subjective dictionary called MPQA [21]. This glossary is characterized by its extensive structure of English words and its translation into Arabic. It has different types:

positive, negative, both and neutral. Moreover, there are two kinds of strong and weak subjectivity.

- Distance and length

The use of these two concepts, help us to count the number of words in each argument and the number of elementary discourse units EDUs between the two arguments. Indeed, we assigned a binary feature to treat the distance in the tree between the connector and the argument. In the case where the argument and the connector are not in the same tree, we get a 0 as a value and 1 in the opposite case. This feature defines the position of the adjacent or non-adjacent attachment.

- Al-masdar

Is a binary feature that attempts to determine whether the first or the second word of each argument contains Al-masdar. This term reposes mainly on a verbal construction. To exploit it, we use the morphological analyzer Al-khalil (see example 6).

Example (6):

[Who re-appointed a new government]_1 [in preparation for the elections this year in the Kingdom] _2

[الذي أعاد تكليفه تأليف حكومة جديدة]_1[تمهيدا لإجراء انتخابات نيابية هذه السنة في المملكة.]_ 2

The process of connection between the two arguments (1) and (2) marked by Al-masdar "preparation" expresses the discourse relation "explanation".

- Textual organization

It indicates the position of each argument in the text (at the beginning, in the middle, at the end). This feature contributes to the attachment between the discourse units. Besides, it inferred certain relations such as interpretation, background, summary and temporal or spatial framework. The example (7) explains the utility of this feature in the attachment task between two discourse units. This connection expresses a relation of temporal frame.

Example (7):

[Later,] _1 [peace envoy Camilo Gomez arrived also in the region] _2

[وفي وقت لاحق،]_1[وصل إلى المنطقة أيضا المبعوث الرئاسي للسلام كاميلو غوميز]_2

- Named entity and anaphora

These two features contribute to solving some of the discourse relations and in the process of linking different discourse relations like parallelism in example (8). The named entity is treated through the ANERGazet Gazetteers [22]. We have also built manually a lexicon of 60 words of pronouns and anaphora.

Example (8):

[I was stylish today for the feast]_1 [and my sister was also stylish for this occasion] _2

[تأنقت اليوم بمناسبة العيد] _1 [وكذلك أختي تأنقت لهذه المناسبة] _2

The process of connection between these two units is done by adopting the same structure in each unit indicating a relation of parallelism.

- Embedded argument

It is a binary feature. Its structural links expresses a set of discourse relations, for example entity elaboration relation example (9) and comment, etc.

Example (9):

[The American aircraft bombed a complex of cave yesterday in eastern of afghansten]_1 [which launched by the fighters of "alqaada" and "Taliban", in the time] _2

[قصفت طائرات أميركية مجمعات كهوف في شرق أفغانستان أمس ضمن الحملة] _1[التي تشنها على مقاتلي تنظيم "القاعدة" وحركة "طالبان" الإسلامية، في الوقت] _2

- Time and negation

The time is an interesting feature and a good indicator for events progression and it can be fast, slow or synchronous in example (10). However, the negation is very useful feature for identifying attachments between discourse units. This is achieved through the construction of a manual lexicon composed of 10 Arabic words express the negation. These two concepts contribute also to the emergence of many discourse relations including condition and result.

Example (10):

[She found a man in civilian clothes pointing a gun at her head,] _1 [suddenly, he pulled her and hit her in the face with shackles] _2

[فوجدت رجلاً في ثياب مدنية يوجه مسدساً إلى رأسها،] _1 [سرعان ما جذبها وضربها في وجهها بالأغلال".]_2

This example is based on the time feature in particular, the rapid progression of the event marked by the adverb "suddenly", which combines units (1) and (2).

Example (11):

[If you do not get rid of yours fears]_1 [you will not be a successful person] _2

[إذا لم تتخلص من مخاوفك] _1 [لن تكون شخصًا ناجحًا] _2

We note that the connection between two discursive units (1) and (2) realized by the negation feature expressing the relation of result.

- The connectors

This feature is considered as a valuable tool with double functions. The first contributes on the automatic segmentation of discourse units and the second promotes the connection process. The implementation of this feature is based on six string characteristics of each connector namely:

the connector, its lemma, the POS, the type (clitic, simple or compound), its syntactic path and the position of the connector (at the beginning, the middle and the end) example (12).

Example (12):

[The international community continues its efforts]_1 [in order to seek peace in this country] _2

[تواصل الاسرة الدولية جهودها]_1[من اجل البحث عن السلام في هذا البلد]_ 2

The first unit of the text is connected directly and locally to the second unit of discourse through the connector “in order to “which indicates the relation of cause.

- The semantic relations

The semantic relations express the semantic meaning of attachments between discourse units. In addition, this feature is mainly based on an Arabic lexicon ArabicWordNet [23]. We have also used a rich and advanced lexicon version proposed by [24]. It consists of 15000 elements and 17 semantic relations (see example 13).

Example (13):

[Alex reads historical books.] _1 [In contrast, his sister write fictional stories] _2

[أليكس يطالع كتبا تاريخية]_1[في المقابل, أخته تألَف قصصا خيالية] _2

We conclude that the attachment between units (1) and (2) is through the semantic relation ”near_antonym” and through the connector “in contrast”, which expresses the contrast relation.

- Lexical cues

This lexical feature contributes efficiently to the structural links between discourse units based on lexicon of discourse connectors. This lexicon contains 174 connectors. Each connector is divided into two different sections. The first section consists of discursive indices such as “but” and the second section consists of a set of indicators (adverbs, particles, etc). In addition, these two features are represented with: strong and weak. If the indicator is strong, the attachment is strong also (but, however, etc) (see example 14). If it is weak, it means that there is an ambiguity. Therefore, the connection process between discourse units is unclear (for, even, and).

Example (14):

[He began a scheduled tour of three states in the Midwest] _1 [while his doctor confirmed that this symptom is not dangerous] _2

[فاستهل جولة مقررة على ثلاث ولايات في الغرب الاوسط]_1 [بينما أكد طبيبه ان هذا العارض ليس خطيراً   ]_2

In this example the first unit (1) is connected to the second unit (2) by a strong connector “while” which expresses a contrast relation.

- The arguments

The arguments are the fundamental constituents for formulating the discourse structure and insuring the attachments. The first three words of each argument include POS and its grammatical class.

3.4 Experiments and results

The input data used in our experiments consists of Arabic texts collected from the ATB corpus. For the ML technique, we exploited the J48, BayesNet, and ZeroR classifier, which are available in the WEKA tool. We used the classical measures of precision, recall and *F*-measure. We discover that the J48 algorithm obtained the best results with a high accuracy 99%. This increase is attributed to one class related to the attachments. This inflation causes the problem of unbalanced data [25] between the two classes of attachments. To deal with this imbalance, we tested three other algorithms (Random Forest, LibSVM and KNN) to handle with imbalanced data as shown in Table 1. But we used the Random Forest algorithm and the technique of under sampling filter, which contributes to the balanced distribution between the two binary classes of attachments.

The algorithm	**F-measure**	**Recall**	**Precision**
KNN	59%	60%	59%
LIBSVM	65.2%	65.1%	65%
Random Forest	%73	73%	73%

Table 1. The metrics of evaluation of the tested algorithms

The results shown in Table 1 indicate that our system achieves an encouraging f-measure to obtain 73% when comparing to other classifiers. In fact, the KNN algorithm achieves a low value of precision 59% and recall 60%. While the LIBSVM provides better performance in terms its precision, which had a 6% improvement, while the recall was slightly lower for this method. This table reports that our system achieves the best improvement to obtain 73% in terms of the precision, recall and *F*-score.

In total, we have 3937 instances incorrectly classified and 1429 instances correctly classified and the absolute average of error is 0.42. At the base of the confusion matrix, we interpret that the number of negative attachments correctly classified is very important as shown in Table 2.

	Negative Prediction (0)	Positive Prediction (1)
Current negative (0)	2154	529
Current positive (1)	900	1783

Table 2. The confusion matrix

In term of error analysis of sources, we can mention that the absence of some punctuation marks makes our system ambiguous to identify the correct discourse units. We mentioned that Arabic language is characterized by irregular punctuation marks. Again, we find that the rate of error in the the non-adjacent attachments is more than adjacent attachments.

Conclusion

This article presents a novel method of automatic discourse parsing of Arabic texts, based on the SDRT. We initially constructed the Discourse Arabic TreeBank corpus. Then, we explained the three stages of our adopted method to build the attachments between the various discourse units. As a result, we obtained a discourse graph with the coherent discourse structure of the text. The proposed method has been included in our tool "Discourse Analysis of Arabic Texts" that obtains encouraged results (73% of F-measure). Indeed, our tool is characterized by its ability to be integrated in several applications of the field of the automatic natural language processing for example the automatic summarization, the automatic translation, the question-answer system, etc.

References

[1] Lewis, Gary Simon and Charles Fennig, 2017. Ethonologue: languages of the world, twentieth edition. Dallas, Texas: SIL international (2017).

[2] Al-saleh, Asma Bader and Mohamed el Bachir Menai, 2017. Automatic Arabic text summarization: a survey. Artificial Intelligence Review (2017).

[3] Farghaly Ali and Shaalan Khaled, 2009. Arabic Natural language processing. ACM transactions on Asian Language Information processing 2009,volume8.

[4] Al-Ayyoub Mahmoud, Nuseir Aya , Alsmearat Khouloud, Jaraweh Yaser and Gupta Brij, 2018. Deep learning for Arabic NLP.Journal of computational science 2018,volume 26.

[5] Mallek Fatma, Belainine Billal and Fatiha Sadat, 2017. Arabic social Media Analysis and Translation. 3rd International conference on Arabic Computational Linguistics, ACLing 2017. Dubai, united Arab Emirates.

[6] Ala addin Sidig, Hamzah luqman, Sabri Mahmoud. Transform-based arabic sign language recognition. 3rd International conference on Arabic Computational Linguistics, ACLing 2017. Dubai united Arab Emirates.

[7] Karaoui Jihen, Zitoune Benamara Farah and Moriceau Véronique, 2017. SOUKHRIYA : Towards an Irony Detection System for Arabic in Social Media. 3rd International conference on Arabic computational Linguistics, ACling 2017. Dubai, Unitd Arab Emirates.

[8] Nan yu, Meishan Zhang and Guohong Fu, 2018. Transition-based neural RST parsing with implicit syntax feature. Proceedings of the 27th International Conference on Computational Linguistics Santa Fe, New Mexico, USA, , 2018.

[9] Xiaomin Chu, Feng Jiang, Yi Zhou, Guodong Zhou, Qiaoming Zhu, 2018. Joint Modeling of structure identification and Nuclearity Recognition in Macro Chinese Discourse Treebank. Proceedings of the 27th International Conference on Computational Linguistics, New Mexico, USA, 2018.

[10] Chuan-An Lin, Hen-Hsen Huang, Zi Yuan Chen and Hsin-Hsi Chen, 2018. A unified RvNN Framework for end to end Chinese Discourse paesing. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, New Mexico, USA, 2018

[11] Asher Nicholas and Lascarides Alex, 2013. Segmented Discourse Representation Theory: Dynamic Semantics with Discours Structure

[12] Chen Chao and Breiman Leo, 2004. Using RandomForest to learn Imbalanced Data.

[13] Maamouri Mohamed, Bies Ann, Kulick Seth, Gaddeche Fatma, Mekki Wigdan, Krouna Sondos and Bouziri Basma, 2008. Arabic TreeBank part 3-v3.0. Linguistic Data Consortium.

[14] Maamouri Mohamed and Bies Ann, 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures and Tools. Proceedings of the workshop on Computational approaches to Arabic script-based language.

[15] Keskes Iskandar, Zitoune Benamara Farah and Bilguith Hadrich Lamia, 2013. Segmentation de textes arabes en unités discursives minimales. Conférence du traitement automatique des langues naturelles TALN 2013.

[16] Boudlal Abdderahim, Lakhouaja Abd-Alhak, Mazrouri Azzeddine, Meziane Abdel Wafi, Bebah Mohamed ould Abdallah and Shoul Mustepha, 2010. Alkhalil morphsys1 : A morphosyntactic analysis system for Arabic Texts. International Aran conference on information Technology. Benghazi Libya.

[17] Boudlal Abdderahim, Lakhouaja Abd-Alhak, Mazrouri Azzeddine, Meziane Abdel Wafi, Bebah Mohamed ould Abdallah and Shoul Mustepha, 2017. Morpho-syntactic analyzer. Journal of King Saud University-Computer and information sciences 2017,volume 29.

[18] Maamouri Mohamed, Graff David, Bouziri Basma, Krouna sondos Bies Ann and Kulick seth, 2010. LDC Standard Arabic Morphological Analyzer (SAMA) version 3.1. Philadelphia : Linguistic Data Consortium 2010.

[19] Keskes Iskandar, Zitoune Benamara Farah and Belguith Hadrich Lamia, 2014. Learning explicit and implicit Arabic discourse relations. Journal of King saud University-Computer and information sciences 2014,volume 26.

[20] Asher Nicholas, 2008. Troubles on the Right Frontier.

[21] Elarnaoty Mohamed, AbdelRahman Samir and Fahmy Aly, 2012. A machine learning Approach for opinion Holder Extraction in Arabic Language. International Jouranl of Artificial Intelligence and Applications (IJATA),volume 3.

[22] Benajiba yassine, Rosso Paolo and BenediRuiz Miguel Jose, 2007. ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. 8th international conference on intelligent text Processing and Computational Linguistics.CICling 2007,volume 4394.

[23] Elkateb Sabri, Black William, Vossen Piek, Rodriguez Horacio, Alkhalifa Musa, Pease Adam and Fellbaum Christiane, 2006. Building a WordNet for Arabic. Proceedingsof the 5th conference on language Resources and Evaluation LREC 2006.

[24] Boudabous Mahdi Mohamed, Kammoun Nadege, Khedher Nacef, Belguith Hadrich Lamia and Sadat Fatiha, 2013. Arabic WordNet Semantic relations enrichment through morpho-lexical patterns. 1st international conference on communications, signal processing and their Applications.

[25] Sun Yanmin, Wong Andrew and Kamel Mohamed, 2009. Classification of Imbalanced data : a review. International journal of Pattern Recognition and Artificial intelligence,volume 23.