Open
Full Length Article

ORIGINAL ARTICLE

Identification of Discourse Boundaries Using Anaphorically Annotated Text

Muhammad Aatif[1], Muhammad Abid Khan[2]

[1]MS Scholar, Department of Computer Science, University of Peshawar, Peshawar
[2]Professor, Department of Computer Science, University of Peshawar, Peshawar

ABSTRACT

For effective and efficient natural language processing (NLP) systems, it is of great significance that a discourse unit must be made a unit of processing. The major obstacle to achieving that goal is the identification of discourse boundaries (DBs). This paper presents an algorithm about the identification of DBs in English text using anaphorically annotated text. The proposed algorithm is based on the structure of anaphoric annotations carried out in Phrase Detectives Corpus 2.1.4 (PD2). It has been tested on anaphorically annotated documents selected at random from PD2 and showed an accuracy of 97.66%.

Keywords: Discourse, Discourse unit, Discourse Boundaries, Boundaries identification, Discourse Analysis, Discourse Boundaries Identification, Anaphorically Annotated corpus, Algorithm

Author`s Contribution
[1]Manuscript writing, Data Collection, Data analysis, interpretation, [2]Conception, synthesis, planning of research, and discussion

Address of Correspondence
Muhammad Aatif
maatif@uop.edu.pk

Article info.
Received: Aug 05, 2019
Accepted: June 09, 2020
Published: June 30,2020

Cite this article: *Aatif, M, Khan MA. Identification of Discourse Boundaries Using Anaphorically Annotated Text. J. Inf. commun. technol. robot. appl.2020; 11(1):47-53*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

The term discourse was used for the first time in 1952 by Zellig Harris. Harris defined it as a joined chunk of text which may consist of one or more sentences connected by different words known as anaphoric devices [1]. Consider the following:

"[Ali is son of Jamila. He is ten years old. Kashif is his elder brother.] [Kashif is a student of MS Computer Science. He studies in University of Peshawar.]"

This is an example which consists of two discourse units delimited from each other by the squared brackets. The brackets represent discourse boundaries (DBs). Related work in the discipline of Discourse Analysis (DA) shows that most of earlier work has focused more on word and individual sentence-based studies. However, when this study is extended beyond sentence level and

discourse unit is introduced as a unit of processing to NLP systems e.g. text understanding, text simplification, text classification, text summarization, Question-Answering systems, machine translation and machine learning more complexities are resolved and performance of these applications can be boosted up to a great extent [2]–[8]. Most of these NLP systems need such a unit of text which is complete *semantically & referentially* and describe the *sub/idea* entirely. When such a unit of text is provided as a unit of processing these NLP systems would boost up in terms of accuracy, efficient processing and getting more effective & useful results. Therefore, to have such a unit of text it is indispensable to first identify DBs. Therefore,

this paper presents an algorithm for the identification of DBs in English text using anaphorically annotated text.

LITERATURE REVIEW

DA has been used for many applications e.g. Reciprocal Anaphora Resolution [9], the representation and identification of discourse markers and their errors correction and possible applications [10]–[13], making the process of group discussion easier and more fruitful [14], and comparing the discourse markers annotations, done both manually and automatically [15]. The authors in [16]–[21] have summarized the text summarization tools and techniques. They have performed the performance analysis of these tools and techniques from discourse and knowledge discovery perspective and exposed great importance of DA to the text summarization techniques. Authors in [22]–[25] have proposed innovative machine learning approaches and reworked on the model of the language for language and dialect identification. In case of language identification, they concluded with experimental justification that the unsupervised approaches which adapts language models should be considered in all languages identification tasks due to its encouraging performance. For the task of dialect identification their experiments preferred the use of derived feature vector over the use of raw feature vectors.

Similarly, the problem of anaphora resolution has also attracted focus of many researchers. These researchers have worked to provide innovative solutions, to improve the available ones and also to make a comparative analysis among them, in different natural languages such as English, Hindi, Spanish and Pashto using different approaches [26]–[36].

Text classification in is also one such problem addressed by many in past and proposed traditional and machine learning-based approaches [37]–[40]. The authors in [37] proposed a method called *compound word statistics* which makes use of feature extraction followed by a ranking function. The features that are considered for extraction are lexical, semantic and syntactic feature. The single ranking function is used to consider the frequencies of the extracted features. This method performs about 6% – 10% better than other 6 well-known approaches in terms of accuracy.

The task of translating the source language to the target language is called machine translation. Different statistical and machine learning based models are proposed for this duty in past while some have proposed hybrid approaches. The models in [41], [42] mainly makes use of a parser to parse the source text, an analyzer to analyze the source text, a mapping function to map the sentences of the two languages and a sentence generator to finally generate the target language sentences.

Sentiment analysis is another popular task in NLP that is usually applied to social media texts for deciding on critical business decisions. In [43] the authors used three methods i.e. *bag* of words, *Word2Vec* and *Doc2Vec* for the first time in Turkish texts to carry out the task of sentiment analysis on Twitter tweets. They justified that Word2Vec is the method which performs better when compared with other methods for text representation while Random Forest is considered the most effective among the many machine learning algorithms. Another recent study on sentiment analysis is done by Mariam Khader et al and Anwar Alnawas et al [44], [45] on social media texts. They used a technique called MapReduce to study big data. A twitter dataset is used in Mariam Khader et al study and enhanced the accuracy of the sentiment analysis. Authors in [46] developed a sentiment analysis system which is applicable to a resource-poor language called Roman Urdu. A dataset is also created followed by its annotations in [46]. The system showed 80.07% accuracy rate and reduced the error rate up to 12%.

Research shows that the earlier work on discourse analysis is very different from each other but one thing which is common in most of them is that, they need such a unit of text which is complete at least from three perspectives i.e. to achieve effective and useful results the unit of text needs to have/include the details regarding the *Semantics, Anaphoric References* & *idea/topic* being discussed in the unit of text. Therefore, it is necessary to identify DBs so that the relevant NLP systems could enjoy such a unit of text.

ALGORITHM

The proposed algorithm is implemented in Python 3.6 using Spyder Integrated Development Environment.

The developed system operates on anaphorically annotated text. It processes the input document sentence by sentence. At the end of each sentence processing, based on some algorithm conditions (given in pseudo code), it is decided to mark or not to mark discourse boundaries. After processing the whole input document, a new textual file is generated which serve as the final output. This file now contains the source text with discourse boundaries being identified.

- **Input to the Algorithm**

The algorithm inputs an anaphorically annotated document at a time. Among many, Phrase Detectives Corpus 2.1.4 is one such latest corpus that has been anaphorically annotated [47]–[50]. In total, phrase detectives corpus 2.1.4 consists of 542 documents having 408153 tokens and 49990 markables (any linguistic expression of interest). Over ten years of efforts and with the help of 1958 annotators, second version of phrase detectives corpus 2.1.4 was released in January 2019 which makes this corpus more reliable. Its current release consists of two subsets: Silver and Gold. The algorithm is tested on documents of the gold subset in the hope of being more accurate than the silver one because the gold subset is additionally annotated by two expert annotators.

- **PSEUDO CODE**

The algorithm processes the sentences of the input document in a loop. The discourse starting and ending boundary is marked by the algorithm before the first sentence and after the last sentence of the input document. The discourse starting boundary is placed to the head of the first sentence while the discourse ending boundary is placed to the tail of the last sentence.

If an anaphoric device is identified in a sentence, it means that it is pointing within a discourse unit that is already started. Therefore, there is neither need to start a new discourse nor to close the already started one. After this, the algorithm just checks the algorithm termination condition i.e. whether the current sentence is the last sentence of the input document then, generate the final output file from the processed input document otherwise, read the next sentence.

If there is no anaphoric device in a sentence, it means that there is no reference to the discourse under process. As a result, it ends here while a new discourse begins from this sentence. In this situation, the algorithm places a discourse ending and starting boundaries to the head of the current sentence. Finally, the algorithm termination condition is checked once again, and appropriate actions are performed as explained in above paragraph. Algorithm is given below, and its flowchart is given in Figure 1.

1. Open an anaphorically annotated document
2. Read next sentence from the anaphorically annotated input document
3. If current sentence being read is the first sentence of the input document
    a. Mark a discourse starting boundary
    b. If this sentence contains one or more anaphoric device for the antecedents which are in previous sentence, then
        i. If current sentence is the last sentence of the input document
            1. Mark a discourse ending boundary
            2. Go to step 5.
        ii. Else, Go to step 2.
4. If there's no anaphoric device in current sentence
    a. Mark a discourse ending boundary
    b. Mark a new discourse starting boundary
    c. If current sentence is the last sentence of input document
        i. Mark a discourse ending boundary
        ii. Go to step 5.
    d. Else, Go to step 2.
5. Copy the processed XML input document into a new textual file that serves as final output file

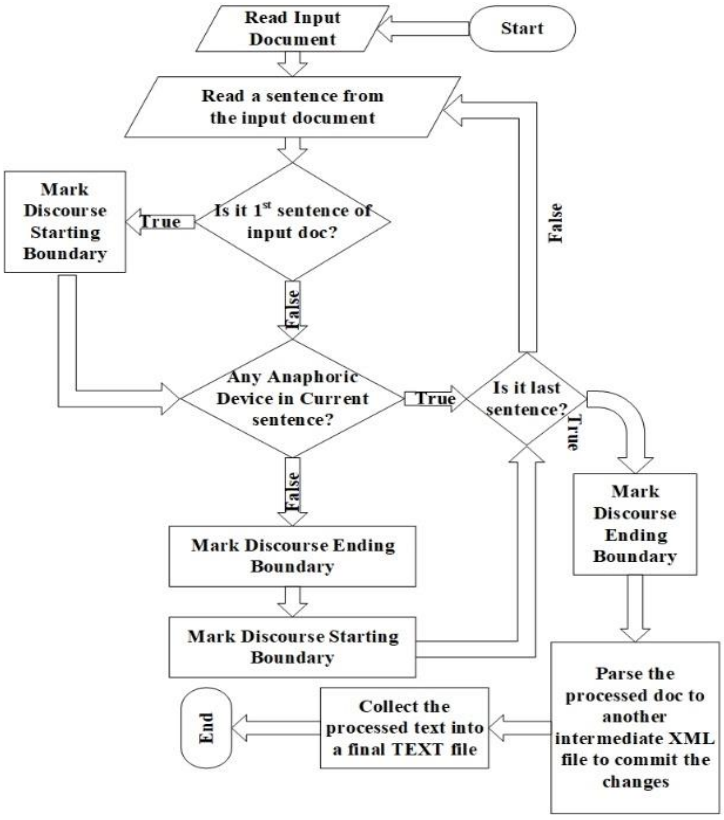

Figure 1: Algorithm Flowchart

EXPERIMENTS AND RESULTS

Our algorithm has been tested on documents/samples selected at random from the gold subset of phrase detectives' corpus 2.1.4 and showed an accuracy of 97.66%. After error analysis, it is revealed that most of the times our algorithm gives incorrect result when there is a fault in the anaphorically annotated input text. Therefore, based on the algorithm evaluation (explained later), it is accomplished that the more error free is the anaphorically annotated input text the better is the algorithm accuracy.

DISCUSSION AND EVALUATION

A similar study has been carried out on discourse boundaries identification using a different approach [51] having accuracy of 96%. The algorithm proposed in [20] can be applied only to English monologue which narrows down its applicability. Moreover, that algorithm design follows a static type approach which makes it very difficult to improve its accuracy.

In this paper, an algorithm for the identification of discourse boundaries using anaphorically annotated text is introduced which is the main contribution of this paper. In the proposed approach, the algorithm makes use of anaphorically annotated text that makes our work unique. Therefore, our work laid the foundation for all those future studies on the same subject that will make use of anaphorically annotated text. Moreover, our algorithm can be applied to both monologue and dialogue instead of only monologue text as reported in [51].

After careful error analysis, it was observed that most of the time the algorithm performance is affected due to faults in the anaphorically annotated input text. All the input faults affecting the accuracy of the algorithm are categorized into three categories as given in the sub-sections below.

- **Missing Annotation**

This category of faults includes all the missing annotations. Missing annotations means that some entities/phrases during annotations have not been annotated by the annotators at all. Missing of this annotation happened due to lack of the knowledge of anaphoric annotations, ambiguity in the anaphoric devices and their corresponding antecedents or some other error in the annotating software e.g. a phrase could not be selected.

- **Incorrect Annotation**

In this category, all those annotations include which are incorrect. Annotators have performed annotations but those are incorrect.

- **Inconsistent Annotation**

This category of faults includes all those annotations which have been carried out in the corpus. Although they are correct, but they violate the annotation rules of the corpus. Each markable/phrase in Phrase Detectives Corpus 2.1.4, if it is a pointing word, must point to its antecedent indirectly through the nearest intermediate phrase/antecedent. When the annotators do not follow this rule, the resultant annotation called an inconsistent annotation due to which the algorithm accuracy is affected.

After corrections of the input faults the proposed algorithm accuracy increased from 88.72% to 97.66%. Therefore, based on this, it can be emphasized that if an anaphorically annotated text is to be used for discourse boundaries identification then accurate anaphoric annotations serve as backbone for achieving high accuracy in discourse boundaries identification systems.

The results we achieved can now be used by other NLP systems such as text understanding, text simplification, text summarization, Question-Answering systems, text paraphrasing, machine translation and other similar systems to get more accurate and highly valuable results.

The implemented algorithm, however, has two limitations as follows:

- The algorithm can process text only if it is anaphorically annotated
- The anaphorically annotated text needs to be in XML format

CONCLUSION

Accurate identification of discourse boundaries is a preprocessing step that once performed can improve the performance of other NLP systems after this barrier is removed. In this paper, an innovative algorithm is designed and implemented for the identification of discourse boundaries which takes anaphorically annotated text as input. The proposed algorithm is tested on anaphorically annotated documents taken from phrase detectives' corpus 2.1.4 which produced accuracy of 97.66%.

In future it is intended to implement the idea of discourse boundaries identification in Pakistani languages having Urdu language as the first priority. The two limitations of this study will also be removed in future. To do that, an anaphorically annotating software will be developed and combined with the present algorithm. This annotating software will be acting as an intermediate processor to input natural text in its original form, anaphorically annotate it and produce the results in XML form that will then become input to the present algorithm. When this intermediate program is combined with the current algorithm then this whole system will be able to input text in natural form and produce the same text as output with discourse boundaries being identified.

REFERENCES

1. Z. S. Harris, "Discourse Analysis," Language, vol. 28, no. 1, pp. 1–30, 1952, doi: 10.2307/409987.
2. A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," Artif. Intell. Rev., vol. 52, no. 1, pp. 273–292, Jun. 2019, doi: 10.1007/s10462-018-09677-1.
3. Y. HaCohen-Kerner, R. Dilmon, M. Hone, and M. A. Ben-Basan, "Automatic classification of complaint letters according to service provider categories," Inf. Process. Manag., vol. 56, no. 6, p. 102102, Nov. 2019, doi: 10.1016/j.ipm.2019.102102.
4. A. Azaria, S. Srivastava, J. Krishnamurthy, I. Labutov, and T. M. Mitchell, "An agent for learning new natural language commands," Auton. Agents Multi-Agent Syst., vol. 34, no. 1, p. 6, Dec. 2019, doi: 10.1007/s10458-019-09425-x.
5. S. Jung, "Semantic vector learning for natural language understanding," Comput. Speech Lang., vol. 56, pp. 130–145, Jul. 2019, doi: 10.1016/j.csl.2018.12.008.

6. S. Romeo et al., “Language processing and learning models for community question answering in Arabic,” Inf. Process. Manag., vol. 56, no. 2, pp. 274–290, Mar. 2019, doi: 10.1016/j.ipm.2017.07.003.
7. O. Kolomiyets and M.-F. Moens, “A survey on question answering technology from an information retrieval perspective,” Inf. Sci., vol. 181, no. 24, pp. 5412–5434, Dec. 2011, doi: 10.1016/j.ins.2011.07.047.
8. P. Nakov, L. Màrquez, A. Moschitti, and H. Mubarak, “Arabic community question answering,” Nat. Lang. Eng., vol. 25, no. 1, pp. 5–41, Jan. 2019, doi: 10.1017/S1351324918000426.
9. R. Ali, M. A. Khan, M. Bilal, and I. Rabbi, “Reciprocal anaphora resolution in Pashto discourse,” in 2008 4th International Conference on Emerging Technologies, Oct. 2008, pp. 1–5, doi: 10.1109/ICET.2008.4777464.
10. P. A. Heeman, D. Byron, and J. F. Allen, “Identifying Discourse Markers in Spoken Dialog,” presented at the AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing, Menlo Park, California, March 1998., pp. 44–51.
11. L. Sepúlveda-Torres, M. Sanches Duran, and S. M. Aluísio, “Automatic detection and correction of discourse marker errors made by Spanish native speakers in Portuguese academic writing,” Lang. Resour. Eval., vol. 53, no. 3, pp. 525–558, Sep. 2019, doi: 10.1007/s10579-019-09467-3.
12. X. Kang, C. Zong, and N. Xue, “A Survey of Discourse Representations for Chinese Discourse Annotation,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 18, no. 3, pp. 26:1–26:25, Jan. 2019, doi: 10.1145/3293442.
13. S. Joty, G. Carenini, R. Ng, and G. Murray, “Discourse Analysis and Its Applications,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Florence, Italy, Jul. 2019, pp. 12–17, doi: 10.18653/v1/P19-4003.
14. K. Tomiyama, F. Nihei, Y. I. Nakano, and Y. Takase, “Identifying Discourse Boundaries in Group Discussions using a Multimodal Embedding Space,” in IUI Workshops, 2018, vol. 2068.
15. P. Furkó, “The Boundaries of Discourse Markers – Drawing Lines through Manual and Automatic Annotation,” J. Sapientia Hung. Univ. Transylvania, vol. 10, no. 2, pp. 155–170, Nov. 2018, doi: 10.2478/ausp-2018-0020.
16. A. Kanapala, S. Pal, and R. Pamula, “Text summarization from legal documents: a survey,” Artif. Intell. Rev., vol. 51, no. 3, pp. 371–402, Mar. 2019, doi: 10.1007/s10462-017-9566-2.
17. C. Orăsan, “Automatic summarisation: 25 years On,” Nat. Lang. Eng., vol. 25, no. 6, pp. 735–751, Nov. 2019, doi: 10.1017/S1351324919000524.
18. K. Kaczmarek-Majer and O. Hryniewicz, “Application of linguistic summarization methods in time series forecasting,” Inf. Sci., vol. 478, pp. 580–594, Apr. 2019, doi: 10.1016/j.ins.2018.11.036.
19. M. Ahmed, “Data summarization: a survey,” Knowl. Inf. Syst., vol. 58, no. 2, pp. 249–273, Feb. 2019, doi: 10.1007/s10115-018-1183-0.
20. P. Verma, S. Pal, and H. Om, “A Comparative Analysis on Hindi and English Extractive Text Summarization,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 18, no. 3, pp. 30:1–30:39, May 2019, doi: 10.1145/3308754.
21. S. Verberne, E. Krahmer, S. Wubben, and A. van den Bosch, “Query-based summarization of discussion threads,” Nat. Lang. Eng., vol. 26, no. 1, pp. 3–29, Jan. 2020, doi: 10.1017/S1351324919000123.
22. N. B. Chittaragi and S. G. Koolagudi, “Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms,” Lang. Resour. Eval., Nov. 2019, doi: 10.1007/s10579-019-09481-5.
23. T. Jauhiainen, K. Lindén, and H. Jauhiainen, “Language model adaptation for language and dialect identification of text,” Nat. Lang. Eng., vol. 25, no. 5, pp. 561–583, Sep. 2019, doi: 10.1017/S135132491900038X.
24. Dr. M. Awais and Dr. M. Shoaib, “Role of Discourse Information in Urdu Sentiment Classification: A Rule-based Method and Machine-learning Technique,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 18, no. 4, pp. 34:1–34:37, May 2019, doi: 10.1145/3300050.
25. A. Alishahi, G. Chrupala, and T. Linzen, “Analyzing and Interpreting Neural Networks for NLP: A Report on the First BlackboxNLP Workshop,” Nat. Lang. Eng., vol. 25, no. 4, pp. 543–557, Jul. 2019, doi: 10.1017/S135132491900024X.
26. M. Palomar et al., “An Algorithm for Anaphora Resolution in Spanish Texts,” Comput Linguist, vol. 27, no. 4, pp. 545–567, Dec. 2001.
27. S. Singh, P. Lakhmani, P. Mathur, and S. Morwal, “Analysis of Anaphora Resolution System for English Language,” Int. J. Inf. Theory, vol. 3, no. 2, pp. 51–57, Apr. 2014, doi: 10.5121/ijit.2014.3205.
28. R. Bunescu, “Associative Anaphora Resolution: A Web-Based Approach,” in Proceedings of the 2003 EACL Workshop on The Computational Treatment of Anaphora, 2003.
29. R. J. Evans and C. Orasan, “NP Animacy Identification for Anaphora Resolution,” J. Artif. Intell. Res., vol. 29, pp. 79–103, Jun. 2007, doi: 10.1613/jair.2179.
30. P. Lakhmani, S. Singh, and S. Morwal, “Performance Analysis of two Anaphora Resolution System for Hindi Language,” vol. 3, no. 3, pp. 576–580, 2014.
31. M. A. Khan and F. T. Zuhra, “Role of Corpus in Anaphora Resolution,” presented at the Corpus Linguistics, ICC Birmingham, Jul. 2011.
32. R. Ali, M. A. Khan, R. Ahmad, and I. Rabbi, “Rule based personal references resolution in pashto discourse for better machine translation,” in 2008 Second International Conference on Electrical Engineering, Mar. 2008, pp. 1–6, doi: 10.1109/ICEE.2008.4553941.
33. R. Iida, K. Inui, and Y. Matsumoto, “Anaphora resolution by antecedent identification followed by anaphoricity

determination," ACM Trans Asian Lang Inf Process, vol. 4, no. 4, pp. 417–434, 2005, doi: 10.1145/1113308.1113312.
34. M. Sen, N. Shah, and L. Kurup, "An algorithm for resolution of Anaphora in English text," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2017, pp. 1–5, doi: 10.1109/ICIIECS.2017.8276078.
35. Y. Zhu, W. Song, X. Liu, L. Liu, and X. Zhao, "Improving Anaphora Resolution by Animacy Identification," in 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Mar. 2019, pp. 48–51, doi: 10.1109/ICAICA.2019.8873499.
36. R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, "Anaphora and coreference resolution: A review," Inf. Fusion, vol. 59, pp. 139–162, Jul. 2020, doi: 10.1016/j.inffus.2020.01.010.
37. A. Adel, N. Omar, M. Albared, and A. Al-Shabi, "Feature selection method based on statistics of compound words for arabic text classification," Int Arab J Inf Technol, vol. 16, no. 2, pp. 1–8, 2019.
38. J. Kaur and J. Saini, "Designing Punjabi Poetry Classifiers Using Machine Learning and Different Textual Features," Int. Arab J. Inf. Technol., vol. 17, no. 1, pp. 38–44, Dec. 2019, doi: 10.34028/iajit/17/1/5.
39. R. Sprugnoli and S. Tonelli, "Novel Event Detection and Classification for Historical Texts," Comput. Linguist., vol. 45, no. 2, pp. 229–265, Mar. 2019, doi: 10.1162/coli_a_00347.
40. M. Martinc and S. Pollak, "Combining n-grams and deep convolutional features for language variety classification," Nat. Lang. Eng., vol. 25, no. 5, pp. 607–632, Sep. 2019, doi: 10.1017/S1351324919000299.
41. E. Alkim and Y. Çebi, "Machine translation infrastructure for turkic languages (MT-Turk)," Int Arab J Inf Technol, vol. 16, no. 3, pp. 380–388, 2019.
42. S. Khan and I. Usman, "A model for English to Urdu and Hindi machine translation system using translation rules and artificial neural network," Int. Arab J. Inf. Technol., vol. 16, no. 1, Jan. 2019.
43. B. Metin and H. Köktas, "Sentiment analysis with term weighting and word vectors," Int Arab J Inf Technol, vol. 16, no. 5, pp. 953–959, 2019.
44. M. Khader, A. Awajan, and G. Al-Naymat, "The impact of natural language preprocessing on big data sentiment analysis," Int Arab J Inf Technol, vol. 13, no. 3A, pp. 506–513, 2019.
45. A. Alnawas and N. Arici, "Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 18, no. 3, pp. 20:1–20:17, Jan. 2019, doi: 10.1145/3278605.
46. K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "Sentiment Analysis for a Resource Poor Language—Roman Urdu," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 19, no. 1, pp. 10:1–10:15, Aug. 2019, doi: 10.1145/3329709.
47. M. Poesio, J. Chamberlain, S. Paun, J. Yu, A. Uma, and U. Kruschwitz, "A Crowdsourced Corpus of Multiple Judgments and Disagreement on Anaphoric Interpretation," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, Jun. 2019, pp. 1778–1789.
48. J. Chamberlain, M. Poesio, and U. Kruschwitz, "Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference.," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016, pp. 2039–2046.
49. O. Uryupina et al., "Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus," Nat. Lang. Eng., vol. 26, no. 1, pp. 95–128, Jan. 2020, doi: 10.1017/S1351324919000056.
50. H. Zafari and M. Hourali, "From Parsed Corpora to Semantically Related Verbs," Comput. Inform., vol. 38, no. 1, pp. 240-264–264, Apr. 2019.
51. M. Abdullah Al-Ghaili, MS Thesis, "Identification of Discourse Boundaries in English Monologue," Department of Computer Science, University of Peshawar, KP Pakistan, 2015.