Sentence Boundary Detection in Colloquial Arabic Text: A Preliminary Result

Article · November 2011

DOI: 10.1109/IALP.2011.38

CITATIONS
4

READS
144

3 authors, including:

Afnan Alsubaihin
University College London
15 PUBLICATIONS **137** CITATIONS
SEE PROFILE

Hend Al-Khalifa
King Saud University
188 PUBLICATIONS **1,261** CITATIONS
SEE PROFILE

Sentence Boundary Detection in Colloquial Arabic Text: A Preliminary Result

Afnan A. Al-Subaihin, Hend S. Al-Khalifa and AbdulMalik S. Al-Salman
College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
{ aalsubaihin, Hendk, Salman }@ksu.edu.sa

Abstract— **Recently, natural language processing tasks are more frequently conducted over online content. This poses a special problem for applications over Arabic language. Online Arabic content is usually written in informal colloquial Arabic, which is characterized to be ill-structured and lacks specific linguistic standardization. In this paper, we investigate a preliminary step to conduct successful NLP processing which is the problem of sentence boundary detection. As informal Arabic lacks basic linguistic rules, we establish a list of commonly used punctuation marks after extensively studying a large amount of informal Arabic text. Moreover, we evaluated the correct usage of these punctuation marks as sentence delimiters; the result yielded a preliminary accuracy of 70%.**

Keywords-Natural Language Processing; Arabic Language; Sentence Boundary Detection; informal Arabic; colloquial Arabic

I. INTRODUCTION

Many natural language processing tasks have converged to regard text contained in the World Wide Web as their main input. This is a natural consequence as the Internet continues to pose itself as an enormous and ubiquitous media vessel. Applications that employ natural language processing techniques, such as machine translation, document classification, opinion and topic mining, have gained special importance as the demand shifts towards creating solutions to analyze and leverage online content.

The task of conducting Natural Language Processing (NLP) over Arabic online content faces a special challenge that is rarely investigated.

Arab societies are known to suffer from a sociolinguistic phenomenon called 'diglossia'. This phenomenon is defined as using linguistic standards in formal media that is different than in every day spoken language. In the Arab region, there are four major types of informal Arabic dialects: Levantine, Egyptian, North African and Gulf Arabic [1]. These dialects are generally known to be non-structured and difficult to standardize and analyze.

We have observed that online Arabic content that is likely to be useful to NLP tasks (e.g. opinion mining) is characterized to be written in informal Arabic language since users generally regard these websites as an informal means of communication. Although most websites use Formal Arabic for the main (static) parts of the website (website description, navigation and so on); however, the integral part of the website which is contributed by the website users, and have the most likelihood to contain significant content, is written using informal Arabic.

Despite this fact, research in the processing of Arabic languages used in informal dialect is known to be scarce.

In this paper, we shed the light on this insufficiently researched problem. Furthermore, we investigate the characteristics of online content written in informal colloquial Arabic to specifically tackle one preliminary problem: sentence boundary identification.

In the upcoming sections we shall introduce the problem of sentence boundary identification and its significance in NLP tasks. Afterwards, we will investigate the usage behavior of online Arabic users of punctuation marks and stop words to delimit the sentences. This is done by conducting a pilot experiment to validate the assumption that users generally provide textual indicators to finalize the ideas contained in sentences.

II. BACKGROUND AND RELATED WORK

While processing text written in any natural language, an important preprocessing step takes place. This step is concerned about isolating independent sentences since NLP regards them as the unit of work [2]. A sentence is the stream of words that conveys a coherent syntactic portion that can be analyzed [3]. Finding the correct sentence boundaries is not a trivial task. This is due to the ambiguity of punctuation marks, and in other cases, the misuse of these marks.

Research over English text that aims to identify sentence boundaries takes into account several features to detect sentences. A widely used method is a rule-based scheme that incorporates specific character patterns that indicate end of sentences. For example, *period-space-capital letter* is a pattern that occurs mostly at the end of a sentence and the beginning of another. An enhancement to this method is investigating the words prior and after the sentence delimiter to determine whether this delimiter actually indicates the end of the sentences. This is done by investigating the part-of-speech of the surrounding words [5], [6], [7].

The usage of machine learning technique has also been applied to solve the problem of automatic boundary detection. These techniques mostly focused on the words surrounding a punctuation mark to disambiguate its usage as a sentence delimiter, and implemented several machine learning algorithms such as regression trees [8], feed forward neural networks [3], [9] and maximum entropy approach [10], [3].

We notice that all of these approaches investigate the problem of disambiguating the usage of punctuation marks in English and other roman character languages.

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.38

This is due to the fact that punctuation marks, more often than not, are used to delimit sentences.

This is also the case with standard formal Arabic text. Published research that investigate sentence boundary detection in Arabic, only investigated its Formal variation. In [11], the problem is solved by segmenting Arabic sentences using known Arabic sentence separators (. ; : ? ،).

However, the usage of punctuation marks in informal written Arabic must be further investigated to establish the research of sentence boundary detection for informal Arabic text.

There are several evaluation schemes to measure the accuracy of sentence boundary detection algorithms. A very simple and successful algorithm, which will be used in our paper, is by benchmarking the accuracy of the algorithm with a certain lower bound percentage. The lower bound is found by marking every potential sentence boundary maker as the end of the previous sentence. Then, detect the accuracy of this assumption by calculating the percentage of sentences that are correctly delimited by the boundary maker [3].

In our experiment, we aim to establish a lower bound of accuracy when using common stop words as sentence delimiter. Thus, we will establish a baseline for evaluating subsequent efforts in detecting sentence boundaries in informal Arabic.

III. PILOT STUDY

Due to the highly informal and ill-standardized nature of colloquial Arabic, a first step is establishing a list of commonly used punctuation marks and white spaces. Upon observing a large amount of informal Arabic text in many websites including Arabic forums, Facebook, and specific evaluative websites such as Qaym.com, we detect the common usage of typical Arabic punctuation marks (Arabic Comma, Period, Exclamation Point and Question Mark). Additionally, we have observed that many users also use English Commas to finalize the sentence; another unexpected sentence delimiter is the usage of *newlines* to finalize a sentence and begin another. In figure 1, we show two user input from two different websites (Facebook and Qaym), we highlight the usage of sentence delimiters (including newlines, periods, Arabic commas and question marks.)

To ascertain the frequent usage of these delimiters, we have investigated their usage frequency over a set of 6364 restaurant reviews that have been extracted from an Arabic Website called Qaym.com. Among these reviews, 34535 sentences have been identified and extracted using the aforementioned delimiters. Table I depicts the extracted sentence delimiters and their usage frequencies.

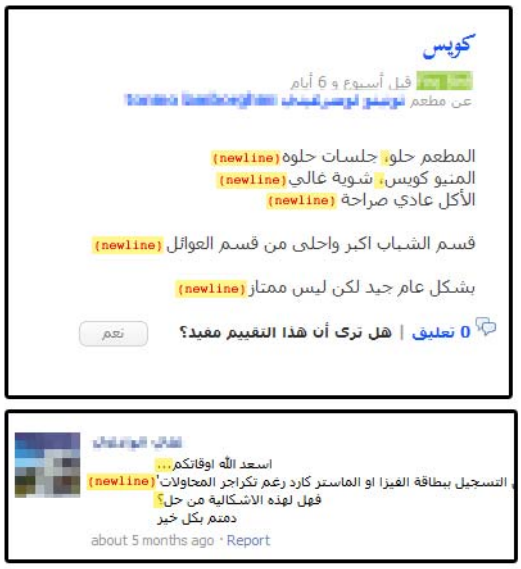

Figure 1. A sample of extracted content of informal Arabic. Used sentence delimiters is highlighted.

TABLE I. EXTRACTED SENTENCE DELIMITERS AND THEIR USAGE FREQUENCY

Delimiter Name	Delimiter Symbol	Usage Frequency	% of usage
Newline	\n or 	18869	54.64
Arabic Comma	،	3125	9.05
Period	.	9524	27.58
Exclamation Point	!	1535	4.44
Question Mark	؟	336	0.97
English Comma	,	1146	3.32

The table shows the frequent usage of newlines that is embedded with a single comment. Although using newlines heavily depends on the website style and users mode of communication in this website, we have observed that websites that provide a free and formatted comment space have the highest likelihood to use newlines as sentence delimiters (e.g. Forums as opposed to Facebook). Another dominant means of delimiting sentences is periods which are more mainstream and commonly identified as a sentence termination symbol. Arabic commas, whenever used, are either a mean of sentence termination or separating list items. This creates certain ambiguity, which will be investigated in the next section.

IV. PRELIMNARY EVALUATION

An issue here is whether this segmentation scheme is reliable, and whether Arabic online users actually use these delimiters in their informal writings to delimit sentences according to the true definition of a sentence. In order to ascertain this fact, we have annotated 300 reviews using the expertise of Arabic linguist to correctly identify sentences. Secondly, we benchmarked

the annotated sentences with our segmentation scheme. This resulted in an accuracy that reached 70%. The results are fully depicted in Table II.

TABLE II. EXTRACTED SENTENCE DELIMITERS AND THEIR USAGE FREQUENCY

# of Reviews	# of sentences	True Positives	False Positives	False Negatives	Accuracy
300	2048	1554	165	494	0.702

In annotating 300 restaurant reviews extracted from Qaym.com, we identified 2048 sentences. Among these sentences, 1719 sentences were identified using the previously mentioned sentence delimiters. However, 9.6% have been falsely classified as independent sentences which show the inclusion of the pre-identified sentence delimiters for usages other than indicating the ending of a sentence.

Additionally, there are 494 instances of false negatives, meaning that the scheme failed to identify 24% of the sentences. This is an indication that writers in informal Arabic may fail to include any punctuation or stop words in indication of ending a sentence which was observed in 24% of the cases.

V. CONCLUSION AND FUTURE WORK

In this paper, we have pointed out the importance of investigating informal colloquial Arabic characteristics to enhance the task of NLP over this text. This is particularly important as NLP applications regard online content as its main input, which in the case of Arabic, is mostly populated with its informal and less structured counterpart. As a first endeavor in this direction, we have investigated the problem on sentence boundary detection. This problem seemed of specific importance since informal Arabic doesn't abide by any punctuation rules to delimit sentences. In order to overcome this problem, we have observed punctuation marks usage behavior of Arabic online users and established a list of commonly used sentence delimiters. To ascertain the accuracy of these sentence delimiters and whether they were used correctly to constitute a sentence, we have compared it to a corpus of text that has been annotated by a human expert.

Upon establishing a list of possible punctuation marks that are frequently used in informal colloquial Arabic text, in addition to, establishing a lower bound for comparison, other efforts in the area of automatic boundary detection of informal Arabic can be conducted.

ACKNOWLEDGMENT

We would like to thank Mr. AbdulAziz Al-Subaihin for his many consultations in his area of expertise as an Arabic linguist; and for taking the time to annotate a large corpus of informal Arabic reviews.

REFERENCES

[1] K. Kirchhoff et al., "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop," presented at the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing., 2003, vol. 1, p. 344 - I-347.

[2] G. Grefenstette and P. Tapanainen, "What is a word, what is a sentence? problems of tokenization." 1994.

[3] D. J. Walker, D. E. Clements, M. Darwin, and J. W. Amtrup, "Sentence boundary detection: A comparison of paradigms for improving MT quality," *In Proceedings of MT Summit Viii: Santiago De Compostela*, p. 18--22, 2001.

[4] D. D. Palmer and M. A. Hearst, "Adaptive multilingual sentence boundary disambiguation," Computational Linguistics, vol. 23, no. 2, pp. 241–267, Jun. 1997.

[5] J. Aberdeen, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain, "MITRE: description of the Alembic system used for MUC-6," Proceedings of the 6th conference on Message understanding, pp. 141–155, 1995.

[6] A. Mikheev, "Tagging sentence boundaries," Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pp. 264–271, 2000.

[7] J. Shim, D. Kim, J. Cha, G. Geunbae Lee, and J. Seo, "Integrated multi-strategic Web document pre-processing for sentence and word boundary detection," *Information Processing & Management*, vol. 38, no. 4, pp. 509-527, Jul. 2002.

[8] M. D. Riley, "Some applications of tree-based modelling to speech and language indexing," in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1989, pp. 339-352.

[9] T. L. Humphrey and F. Q. Zhou, "Period disambiguation using a neural network," in *Proceedings of International Joint Conference on Neural Networks.*, Washington, DC , USA, 1989, vol. 2, p. 606.

[10] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," *Proceedings of the fifth conference on Applied natural language processing*, pp. 16–19, 1997.

[11] R. Ouersighni. "A major offshoot of the DIINAR-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts," *proceeding of Arabic NLP Workshop at ACL/EACL*, 2001.