ArabicSeg: An Arabic News Story Segmentation System

3 authors, including:

Michael A. El-Shayeb
Cairo University
3 PUBLICATIONS **15** CITATIONS

SEE PROFILE

Samhaa R. El-Beltagy
Nile University
95 PUBLICATIONS **551** CITATIONS

SEE PROFILE

ArabicSeg: An Arabic News Story Segmentation System

Michael A. El-Shayeb & Samhaa R. El-Beltagy
Computer Science Department,
Faculty of Computers and Information,
Cairo University, Giza, Egypt
mikeazmy@yahoo.com, samhaa@computer.org

Ahmed Rafea
Computer Science Department,
American University in Cairo
Cairo, Egypt
rafea@aucegypt.edu

Abstract

Text segmentation is a very critical step to many applications and while it has been addressed extensively for the English language, work on text segmentation for other languages is still lagging behind. In this paper the ArabicSeg system for segmenting Arabic news stories is presented. The developed system is based on a linguistic technique called lexical chaining which measures the cohesion between textual units. In conjunction with this technique, a set of error reduction filters have been introduced and were found to significantly reduce segmentation errors in the detection of borders in Arabic based news stories. The results of evaluation experiments carried out on an Arabic Reuters news story dataset are presented. An analysis of the effect of introducing each of the proposed error reduction filters is also provided. The evaluation shows that the segmentation results produced by the presented system are not only comparable to the results reported by English based segmentation algorithms, but also outperform them.

1. Introduction

Text in long documents or that obtained from continuous text streams, needs to be separated into topically coherent units in order to enable effective querying, analysis, and usage. In information retrieval for example, having topically segmented documents can result in the retrieval of short relevant text segments that directly correspond to a user's query instead of long documents which the user has to examine carefully in order to find the object of his/her interest. Having topically segmented documents also benefits the task of text summarization as a better summary can be obtained from the various segments constituting a document [15]. Recently, unsegmented or continuous news streams have also become an important area where text segmentation can be applied since the success of tasks such as topic tracking depends heavily on the accuracy of the detection of distinct news stories.

While extensive research has targeted the problem of determining boundaries in English news streams, few have studied the problem in other languages and almost no one has addressed it for the Arabic language. In this paper a system for segmenting Arabic news stories is presented. The algorithm adopted by this system is based on the lexical chaining technique [5] which is a linguistic technique that measures the cohesion in text by using the semantic relationship between terms in the text. The technique builds chains of related terms which span the entire text to relate its sentences.

The presented system was evaluated using an Arabic Reuters newswire dataset which is composed of one thousand unbounded news stories. The evaluation experiments demonstrate that adding some error reduction features to the basic lexical chaining approach is very effective in improving performance. The performance of the system is also compared to that of three well known English segmentation systems (one of which is also based on lexical chaining) and the results show that the performance of the developed system is comparable to state-of-the-art.

This paper is organized as follows: section 2 presents related work, section 3 presents an overview of the proposed Arabic news story segmentation system; experimental results are reported in section 4 and finally section 5 concludes the paper and presents future research directions.

2. Related Work

Existing approaches to text segmentation fall into two main groups: lexical cohesion based approaches and feature based approaches. Lexical cohesion based approaches depend on the tendency of topic units to hang together. Approaches to measure this type of cohesion, can be divided into two categories: similarity based approaches where patterns of syntactic repetitions are

used to indicate cohesion, and lexical chaining based approaches where other aspects of lexical cohesion (like relationships between terms) are also analyzed.

An example of the latter approach is the TextTiling system [12] which uses the cosine similarity metric between term vectors to measure cohesion strength between adjacent blocks. Another example is that of the C99 algorithm which also uses the cosine similarity metric to determine similarities among sentences and then projects these graphically and then applies image-processing techniques to determine topic boundaries. Another system that utilizes a similarity based approach is presented in [16]. This system uses the probabilistic latent semantic analysis (PLSA) model along with the clarity-based similarity metric to detect boundaries. The work measures the similarity using the probability distribution of words that are calculated using the PLSA model instead of using term counts.

The application of Lexical chaining based approaches to text segmentation was first attempted in [14] [9] and, [7]. In these works, segmenting a single document to its sub topics was the major goal. Recently lexical chaining has also been used in news story segmentation [13] [11]. Work in [13] uses lexical chaining technique for determining distinct news stories in spoken and written broadcast news streams. The work analyzes the cohesion in text by examining term repetitions and three other basic types of cohesion (synonymy, generalization/ specialization and part-whole/ whole-part relationships) provided by the WordNet online thesaurus [4]. In [11] the lexical chaining based approach is used in conjunction with the similarity based approach. In this work, lexical cohesion between two adjacent blocks is determined by computing the cosine similarity between the two blocks through analyzing the lexical chains that overlap with the two blocks instead of using word counts. Evaluation of this work was based on the topic detection and tracking (TDT) corpora.

The second main category in text segmentation is that of feature based approaches in where features like cue phrases, full proper nouns and named entities are used to detect boundaries between topics. An example of a system that uses that approach is presented in [2]. Feature based approaches can be domain dependent (as in news transcripts) if they depend on very specific domain features. Lexical cohesion can also be added as a feature in the feature based approach as exemplified by work presented in [1] and [6].

3. The Proposed System

In this section the proposed Arabic news story segmenter is described the system takes in a long text document or a stream of text and outputs segments consisting of individual news stories. For this system to work, an indicator of the discourse structure of the underlying text must exist. In the proposed system lexical chaining is used for representing this type of discourse. Essentially, this is carried out by building a set of lexical chains that capture the cohesive structure between sentences of the underling stream of text indicated by the use of semantically related vocabulary.

Overall, four phases are employed for carrying out the segmentation task:

1. Preprocessing.
2. Lexical chains building.
3. Segment boundary detection.
4. Error reduction filtering.

In the first phase, all potentially chainable terms are collected and stored along with their location in the text (sentence numbers). In the second phase, a lexical chain is built for each repeated term beginning from the first occurrence of the term and ending when a gap exists between one occurrence of the term and the following occurrence of that term. In the third, a score is assigned to each boundary between two sentences and a segment boundary is determined based on a score threshold. In the fourth, error filters are applied to reduce errors and hence enhance the results. Each of these phases is described in details in the following four subsections.

3.1. The Preprocessing Phase

In this work, sentences are considered to be the basic textual units on which work can be carried out. So an important step in the preprocessing phase is to identify sentence boundaries through the use of ".", "?" and "!". Exceptions to this rule include abbreviations. Sentence boundary identification is thus the initial step in the preprocessing phase.

For the lexical chaining technique to work, nouns and adjectives need to be extracted from the input sentences.. So in the second step of the preprocessing phase, part-of-speech (POS) tagging is carried out in order to tag required terms and to also determine conjunctions as they are needed in the error reduction phase. Determining full proper nouns is also needed in the error reduction phase, so the third step in the pre-processing phase is the application of a chunk parser. For carrying out steps two and three, the Arabic POS tagger and chunk parsing tool presented in [10] were used. After these initial three steps are carried out, stop words are removed and all remaining words are stemmed. Finally n-grams of nouns and adjectives are formed and saved with their locations in the text (represented as the sentence numbers in which they appear) for usage in the next phase of lexical chains building.

3.2. The Lexical Chain Building Phase

The main goal of the lexical chaining technique is to connect textual units that are related to each other through the detection of the existence of semantic relationships between vocabularies appearing within these textual units. Specifically, a lexical chain is defined as a cluster of semantically related terms. Each chain is tagged with the locations of the sentences where it begins and ends. Work on text segmentation using lexical chaining has almost always used an external resource such as Wordnet, to capture relationships between terms in the text. However, recent work has shown that the best performance achieved for segmenting text with a lexical chainer was obtained when only patterns of repetitions were used [11] [13]. As no general purpose Arabic thesaurus such as Wordnet is currently available, and based on these recent findings, only repletion patterns (of both single terms and n-grams) across sentences are used in the chaining process thus ignoring synonymy and other semantic relationships. Starting from the first sentence in the input text, whenever a term (singular or compound) is found to repeat in adjacent sentences, a chain for that term is built. The chain start tag is always assigned to the sentence location where the term first occurs. The chain is said to continue until a gap that exceeds an allowable gap length exists between the last occurrence of the term in the chain and its next occurrence in the text, or if the last term in the chain is also the last term in the document. The best allowable gap length was determined experimentally as 11 sentences [11]. In case of an occurrence of a gap, a new chain is created for the same term when it appears at a later point in the document. The reason behind cutting the chain when a gap occurs between two occurrences of the same term is because the probability for these two occurrences belonging to the same topical entity decreases as the distance between the two occurrences increases. So, termination of a chain in such a case avoids the creation of weak chains. The algorithm for chain building is given in Figure 1.

3.3. The Segment Boundary Detection Phase

The aim of this phase is to determine the boundary points between the distinct news stories in the text. The input to this phase is the lexical chains that were constructed from the previous phase. Each chain connects a group of sentences that share a term indicating that they belong together. The chain-begin point indicates that there is no relation between the chain sentences and the previous sentences in the text.

```
FOR EACH Repeated_Term_List IN Text
  Create_New_Chain (firstTermOccurence);
  //for first occurrence of the repeated term
   Tag_Chain_Begin(firstTermOccurence.Location)
  prevT = firstTermOccurence;
  FOR EACH curT IN Repeated_Term_List
 //Sorted based on its location in the text
    IF (curT.Location - prevT.Location <
    avgStorySentLen)
        Add_Term_To_Curr_Chain (curT);
    ELSE
        Tag_Chain_End (prevT.Location);
        Create_New_Chain (curT);
        Tag_Chain_Begin (curT.Location);
```

Figure 1. Pseudo-code for the chain building algorithm

The chain-end point indicates that there is no relation between the chain sentences and the following sentences in the text. So the existence of a high number of chains that end at a sentence n and a high number of chains that begin at the following sentence n+1 is an indicator of a possible boundary between the text segment ending with sentence n and the text segment beginning at sentence n+1. Based on this hypothesis, the boundary strength between each two adjacent sentences in the text is calculated. In this work we follow the same approach adopted for boundary calculation in [13]. Basically, the boundary strength between two adjacent sentences is the sum of the number of chains that end at sentence n and the number of chains that begin at sentence n+1. After calculating boundary strengths between each two adjacent sentences in the entire text, the mean value of these boundary strengths is then used as a threshold for determining the boundaries between the distinct news stories.

3.4. The Error Reduction Filtering Phase

The aim of this phase is to "correct" the news story boundaries that were determined from the previous phase by eliminating noisy boundaries and thus increasing boundary precision. In the context of this work, three cases define noisy boundaries.

In the first case, an identified boundary point only separates a small number of sentences the individual collection of which is unlikely to form a distinct news story. To clean this type of noisy boundary, a distance filter is utilized. This filter removes boundaries that are close to a boundary with a higher boundary strength as has been done in [13]. By experimenting in this work, a distance boundary of 80 words was found to yield the best results.

In the second case, a boundary point is followed by a conjunctive sentence, i.e. a sentence that begins with a conjunction. In Arabic news stories, examples of conjunctions include "و", "كذلك", "كما", "فيما", "من جانبه", "من جهة أخري", "الي ذلك" [1], etc. A conjunctive filter is used

[1] Translated to English these would be: and, also, in addition, with, for his part, on the other hand, & so.

to eliminate these boundaries as they separate a segment that begins with a sentence related to the previous segment with a conjunction. As will be shown, this type of filtering has been found to result in significant error reduction in Arabic news stories. Two complementary methods are proposed in this work to determine conjunctions. The first method makes use of terms that have been tagged as conjunctions through POS tagging performed in the preprocessing phase. The second, uses a list of conjunctions that seem to occur with a high frequency in the beginning of the inner sentences of each news story. The second method is used as a fail safe when POS tagging can not tag a conjunction appropriately or when the conjunction takes the shape of a noun phrase like, "الموقف الذي يقضي"[2].

In the third and last case, a boundary point separates a sentence that has a full proper noun from a sentence containing a short form of this full proper noun. For example, Ludwig van Beethoven is a full proper noun, while Beethoven, is the short form for this full proper noun. Similarly، "الأمين العام للأمم المتحدة كوفي أنان"[3], is a full proper noun which usually only appears in its full form the first time it is referenced, and is then referred to using short proper noun instances such as "أنان"[4]. To detect this kind of occurrence, it is assumed that first appearance of a proper noun is a full appearance of this proper noun. A rule is then used to connect the full source proper noun with its referrers. This rule specifies that if after the appearance of a full proper noun, a term is found which is a substring of the full proper noun after removing the first word in that full proper noun, then it is a reference or a short form of the original full proper noun. A proper noun fi1lter that applies this rule was implemented. Full proper nouns were identified in this work using the chunk parsing tool [10].

4. Evaluation

In this section the results of evaluating the ArabicSeg system using an Arabic Reuters news story dataset, are presented. The obtained results are compared to the results of three state-of-the art segmentation systems one of which is also based on lexical cohesion. The systems with which the evaluation results are compared against are SeLeCT [13], C99 [3], and TextTiling [12]. These systems were evaluated using an English Reuters dataset. In addition, an analysis of the effect of introducing each of the proposed error reduction filters is also provided.

4.1 Evaluation Metrics

In the evaluation experiments, four segmentation evaluation metrics were used. The first two of these are the standard precision and recall metrics. In the context of text segmentation, precision is defined as the number of correctly system detected boundaries divided by the total number of system generated boundaries, while recall is defined as the number of correctly system detected boundaries divided by the total number of actual boundaries in the used dataset. The precision and recall metrics have been criticized by a number of researchers [1] [13] for their failure to take into account near-boundary misses. As a result, Beeferman [1] proposed a new metric Pk, which is a probabilistic evaluation metric that tries to address the inadequacies of precision and recall. However Pevzner and Hearst [8] recently criticized this metric for being biased, because it penalizes false negatives more than false positives and over penalizes near-misses. Pevzner and Hearst [8] also suggested a new metric called WindowDiff which handles the criticisms of all previous metrics. So our third used metric is the WindowDiff metric which uses a sliding window over the text and measures the difference between the number of hypothesized boundaries and the actual boundaries within the window. An additional metric called RSeg is proposed by this work. The RSeg metric is used to calculate the exact number of correctly extracted news stories (not boundaries) which reflects the accuracy of the text segmentation algorithm. It should be noted however, that the RSeg metric is a very strict measure of accuracy so it should never be used on its own, but rather used in conjunction with other metrics like those presented here.

4.2. The News Story Segmentation Test Corpora

The ArabicSeg segmentation system was evaluated using the Arabic Reuters corpus in order to compare it to previously published systems that used the English Reuters corpus. Of course an exact comparison is by definition not possible, because the Arabic Reuters dataset and the English Reuters dataset are essentially different, but this is as close as possible a match between two datasets that are represented in different languages. To construct the test dataset, a collection of 1000 randomly selected news stories were obtained form a huge Arabic Reuters dataset and concatenated into one big file (this is the same size as the English test dataset). The advantage of using a single file of concatenated news stories as a dataset for topic detection evaluation is that it does not require human intervention for judging topic shifts as a segment in this context refers to a distinct news story. Our Arabic Reuters dataset is available online for future

[2] The translation of this is: 'the position whereby'
[3] Translation: Secretary-General of the United Nations Kofi Annan
[4] Translation: Annan

comparison by any one who develops an Arabic text segmentation system [17].

4.3. Segmentation results

As mentioned before, the implemented ArabicSeg was evaluated against three English segmentation systems: SeLeCT, C99, and TextTiling. The results of evaluating these three systems using the English Reuters dataset were reported in [13]. Table 1 shows a comparison between those results and those of ArabicSeg using the precision, recall and WindowDiff metric.

Table 1. Results for segmentation on concatenated Reuters news stories (English and Arabic)

System	Recall	Precision	WindowDiff
ArabicSeg	88.5%	83.8%	0.127
C99	70.0%	74.9%	0.121
SeleCT	60.6%	79.1%	0.209
TextTiling	32.1%	41.0%	0.247

From this table, it can be observed that ArabicSeg outperforms both the SeLeCT and TextTiling systems. However, C99 outperforms ArabicSeg with respect to the WindowDiff metric, but only by a 0.006 margin. Optimal performance for the ArabicSeg system was achieved by using n-gram repetitions for building the lexical chains and employing conjunction and proper nouns error reduction filters.

To examine the effect of the different error reduction filters on the performance of our segmentation system, the results of the system before and after using every filter are presented in tables 2, 3, and 4. Table 2 shows that the use of the distance error filter reduces the WindowDiff error value and increases the number of correctly extracted distinct news stories compared with the results of the system when using n-gram repetitions only. The table also shows that recall decreases after using the distance filter. This is because some correct boundaries can be falsely removed as the size of the news stories in the used data set vary from very short to very long.

Table 2. Improvements in the system performance as a result of using a distance error filter.

	Recall	Precision	WindowDiff	RSeg
Before	90.7%	23.2%	0.988	14.7%
After	69.0%	40.6%	0.497	21.1%

Table 3 shows a significant enhancement in the segmentation results when the conjunction error filter is used. By introducing this filter, the error (indicated by WindowDiff) decreases from 0.988 to 0.134 and the number of correctly extracted news stories increases from 14.7% to 69.4%. The advantage of this filter is that it rarely removes correctly boundaries compared with the distance error filter.

Table 3. Improvements in the system performance as a result of using the conjunction error filter

	Recall	Precision	WindowDiff	RSeg
Before	90.7%	23.2%	0.988	14.7%
After	89.3%	82.2%	0.134	69.4%

Table 4 shows that proper noun filtering achieves a slight improvement in approximately all the metrics except of the recall value which is slightly reduced. It was also observed that the used POS tagger and chunk parser do not always tag proper nouns appropriately which may have led to this very slight improvement when using this filter.

Table 4. Improvements in the system performance as a result of using the proper noun error filter

	Recall	Precision	WindowDiff	RSeg
Before	90.7%	23.2%	0.988	14.7%
After	89.9%	25.7%	0.809	17.5%

When trying to combine the various error reduction filters to capture the highest performance, we found that the best choice is to use the conjunction filter along with the proper noun filter as shown in table 5.

Table 5. Arabic Reuters segmentation results. (D= distance filter, C= conjunction filter, P= proper noun filter)

	Recall	Precision	WindowDiff	RSeg
D+C+P	76.8%	90.5%	0.137	55.8%
D+C	76.8%	89.5%	0.139	55.5%
D+P	69.7%	43.1%	0.461	23.7%
C+P	88.5%	83.8%	0.127	69.4%
C	89.3%	82.2%	0.134	69.4%
D	69.0%	40.6%	0.497	21.1%
P	89.9%	25.7%	0.809	17.5%

5. Conclusion and Future Work

In this paper we have presented a lexical chaining based approach to segment news stories. An interesting result reported in this paper was the significant improvement in the results when using conjunctions for error reduction filtering. Although the SeLeCT system and the presented ArabicSeg system both use the lexical chaining based approach, conjunctions have made a huge improvement on the performance of the ArabicSeg system compared with the decrease in performance in the SeLeCT system. Another conclusion that can be made by the presented work, is that with very little modification to the English based text segmentation based on lexical

chaining and presented in [13], results comparable to those obtained in English based systems, are possible. In future we plan to investigate the effect of introducing a different part-of-speech tagger and chunk parser. We also plan to enhance the boundary scoring method by experimenting with combining the lexical chaining approach with other approaches.

Acknowledgments

Work presented in this paper has been supported by the Center of Excellence for Data Mining and Computer modeling within the Egyptian Ministry of Communication and Information (MCIT).

References

[1] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, pp. 177 - 210, 1999.

[2] D. Kauchak and F. Chen, "Feature-based segmentation of narrative documents," presented at the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing, Ann Arbor, MI, USA, 2005.

[3] F. Choi, "Advances in domain independent linear text segmentation," presented at the first conference on North American chapter of the Association for Computational Linguistics (NAACL), Seattle, Washington, 2000.

[4] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Five papers on wordnet," Cognitive Science Laboratory, Technical report 1990.

[5] J. Morris and G. Hirst, "Lexical cohesion by thesaural relations as an indicator of the structure of text," *Computational Linguistics*, vol. 17, pp. 21- 48, 1991.

[6] J. Reynar, "Topic Segmentation: Algorithms and Application," in *Computer and Information Science*. Pennsylvania: UPenn, 1998.

[7] Kan, Min-Yen, J. L. Klavans, and K. R. McKeown, "Linear segmentation and segment relevance," presented at the International Workshop of Very Large Corpora (WVLC 6), Montreal, 1999.

[8] L. Pevzner and M. A. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Computational Linguistics*, vol. 28, pp. 19 -36, 2002.

[9] M. A. Stairmand, "A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval," in *Department of Language Engineering*. Manchester, England: UMIST, 1996.

[10] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks," presented at HLT/NAACL, Boston, 2004.

[11] M. Galley, K. McKeown, E. Fosler-lussier, and H. Jing, "Discourse segmentation of multi-party conversation," presented at 41st Annual Meeting of ACL, Sapporo, Japan, 2003.

[12] M. A. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, pp. 33-64, 1997.

[13] N. Stokes, J. Carthy, and A. F. Smeaton, "SeLeCT: a lexical cohesion based news story segmentation system," *AI Communications*, vol. 17, pp. 3 - 12, 2004.

[14] O. Manabu and H. Takeo, "Word sense disambiguation and text segmentation based on lexical cohesion," presented at The International Conference on Computational Linguistics, Kyoto, Japan, 1994.

[15] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," presented at Intelligent Scalable Text Summarization Workshop, Madrid, Spain, 1997.

[16] T. Brants, F. Chen, and I. Tsochantaridis, "Topic-based document segmentation with probabilistic latent semantic analysis," presented at CIKM, McLean, Virginia, USA, 2002.

[17] http://www.claes.sci.eg/coe_wm/ar_seg/segdataset.zip