Agglomerative Sentence Clustering Approach for Discourse Segmentation

Christopher Augustine
M.Tech (Computational Linguistics),
Govt: Engineering College, Palakkad,
Kerala, India, 678633

P. C. Reghuraj
Professor, Dept. of Computer Science and Engineering,
Govt: Engineering College, Palakkad,
Kerala, India, 678633.

Abstract

Automatic recognition of discourse which describes about a situation or set of entities is important natural language tasks like summarization, information retrieval, etc. A text can be viewed as a collection of discourses that describe a set of nouns. This paper presents an agglomerative sentence clustering approach for text segmentation based on nouns. This method considers both cohesion and coherence relationships among sentences. The nouns are the best representatives of the sentence in discourse. We find that clustering of the sentence by considering these nouns gives better segmentation strategy. The output of the clustering process is further refined with named entities and WordNet for better accuracy.

1. Introduction

Discourse analysis deals with a group of sentences that are cohesive or coherent. The coherent processing of this text segments requires something more than interpretation of the individual sentences. While syntax and semantics work with sentence-length units, the discourse level of NLP works with units of text longer than a sentence. That is, discourse focuses on the properties of the text as a whole that convey meaning by making connections between component sentences [10]. The important parameters necessary for the text segmentation are: *cohesion, coherence* and *lexical chain. Cohesion* and *coherence* are terms used in discourse analysis and text linguistics to describe the properties of written texts. *Cohesion* is used to specify linking of sentences and paragraphs and their relationships. In English, *cohesion* is established through the use conjunctions, pronouns, conjuncts, etc. A cohesive text may not be coherent. Lexical and grammatical relationships among sentences are *cohesion*, whereas *coherence* is based on semantic relationships. Often, lexical cohesion occurs not simply between pairs of words but over a succession of a number of nearby words spanning a topical unit of the text. These sequences of related words are called *lexical chains*. There is a distant relation between each word in the chain, and the words co-occur in the given span. *Lexical chains* do not stop at sentence boundaries. They can connect a pair of adjacent words or can range over an entire text.

More subtle distinctions are sometimes made. One can distinguish between discourse-as-product (the linguistic construct) and discourse-as-process (the communicative event). *Coherence* can be reserved for the conceptual relationships that comprehenders use to construct a coherent mental representation accommodated by what is said in the discourse. *Cohesion* is limited to the linguistic markers that cue the comprehender on how to build such coherent representations. *Cohesion* emphasizes discourse-as-product, and coherence emphasizes discourse-as-process [5].

There are a number of theories of natural-language discourse processing, and computational models of these theories exist. Theories concentrate on the themes of semantics, structure, and intention. A common principle of all approaches is that they provide a model of the coherence of discourse. Semantic theories argue that the coherence of a discourse is a feature of its meaning and if the meaning modelled, the coherence falls out of that [9]. For practical purpose, it is currently difficult to develop a comprehensive discourse analyser with full text understanding capability. On the one hand, a model to understand text is still being developed [7].

In the following section some of the related works are briefly explained, Section 3 describes the proposed method of this paper and in section 4, results of this method are discussed and limitations of this method and suggestion on future work are given in section 5.

2. Related Work

Marti A. Hearst proposed a famous method, TextTiling, for partitioning full-length text documents into coherent multi-paragraph units. TextTiling illustrates, a computational approach to segment written expository text into contiguous, non-overlapping discourse units that correspond to the pattern of subtopics in a text [6].

The model described in [3] fall into two broad classes that capture orthogonal dimensions of entity distribution in the discourse. The first class is the syntactic aspects of the text coherence and characterizes how mentions of the same entity in different syntactic positions are spread across adjacent sentences. The key assumption is that certain entity transitions are likely to appear in locally coherent discourse. The second is the semantic class that quantifies local coherence as the degree of connectivity across the sentences. A number of linguistic devices-entities such as repetition, synonymy, hyponymy and meronymy are considered.

The coherence of a text based on statistical distribution of the discourse structure and relation is described in [4]. This model specifically focuses on the discourse relation transitions between adjacent sentences, modelling them in a discourse role matrix. When a term appears in a discourse relation, the discourse role of this term is defined as the discourse relation type and the arguments span in which the term is located.

John Morris and Graeme Hirst describe a model based on lexical cohesion [8]. Lexical cohesion arises from semantic relationship between words. Here, the main requirement is that there be some recognizable relation between words. The sequence of words is called lexical chains. Lexical chains do not stop at sentence boundaries. It provides a clue for the determination of coherence and discourse structure, and hence larger the meaning of the sentence.

Text collections are increasingly heterogeneous. An important aspect of heterogeneity is length. On the World Wide Web, document sizes range from home pages with just one sentence to server logs of half a megabyte.

In order to capture discourses with various sizes, there should be a system which works at sentence level rather than fixed block size. A system which considers both cohesion and coherence can give good results. Next section descries the proposed method which considers both cohesive and coherent relation among sentences.

3. Proposed Method

Nouns refer to people, things, concepts, and other objects. They are the original and central building blocks of language [2]. The analysis of text based on the nouns gives a light to the discourse segmentation. There are lot of nouns from the first sentence to last one in the text. Nouns have important characteristics, since other particles are used to complement them. These nouns can be used to find the cohesive relationship between the sentences. Based on this relationship, the sentences in the text can be grouped.

A discourse is meant to describe a particular topic. So, there is a set of nouns and named entities specific to a discourse. This paper proposes a method includes three modes of operations which can segment a text based on cohesion of nouns and named entities and coherence as shown in Fig. 1. Three operations are: (i) Sentence clustering based on nouns, (ii) Boundary adjustment with named entities (NEs) and (iii) Boundary adjustment with WordNet.

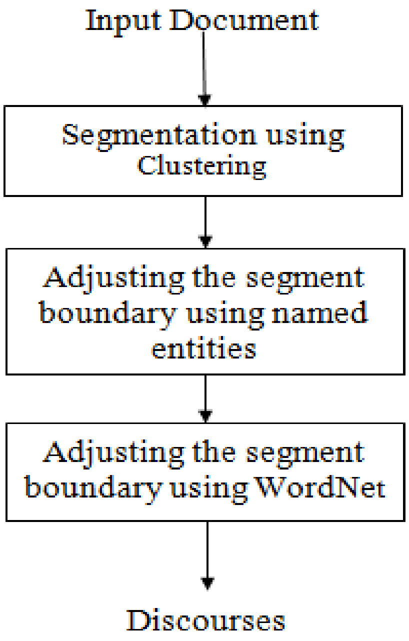
Input Document
Segmentation using Clustering
Adjusting the segment boundary using named entities
Adjusting the segment boundary using WordNet
Discourses

Figure. 1. Block Diagram of the Proposed Method

a. Sentence clustering based on nouns

This clustering ensures that each cluster of sentences contains sentences about a particular set of nouns. The proposed algorithm has three main parts:

1) Splitting of text into sentences.
2) Clustering of sentences.
3) Combining adjacent clusters.

Clustering algorithm for sentence clustering based on nouns is given in Algorithm-1.

```
Algorithm-1: Clustering Algorithm
Input: Input Text
Output: discourse_CL

Read the input text as raw_text
Split the raw_text into sentences S
Find POS tags for the S as pos_tag
S_group= groups of three consecutive sentences in S
P_group= POS tags for each S_group
old_S_group=[]
while old_S_group $\neq$ S_group
    for each group g in S_group
        noun1_bag=nouns in g
        noun2_bag=nouns in next(g)
        if noun1_bag $\cap$ noun2_bag $\neq$ null
            combine g and next(g) in S_group
            combine p and next(p) in P_group
        for each group g in S_group
            s=boundary sentence of g
            if s is cohesive to adjacent(g)
                 remove s from g and   add to adjacent(g)
        old_S_group=S_group
```

```
discourse_CL=S_group
return discourse_CL
```

After the above operations, the clusters are again checked to see whether any clusters with only one or two sentences exist. If there are such clusters, attach them to the appropriate adjacent clusters. In practice, setting initial cluster size to 3 sentences works best for many texts. The clustering algorithm ensures that a cluster of sentence will not contain any sentence that is not cohesive to the cluster. The clustering algorithm itself gives better results compared to other approaches. The segment boundary of the output of clustering algorithm can be corrected with the help of two approaches specified with next section.

b. Boundary Adjustment with NEs

The output of the clustering algorithm may contain lot of clusters or sometimes the boundary may not accurate one. This output is to be further refined for better accuracy. One of the effective methods is to re-arrange the boundary of previous groups and/or merge the clusters based on the presents of named entities. The clustering algorithm result is a basic one and it helps to identify the named entities which can come together. The boundary adjustment with named entity is explained in Algorithm-2.

```
Algorithm-2: Boundary-NE Algorithm
Input: discourse_CL
Output: discourse_NE

Read discourse_CL
discourse_NE=discourse_CL
tag discourse_NE
old discourse_NE=[]
while old_discourse_NE ≠discourse_NE
    for each d in discourse_NE
        disc_bound=boundary sentences of d
        if disc_bound contains Named Entity in next(d)
            move sentences to next(d) at the break point
    old_discourse_NE=discourse_NE
return discourse_NE
```

Boundary sentences are the candidate sentences for boundary correction. The breaking point for boundary correction is obtained by maintaining a *score vector*. The score vector is meant to store boolean values for indicating the presents/absence of NEs. Size of the vector is the number of candidate sentences both adjacent clusters. *Score vector* value is true if the intersection of named entities in the respective candidate sentence and the discourse excluding candidate sentence is not null. Now the *score vector* value is scanned from one side. If it finds a true in the vector, that is the breaking point to adjust the segment boundary in the discourse. Remove those sentences from the respective discourse and add them to the other discourse. The *score vector* is again scanned from other side and do the operations as specified in previous two sentences. The number of candidate sentences taken for this method is four which gives the best result.

This algorithm gives perfect result based on the Named Entities. The discourse boundaries can again be corrected based on lexical continuity of the sentences at the boundary. WordNet can be used for analysing lexical continuity as specified in the next section.

c. Boundary adjustment using WordNet

WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets [1]. By using WordNet, it is possible to find the coherence relationship in text. Here, WordNet is used for finding a position in segment boundary where lexical chain ends. Algorithm-3 implements it.

```
Algorithm-3: Boundary-WordNet Algorithm
Input: discourse_NE
Output: discourse_WN

Read discourseNE
discourseWN=discourseNE
Tag discourseWN
for each d in discourseWN
    find disc_bound of d
    for each sentence in disc_bound
        find the lexical_cohesion of the sentence using
        hyperrnymy
    find a break_point at which lexical_cohesion has large
    valley point
     move the sentences from one d to another at the
    break_point
return discourseWN
```

Here *lexical_cohesion* is calculated to find the end point in the lexical chain. *disc_bound* is the candidate sentences from the discourse boundary. A *score vector* is maintained for candidate sentences with lexical similarity between them. The lexical similarity is calculated between discourse d_i and sentence s_j using the *hypernymy* relation of the WordNet as given below.

$$sim(di, sj) = \frac{\mathrm{hyper\_sim(di, sj)}}{\sqrt{len(di)^2 + len(sj)^2}}$$

where $hyper\_sim(d_i,s_j)$ returns the cardinality of the intersection set of hypernyms of both d_i and s_j and $len(d_i)$ and $len(s_j)$ are number of tokens in discourse d_i and sentence s_j respectively. In order to determine where the lexical chain ends, find a valley point at which the depth is maximum among three adjacent scores. If the selected valley point very much lower than other valley points, change the boundary of discourse at that valley point by moving the sentences from one discourse to another. Thus, we can ensure lexical coherence in the discourse.

4. Results

This method is experimented on a number of text inputs. The text inputs are collected set of news from various news sites and electronic papers. While comparing with other models described in the literature survey, the output of the proposed method is far better and domain independent. But, in other models like TextTiling, the output changes tremendously with token sequence size. As specified earlier, the discourse segmentation is carried out through 3 stages. After first stage, ie., the output of the clustering algorithm gives more number of clusters. Among them some of them are to be combined. While cross checking the output of the clustering algorithm with the human judged discourse, almost all discourse starting points are somewhat aligned. When the clustering algorithm output is given to the next level, i.e. boundary correction using named entities, the output is changed from the previous level. Boundaries are adjusted at required places and some clusters are merged. The boundary is again adjusted when it is cascaded with next level of algorithm using WordNet. Most cases, it gives best results. The Table 1 shows precision and recall of some of the outputs while comparing with the human judgement.

Table 1. Precision and Recall of Discourse Identification

Discourses (Human)	Discourses (System)	Correctly Identified	Precision	Recall
8	14	6	0.75	0.42
15	12	7	0.58	0.46
5	5	2	0.4	0.4
17	20	9	0.45	0.52
12	8	3	0.375	0.25
25	21	18	0.85	0.72
16	14	10	0.71	0.62
10	13	8	0.61	0.8

The output of this system is compared with the human judgement of the same text in order to assess the accuracy. The overall accuracy is calculated as 70%.

5. Limitations and Future Scope

This system mainly depends on the nouns and named entities for discourse segmentation. Therefore, it should have handled all the nouns and Named Entities in the text input. The main hurdle for this system is the presence of anaphora. While examining any text, it is clear that, there are lot of third person pronouns. Since pronoun is the replacement of the nouns, this method cannot process every occurrence of nouns and Named entities. In rare cases some discourses may be swallowed by others, since the pronouns are not considered. The anaphora resolution can change this method drastically to the best result. So, this system must be enhanced with anaphora resolution in the future. Derived nouns may be excluded, since exclusion of derived nouns can also improve the accuracy of clustering algorithm.

6. Conclusion

The method described in this paper is characterized by the combination of cohesion and coherence relationships among the sentences in the text. The cohesive relationship is revealed by agglomerative sentence clustering and boundary adjustment with named entities. The lexical continuity is obtained by considering the hypernymy relation of the WordNet. The cascading of these two approaches refines the output in a better way. It gives better result of 70% accuracy for discourse segmentation. The accuracy can be further improved by incorporating anaphora resolution. The important feature is the introduction of a method for discourse segmentation with nouns. The main advantage is that this model is easy to understand. Each discourse segment is assured to be concentrated on particular set of nouns. Since nouns are used as main strategy for this method, it is possible for information retrieval based on noun query terms.

7. Reference

[1] George A. Miller, ”WordNet: A Lexical Database for English”, in Magazine Communications of the ACM CACM Homepage archive Volume 38 Issue 11, Pages 39-41, Nov. 1995

[2] John F, ”Use of signalling nouns in a learner corpus”, International Journal of Corpus Linguistics, Dec 2005

[3] Lapata M and Barzilay R, ”Automatic Evaluation of Text Coherence: Models and Representations”, in Proceeding IJCAI’05 Proceedings of the 19th international joint conference on Artificial intelligence,Pages 1085-1090, 2005

[4] Lin Z, NG HT and Kan MY, ”Automatically Evaluating Text Coherence Using Discourse Relations”, in HLT ’11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Pages 997-1006, 2011

[5] Louwerse M.M. and Graesser AC, ”Coherence in Discourse”, in In Strazny, P. (ed.),Encyclopedia of linguistics. (pp. 216-218) Chicago, Fitzroy Dearborn.

[6] Marti A. Hearst, PARC X, ”TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages”, in Computational Linguistics archive Volume 23 Issue 1,Pages 33-64, March 1997

[7] Matthiessen C, ODonnell M and Zeng L, Discourse Analysis and the Need for Functionally Complex Grammars in Parsing, in Proceedings of the Second Japan-Australia Joint Symposium on Natural Language Processing, October 2-5, Kyushu Institute of Technology, Japan, 1991

[8] Morris J and Hirst G, ”Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text”, in Journal Computational Linguistics archive Volume 17 Issue 1, Pages 21-48, 1991

[9] Paul Mc Kevitt, Partridge D and Wilks Y, ”Approaches to Natural Language Discourse Processing”, in Artificial Intelligence Review, Volume 6, Issue 4, pp 333-364, 1992

[10] Thomas C. Rindflesch, ”Natural Language Processing”, in Annual Review of Applied Linguistics, Volume 16, pp 70-85, March 1996