A Realistic Success Criterion for Discourse Segmentation

Meltem Turhan Yöndem and Göktürk Üçoluk

Dept of Computer Eng.,
Middle East Tech. Univ. 06531
Ankara, Turkey
{mturhan,ucoluk}@ceng.metu.edu.tr

Abstract. In this study, compared to the existing one, a more realistic evaluation method for discourse segmentation is introduced. It is believed that discourse segmentation is a fuzzy task [Pas96]. Human subjects may agree on different discourse boundaries, with high agreement among them. In the existing method a threshold value is calculated and sentences that marked by that many subjects are decided as real boundaries and other marks are not been considered. Furthermore automatically discovered boundaries, in case of being misplaced, are treated as a strict failure, disregarding the proximity wrt to the human found boundaries. The proposed method overcomes these shortcomings, and credits the fuzziness of the human subjects' decisions as well as tolerates misplacements of the automated discovery. The proposed method is tunable from crisp/harsh to fuzzy/tolerant on human decision as well as automated discovery handling.

1 Introduction

This study is about evaluation of the success of programs that divides large volumes of text into a certain number of utterances that form coherent units called *Discourse Segments*. It is widely accepted that discourse is structured. Many experiments support this statement [Gro92, Gro86, Hir93, Lit95, Man88, Pol88, Web91]. Naive subjects also agree on the discourse boundaries of a given text.

The first step is to represent the segment boundary information decided by human subjects in a natural way. It is quite a common technique to ask a group of subjects to read a given text and then indicate the sentences where a new context (discourse) starts. Following this, usually a statistical analysis of the results is carried out to decide whether there is an agreement among the subjects about each candidate boundary. Then, having the boundaries at hand, computer generated boundaries are tested against them, and by this way a decision about the quality of the computer segmentation is made.

Most of the current works use crisp methods in deciding about the success of the boundary discovery. This, of course leads to non realistic evaluations where near-miss cases receive the same discrediting as complete misses. Furthermore, the fuzziness steaming from disagreements among human subjects are also not handled. There exist a minor number of previous work that pinpoint this issue and propose solutions [Bee97, Pev94]. This work proposes a new evaluation scheme claiming to be a solution to all such aspects.

2 Discourse Segmentation

The hypothetical line between the discourse segments is named as *discourse boundary*. The following example is a portion of some Turkish corpus used and the human subjects' decision about the boundaries. *(Leading smilies before each sentence reflect the count of agreement and intend to aid the visualization of where the boundaries are.)*

Emine bir aralık bu oyundan usandı. (marked as a boundary by 0 subjects)
After some time Emine got bored of this game.
Kamer'in kucağından inip taşlığın sonundaki bahçeye koştu. (0 subject)
Getting from Kamer's lap, she run to the garden at the end of the stone courtyard.
☺☺☺☺ İstanbul'un Kıztaşı taraflarında dededen kalma bir konaktı. (4 subjects)
It was a mansion in İstanbul by Kıztaşı, inherited from grandfather.
Annesi, babası, ağabeyi bir tarafta, teyzesi, eniştesi,ablası öbür tarafta otururlardı. (0 subjects)
On one side her mother, father and brother, on the other side her aunt, uncle in law and sister used to lie.
Bahçe iki ailenindi. (0 subjects)
The garden belonged to the two families.
☺☺☺☺ Fakat hikayemizi anlattığımız sıralarda kadınlar erkekten kaçardı. (4 subjects)
But at the time we were telling our story, women used to run away from men.
☺ Bu, ne demek diyeceksin. (1 subjects)
what does this mean, you ask?
☺ Evet, sevgili küçük okuyucum, kadınlarla erkeklerin hayatları ayrı geçer, babalar, kardeşler ve çok yakınlarla ancak akşamları buluşulurdu. (1 subjects)
Yes, dear fellow reader, women and men lived life separately; it was only in the evenings that fathers, brothers and close relatives were met with.
Erkekler çoğu vakitlerini evin selamlık denilen bir bölümünde geçirirlerdi. (0 subjects)
Men used to spend most of their time in a part of the house called the selamlık.
Kadınların oturdukları bölümün adı haremdi. (0 subjects)
The part in which women used to live was called the harem.
Mutfak kapısı bahçeye açıldığından, üstelik aşçı da erkek olduğundan, kadınlardan pek oraya çıkan olmazdı. (0 subjects)
Since the kitchen door opened to the garden and also because the cook was a man, women did not usually enter there.
☺☺☺☺☺☺☺ Emine taflanlı yollarda koştu, bahçenin ortasındaki uzun servinin çevresini birkaç kere dolandı. (7 subjects)
Emine ran on the flowered roads, went around the tree in the middle of the garden a couple of times.
☺ Hava oldukça serindi. (1 subject)
It was quite chilly.
Aşçı Yaver ağa, mutfak kapısından başını uzatıtı. (0 subjects)
The cook Yaver ağa looked through the kitchen door.

3 Existing Evaluation Method

The method can be summerized as:

- Statistically determine what makes a human subjects' boundary decision reliable.
- Nominate all such reliable boundaries to be the *real boundaries*.
- Consider all boundaries discovered by the automated process that match exactly with real boundares as success points. Consider all other automaticaly discovered points as failures.

As the concept of reliable boundary is major to this method, it is worth to look into it in greater details. In the following subsection you will find the statistical procedure used so far in the existing method to discover the reliable boundaries.

3.1 Statistical Analysis: Cochran's Q method

This section is a brief summary of the Cochran's Q method used to judge the agreement between the subjects on segment boundaries. Passonneau & Litman [Pas96, Pas97] used this method to judge their subject's agreement. The advantage of this method is that it gives the number of agreements necessary for a variable to be considered as the needed class, in this case a boundary. In order to find the number of agreements necessary, usually a partial Cochran's Q method is used.

In this kind of agreement calculation, the distribution of 1's (indicating there is a boundary before the sentence) and 0's (indicating there is no boundary before the sentence) in the subjects' responses are tested against the null hypothesis that the agreement was by chance. A good example for such a calculation of an English corpus can be found in [Pas96].

$$Q = \frac{k(k-1)\sum_{j=1}^{k}\left(G_j - \sum_{j=1}^{k} G_j^*\right)^2}{k\sum_{i=1}^{N} L_i - \sum_{i=1}^{N} L_i^2} \tag{1}$$

where

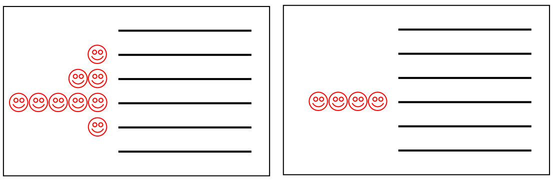
G_j : total number of boundaries in the j^{th} column
G^* : mean of the G_j
L_i : total number of boundaries in the i^{th} row

As an example, the Q calculated for a human segmentation of a corpus of 3691 sentences from which the above given extract was taken yielded a Q value of 16931. Chi-square values indicated that the result would be significant when Q had a value greater than 45. In this case, the Q value was greater than this value for $p = 0.001$, which means that the probability of accidental agreement among subjects is less than 0.001 for this narrative; that is, the agreement is so high that it could not occur by chance.

The next issue is to determine the required number of subject agreements to consider a sentence as a boundary. Similar to Passonneau [Pas96], and other similar studies do this by partitioning Q into distinct components for each possible G_i (0 to *total number of subjects*). The boundaries of any discourse can be decided by partially calculating Cochran's Q. By this way, it is statistically possible to determine the required minimal number of subjects to agree on a single sentence with accidental occurrence probability lesser than the calculated p value. All of the sentences marked as boundary by that minimal number or more human subjects will be considered as boundaries in subsequent analyses of the corpus.

3.2 Short Commings of the Existing Evaluation Method

The following cases explains the need for more realistic discourse segmentation evaluation method.



(a) (b)

Fig. 1. A case of more information content

The existing evaluation method would consider sentence 4 (third bar in the figure) as a boundary in both figure 1 (a) and figure 1 (b). There will be no difference in the credits received for both cases if the automated boundary discovery hits/misses this boundary. But we believe that there is more clue about the boundary in (a) with respect to (b) since much more human subjects agreed that there is a boundary in that vicinity.

Consider the human subjects' agreements are as in the figure 2. If sentences 3 is accepted as a boundary then 4 is also a boundary. This is so because the existing system cares only about the number of subjects that marked a sentence as boundary to be above a constant threshold in order to accept it as a boundary. But, obviously there is only one boundary which the subjects could

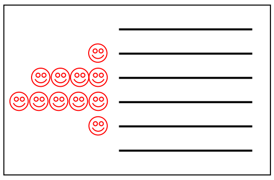
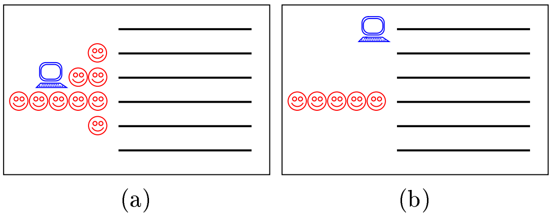


Fig. 2. A case of two peeks in the same vicinity



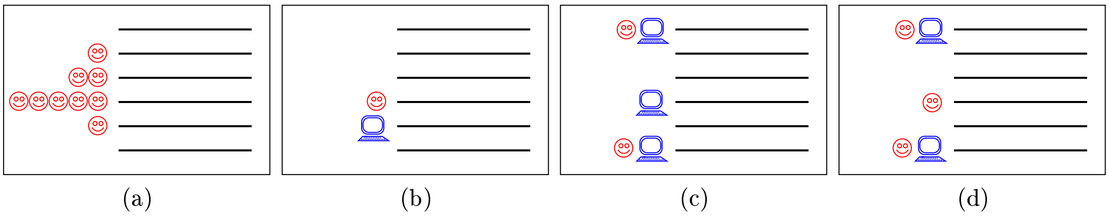
(a) (b)

Fig. 3. A case of week vs. strong disagreement of human and computer

not come to a unified agreement about its position. The group has split into two groups where the count of each group has passed the threshold value.

We also believe that if the program finds a boundary at a sentence that no subject marks a boundary as it is the case in figure 3 (b), that should be considered different than finding a boundary at a place that at least some of the subjects put a boundary mark. Existing systems do not consider figure 3 (a) and Figure 3 (b) as different.

Summing up, we are basicly facing the following problems which are ill-treated in the existing systems:



(a) (b) (c) (d)

Fig. 4. Currently ill-treated cases

- The human segmenters do not always agree on a boundary (figure 4 (a))
- The automated segmentation process does not always hit the boundary even if it is crisp. (figure 4 (b))
- Occasionally the automated segmentation process may
  - create additional boundaries (we will name these as *fantom*) (figure 3) (figure 4 (c))
  - or omit the existence of a boundary. (figure 4)(figure 4 (d))

4 Proposed Evaluation Method

The idea is to set up a system which considers boundaries as islands of the distributions. We name these islands as *clusters*. A clustering algorithm based on statistical values determines the cluster formations in the results of the human subjects segmentation. The boundary decisions of the automated system which fall into the close vicinity of those clusters are considered to be a distribution about a single boundary (similar to the case of human subjects) and the quality of the judgment of the automated process is assessed by an distribution versus distribution evaluation.

Placements of phantom boundaries (claimed boundaries of the automated system which are not in the close vicinity of a cluster) are punished proportional to their distance from the nearest cluster.

The proposed system has adjustable parameters which control the crispness of this treatment. These are explained below, along with the algorithm.

4.1 Evaluation Algorithm

Let us assume the following:

- S human subject participated the segmentation.
- $h(i)$ is the discrete distribution value of the count of human subjects that agree there is a segmentation boundary at the end of line i of a given text.
  If $h(3)$ *is 5 this will mean that 5 individual have agreed that there is a segmentation boundary before the 5 th sentence.*
- $c(i)$ is the discrete boolean distribution, generated by a computer segmentation system, that yields $true$ if the existence of a segmentation boundary claimed before line i of the text, and a $false$ if no boundary there exists.
  $c(3)$ *is true if the computer segmentation system is claiming the existence of a boundary before the 3rd sentence.*
- Let N number of sentence exist in the text.

4.2 Step 1: Determine Clusters of *h*

Clusters are subranges of [1,N]. We define

$$cluster_k \overset{\triangle}{=} \langle cluster_k.lower, cluster_k.upper \rangle$$

So that

- *cluster* denotes the count of clusters.
- $cluster_k.lower \leq cluster_k.upper$
- $cluster_k.upper < cluster_{k+1}.lower$ for $k = 1, \ldots, cluster$
- $(x \in cluster) \overset{\triangle}{=} (cluster_k.lower \leq x \leq cluster_k.upper)$
- $\forall s,\ s \in cluster_k,\ similarity\_measure(i, cluster_k, h) < expected\_similarity(cluster_k, h)$

The last item, which defines the criterion of being a member of a cluster is based on two functions: *similarity_measure* and *expected_similarity*. Literally the second imposes a quantified level of similarity based on the portion of h falling into the cluster, and the first is a quantified measure of similarity of a candidate point with respect to a cluster.

This recursive definition of a cluster can be implemented in various ways. We propose a very simple $\mathcal{O}(n)$ algorithm based on two parameters:

chain_strength A maximum allowed value of $\mid s_1 - s_2 \mid$ such that $\forall s,\ s_1 < s < s_2,\ h(s) = 0$ and $s_{1,2} \in cluster_m$

minimal_peak An empirical value such that $\forall k \exists s, s \in cluster_k,\ s \geq$ minimal_peak.

Clustering algorithm

```
$find\_clusters(cluster, h, N, chain\_weekness, minimal\_peak) \leftarrow$
{
  $k \leftarrow 0$
  $s \leftarrow 1$
  $candidate\_exists \leftarrow false$
  for $s \leftarrow 1 \ldots N$ do
    {
      if $h(s) > 1$ then
                      if $\neg candidate\_exists$
                        then { $candidate\_exists \leftarrow TRUE$
                               $candidate\_start \leftarrow s$
                               $weekness \leftarrow 0$
                               $has\_minimal\_peak \leftarrow (h(s) \geq \mathsf{minimal\_peak})$ }
                       else
                           if $\neg has\_minimal\_peak \ \wedge \ h(s) \geq \mathsf{minimal\_peak}$
                             then $has\_minimal\_peak \leftarrow TRUE$
                              else
                                 if $weekness = \mathsf{chain\_strength}$
                                   then
                                       if $has\_minimal\_peak$
                                         then { $k \leftarrow k + 1$
                                                $cluster\_k.lower \leftarrow candidate\_start$
                                                $cluster\_k.upper \leftarrow s - weekness$ }
                                          else $candidate\_exists \leftarrow FALSE$
    } }
```

This algorithm is based on the assumption that at least a gap of size chain_strength exists between successive clusters. If that is not the case then a more elaborated clustering algorithm must be employed.

Note that is is quite possible that some of the human data gets ignored and does not become a part of a cluster.

4.3 Step 2: Evaluating performance of *c*

Now having to hand the clusters we propose the following evaluation formula for the performance of a computer segmentation c:

$$success(c) = \mathcal{P}(c) - \mathcal{N}(c)$$

where $\mathcal{P}(c)$ is some calculated positive value awarding the success points and $\mathcal{N}(c)$ is some subtracted positive value punishing the failure points.

The following is proposed for $\mathcal{P}(c)$

$$\mathcal{P}(c) \triangleq \frac{1}{|cluster| \cdot S} \cdot \sum_{k=1}^{|cluster|} \sum_{s=low(k)}^{up(k)} \sum_{q=low(k)}^{up(k)} \frac{h(s) \cdot c(q)}{|s - q + 1|^{\alpha}}$$

Here $up(k)$ and $low(k)$ are either the limits of $cluster_k$ or a narrowed down version of them based on the statistical properties of the h distribution in the range of $cluster_k$. α is an exponent value in the range of $1 \leq \alpha \leq 2$ that we will name as *hardness*.

$$low(k) \triangleq \max(\mu_k - \beta\sigma_k, cluster_k.lower)$$

$$up(k) \triangleq \min(\mu_k + \beta\sigma_k, cluster_k.upper)$$

Where μ_k and σ_k are defined as the mean and standard deviation of the distribution in the range of $cluster_k$, respectively. β is a parameter that controls the *crispness* of the success evaluation. The bigger β is the fuzzier the evaluation gets.

For the punishment term that is going to be introduced for the cases of phantom boundaries the fallowing is proposed. This is a linear function of the distance of all phantom boundaries from the mid point of their nearest clusters. Assuming a minimal punishment of value P_{min} and a maximal of P_{max} the $\mathcal{N}(c)$ term can be expressed as

$$\mathcal{N}(c) \triangleq \sum_{k=1}^{|cluster|-1} \sum_{q=low(k)}^{up(k)} \frac{P_{min} - P_{max}}{2up(k)} \cdot |2q - low(k+1) - up(k)| + P_{max}$$

5 Conclusion

We are proposing a more realistic approach to the evaluation of an automated discourse segmentation. The following are the problems of discourse segmentation which are not handled properly and delicately in the existing approach:

- Discourse segmentation is a fuzzy task and human responses should be kept as close to natural as possible without forcing this information to harsh and crisp representations. Not doing so causes information loss and extremely oversimplifies the evaluation.
- Automated segmentation is just an approximation to human intelligence. It is easily misled in weak circumstances especially in cases where human segmenters have not come to a full agreement, either.

A new method is proposed to overcome the injustice of the existing method. The new method has the following advantages:

- It is cognitively more similar to the human way of success evaluation.
- Almost all the boundary information marked by human subjects are kept and considered.
- A tunable metric for closeness of two distribution, namely the ones of human subjects and the automated process, is introduced. This is believed to increase the fairness of the evaluation.
- Phantom boundaries claimed by an automated process is punished by a severeness criterion which is based on the distance from the nearest real boundary.

References

[Bee97] , Beeferman, D., 'A Probabilistic Error Metric for Segmentation Algorithms', *Unpublished notes*, `http://www.dougb.com/research.html`, 1997.

[Gro86] Grosz B., Sidner, C., 'Attention Intention and Structure of Discourse: A Review', *Computational Linguistics*, Vol. 12, No. 3, pp. 175-204, 1986.

[Gro92] B. Grosz, J. Hirschberg, 'Some Intentional Characteristics of Discourse Structure', *Proceedings of the International Conference on Spoken Language Processing*, pp. 429-432, 1992.

[Hir93] Hirschberg, J., Litman, D., 'Empirical Studies on the Disambiguation', *Computational Linguistics*, Vol. 19, No. 3, pp. 501-530, 1993.

[Lit95] Litman, D., J., Passonneau, R., 'Combining Multiple Knowledge Sources for Discourse Segmentation', *33rd Annual Meeting of Association for Computational Linguistics*, pp. 108-115, 1995.

[Man88] Mann, W.C., Thompson, S.A., 'Rhetorical Structure Theory: Towards a Functional Theory of Text Organization', Text, Vol. 8, pp. 243-281, 1988.

[Pas96] Passonneau, R., Litman, D.J., 'Empirical Analysis of Three Dimensions of Spoken Language: Segmentation, Coherence and Linguistic Device', ed. Hovy, E., Scott, D., Computational and Conversational Discourse, pp. 161-194, Springer Verlag, 1996.

[Pas97] Passonneau, R., Litman, D.J., 'Discourse Segmentation by Human and Automated Means', Computational Linguistics, Vol. 23, No. 1, 1997.

[Pev94] Pevzner, L., Hearst M., 'A Critique and Improvement of an Evaluation Metric for text Segmentation' *Computational Linguistics*, Vol. 21, No. 1, pp. 19-36, 1997.

[Pol88] Polanyi, L., ‘A Formal Model of the Structure of Discourse’, Journal of Pragmatics, Vol. 12, pp. 601-638, 1988.

[Web91] Webber, B. L., ‘Structure and Ostension in the Interpretation of Discourse Deixis’, Language and Cognitive Processes, Vol. 6, No. 2, pp. 107–135, 1991.