Discovering Utterance Fragment Boundaries in Small Unsegmented Texts

László Drienkó

dri@t-online.hu

Abstract: We propose an algorithm for inferring boundaries of utterance fragments in relatively small unsegmented texts. The algorithm looks for subsequent largest chunks that occur at least twice in the text. Then adjacent fragments below an arbitrary length bound are merged. In our pilot experiment three types of English text were segmented: mother-child language from the CHILDES database, excerpts from *Gulliver's travels* by Jonathan Swift, and *Now We Are Six*, a children's poem by A. A. Milne. The results are interpreted in terms of four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability. We find that i) Inference Precision grows with merge-length, whereas Alignment Precision decreases – i.e. the longer a segment is the more probable that its two boundaries are correct; ii) Redundancy and Boundary Variability also decrease with the merge-length bound – i.e. the less boundaries we insert, the closer they are to the ideal boundaries.

1. Introduction

The problem of how to segment continuous speech into components dates back at least to Harris [2]. Harris used "successor frequencies", i.e. statistics, to predict boundaries between linguistic units. [8], using syllable-based artificial languages, demonstrated that statistical information is indeed available for infants acquiring language. Results in language acquisition research indicate that speech segmentation is affected by various lexical and sub-lexical linguistic cues (see e.g. [4]). Computational models of speech segmentation typically seek to identify the computational mechanisms underlying children's capacity to segment continuous speech (see [1] for a review). [6] outlines an integrated theory of language acquisition where the learner uses various cognitive heuristics to extract large chunks from the speech stream and the 'ultimate' units of language are formed by segmenting and fusioning the relevant chunks. The philosophy behind our boundary inference algorithm is, broadly speaking, similar in that we first identify "large" utterance fragments in unsegmented texts, i.e. character sequences, and then apply 'fusion' – 'merging', in our terminology – to see how precision changes.

Our heuristic for identifying utterance fragments is based on the assumption/intuition that recurring long sequences are more informative of segment boundaries than recurring shorter ones. Individual characters, for instance, can be followed

by, practically, any other character even in a short text. Reoccurring words or, notably, word combinations, on the other hand, are less likely to be followed/preceded by a word beginning/ending with the same letter as on the first occurrence. If a word combination still happens to be followed/preceded by a word beginning/ending with the same letter as on a previous occurrence, this reoccurring word can be considered as being part of an even larger word combination which, in turn, is less likely to reoccur in the same context. Thus we intuit that largest chunks represent sequence boundaries more reliably than shorter ones. Naturally, the ultimate largest chunk is the whole text itself with its two 100%-certain boundaries.

In our pilot investigation three types of English text were segmented: mother-child language from the CHILDES database, excerpts from *Gulliver's travels* by Jonathan Swift, and *Now We Are Six*, a children's poem by A. A. Milne. The texts are relatively short, with 60 - 7741 word tokens, 186 - 32,859 characters.

2. Description of the algorithm

The basic, CHUNKER, module of our algorithm looks for largest character sequences that occur more than once in the text. Starting from the first character, it concatenates the subsequent characters and if a resultant string s_i only occurs in the text once, a boundary is inserted before its last character in the original text since the previous string, s_{i-1}, is the largest of the i strings. Thus the first boundary corresponds to s_{i-1}, our first tentative speech fragment. The search for the next fragment continues from the position after the last character of s_{i-1}, and so on. As can be seen from our results, in terms of sequence length, the fragments output by this module broadly correspond to words.

The MERGE component of the algorithm concatenates fragments s_i and s_{i+1} if s_{i+1} consists of less than k characters. In other words, the boundary between s_i and s_{i+1} is deleted if s_{i+1} is shorter than k, an arbitrary length bound. In our experiments we had $1 \leq k \leq 11$.

The EVALUATE module computes four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability.

Inference Precision (IP) represents the proportion of correctly inferred boundaries (cib) to all inferred boundaries (aib), i.e. IP = cib / aib. The maximum value of IP is 1, even if more boundaries are inferred than all the correct (original) boundaries (acb).

Redundancy (R) is computed as the proportion of all the inferred boundaries to all the correct (original) boundaries, i.e. R = aib / acb. R is 1 if as many boundaries are inferred as there are boundaries in the original text, i.e. aib=acb, R is less than 1 if less boundaries are inferred than acb, and R is greater than 1 if more boundaries are inferred than optimal. Note that 1/R = acb\aib specifies how many words are grouped together on average in an inferred segment, i.e. the average fragment length in words.

Alignment Precision (AP) is specified as the proportion of correctly inferred boundaries to all the original boundaries, i.e. AP = cib / acb. Naturally, the maximum value for AP is 1.

Boundary Variability (BV) designates the average distance (in characters) of an inferred boundary from the nearest correct boundary, i.e. BV = (Σdf_i)/aib.

The above measures are not totally independent, since Inference Precision × Redundancy = Alignment Precision, but emphasise different aspects of the segmentation mechanism. Obviously, IP =AP for R=1.

3. The experiments

3.1. Experiment 1

In this experiment the first Anne file, *anne01a.xml*, of the Manchester corpus, [9], in the CHILDES database, [3], was investigated. The files were converted to simple text format, annotations were removed together with punctuation symbols and spaces. Mother and child utterances were not separated, so the dataset constituted an unsegmented (written) stream of 'mother-child language'. The original text consisted of 1815 word tokens and the average word length was 3.75 characters. The unsegmented version of the text consisted of 6801 characters. Initially, *k=1*, the CHUNKER module of our algorithm inserted 1129 boundaries, i.e. 1129 segments were identified with average segment length 6.02 characters. This means that the inferred fragments were, on average, 2.27 characters longer. The precision values were as follows: Inference Precision = 0.66, Redundancy = 0.62, Alignment Precision = 0.41, Boundary Variability = 0.53. In the second part of the experiment we let the merge-length bound *k* change from 2 to 11. For instance, *k = 3* means that, given the segmentation as provided by the CHUNKER module (the *k = 1* case, with no merging), fragment f_{i+1} is glued to the end of f_i if f_{i+1} consists of less than 3 characters, i.e. if f_{i+1} is one- or two-character-long. That is, the maximum merge-length is 2 for *k=3*. Figure 1 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 2 plots how the precision values change. For *k=11*, the values were IP=0.78, R=0.07, AP=0.05, BV= 0.3, and 121 boundaries were inserted.

3.2. Experiment 2

In this experiment the first part of Chapter 1 from *Gulliver's travels* by Jonathan Swift, [7], was investigated. The original text consisted of 1634 word tokens and the average word length was 4.05 characters. The unsegmented version of the text consisted of 6621 characters. The CHUNKER module inserted 1565 boundaries. The average segment length was 4.23 characters, which is quite close to the 4.05 average for the original text. The precision values were the following: IP = 0.5, R = 0.96, AP = 0.48, BV = 0.9. Figure 3 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 4 plots how the precision values change. For *k=11*, the values were IP=0.86, R=0.02, AP=0.02, BV= 0.17, and 30 boundaries were inserted.

3.3. Experiment 3

In this experiment Chapter 1 from *Gulliver's Travels* was investigated. The original text consisted of 4034 word tokens and the average word length was 4.17 characters. The unsegmented text consisted of 16,821 characters. The CHUNKER module inserted 3307 boundaries. The average segment length was 5.09 characters, about 1 character longer than the 4.17 value for the original text. The precision values were the following: IP = 0.53, R = 0.82, AP = 0.43, BV = 0.84. Figure 5 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 6 plots how the precision values change. For *k=11*, the values were IP=0.71, R=0.03, AP=0.02, BV= 0.35, and 133 boundaries were inserted.

3.4. Experiment 4

In this experiment Chapters 1 and 2 from *Gulliver's travels* were merged into a single text. The two chapters contained 7742 word tokens and the average word length was 4.24 characters. The unsegmented text consisted of 32,859 characters. The CHUNKER module inserted 5802 boundaries. The average segment length was 5.66 characters, about 1.5 characters longer than the 4.24 value for the original text. The precision values were the following: IP = 0.53, R = 0.75, AP = 0.4, BV = 0.8. Figure 7 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 8 plots how the precision values change. For *k=11*, the values were IP=0.75, R=0.04, AP=0.03, BV= 0.29, and 323 boundaries were inserted.

3.5. Experiment 5

Finally, we examined the children's poem *Now We Are Six* written by A. A. Milne, [5]. The poem consists of 60 word tokens and the average word length was 3.1. The unsegmented poem consists of 186 characters. The CHUNKER module inserted 79 boundaries, cf. Box 1. The average segment length was 2.35 characters, about 0.7 character shorter than the 3.1 value for the original text. The precision values were the following: IP = 0.45, R = 1.32, AP = 0.6, BV = 0.77. Figure 9 shows how the number of inferred boundaries changes with the maximum merge-length. Figure 10 plots how the precision values change. For *k=11*, the values were IP=1 R=0.0167, AP=0.0167, BV= 0, i.e. the original sequence was restored with a 100% inference precision due to the single boundary inserted at the end of the text.

4. Discussion and conclusions

For all the texts that we looked at, the following pattern could be observed:

Inference Precision (the proportion of correctly inferred boundaries of all inferred boundaries) grows (45-66% to 70-100%) with maximum merge-length (0 to 10), whereas Alignment Precision (the proportion of correctly identified boundaries of all

the original, correct boundaries) decreases: i.e. the longer a segment is the more probable that its two boundaries are correct.

Redundancy (the proportion of all the inferred boundaries to all the correct boundaries) and Boundary Variability (the average distance from the closest correct boundary) also decrease with the merge-length bound: i.e. the less boundaries we insert, the closer they are to the ideal boundaries.

Our data suggest, most explicitly in Experiment 5, that, as the merge-length bound grows, Inference Precision approaches 1, Boundary Variability 0, Redundancy and Alignment Precision 1/n, where n is the number of word tokens in the original text.

The utterance fragments that our algorithm can detect are not necessarily individual words or syntactic phrases. The possible strengths of our method lie, on the one hand, in its potential to provide empirical insights into the statistical structure of natural language i) on the basis of small texts ii) without previous training corpora or iii) explicit probability values. On the other hand, the utterance fragments detected by our algorithm can serve as input for subsequent segmenting mechanisms to break down text into ultimate components, practically, into words.

Our results also suggest that looking for largest recurring chunks may be a powerful cognitive strategy. Statistically, the lengths of the fragments that our CHUNKER identified are quite close to the original word lengths. Note also, that all BV values were less than 1, which means that, for a given R value, a learner could obtain an optimal segmentation – i.e. where all inferred boundaries are correct – by shifting the inferred boundaries less than 1 character, on average, to the right or to the left. In other words, language learning could be based on memorizing tentative chunks that could be "finalised" later, as cognitive development progresses.

Figures and boxes

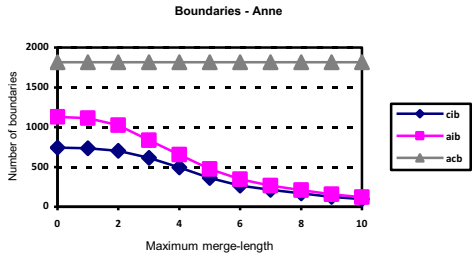

Fig. 1. Number of boundaries as function of maximum merge-length – Anne file from CHILDES (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

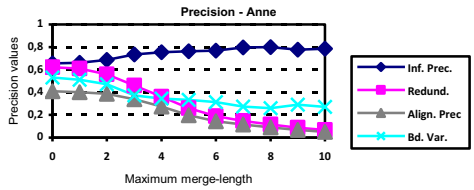

Fig. 2. Precision values changing with maximum merge-length – Anne file from CHILDES.

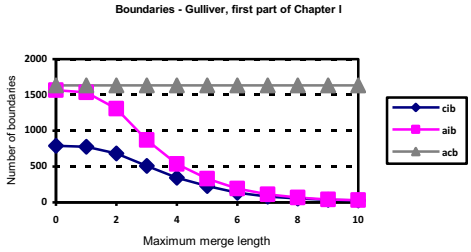

Fig. 3. Number of boundaries as function of maximum merge-length – first part of Chapter 1, *Gulliver's travels*. (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

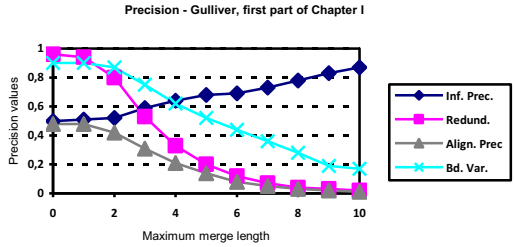

Fig. 4. Precision values changing with maximum merge-length – first part of Chapter 1, *Gulliver's travels*.

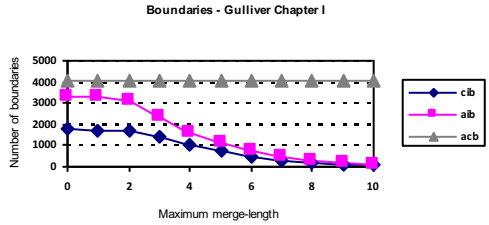

Fig. 5. Number of boundaries as function of maximum merge-length – *Gulliver's travels*, Chapter 1. (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

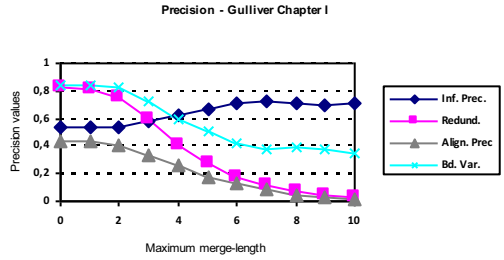

Fig. 6. Precision values changing with maximum merge-length – *Gulliver's travels*, Chapter 1.

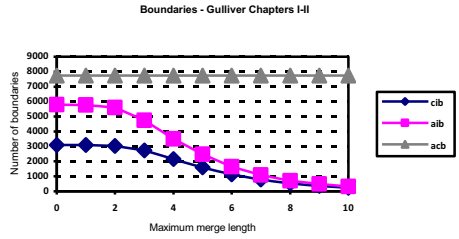

Fig. 7. Number of boundaries as function of maximum merge-length – *Gulliver's travels*, Chapters 1 and 2. (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

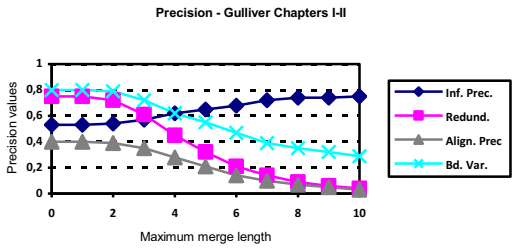

Fig. 8. Precision values changing with maximum merge-length – *Gulliver's travels*, Chapters 1and 2.

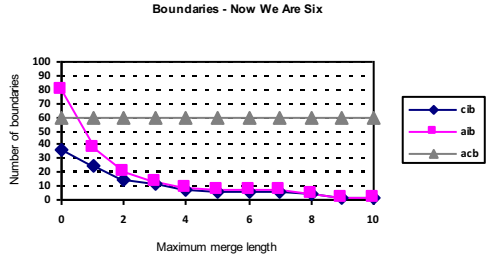

Fig. 9. Number of boundaries as function of maximum merge-length – *Now we are six.* (cib: correctly inferred boundaries, aib: all inferred boundaries, acb: all correct boundaries).

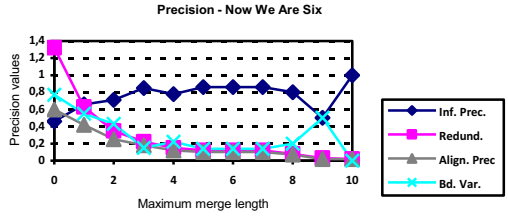

Fig. 10. Precision values changing with maximum merge-length – *Now we are six*.

Box 1

WHENIWAS:O:NE:I:HA:D:JUST:BE:G:U:N:WHENIWAST:W:OI:WASN:E:AR:LY:NE
:W:WHENIWAST:H:RE:EIWAS:HA:R:D:LY:M:EWHENIWASF:O:U:R:IWASN:O:T:M
:U:C:H:M:ORE:WHENIWASF:IVE:IWAS:JUST:A:L:IVE:B:U:T:NOW:I:A:M:SIX:I':M:
ASCLEVER:ASCLEVER:SO:I:TH:I:N:K:I':L:L:BE:SIX:NOW:A:N:D:FO:RE:VER:

References

1. Brent, M. R.: Speech segmentation and word discovery: a computational perspective. Trends Cognitive Sciences 3(8) (1999) 294–301
2. Harris, Z. S.: From phoneme to morpheme. Language 31 (1955) 190–222
3. MacWhinney, B.: The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates (2000)
4. Mattys, S. L, White, L., Melhorn, J.F.: Integration of multiple speech segmentation cues: a hierarchical framework. Journal of Experimental Psychology: General 134(4) (2005) 477–500
5. Milne, A.A. Now we are six (poem). Source: http://www.familyfriendpoems.com/poem/now-we-are-six-by-a-a-milne#ixzz3lkEs6IVU
6. Peters, A. (1983). The units of language acquisition. Cambridge, Cambridge University Press.
7. Swift, J.: Gulliver's Travels. The Project Gutenberg eBook. http://www.gutenberg.org/files/829/829-h/829-h.htm
8. Saffran, J. R., Aslin, R. N., Newport, E.L.: Statistical learning by 8-month-old infants. Science 274(5294) (1996) 1926–8
9. Theakston, A. L., Lieven, E. V., Pine, J. M., Rowland, C. F.: The role of performance limitations in the acquisition of verb-argument structure: an alternative account. J. Child Lang. 28(1) (2001) 127–52