AUTOMATIC DETECTION OF SEMANTIC BOUNDARIES

M. Cettolo and A. Corazza

IRST - Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy

{cettolo,corazza}@itc.it
http://poseidon.itc.it:7117

ABSTRACT

In spoken language systems, the segmentation of utterances into coherent linguistic/semantic units is very useful, as it makes easier processing after the speech recognition phase. In this paper, a methodology for semantic boundary prediction is presented and tested on a corpus of person-to-person dialogues. The approach is based on binary decision trees and uses text context, including broad classes of silent pauses, filled pauses and human noises. Best results give more than 90% precision, almost 80% recall and about 3% false alarms.

1. INTRODUCTION

This work focuses on the automatic segmentation of dialogue turns into homogeneous Semantic Units (SUs) [7]. The approach described below is evaluated in the domain of appointment scheduling, where a system able to deal with this kind of interaction between two persons speaking different languages is being developed [1].

As a working hypothesis, it is assumed that each turn can be represented as a "flat" sequence of concepts, i.e. no nesting or recursion is allowed. Under such an assumption, the problem becomes that of locating in the word sequence the Semantic Boundaries (SBs) that divide consequent SUs, or better, for every pair of adjacent words, to decide if they are separated by a SB or not. If such segmentation could be performed before linguistic processing (parsing), the ambiguity of this latter step would be greatly reduced, together with its computational complexity. On the other hand, in this case only the information given by the acoustic recognizer can be used. In conclusion, the main goal of this work is to evaluate the possibility of an automatic SB detection based only on the recognizer output.

From the recognizer output, information that is mainly related to acoustics or to word or class n-grams can be extracted. Among the acoustic features that can be easily obtained as a by-product of the speech recognition process, the presence of a pause and its length is clearly very important for the task being studied. Moreover, other types of acoustic information can be used, including the presence of spontaneous speech phenomena, such as hesitations, pause fillers, vowel lengthening, speaking rate and energy variations.

On the other hand, the presence of given sequences of words also can be very effective in identifying the presence of an SB. In this case, the number of words that need to be considered before and after the SB candidate has to be determined.

In this paper, a methodology for the semantic boundary prediction based on the binary decision tree algorithm is presented. The classifier is required to hypothesize an SB between each pair of successive words, on the basis of context. This can include silent pauses, quantized in ten classes depending on their length, filled pauses, all mapped into a same class, and human noises, also collapsed into a unique class.

The next section gives an overview of some works dealing with the problem of semantic boundary detection. In Section 3 a few words are spent for introducing binary decision tree algorithm used in the experiments, while Section 4 describes the corpus utilized. In Section 5 the experimental set-up and results are presented. Section 6 discusses about the work done, while Section 7 about that to do.

2. RELATED WORKS

The problem of semantic segmentation is in fact a subproblem of the wider question of extracting from an utterance its *discourse structure*. The literature shows how difficult it is to give a formal definition of discourse structure. One of the best known attempts is probably that proposed in [5], which does not consider any acoustic feature. On the other hand, prosodic information is very useful in detecting conjunctions between successive SUs inside turns. This explains why many papers are devoted to the analysis of the correlation between acoustic and prosodic features and discourse structure. For example, this correlation is analyzed on read texts in [6] where the theoretical framework of [5] is assumed.

A data-driven approach is presented in [9] and applied on unrestricted speech. Instead of deriving a technology from an abstract theory, a comparative analysis of the segmentations performed by several human labelers is shown. One interesting result of this work is that segmentations performed only on the basis of texts are very similar to those obtained by also listening to the audio signal, but

less reliable. In addition, correlation existing between human detected SBs and some prosodic features is evaluated. The results point out a limited but significative correlation between pause length and SB presence; moreover, the pitch is lowered from the beginning towards the end of the utterance, but a reset can be depicted at the beginning of a new segment. Prediction of SBs based on these prosodic parameters is eventually considered and experimentally evaluated.

The previous work is completed by [10], where Filled Pauses (FPs) are especially considered. They are shown to occur at SBs, but in a different way compared to how they occur in the segment interior. In fact, in correspondence of SBs they are usually surrounded by silent pauses.

A rule-based approach is discussed in [12], obtaining good results, while a completely statistical approach is presented in [8], where only lexical information is considered. Interesting results in SB prediction are obtained by using a n-gram LM also including some extra-linguistic phenomena.

Finally, an approach based on multi-layer perceptron is presented in [2], for a similar problem, the prediction of syntactic boundaries. The neural network is trained using prosodic elements and its results are combined with a lexical model, based on trigrams. Comparing the results independently obtained by each model and by the combination of the two, it can be seen that their integration gives the best results, even if the trigram model performance also is good.

3. CLASSIFIER

The main advantage of statistical approaches over rule-based ones is that automatic training algorithms allow the construction of the classifier from problem related data. Therefore, they are easier to use on new domains. The results presented in [8, 2] show how such statistical approaches perform well.

Nevertheless, when a statistical approach is used, it is often very difficult to find out which kind of information is useful in some situation, and then to decide which information source is better. On the contrary, Binary Decision Trees (BDTs) [3, 11] allow different information sources to be put in competition in order to find out which one is more relevant for the task. In fact, a simple analysis of the internal nodes of the tree shows which information is tested on the input and in which order.

The drawback that BDTs share with other supervised learning algorithms is that they need to be trained on labelled data. Data labelling needs to be done manually, and is very expensive. In our case, about 200 dialogues were manually segmented in order to perform the experiments that are described in the following. Another problem with manual segmentation is that it is often error prone. On the other hand, statistical methods that are based on average behaviour, are robust in respect to such (unavoidable) labelling errors.

The BDTs used for the here presented experiments are implemented by following [3], with the only exception being the pruning algorithm, which is that presented in [4]. The set of all possible questions that can be associated to each internal node is represented by *keywords*: every word in the vocabulary can be a keyword, with the constraint that it must appear a minimum number of times in the training corpus. Note that this is not a restriction, because rare words are not significant to the task.

	Training	Test	Whole Corpus
# dialogue	169+12/2	20+12/2	201
# speaker	50	11	61
# male	32	7	39
# female	18	4	22
# turn	2680	406	3086
# segment	3086	462	3548
minutes of speech	240.4	38.2	278.6
|W| (non-noise)	27786	4683	32469
|V| (non-noise)	1291	627	1433

Table 1: Training and test set statistics.

Given an input sentence and a candidate position inside it, the BDT is requested to decide if that position actually corresponds to an SB. Each node in the tree is associated to a question regarding the presence of a keyword around the candidate position.

In our case, the word vocabulary (possible keywords) also includes some kind of extra-linguistic phenomena, among which FPs as suggested by the analysis performed in [10], and in analogy with the approach presented in [8].

Preliminary tests showed that it is better to associate to each keyword its relative position with respect to the candidate SB position: that is, each node question is associated to a pair *(word, position)*, where *word* is a keyword, while *position* is an integer identifying the word position: -1 indicates the word preceding the current candidate position, $+2$ the second word following the candidate position, and so on. Therefore, the question associated to *(word, position)* is: "Is the word *word* in position *position* in respect to the candidate SB position?".

4. DATA

The experiments were performed by using a dialogue corpus collected at IRST, which is composed of 201 monolingual person-to-person Italian conversations for which acoustic signal, word transcriptions and linguistic annotations are available. The two speakers were asked to fix an appointment, observing the restrictions shown on two calendar pages they were given; they did not see each other and could hear the partner only through headphones. The conversations took place in an acoustically isolated room and were naturally uttered by the speakers, without any machine mediation.

The dialogues were transcribed by annotating all extra-linguistic phenomena such as mispronunciations, restarts and human noises, with the exception of pauses. The latter were located and their length evaluated by means of a series of recognition experiments in which only the correct sequence of words was used in the search, but

a pause introduction was admitted between each pair of adjacent words.

SBs were marked by hand in the transcriptions, according to the linguistic labeling of the corpus. All that gives 3548 SUs and, given that turns are 3086, in 462 internal (intra-turn) SBs and 6172 border SBs.

The whole corpus was then divided into training and test sets (see Table 1). The latter consisted of all the sentences uttered by 11 speakers, resulting in 20 complete dialogues and 12 half dialogues, for a total of 406 turns and 462 segments, corresponding to 56 internal SBs and 812 border SBs.

5. EXPERIMENTAL RESULTS

As described in Section 3, questions at each BDT node regard the presence of a keyword around the candidate SB position. Because of the limited size of the training corpus and for efficiency reasons, the range of possible questions has to be bounded. This is equivalent to only considering a window of words of a given size around the candidate. Inside such a window, the algorithm is free to choose the most significant keywords in any position. For practical reasons, a window size going from 4 to 8 was chosen. The words preceding the candidate SB in the window form the *left context*, while the words following it are part of the *right context*.

Moreover, in addition to the word vocabulary, a set of extralinguistic phenomena classes was considered. The first of these classes includes all various kind of filled pause (eh,mmm,ehm,ah...). Another class was considered for inhalations, exhalations and mouth noises, which are all connected with breathing. Silent pauses were divided into 10 different classes depending on their length. This results in a total of 12 classes for extra-linguistic phenomena.

Note that it is possible that a SB can fall inside the context window: in this case it was ignored. In order to be completely coherent, all the text occurring at the exterior of this SB should have been ignored, being part of its context. Nevertheless, in the authors' opinion such incorrectness is statistically irrelevant.

As only a fixed length window is considered, the training set is transformed into a collection of labelled patterns, where each pattern is the set of pairs *(word, position)* falling in the context window; the pattern labels are YES or NO, depending on whether the pattern corresponds to an actual SB or not. The BDT is required to learn from such a collection how to correctly classify this type of patterns.

Note that every position between two different words in the corpus (where extra-linguistic phenomena are considered as words) can be a SB candidate position. Clearly, positions that are *not* SBs are much more frequent than position that *are* SBs: therefore, the number of negative examples is much higher than that of positive examples. Since the BDT training algorithm is affected by this type of biasing, positive samples were repeated in the training set until their number resulted similar to that of negative samples. In the test set, on the contrary, the actual proportion was kept.

Moreover, a special class of SBs needs to be considered: those at the borders of every turn. One working hypothesis was that the first and last word in a segment need be a real word, not an extra-linguistic phenomena. Therefore, the initial SB can be preceded only by extralinguistic phenomena, and the final one can be followed only by extralinguistic phenomena. However, as they are usually close to the border, the window can be partly empty. In these cases, it was filled by *dummy* words (indicated by # in Table 2), even if this was not required by the BDTs, which can easily deal with variable context length. Nevertheless, some preliminary tests showed that this choice did not affect performance.

In Table 2, the patterns associated with the sentence fragment "the first is right for me SB while the second ..." are given.[1] A window of three words is considered around each candidate, with a left context of two items, and a right context of one.

left context		candidate SB	right context	label
(#,-2)	(#,-1)	??	(the,+1)	N
(#,-2)	(the,-1)	??	(first,+1)	N
(the,-2)	(first,-1)	??	(is,+1)	N
(first,-2)	(is,-1)	??	(right,+1)	N
(is,-2)	(right,-1)	??	(for,+1)	N
(right,-2)	(for,-1)	??	(me,+1)	N
(for,-2)	(me,-1)	??	(while,+1)	Y
(me,-2)	(while,-1)	??	(the,+1)	N
(while,-2)	(the,-1)	??	(second,+1)	N

Table 2: *Patterns associated to the sentence "the first is right for me SB while the second ..." together with their label (Y = YES, N = NO).*

A set of experiments was performed on the corpus described in Section 4 for different window sizes. Results are reported in Table 3 by using various metrics, *precision*, *recall* and *false alarms* (FA), defined as follows:

$$\texttt{precision} = \frac{a}{a+c} \times 100$$

$$\texttt{recall} = \frac{a}{a+b} \times 100$$

$$\texttt{FA} = \frac{b}{a+b+c+d} \times 100$$

hypothesis	actual	
	Y	N
Y	a	b
N	c	d

confusion matrix

6. DISCUSSION

Two major problems need to be taken into account when evaluating the experimental results presented in the previous section. First of all, the experiments were performed using the transcriptions of acoustic signals instead of the recognizer outputs. More realistic experiments with the acoustic recognizer will be done in the near future.

[1] Note that this example has been translated from Italian to English for the sake of clarity.

window length	context left	context right	precision (internal SB)	recall	FA
4	3	1	94.0 (67.9)	68.3	5.1
4	2	2	89.7 (67.9)	65.1	5.6
5	4	1	93.7 (66.1)	75.4	3.6
5	3	2	89.1 (69.6)	69.5	4.6
5	2	3	89.1 (67.9)	64.6	5.7
6	5	1	93.7 (62.5)	78.2	3.1
6	4	2	93.9 (66.1)	76.9	3.3
7	5	2	90.0 (62.5)	78.4	2.9
7	4	3	93.5 (64.3)	78.2	3.1
8	5	3	92.7 (55.4)	82.0	2.4

Table 3: Segmentation performance using different context lengths.

A less obvious problem involves the distinction between internal SBs and those at the turn borders (initial and final SBs). In fact, initial and final SBs are automatically defined by the starts and stops of acoustic signals of turns, and therefore their detection is not necessary. On the other hand, it could be advantageous for the prediction of internal SBs to exploit the syntactic and acoustic similarity between the right contexts of initial SBs and of internal SBs, and between the left contexts of final SBs and of internal SBs. To completely use such a large amount of available information, final and initial SBs also were used for training and test.

In term of the three given metrics, the results show a general improvement in the prediction of SBs as the context window size increases. Actually, the trend of the precision is unstable, and even negative when only the internal SBs are considered. On the contrary, the improvement of both recall and FA obtained by enlarging the context and shifting it towards right, is quite stable.

It is difficult to compare the obtained results with those presented in [8], because the experiments were performed on different corpora, in different languages, and further differences can probably be found in the segmentation style. Nevertheless, as the results obtained in the two cases are similar, a comparison of the two approaches can be interesting. In [8], a trigram LM was trained on transcriptions including SBs; SB hypotheses inside a test/input sentence were done on the basis of a Viterbi search in which an SB could be inserted between each pair of successive words. The selection of the Viterbi-best segmented input, assures that SBs are globally hypothesized inside the sentence, and not only on the basis of a local context. Nevertheless, BDT uses items of the (local) context that are more useful for the classification, and the optimal context size can be experimentally established. Moreover, BDTs are suitable to process information of different nature (numerical and symbolic), making the integration of syntactic and prosodic information for SB prediction easier.

7. FUTURE WORK

An important point to be investigated is the performance degradation associated to the introduction of the acoustic recognizer. Therefore, the experiments will be soon repeated using recognizer outputs and the results will be compared with those obtained on signal transcriptions.

Moreover, as many typical (sequences of) extralinguistic phenomena can be observed in the collected data, further investigation will be devoted to design classes of extralinguistic phenomena. In addition, experiments will be performed with a more accurate quantization of pauses, together with new prosodic features, such as pitch and variation of energy and speaking rate. Finally, a problem to be examined is that part of the FAs are spurious because of SBs hypothesized in close positions.

8. REFERENCES

[1] B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, and G. Lazzari. Multilingual Person to Person Communication at IRST. In *Proc. of ICASSP*, Munich, Germany, 1997.

[2] A. Batliner, R. Kompe, A. Kiessling, H. Niemann, and E. Noeth. Syntactic-Prosodic Labeling of Large Spontaneous Speech Data-Bases. In *Proc. of ICSLP*, Philadelphia, USA, 1996.

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth Inc., 1984.

[4] S. Gelfand, C. Ravishankar, and E. Delp. An Iterative Growing and Pruning Algorithm for Classification Tree Design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):163–174, February 1991.

[5] B.J. Grosz and C.L. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[6] J. Hirschberg and B. Grosz. Intonational Features of Local and Global Discourse Structure. In *Proc. of the Speech and Natural Language DARPA Workshop*, pages 441–446, Harriman, NY, 1992.

[7] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J.J. Qantz. Dialog Acts in Verbmobil. Verbmobil Technical Report 65, Hamburg University, DFKI Saarbrucken, Erlangen University and TU Berlin, Germany, April 1995.

[8] A. Stolcke and E. Shriberg. Automatic Linguistic Segmentation of Conversational Speech. In *Proc. of ICSLP*, Philadelphia, USA, 1996.

[9] M. Swerts. Prosodic Features at Discourse Boundaries of Different Strength. *Journal of Acoustical Society of America*, 101(1):514–521, 1997.

[10] M. Swerts, A. Wichmann, and R.-J. Beun. Filled Pauses as Markers of Discourse Structure. In *Proc. of ICSLP*, Philadelphia, USA, 1996.

[11] M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech and Language*, 6(2):175–196, 1996.

[12] S. Wermter and M. Loechel. Learning Dialog Act Processing. Verbmobil Technical Report 139, Hamburg University, Germany, July 1996.