

Identifying Segments for Routing Emergency Response Dialogues

Niels Netten

Human-Computer Studies Laboratory
University of Amsterdam
netten@science.uva.nl

Maarten van Someren

Human-Computer Studies Laboratory
University of Amsterdam
maarten@science.uva.nl

ABSTRACT

In crisis management situations information is exchanged in different ways. In general, information is exchanged through spoken dialogues or text messaging conversations. Part of this exchanged dialogue information is often relevant to other actors involved in managing the crisis. Due to the dynamic character of the situation, dialogue partners may not be aware of who else needs the exchanged information.

We present a coarse-grained segmentation method for automatically recognizing coherent dialogue segments which are then used for routing. We investigate the effectiveness of our features for recognizing boundaries of segments on transcribed emergency response dialogues and we compare classification by relevance of the identified information segments to the ideal topic segments.

Keywords

Dialogue segmentation, intelligent text routing, emergency response dialogue.

INTRODUCTION

Task or situation specific information is a precious asset for any organizational member operating in a fast-changing environment. It is the source of knowledge for dealing with current and future situations. Delivering the right information to the right people at the right time does not only help to improve decision-making. This can also dramatically reduce the associated costs due to false or missed information. In addition, it reduces the information load of an individual actor assigned to route information to others. For those reasons targeted information routing, or information push, will become more and more an important part of many organizational information systems.

Rescue actors involved in crisis situations are faced with managing the flow of information that accumulates during such an event. In such a scenario an information push system could be used to support the optimal sharing of information between the rescue actors. Beside the normal way of communicating and sharing information the information push system sends copies of information to others involved in the crisis for which this also might be useful. The receiver of a message will then decide what to do with it. Ignoring it, directly accepting it, or contacting others for confirmation before continuing his/her activity. In that case negative effects are limited to reading and deleting the message. This kind of automated support will cause rescue actors to be able to work more effectively, see Netten et al. (2006). Information is now able to reach actors who otherwise would not have received this potentially important information or in a much later stadium of the crisis.

During crisis management situations information is communicated via speech (e.g. via telephone or portophone devices) or text message applications (e.g. email or text messaging applications) or other systems, for example icon-based geographical systems (Fitriani et al, 2007). In order to acquire comprehensible information from dialogue conversations is to group the dialogue fragments to a coherent message. Therefore, identifying coherent segments from spoken or textual dialogues is an important task for an information push system.

Recognising relevant information in dialogue involves a combination of segmentation, identifying coherent segments in a dialogue, and classification of these segments as relevant or not for an emergency response actor. Text Classification methods (Sebastiani, 2002) can be used for classifying segments. The classification task is to assign natural language texts to predefined categories based on their content. The problem of segmentation is not straightforward. Dialogues are continuous streams that can involve multiple and changing participants introducing multiple topics. Segmenting spoken or text dialogues is about the same. However, segmenting spoken dialogues requires a translation step from the speech to text. Currently, fairly good transcriptions are achievable by speech recognizers if all speakers that take part are identified beforehand and have individually trained the recognizer.

The information push system needs to find segments that cover one piece of information and that are comprehensible for another emergency response actor. To identify these segments in the dialogues we propose a machine learning method to build a boundary recogniser. The topic of this paper is segmentation of transcribed dialogues into meaningful coherent coarse dialogue segments, which we will call ‘*information segments*’. We focus on linguistic techniques using boundary markers to approach this problem. Finally, we evaluate the effect on classification performance of text segment classification using automatically detected topic segments by the boundary recogniser compared to using the ‘ideal’ topic segments.

SPEAKER	RECIPIENT	UTTERANCE
6131	AC	AC, HERE THE 6131
AC	6131	6131, WE ARE RECEIVING MULTIPLE REPORTS AT THE INCIDENT SCENE THERE ARE ROAD CONSTRUCTIONS MULTIPLE VEHICLES ARE INVOLVED AT THE SCENE OVER
BOUNDARY	→	----- DO YOU TAKE INTO ACCOUNT THE LOCAL THICK FOG
6131	AC	YES, YOU REPORT THAT THERE ARE ROAD CONSTRUCTIONS THERE ARE MULTIPLE REPORTS CONCERNING MULTIPLE VEHICLES AND WEATHER CONDITIONS WITH THICK FOG THAT IS UNDERSTOOD
BOUNDARY	→	----- THE 6131 REPORTS MEDIUM ACCIDENT
AC	6131	YES, 6131 REPORTS MEDIUM ACCIDENT UNDERSTOOD,OVER
BOUNDARY	→	-----

Figure 1. Translated excerpt of a recorded Dutch conversation involving the control room operator (AC) and a fire fighter of vehicle 6131 during a response operation.

RELATED WORK

The problem of automatic segmentation of dialogue has different approaches. The problem is often considered as similar to the problem of text segmentation; the process of dividing written text into words or other similar meaningful units, such as sentences or topics. Therefore, techniques previously developed to segment textual documents have also been adopted and applied to segment transcribed read speech, e.g. broadcast news (Allen, 2001) and spoken dialogue (Ballantine, 2004). Some of these techniques used, segment text based on *lexical-cohesion* models. For example, the TextTiling algorithm (Hearst, 1997) assumes that slightly different vocabulary sets are required for discussing different topics. Local changes in word distributions (i.e. word frequencies) may indicate topic shifts in the text. The semantic network approach of Kozima (1993) determines the similarity of words and groups areas of text into topics. A low calculated similarity score of consecutive sentences indicates a shift in topic. For short dialogues as is the case in emergency response situations these statistical techniques often fail to adequately locate topic changes.

Another approach focuses on discourse markers. Discourse markers, also termed cue words or cue phrases in the computational linguistic and conversational analysis literature, are words and phrases such as for example *now* and *well* which serve primarily to indicate transitions in discourse structure or flow, rather than to impart information about the current topic. The idea is that people use cue words to indicate that they are shifting topic. The problem however is that cue words are often ambiguous and need to be placed into context. However, research shows that certain cue phrases can be relatively accurately determined from transcriptions of speech and be used for segmenting dialogue into topics (Hirschberg and Litman, 1993). More recent research work focuses on a hybrid approach of statistical and linguistic topic segmentation techniques (Arguello et al, 2006). This hybrid approach seems to work well on segmenting spontaneous chat dialogues.

Boufaden et al (2001) approach the transcribed spoken dialogue segmentation problem by constructing a first-order Hidden-Markov Model (HMM) based on multi-knowledge source to build a language model of transcribed search

and rescue dialogue data. They assume availability of perfect transcriptions of the spoken dialogues, i.e. recognition of correct utterances including punctuation. Thereby, transforming the task of segmenting the dialogue stream to the task of sentence classification, where a sentence is a boundary or not. The HMM models the conversational structures based on 5 conversational states (begin- and end of conversation, new topic, continuing topic and end of topic) using several extracted discursive and content features. Identified topic segments are subsequently used to extract relevant information at the level of predefined information templates. An advantage of this approach is that training a HMM model works fast but the drawback is that a lot of training data is required for such a language model to become effective for topic segmentation. This becomes clear since their approach has problems identifying utterances beginning new topics.

An alternative approach to transcribed spoken dialogue segmentation is audio topic segmentation (Hirschberg and Nakatani, 1998). Instead of transcribing dialogues to text the audio is used directly. This approach makes use of information not available in a transcript, such as prosodic and pitch-change cues available in the recorded voice signal. However, our focus is on text or transcribed output by means of an automatic speech-to-text system leaving the audio segmentation approach outside the scope of this paper.

DATA

Our spoken dialogue data were collected from real field exercises that were held as part of standard fire training exercises by multiple Dutch fire departments in the region of Twente in the Netherlands. All communication between responders - the commanders and the control room operators - was recorded from portophone communication. Afterwards, these audio recordings were manually transcribed (SET1 and SET2). Each recording session contained approximately 3000 words. In addition, we use a dialogue corpus (SET3) which was extracted manually from a detailed written report of a fire, the '*Koningskerk disaster*', which includes detailed conversations between participating emergency responders (Scholtens et al, 2004).

Besides the speech we were also able to extract other data from the recordings. We extracted start, end and duration of the uttered information, which communication channel was used, as well as the initial speaker and recipient of the information. At the moment, we only focus on two-party dialogues. Therefore, we only extracted the speaker and explicitly targeted recipient. Automatically detecting a change in speaker or hearer from the audio recordings within this domain will not be difficult since in general identity cues are given during communications. Acquiring meta semantics of utterances (e.g. dialogue act recognition) provides supplementary information about the conversational structure and more fine-grained topic boundaries in discussions (Ivanovic, 2005). For example, if an utterance is a question. Question-answer pairs are often good semantic information segments. Currently, we focus more on a coarse-grained approach. In text messaging punctuation can be easily added and recognized. In case of automatic transcribed speech this would require an additional recognition step. Therefore, currently the '?' has been omitted as part of the data. Figure 1 shows an example dialogue excerpt of our data. The left column gives the speaker, the second column the addressee or recipient and the third an utterance. The manually determined segment boundaries are indicated. Omitted from Figure 1 excerpt are the timestamps of the time between phrases.

APPROACH

Our aim is to acquire '*information segments*' of dialogues by building a boundary recogniser using machine learning methods. These identified segments should represent a coherent comprehensible message. Our approach to the segmentation task is to define a number of features of boundaries and combine these into a boundary recogniser. We will use a decision tree algorithm to construct the recogniser. This recognizer takes a transition between two utterances and decides if at that point there is a segment boundary or not. To be used in a crisis management setting a single off-line training period is necessary. Based on dialogue exchanges someone has to indicate the right boundary locations to train the boundary recognizer. Hereafter, the recognizer automatically identifies boundaries.

Communication in crisis management is characterised by the use of protocolled communication. Specific words are used to signal to the other in the dialogue a shift in discourse or to convey certain status information. For example, the word '*over*' is used to signal the end of a dialogue-turn and for example the phrase, '*at the scene*' is used to indicate that a particular unit arrived at the designated location. This structured communication style may make it easier to detect segment boundaries than in less structured communication.

FEATURES OF BOUNDARY

We choose several boundary features that can be distributed into:

Dialogue-turn: when the turn in a dialogue to speak shifts from speaker.

Speaker-hearer change occurs when a new speaker or hearer compared to previous dialogue utterance is detected. The rationale for this boundary feature is that within the emergency response domain a single topic discussed in general coincides with a conversation (i.e. covers only one single topic).

Elapsed time is another indicator of a boundary location in a dialogue. If a pause between words is more than 5 seconds we assume a meaningful block of information has ended. This aligns with the work of Litman (1993) identifying that topic shifts often occur after a pause of relatively long duration.

Lexical and syntactic cues. Lexical cues are repetitions and acknowledgment words such as ok, yeah, thank you, sure . . . , continuers like *hum, I mean*. Lexical cues in this domain are the words ‘understood’ and ‘clear’ which confirm that the previous communicated information was understood and in many cases signals the end of a conversation. Also, ‘thanks’, ‘moment’, ‘yes’, and the two word collocation ‘further notice’ are lexical cues which mark the end of a message. Syntactic cues are conjunctions (and, or, but . . .), questions marks and temporal adverbs (then, before). We selected the most common lexical and syntactic cues observed from the data.

Dialogue-turn, elapsed time and speaker-hearer change are features which can be generally identified. On the other hand, lexical and part of the syntactic cues are of course more language and domain bound. Also, the availability of certain syntactic cues (e.g. ‘?’) is dependent on speech recognition output.

EXPERIMENTS AND RESULTS

The three data sets (SET1, SET2 and SET3) containing the transcribed emergency response utterances and additional information will be used as input to a segmenter. Our data sets have been annotated manually with topic boundary locations. The automatically found segmentations can then be compared to the boundaries as assigned by the human annotator. The presence of features is extracted from the input stream. To build our boundary recognizer we used the Weka toolkit and choose the J48 decision tree algorithm. SET1 contains 110 boundaries locations, SET2 117 boundary locations and SET3 contains 77 boundary locations. We train boundary classifiers for each set and use cross-validation to evaluate the built classifiers.

We first evaluate the features individually and measure how well they identify boundary locations. Second, we evaluate combinations of features to see which combination best identifies the boundary locations from our data. In the next Section our performance metrics are described.

Performance metric

To measure performance of the boundary recognizer to identify boundary locations we use accuracy, recall, precision and the F₁-measure.

$$\text{Accuracy} = \frac{\text{nr of true boundaries} + \text{nr of true non-boundaries}}{\text{size of set}}$$

Accuracy is the percentage of correct boundary locations and non-boundaries locations identified of the total set.

$$\text{Recall} = \frac{\text{nr of true boundaries}}{\text{nr of true boundaries in the ideal case}}$$

Recall is the percentage of correct boundary locations divided by the number of boundaries that should in the ideal case have been found. To measure how many boundary locations of all are actually identified.

$$\text{Precision} = \frac{\text{nr of true boundaries}}{(\text{nr of true boundaries} + \text{nr of false pred. non-boundaries})}$$

Precision is the percentage of correct boundary locations divided by the amount correct boundaries plus wrongly assigned non-boundary locations.

As a trade-off between the influence of recall and precision on the overall result the F_1 measure is used. F_1 is the weighted average of recall and precision.

$$F_1 = \frac{2 * (\text{recall} * \text{precision})}{(\text{recall} + \text{precision})}$$

In this setting we require high recall as well as high precision since we want to identify all the topic boundaries as well as reduce wrongly assigned boundary locations. Furthermore, we would like to know what is missed or wrongly predicted as boundary location.

Segmentation Experiments

We start out with evaluating how well individual features detect boundaries. Next, we investigate the best combination of features that best identify the boundary locations.

Elapsed Time

	SET 1	SET 2	SET 3	Average
Accuracy	0.81	0.83	0.76	0.80
Recall	0.54	0.65	-	0.40
Precision	0.73	0.81	-	0.51
F_1	0.63	0.72	-	0.45

Table 1. Elapsed time

Results in Table 1 show that the elapsed time feature has a relative high accuracy value for identifying boundaries and non-boundaries. The amount of boundary locations identified is approximately 50 %. Precision is high in both SET1 and SET2 results because of only a few wrongly identified non-boundary locations. We left out the recall, precision and F_1 results of SET3 because the elapsed time feature is unable to identify any correct boundary.

Dialogue Turns

	SET 1	SET 2	SET 3	Average
Accuracy	0.72	0.81	0.76	0.76
Recall	-	0.45	-	0.15
Precision	-	0.93	-	0.31
F_1	-	0.61	-	0.20

Table 2. Dialogue-turn

Table 2 shows the results of dialogue-turn detection. SET1 and SET3 yield no identified true boundary locations. All boundary locations are predicted as non-boundary in those sets. The dialogue-turn feature is more associated with non-boundary locations. Accuracy over all three sets of is relatively high. This is caused by the large percentage of correctly identified non-boundary locations. In SET2 almost 50% of the correct boundary locations are detected. The problem is that the recognizer marks too many locations as a boundary when they actually are not. Precision is high due to low number of wrongly marked non-boundary locations. Dialogue-turn identification groups all consecutive words of one speaker within his/her turn of dialogue as one piece under the assumption they discuss the same topic. In many cases this is a rather coarse and error prone method to find a coherent message since it focuses only on the information from one speaker.

Speaker Hearer change

The introduction of a new speaker or hearer with regard to the previous utterance is a good indicator of the end of a conversation. Many dialogue conversations in this setting discuss only a single topic. Hence, the results in Table 3 show very good performance on accuracy and precision. Boundary locations missed by this recogniser are long pauses in communication. Whereupon, the same speaker-hearer combination start discussing a new topic. Also, the

more subtle topic shifts within a discussion marking a boundary are not recognized. For instance, speaker hearer party end a dialogue but one of them continues with communicating other information.

	SET 1	SET 2	SET 3	Average
Accuracy	0.87	0.83	0.89	0.86
Recall	0.64	0.51	0.68	0.61
Precision	0.85	0.97	0.84	0.89
F ₁	0.73	0.67	0.75	0.72

Table 3. Speaker-hearer change

Lexical and Syntactic Cues

The data contain domain-specific lexical cues such as '*understood*', '*clear*', '*okay*', '*thanks*' or '*further notice*' as well as syntactic cues such as '*yes*', '*moment*', '*then*', '*but*', '*and*' which are words that may mark a shift in topic.

	SET 1	SET 2	SET 3	Average
Accuracy	0.77	0.71	0.79	0.76
Recall	0.18	0.10	0.09	0.12
Precision	1.00	1.00	1.00	1.00
F ₁	0.31	0.19	0.17	0.22

Table 4. Lexical and Syntactic features

The results in Table 4 show that although accuracy is relatively good recall is very low. These lexical and syntactical words not capture many of the boundary locations. The lexical words '*understood*', '*clear*' and '*thanks*' often coincide with a boundary location. However, we miss boundary locations because these words are not always used at a boundary location. Lexical and syntactic cues can also cause many mistakes because of word ambiguity. For example, the word *moment* is used as an temporarily end (i.e. pause) of a discussion or means mean a time indication (e.g. *in a few moments we arrive*). Additional context information would be required to solve this problem. Furthermore, inconsequent use of words that denotes flow of conversation makes it difficult to recognize those boundaries.

Feature combinations

Now, we combine previous features to obtain the optimal boundary recognizer. After the individual evaluation of the features of boundary the speaker hearer change detection boundary provided very good results. We use this feature and supplement other features to see if we can improve boundary recognition. We evaluate the following combinations:

- Speaker-hearer change and Dialogue-turn (**SphD**)
- Speaker-hearer change, Dialogue-turn and Elapsed time (**SphDE**)
- Speaker-hearer change, Lexical and Syntactic features (**ShpLS**)
- Speaker-hearer change, Dialogue-turn , Elapsed time and Lexical and Syntactic features (**SphDELS**)

	SET 1	SET 2	SET 3	Average
Accuracy	0.87	0.92	0.89	0.89
Recall	0.64	0.80	0.68	0.71
Precision	0.85	0.94	0.84	0.88
F ₁	0.73	0.87	0.75	0.78

Table 5. SphD features

Table 5 shows the performance results of the decision tree built using the speaker hearer change detection feature as well as dialogue-turn feature. We only observe an increase in performance in SET2. The reason for this in SET2 is that certain overlooked boundary locations by speaker-hearer change feature now become identified by dialogue-turn detection. Therefore, the number of identified boundaries increases and the number false assigned non-boundary locations decreases. In SET 1 and SET3 the dialogue turn feature makes no contribution at all. It does not provide any additional information for identifying the correct boundary locations due to coinciding with the boundary identification of the more informative speaker hearer change boundary.

	SET 1	SET 2	SET 3	Average
Accuracy	0.87	0.92	0.84	0.88
Recall	0.75	0.80	0.68	0.74
Precision	0.79	0.94	0.84	0.86
F ₁	0.77	0.87	0.75	0.80

Table 6. SphDE features

The results in Table 6 show the performance results using the three domain generic boundary features. We observe an increase in performance in SET1. In this set previously overlooked boundary locations are now identified by the elapsed time feature increasing recall, precision and F₁. In SET 2 and SET3 the elapsed time feature makes no contribution due to coinciding with boundaries already identified by speaker-hearer changes and dialogue-turn.

	SET 1	SET 2	SET 3	Average
Accuracy	0.88	0.83	0.90	0.87
Recall	0.67	0.51	0.71	0.63
Precision	0.85	0.97	0.85	0.89
F ₁	0.75	0.67	0.78	0.73

Table 7. SphLS features

In Table 7 we show performance results of the decision tree built using the speaker hearer change feature as well as lexical and syntactic features. Here, we observe a slight increase in performance in SET1 and SET3. In those sets the words ‘clear’ and ‘understood’ identify previously missed boundary locations. These are locations within a dialogue between the same speaker and hearer that continue the dialogue. In SET2 the results are the same as with speaker-hearer change detection individually. We also notice that the syntactic cues have no contribution in identifying boundary locations in our data.

Finally, we construct a boundary recogniser using all three sets and our boundary features to build the model from. Figure 2 shows the resulting decision tree. Each leaf node tells how many instances in the training set are correctly classified by this node and the number of instances incorrectly classified by the node. For example, when speaker-hearer change is detected 208 of the training instances are correctly classified and 25 not. Again, syntactic cues (*yes, and, then, but*) not present in the tree do not contribute to boundary identification. The most informative feature to determine a boundary location, located at the root node of the tree, is again the speaker-hearer change. If detected than we identify a boundary location. If not than we check if the lexical cue ‘understood’ is present in the content phrase to see if we are at a boundary location. If still not the case we check on elapsed time (> 5 seconds). If there is a time difference larger than 5 seconds we check on a dialogue turn presence and on the lexical cues presence to determine boundary location or not. The boundaries still missed are ones were we detect elapsed time difference a dialogue turn and the lexical cue ‘clear’. The recognizer marks the transition as a non-boundary while it actually is a boundary. Mistakes in marking non-boundaries as boundary locations are still made by the recognizer when a response of the hearer in our setting takes larger than 5 seconds. Actually there is no boundary at that transition but because of our elapsed time constraint of 5 seconds the recognizer falsely decides there is.

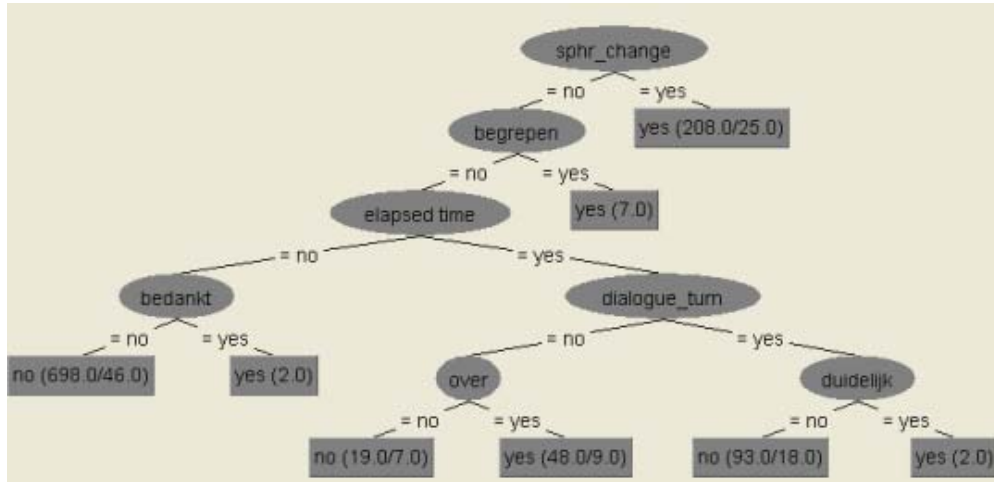


Figure 2. Decision tree

In Table 8 we compare the average results of all combinations. The results of optimal decision tree are close to those of speaker-hearer change, elapsed time and dialogue turn (**SpDE**). Thus, using only the domain generic features of boundary will identify most of the boundary locations. In this setting we do see that the lexical cues improve the results slightly. Still most of them coincide with speaker-hearer change detection.

	SpD	SpDE	SphLS	SphDELS
Accuracy	0.60	0.88	0.87	0.90
Recall	0.71	0.74	0.63	0.75
Precision	0.88	0.86	0.89	0.87
F ₁	0.78	0.80	0.73	0.80

Table 8. Average results comparison

To improve the current results any further more domain specific lexical cues could be used. Furthermore, we could stretch the time constraint a little further than 5 seconds to reduce false predicted boundary locations where the response between speaker and hearer takes longer than 5 seconds.

Segment Classification

Now, we investigate if routing performance deteriorates when using identified segments by comparing the most 'ideal' topic segments (i.e. 'gold standard') of the transcribed spoken dialogues versus the identified topic segments found with our boundary recogniser. First, we use the initial SET1 and SET3 which are different emergency response situations and manually segment them into the most ideal topic segments; 101 segments for SET1 and 66 for SET3. SET1 initially having 2790 words over 394 dialogue turns, an average of ± 7 words. SET3 consists of 2049 words and 266 dialogue turns, an average of ± 8 words. After applying the boundary recogniser we obtain 155 information segments for SET1 and 84 for SET3. Second, we manually label the ideal segments and the identified segments datasets with the target class label(s). Labels indicate for which emergency response actor the information segments are relevant. An information segment can be relevant for multiple emergency response actors. Therefore, the learning task is a multi-label classification problem, i.e. instead of only one label a subset of labels can be assigned to the same information segment (Mitchell, 1997). Multinomial Bayes distribution has been chosen to deal with a relatively large dictionary size (number of dialogue content word features). Furthermore, we know that in segments word repetitions occur often. Using a multinomial distribution we are able to determine the class of certain information not only by the presence of a certain word also by number of times that word occurs. In general this performs better than the ordinary approach (Witten and Frank, 2005).

Class labels that only occur a few times in the segmented data were left out of the classification results. Below we list 6 target classes of the classifier with enough examples (number of segments) to be used for building the classifier.

- Alarm Center (AC)
- Alarm Center Police (AC-P)
- Fire Officer on Duty (OvD)
- Fire Truck team: 741, 742, 743, 6481, 6332, 6131, 6232

Since the actors involved in both sets are partially different we have to learn and evaluate classification results separately. The labels of SET1 are AC, OvD, 6481, 6332, 6131 and 6232. For SET3 they are AC, OvD, 741, 742, 743 and AC-P.

Classification Results

Here we present the classification results. We used Weka's Multinomial Naïve Bayes algorithm to construct the classifier. Ten-fold cross-validation was used to evaluate the classifiers. In Table 10 the average classification results of the ideal dialogue segments are given. The results on recall and precision are low. There are two reasons for that. First, the segmented datasets are small which influences the training of the classifier. Second we classify and evaluate on content only. Using additional actor context features like task or location information will provide better results (Netten et al, 2006).

	SET 1	SET 3
Accuracy	0.75	0.78
Recall	0.30	0.58
Precision	0.23	0.38
F ₁	0.26	0.46

Table 10. Gold standard segment classification

	SET 1	SET 3
Accuracy	0.82	0.79
Recall	0.30	0.52
Precision	0.23	0.35
F ₁	0.26	0.41

Table 11. Identified segment classification

Table 11 shows the classification results of the identified segments. The evaluation scores of both classifications do not differ that much. This means that identified information segments from dialogue do not deteriorate information classification performance. The classification results of the identified segments are comparable to that of classifying the ideal segments.

CONCLUSION

The paper addressed the problem of identifying coherent information segments from emergency response dialogues. This is an important part of information push system to extract and route relevant dialogue information to others involved for which it also might be useful. A decision tree algorithm was used to construct a boundary recogniser that takes transitions between utterances and based on the presence of certain boundary markers decides if there is a boundary or not.

Transcribed audio recordings from real exercise emergency response situations have been used to build and evaluate the boundary recogniser. The coarse recogniser segments our emergency response dialogues with a high accuracy of 90% and F₁ score of 0.80. Dialogue-turn, speaker-hearer change detection and elapsed time markers identify most of the boundary locations. The most informative feature for this type of data is speaker-hearer change. Finally, we compared performance of classification by actor relevance of the automatically identified segments with that of the most ideal segments of the dialogue data. Results showed that classification of identified segments does not deteriorate much and is comparable to that of the ideal segments.

FURTHER WORK

For the experiments we used hand-made transcriptions of audio recordings. These transcriptions contain no errors to what has actually been spoken. However, automatically transcribing speech using automatic speech recognition (ASR) would not directly provide such nice transcripts or any structure. An interesting next step would be to investigate how well ASR would work in a crisis setting and if the ASR output deteriorates segment classification performance.

Proceeding to a more fine-grained approach would require meta semantics recognition of utterances which would provide extra information about the conversational structure. Especially, identifying and grouping of question-answer pairs utterances from the data would identify shifts within a conversation.

ACKNOWLEDGMENTS

The authors like to thank the fire departments of the region Enschede (The Netherlands) for their hospitality for letting us record during their emergency response exercises. This research is funded by SenterNovem under project nr: MMI04006, by the Dutch Ministry of Economical Affairs. The content of this article does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

1. Allan, J., G. Doddington, J. C., Yamron, J. and Yang, Y. (2001). Topic detection and tracking pilot study. *Topic Detection and Tracking Workshop Report*.
2. Arguello, J. and Rose, C. (2006). Topic segmentation of dialogue. *Proceedings of Workshop on Analyzing Conversations in Text and Speech*, 42-50.
3. Ballantine, J. (2004) Topic Segmentation in Spoken Dialogue. Bachelor Thesis, Department of Computing, Division of ICS, Macquarie University, Sydney Australia.
4. Boufaden, N., Lapalme, G., and Bengio, Y. (2001). Topic segmentation: A first stage to dialogue-based information extraction. *Natural Language Processing Rim Symposium*, 273-280. Newark, New Jersey: Morgan Kaufmann.
5. Fitrianie, S., Dacu, D. and Rothkrantz, L. J. M. (2007). Human communication based on icons in crisis environments. *HCI 11*, 57-66. Springer.
6. Hearst, M. A. (1993). TextTiling: A quantitative approach to discourse segmentation. Technical Report Sequoia 93/24, Computer Science Division, University of California, Berkeley.
7. Hirschberg, J. and Litman, D. J. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19, 501-530
8. Hirschberg, J. and Nakatani, C. H. (1998) Acoustic indicators of topic segmentation, In *ICSLP-1998*.
9. Ivanovic, E. (2005). Automatic Utterance Segmentation in Instant Messaging Dialogue. *Proceedings of the Australasian Language Technology Workshop*, Sydney Australia, 241-249
10. Kozima, H. (1993) Text Segmentation Based on Similarity between Words, *Proceedings of the ACL*
11. Mitchell, T. (1997). *Machine learning*. New York: McGraw Hill.
12. Netten, N., Bruinsma G., Someren, M. van, and Hoog, R., de (2006). Task-Adaptive Information Distribution for Dynamic Collaborative Response, *Special Issue on Emergency Management Systems of the International Journal of Intelligent Control and Systems (IJICS)*. Vol. 11, No. 4 December 2006, 237-246
13. Scholtens, A. and Drent, P. (2004). Brand in de Koningkerk te Haarlem (Technical Report). Inspectie Openbare Orde en Veiligheid.
14. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1-47.
15. Witten I. A. and Frank E. (2005) *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.