Identifying Relevant Phrases to Summarize Decisions in Spoken Meetings*

Raquel Fernández[1], *Matthew Frampton*[1], *John Dowding*[2], *Anish Adukuzhiyil*[1],
Patrick Ehlen[1], *Stanley Peters*[1]

[1]Center for the Study of Language and Information
Stanford University, California, USA
{raquel,frampton,ajohna,ehlen,peters}@stanford.edu

[2]University of California/Santa Cruz
Santa Cruz, California, USA
jdowding@ucsc.edu

Abstract

We address the problem of identifying words and phrases that accurately capture, or contribute to, the semantic gist of decisions made in multi-party human-human meetings. We first describe our approach to modelling decision discussions in spoken meetings and then compare two approaches to extracting information from these discussions. The first one uses an open-domain semantic parser that identifies candidate phrases for decision summaries and then employs machine learning techniques to select from those candidate phrases. The second one uses categorical and sequential classifiers that exploit simple syntactic and semantic features to identify words and phrases relevant for decision summarization.

Index Terms: phrase extraction, human-human meetings, decision detection and summarization

1. Introduction

Human-human meetings are routine in professional and academic environments, and the demand for automatic methods that process, understand and summarize information encoded in audio and video recordings of meetings is growing rapidly, as evidenced by on-going projects which are focused on this goal [1, 2]. Our research is part of a general effort to develop a system that automatically extracts and analyzes the information content of meetings in various ways, including automatic transcription, targeted browsing, topic detection and segmentation, and action item and decision identification [3].

In our current research, one of our main concerns is the automatic extraction of decision discussions from multi-party meetings. We tackle this problem in two stages. The first stage involves *detecting* the dialogue regions and the dialogue acts within those regions that contain important decision-related information. In recent work [4], we have addressed this problem by applying a simple notion of dialogue structure that takes account of the roles that different utterances play in the decision-making process, (more on this in Section 3.2). The second stage is to then zoom into the decision-related dialogue acts and identify words and phrases that can be used to produce concise, descriptive *summaries* of the decisions.

In this paper we concentrate on the second stage, and investigate two different summarization approaches. The first approach uses an open-domain semantic parser to parse decision-related dialogue acts. This produces multiple candidate phrases and a Support Vector Machine (SVM) is then used to select the phrase which is most likely to appear in a decision discussion summary (according to manual, gold-standard annotations). We compare this to a second approach where non-sequential (SVM) and sequential (Hidden Markov Model (HMM)) classifiers are trained to extract relevant words from decision-related dialogue acts using a variety of features that exploit syntactic, semantic and dialogue information.

The paper proceeds as follows. In the next section, we briefly describe previous research on dialogue summarization, and in Section 3, give details about the data used for our experiments, including the corpus and our set of annotations. In Section 4, we then present an experiment with the summarization approach based on semantic parsing. The alternative word-level based approach is reported on in Section 5. Finally, in Section 6, we conclude and outline some directions for future work.

2. Dialogue Summarization

Although previous work on automatic summarization mostly deals with written text, (see [5] for an overview), in recent years, there has been a growing interest in summarization of spoken dialogue [6, 7, 8, 9]. Besides dealing with aspects common to text summarization (e.g. topic segmentation, determination of the salient information, anaphora resolution), dialogue summarization poses new challenges including the detection and removal of speech disfluencies and the detection and linking of cross-speaker information units such as question-answer pairs. While most approaches use manual transcriptions, recent attempts (e.g. [8, 10]) use ASR output as well. There is a wide range of methods employed, with many of them relying on notions of salience and semantic similarity, such as MMR (Maximum-Marginal Relevance), LSA (Latent Semantic Analysis) or term-weighting methods borrowed from work on information extraction (see [7, 9] for overviews of these methods).

While the vast majority of the work cited above deals with the summarization of full dialogues, Purver et al. [10] attempted to generate targeted summaries of specific dialogue events, in this case *action items*—a particular kind of decision where one or more individuals commit to undertake a task. They identified utterances that contain action item-related information, such as description of the task, its timeframe, or the person(s) responsible, and then focused on summarizing the *task descriptions* and *timeframes*. Their approach involved first parsing the Word Confusion Network (WCN) for each relevant utterance using a general rule-based parser [11], which produced multiple short

*This work was supported by the CALO project (DARPA grant NBCH-D-03-0010).

fragments rather than one full utterance parse. An SVM classifier was then trained to learn a model which ranked these phrases according to their likelihood of appearing in a gold-standard extractive action item summary. Various features were used including lexical and temporal expression tags, as well as properties of the WCN paths and of the parsed phrases. For *timeframes*, the results were generally higher than a baseline that took the entire 1-best utterance transcription, while for *task descriptions*, precision was higher, but the F-score lower. In Section 4 we investigate how well this approach performs when applied to more general decisions.

3. Data

3.1. Corpus

For the experiments reported in this study, we use 17 meetings from the AMI Meeting Corpus [12], a freely available corpus of multi-party meetings containing both audio recordings and manual transcriptions. Each meeting lasts around 30 minutes, and is scenario-driven with four participants playing different roles in a company's design team. The overall sub-corpus of 17 meetings makes up a total of 15,680 utterances/dialogue acts—approximately 920 per meeting.

3.2. Modelling Decision Discussions

As noted earlier, our recent work [4] has been concerned with detecting regions of dialogue where decisions are made. For this we have taken an approach that models the structure of decision discussions as consisting of four main components: (a) initially, a topic/issue is raised for which a decision is required; (b) one or more proposals/possible resolutions are then considered; (c) once agreement is reached upon a particular possible resolution, this becomes the decision; (d) optionally, this resolution/decision is summed up or restated.

In line with these observations, we designed an annotation scheme that distinguishes between three main decision dialogue acts (DDAs): *Issue* (*I*), *Resolution* (*R*) and *Agreement* (*A*), with class *R* being further subdivided between *Resolution Proposal* (*RP*) and *Resolution Restatement* (*RR*). Utterances in the transcriptions of our sub-corpus were annotated with these DDA classes. We then took a hierarchical classification approach in which SVMs hypothesized occurrences of each of the DDAs, and then based on these hypotheses, a further SVM decided which regions of dialogue were decision discussions. This approach proved to be advantageous over a flat classification approach where the different roles that utterances play in the decision-making process are not taken into account [13]. More details on our approach and on how it compares to [13] can be found in [4].

We hypothesize that the advantage of our 'structure-based' approach is two-fold—not only does it help in detecting regions of dialogue which are decision discussions, but also, by identifying important constituents within these decision discussions, it opens the way to better-targeted summaries.

3.3. Experimental Data for Decision Summarization

Although agreements help in detecting decision discussions, the semantic gist of a decision is expressed in more contentful DDA classes such as *Issue* and *Resolution*. Hence, for summarization purposes we only consider utterances annotated with classes *I* and *R*. Note that an utterance can be tagged with more than one DDA class and that a decision discussion can contain more than one utterance tagged with each of the DDA classes. The sparseness of DDAs is a principal reason for why their detection is difficult. In our 17-meeting sub-corpus there are a total of 118 utterances tagged as *Issue* (0.9%), and 179 tagged as *Resolution* (1.4%). The average length in words of *Issues* and *Resolutions* is 12.2 and 11.9 respectively, and so this gives a total of 1,440 words for *Issues* and 2,131 words for *Resolutions*.

In order to provide a gold-standard for training classifiers and evaluating their performance, phrases from *I* and *R* utterances were manually annotated as summary-worthy. The aim was to select those phrases in the manual utterance transcriptions that should appear in an extractive summary, or that could be the basis of a generated abstractive summary. In practice, this meant selecting the phrase(s) which describe the issue/resolution as succinctly as possible—hence this does not include phrases which express the speaker's attitude towards the issue/resolution, nor, clearly, phrases which contain any other information that is not directly relevant. (1) shows an example of an utterance tagged as *Issue* and another tagged as *Resolution* within the same decision discussion. The phrases that were selected as summary-worthy are indicated in square brackets.

(1) A:(*I*) So we we're looking at [*sliders for both volume and channel change*]
B:(*R*) I was thinking kind of [*just for the volume*]

3.4. Baseline

The utterances singled out to express DDAs contain a great deal of material that can be worth extracting for a decision summary. Indeed, for *Issues*, on average 49% of words in the utterance transcription corresponds to gold-standard summary-worthy words. For *Resolutions* the overlap is even more pronounced: on average 59% of words in the utterance transcription are considered summary-worthy. This indicates that taking the entire utterances to create an extractive summary could, to some extent, be already useful (indeed this seems to be the approach taken in [13]). Our aim is however to provide more concise and targeted summaries by extracting the most relevant information contained within decision-related utterances.

We will evaluate our results against a baseline system that always takes the full utterance transcription (as in [10]). Recall is computed as the proportion of the gold-standard phrase covered by the full utterance, and precision as the proportion of the full utterance that overlaps with the gold-standard phrase. Given the percentages mentioned above, this will be a challenging standard to aim for. As the baseline obviously yields a recall of a 100%, we will be especially concerned with improving precision, in order to create a more targeted basis for an extractive decision summary.

4. Parse-based Summarization

Here we follow the summarization approach of Purver et al. [10], but use manual transcriptions, (provided by AMI), as opposed to ASR transcriptions. We first parse the relevant utterances using the Gemini parser [11] to produce multiple short fragments, and then train an SVM classifier using various features to select among these phrases.

4.1. Open-Domain Semantic Parser

Since we expect that human-human conversational dialogue, after being processed by an imperfect recognizer, will be highly ungrammatical, we focused on developing a semantic parser that only attempts to find basic predicate-argument structures of the major phrase types (S, VP, NP, and PP) but has access to a broad-coverage lexicon. Our approach to building a broad-coverage lexicon has been to make use of publicly available lexical resources for English, including COMLEX, VerbNet, WordNet, and NOMLEX.

COMLEX [14] provides detailed syntactic information for the 40K most common words of English. VerbNet [15] provides detailed semantic information for verbs, including verb class, verb frames, thematic roles, mappings of syntactic position to thematic roles, and selection restrictions on thematic role fillers. From WordNet [16] we extract another 15K nouns, and the semantic class information for all nouns. These semantic classes are hand-aligned to the selectional classes used in VerbNet, based on the upper ontology of EuroWordNet [17]. NOMLEX [18] provides syntactic information for event nominalizations, and information for mapping the noun arguments to the corresponding verb syntactic positions.

These resources are combined and converted to the Prolog-based format used in the Gemini framework [11], which includes a fast bottom-up robust parser in which syntactic and semantic information is applied interleaved. Gemini can compute parse probabilities on the context-free skeleton of the grammar. In the experiments described here these parse probabilities are trained on Switchboard tree-bank data.

4.2. Experiments & Results

Since our eventual goal is to parse speech recognition output, transcriptions are modified to remove text-specific characteristics, such as punctuation and capitalization, and cleaned up by removing disfluency and filled pause markers. The cleaned-up transcriptions are then converted to WCN format.

For each phrase returned by the parser we extract several types of features: properties of the raw WCN paths, properties of the parsed phrases including semantic class features, and lexical features reflecting the identity of the main verb and head noun—a list is given in Table 1. As lexical features are likely to be more domain-specific, and increase the size of the feature space dramatically, we prefer to avoid them if possible.

WCN	phrase length (WCN arcs) start/end point (absolute & percentage)
Parse	parse probability phrase type (S/VP/NP/PP)
Semantic	main verb VerbNet class head noun WordNet synset noun class of *agent* thematic role (if any)
Lexical	main verb, head noun

Table 1: Features for parse fragment ranking

We then trained *SVMlight* [19] on these features to rank the phrases obtained for each utterance according to their probability of matching the gold-standard summary. The phrase ranked highest is then selected as the automatically-generated summary. To evaluate performance, we use the same evaluation metric as Purver et al. [10]: Recall corresponds to the total proportion of the gold-standard extractive summary covered by the selected phrase; precision, to the total proportion of the chosen phrase which overlaps with the gold-standard summary. As discussed in Section 3.4, the baseline corresponds to using the entire transcription. We also compare to an oracle that always chooses a phrase with the highest F-score. Results obtained using 10-fold cross-validation are given in Table 2.

None of the feature sets that we experimented with were able to outperform the baseline's F-score, as higher precision failed to fully compensate for the baseline's perfect recall. Using semantic features in addition to syntactic features was found to improve performance. The comparatively higher recall and precision of the Oracle suggests that high quality phrases are available in the output of the parser, and that further investigation of phrase selection is warranted.

	Issue			*Resolution*		
	Re	Pr	F1	Re	Pr	F1
Baseline	1.0	.49	.66	1.0	.59	.74
Oracle	.76	.96	.84	.73	.98	.84
WCN + parse	.53	.64	.57	.57	.71	.63
+ semantic	.56	.67	.60	.59	.73	.65
+ lexical	.55	.68	.60	.58	.78	.67

Table 2: Parse-based results for *I* & *R* Utterances

5. Word-level Summary Identification

We now turn to our second experiment, which uses a different methodology based on extracting summary-worthy information at the word level.

5.1. Methodology

For both *I* and *R* utterances, we trained two different types of classifier for distinguishing summary-worthy words from non-summary-worthy words: the first was an SVM, (produced using *SVMlight*), and the second an HMM, (produced using *SVMhmm* [20]). SVMhmm trains models that are isomorphic to HMMs. After trial classifier experiments, the labelling scheme that we settled on for the HMM distinguishes between the word at the beginning of a sequence of summary-worthy words (labeled B), all other words inside the sequence (I), and words outside of the sequence (O).

For each utterance transcription, we extracted a small set of simple features: Lexical features reflecting the identity of the words themselves as well as the immediately preceding and following words; POS tags as generated by the Stanford POS tagger [21]; and a semantic similarity feature (Sim) at the level of the decision discussion. This feature records occurrences of the same word in other utterances tagged as DDAs within the same decision discussion. This applies only to words that have been tagged as either nouns or adjectives by the POS tagger, since these are the word classes that are more prominent in our data. For instance, for a given noun or adjective in an *I* utterance, this feature records whether it also appears in the *Resolution Proposal* and/or the *Resolution Restatement*. Distinguishing between *RP* and *RR* yielded slightly better results than taking the common class *R*.

5.2. Experiments & Results

All experiments were performed using 10-fold cross-validation and were evaluated using the same evaluation metric as in the previous experiment (see Section 4.2). We again took the entire utterance transcription as our baseline. Note that when evaluating the HMM's classifications, labels B and I were collapsed into a single class. Table 3 shows the results obtained using different sets of features.

Using the SVM classifier, we are able to improve slightly

	Issue			*Resolution*		
	Re	Pr	F1	Re	Pr	F1
Baseline	1.00	.49	.66	1.00	.59	.74
SVM lexical	.78	.58	.67	.86	.67	.75
+ POS	.80	.61	.68	.86	.68	.76
+ Sim	.77	.60	.67	.85	.68	.75
HMM lexical	.54	.63	.58	.76	.75	.76
+ POS	.56	.65	.61	.80	.75	.77
+ Sim	.56	.71	.63	.81	.75	.78

Table 3: Results for word-level summary identification

over the baseline's F-score. Precision is considerably higher than the baseline, although note that it is not as high as that achieved by the parse-based approach (see Table 2). Recall, however, is over 20 points higher with this approach. In general, the HMM classifier yields lower recall than the SVM, and better precision (but not as good as the parse-based approach). For *Issues*, the drop in recall results in F-scores that are lower than the baseline. For *Resolutions* the highest F-scores are obtained with the sequential model, which in this case produces relatively high scores for both recall and precision.

Regarding the contribution of different features, POS tags give small improvements whichever classifier is used, while the semantic similarity feature only seems to improve the performance of the sequential model, in particular, by boosting precision for *Issues*.

6. Conclusions & Future Work

We have investigated two different approaches for extractive summarization of decisions in human-human meetings—a parse-based approach and a word-based approach. While the parse-based approach yields higher precision, the word-based approach gives better recall resulting in higher F-scores. The high scores for the parse-based Oracle indicate that the parse-based approach has the potential to yield very good results, and so motivates looking at how the performance of the classifier can be improved in selecting the best parses for summarization. We believe that the present results are promising and that both approaches warrant further investigation.

One of the first items on our research agenda is to investigate methods that assume a stronger coupling between the issue and resolution of a decision discussion, for example, by training classifiers which consider pairs of possible issue and resolution summaries. In this effort, we intend to use tools like WordNet which can provide us with more complex semantic features, and also, to explore techniques used in Question-Answering.

In the present study, we have used semi-automatic evaluation methods based on manual annotations. Our ultimate goal is to evaluate decision summaries in the context of our meeting browser [3]. Immediate future work towards this end will be to evaluate our summaries on the basis of human judgements. Modelling decisions as consisting of issues and resolutions can be particularly helpful here, since humans can be asked to judge the extent to which the resolution in the extractive summary resolves the corresponding issue. We also plan to investigate how the use of WCNs from ASR output affects summary quality.

7. References

[1] A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, R. Stiefelhagen, and J. Yang, "SMaRT: The smart meeting room task at ISL," in *ICASSP*, 2003.

[2] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Marcías-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*, 2004.

[3] L. Voss, P. Ehlen, and the DARPA CALO MA Project Team, "The CALO Meeting Assistant," in *Proceedings of NAACL-HLT*, Rochester, NY, USA, 2007.

[4] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue," in *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008.

[5] I. Mani and M. T. Maybury, Eds., *Advances in Automatic Text Summarization*. MIT Press, 1999.

[6] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, 2002.

[7] I. Gurevych and M. Strube, "Semantic similarity applied to spoken dialogue summarization," in *Proc. of the 20th Annual Meeting of the ACL*, 2004.

[8] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. of INTERSPEECH*, 2005.

[9] G. Murray and S. Renals, "Term-weighting for summarization of multi-party spoken dialogues," in *Proc. of MLMI*, 2007.

[10] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multiparty dialogue," in *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[11] J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "GEMINI: a natural language system for spoken-language understanding," in *Proc. ACL*, 1993.

[12] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus," in *Proc. of Measuring Behavior, the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[13] P.-Y. Hsueh and J. Moore, "Automatic decision detection in meeting speech," in *Proc. of MLMI*, ser. Lecture Notes in Computer Science. Springer-Verlag, 2007.

[14] R. Grishman, C. Macleod, and A. Meyers, "COMLEX syntax: Building a computational lexicon," in *Proc. of the 15th Int'l Conference on Computational Linguistics (COLING)*, 1994.

[15] K. Kipper, H. T. Dang, and M. Palmer, "Class-based construction of a verb lexicon," in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, Austin, Texas, July 2000.

[16] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[17] P. Vossen, "EuroWordNet: a multilingual database for information retrieval," in *Proc. of the DELOS Workshop on Cross-language Information Retrieval*, 1997.

[18] C. Macleod, R. Grishman, A. Meyers, L. Barrett, and R. Reeves, "NOMLEX: A lexicon of nominalizations," in *Proceedings of EURALEX*, Liege, Belgium, 1998.

[19] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods – Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999.

[20] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov support vector machines," in *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 2003.

[21] K. Toutanova and C. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proc. of EMNLP/VLC*, 2000.