Emotion and Behaviour in Automatic Dialogue Summarisation

Norton Trevisan Roman
Institute of Computing
University of Campinas
Campinas, Brazil
nortontr@gmail.com

Paul Piwek
Centre for Research in Computing
The Open University
Milton Keynes, UK
p.piwek@open.ac.uk

Ariadne Maria Brito Rizzoni Carvalho
Institute of Computing
University of Campinas
Campinas, Brazil
ariadne@ic.unicamp.br

ABSTRACT

This paper presents an overview of a six-year research project on automatic summarisation of emotional and behavioural features in dialogues. It starts by describing some evidence for the hypothesis that whenever a dialogue features very impolite behaviour, this behaviour will tend to be described in the dialogue's summary, with a bias influenced by the summariser's viewpoint. It also describes the role some experiments played in providing useful information on when and how assessments of emotion and behaviour should be added to a dialogue summary, along with the necessary steps (such as the development of a multi-dimensional annotation scheme) to use these experimental results as a starting point for the automatic production of summaries. Finally, it introduces an automatic dialogue summariser capable of combining technical and emotional or behavioural information in its output summaries.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Human Factors, Languages

1. INTRODUCTION

Although emotions are increasingly drawing the attention of much research on the designing of computer interfaces, they seem not to have raised the same interest in the field of automatic dialogue summarisation, where producing summaries that completely ignore this human facet is the rule rather than the exception (*e.g.* [6, 2, 18, 12, 5]). In this paper, we both make the case for the importance of emotional information in dialogue summarisation and work out how such information can actually be automatically incorporated in dialogue summaries.

As a starting point, consider the following dialogue[1], in which a buyer interacts with a vendor, in a car sale scenario:

Vendor:	Hey you there! I'm Ritchie.
Client:	Can you tell me something about that silver car?
Vendor:	That silver car is not terribly cheap. It costs 29,000 Euros.
Client:	Does it have power windows?
Vendor:	Don't ask me?
Client:	No problem. Does it have leather seats?
Vendor:	Silly question! Of course!
Client:	Great! What kind of interior does it have?
Vendor:	It has a cramped interior.
Client:	Interesting. How fast does it go?
Vendor:	It goes up to 133 miles per hour.
Client:	Fabulous! How much horsepower does it have?
Vendor:	It has 165 horse power.
Client:	Fabulous! Thank you for your help. I have to think a bit more about this.
Vendor:	I should have guessed ! Well thanks for wasting my time.

In this dialogue, it is practically impossible not to notice the vendor's extreme rudeness when addressing the client. Moreover, when this dialogue is summarized it seems intuitive that the improper behaviour of the vendor should be mentioned somehow. Or maybe not? This is one of the questions that have so far not been answered in the literature: is it really important (as judged by humans) to report, in a dialogue summary, behavioural or emotional features of the dialogue? More specifically, should polite or impolite behaviour be mentioned?

As far as automatic text summarisation is concerned, the few systems that do account for emotional features and politeness (*e.g.* [16, 1, 4]) refrain from answering these questions, apparently basing all their decisions on the intuition of the researchers rather than on empirical findings. Within the above context, our main contributions (described in depth in [14, 13]) are:

1. An experiment with human summarisers, resulting in empirical evidence about how important it is to take

[1]Taken from NECA (see Section 2).

into account emotional or behavioural features when producing dialogue summaries;

2. Determination of the circumstances, within a car sales set-up, in which such features should and should not be included in the summary;

3. A description of *how* emotional and behavioural features should be reported in a summary, according to the point of view under which it was written;

4. A categorical multi-dimensional annotation scheme for summaries, designed to identify judgements of the emotional features that arise from the way the dialogue participants interact with each other;

5. A computational algorithm for the automatic production of dialogue summaries, in order to verify the computational applicability of the empirical results.

As for this last contribution, the developed automatic dialogue summariser defines when and how judgements of emotional features, arising from the interaction between the dialogue participants, should be included in the dialogue's summary, thereby producing summaries where the non-emotional information presented in the dialogue goes hand in hand with its emotional or behavioural content.

To do so, the system takes into account not only the dialogue text, but also the politeness degree of each participant of the interaction, along with the viewpoint under which the user wants the summary to be written. This system, however, given the broad coverage of the subject, does not cover all the ways that emotions can influence a summary, focusing mainly on the emotional features that come up as a consequence of the interaction between the dialogue participants. Such a system could be used, for example, to evaluate the quality of the interaction between clients and attendants in call centres, or even to generate summaries in internet support environments, in which both participants might watch the main issues they have discussed from each other's point of view. Besides, the system has a joyful side too. A system such as NECA (*Net Environment for Embodied Emotional Conversational Agents*) [17] – a platform for conversational agents which is intended, among other things, to entertain users by playing humorous videos of interactions between computer-animated characters, could be extended with a facility allowing the characters to subsequently recount their dialogue experience to the user from their personal and biased point of view.

The remainder of this paper is organised as follows. Section 2 presents the empirical foundations for the conclusions we arrived at. Next, in Section 3, we briefly describe the data necessary to build the automatic dialogue summariser presented in Section 4. Finally, Section 5 presents our conclusions.

2. EMPIRICAL FOUNDATIONS

Determining **how**, **when** and **if** emotions and behaviour must be taken into account when producing a summary required an experiment [14]. We had 30 volunteers summarise a set of dialogues that were automatically generated by the NECA system. Within NECA, the user can specify a pair of characters, defining their roles in the dialogue, their personalities and their interests [17]. Based on these values, the system can then automatically generate dialogues between these characters. The generated dialogues take place in one out of two possible domains: either they portray the interaction between a client and a vendor in a car shop (*eShowRoom*), or represent a snapshot in the life of the inhabitants of a student district in Vienna, Austria (*Socialite*).

To carry out the experiments, four dialogues were taken from the *eShowRoom* domain, and given in sequence to the experiments' volunteers, who were asked to summarise them according to one of three different points of view: observer (a neutral viewpoint), client or vendor. One month after the first experiment, the same set of volunteers had to undertake the same task once again. This time, however, they had their summaries limited to as few as 10% of the number of words in the corresponding dialogue. The experimental results demonstrate that (i) people do report the dialogue participants' emotion and behaviour whenever they produce very impolite behaviour; (ii) this report varies considerably depending on the summariser's viewpoint; and (iii) constraints on the maximum summary size have no influence on items (i) and (ii). These results were confirmed later on by independent annotations carried out by nine independent volunteers [13].

The choice of automatically generated dialogues was motivated by an absolute lack of sources for naturally occurring sales dialogues in which some party presents an improper behaviour. Also, using an automatic dialogue generator allowed for some variables (like the participants' politeness degree and the dialogue length, for example) to be changed systematically.

3. THE AUTOMATIC SUMMARISER

In order to build an automatic dialogue summariser capable of taking into account both emotional and behavioural information, we found it important to have (1) the semantic representation of the source dialogue, (2) a way to detect which dialogue participant displayed improper behaviour, (3) some means to determine where in the source dialogue this behaviour was demonstrated, (4) the semantic meaning of each clause in the human produced summaries, so that a link can be established between the information within the summarised dialogue and its counter-part in the summary, and (5) a way to determine what kind of interaction the clauses in the human generated summaries convey, so they can serve as templates for the automatically generated summaries.

Items (1) and (2) can be taken directly from NECA, since this system delivers, alongside the dialogue text, its semantics and the politeness degree of each participant, codified according to a representation language called RRL (*Rich Representation Language*) [10, 17]. Item (4) was obtained from the semantic annotation of the human produced summaries [14], manually annotated by one of the authors. For this purpose, a *Summary Act*[2] was assigned to each clause in the summaries, in order to identify the basic action the

[2]Based on work by Searle [15].

Table 1: Summary Acts used in this research.

Summary Act:	*The summariser...*
Advice	advises the reader to so something
Closure	describes the way the dialogue finished
DescrSituation	describes the overall situation in the dialogue
Evaluation	directly or indirectly assesses something or someone's behaviour or emotional state
Inform	mentions some characteristics of an object
InformAction	reports an action by some participant
Opening	describes the way the dialogue started
Opinion	presents a personal opinion

summariser executed when presenting a given information. Table 1 summarises the set of *Acts* used in this research.

Along with the *Summary Act*, each clause was assigned a corresponding semantic meaning, codified as a predicate-arguments pair, in first order logic. As an example, consider the predicate *take(tina,car,)*, meaning that the customer – Tina – took the car. In this example, the predicate-arguments pair is responsible for capturing information about (a) who executed the action or is the bearer of some attribute (Tina); (b) to whom the action was directed (implicitly, the vendor); (c) what object is involved (a car); and (d) how the action was executed (left undetermined). Additionally, a predicate was attached to this semantic codification, in order to account for the identification of the clause's polarity, as well as of its bearer (for more details see [13]).

Just like item (4), item (5) was also obtained from the experimental data. This time, however, instead of sticking to the original annotation [14], carried out by a single person, we relied on the results coming from applying the multi-dimensional scheme described in [13] by nine independent annotators. From the resulting annotation, it was possible to verify whether a clause contained some remark about emotion or behaviour and, if so, what was its polarity (a positive or negative report), along with the dialogue participant whose behaviour or emotion was reported.

Finally, the definition of item (3) turned out to be the hardest of all. The problem was that, even though NECA does associate a semantic meaning to most of its utterances (although not to all of them), it is not concerned with identifying in which clauses the dialogue participants produced a polite or impolite behaviour. We worked around this drawback by manually building a mapping between the *Summary Acts* in the human generated summaries, and the *Dialogue Acts* assigned by NECA to the dialogue utterances[3], as illustrated in Figure 1 (in this figure, links represented by a —• refer to *Summary Acts* with no corresponding *Dialogue Act* in the source dialogue). Thus, by identifying that some clause had a report on some participant's behaviour or emotional state, as assessed by the human summariser that produced that clause (step (5)), it was possible to follow this mapping to the utterance in the source dialogue with the highest chance of giving rise to such a remark.

[3]For details, see [9].

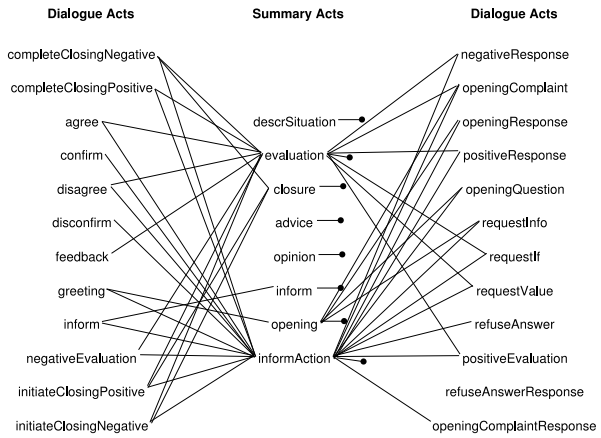

Figure 1: Possible mappings between Dialogue Act/Summary Act pairs.

4. SYSTEM DESCRIPTION

Our system was designed as a pipeline which, from input data, follows a non-deterministic algorithm to pick, from the 240 human made summaries, a candidate (Figure 2). This candidate, consisting of an almost-empty template, is run through the pipeline, being refined over and over at each of its stages, until it comes out as the final summary in the form of a set of semantic predicates that represent each of the summary clauses. Based on the experimental data corresponding to the desired viewpoint, the summary's maximum length, and the source dialogue, the first step taken by the system is to pick a random template for the summary, containing only enough information to tell apart those clauses presenting emotional or behavioural information ('E') from the rest ('r').

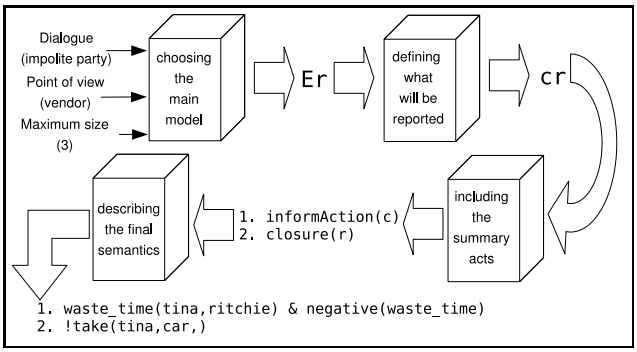

Figure 2: The summary construction pipeline.

In the next stage – *defining what will be reported* – the emotional or behavioural information is further detailed, so that it presents which entity or whose behaviour was reported in each clause, along with the polarity of this description. In the example given in Figure 2, the template tells us that the first clause of the summary must be a negative report about the client (c), whereas the remaining clauses must be kept neutral (r). Next (*including the summary acts* in the figure), the system defines a sequence of *Summary Acts* for the summary clauses, but without loosing sight of the information that came from the previous stage.

Roughly, the choice of such a *Summary Act* is made by randomly picking, from all human generated summaries under the same viewpoint and with approximately the same maximum length as the system's input, some *Summary Act* used by the summarisers at that approximate position. In this example, the summary must describe a negative action executed by the client (*informAction(c)*), followed by the way the dialogue finished (*closure(r)*). In doing so, we rely on the assumption that it would be safe for the system to emulate the way people start and finish the dialogues, as well as the order they present the summary acts, as long as we work on the same domain as people did.

Finally, and by following the mapping between *Summary Acts* and *Dialogue Acts* described in Section 3, the semantic content of the summary is determined, resulting in a sequence of logical predicates [13], representing the semantics of each of its clauses. In this example, the generated summary can be realised as "The client only wasted my time and didn't take the car". The output predicate sequence can then be picked up by an automatic natural language generator (*cf.* [11, 3]) and turned into a text, or even translated back into NECA's RRL. At this point, a very interesting feature of this system is the fact of its final product being a set of semantic representations, which can hence be realised in whatever language, provided that a corresponding natural language generator is attached to the system[4]. It is also worth noticing that, should any of the pipeline stages fail, the system starts the whole process over, so the conditions that caused the failure can change.

By changing the input to reflect the client's point of view, the system produces the summary shown in Figure 3, *i.e.*, "I asked a vendor about a car and got badly treated.". In line with the summary on Figure 3, the summary in Figure 4 illustrates one of the possibilities under the observer's viewpoint, which might be realised as "Tina wanted to buy a car and the vendor rudely answered to her". In both figures, an E stands for an emotional clause, whereas an r represents a non-emotional clause (*i.e.*, a neutral report) and a v represents a negative report about the vendor.

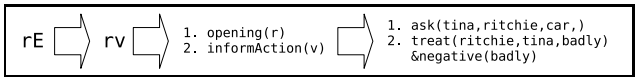

Figure 3: A summary by the client (maximum of 3 clauses).

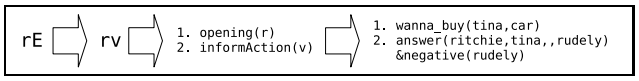

Figure 4: A summary by the observer (maximum of 3 clauses).

To illustrate the fact that non-emotional summaries can come out of the system too, Figure 5 illustrates a neutral summary, built under the observer's viewpoint, and which can be realised as "The customer asked the vendor about a car which she did not buy".

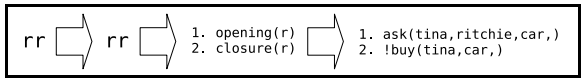

Figure 5: A neutral summary (up to 3 clauses).

As a last example, and also to show that the system is actually capable of producing longer summaries, Figure 6 shows a summary restricted to at most 14 clauses, which can be realised as "Ritchie was not respectful and was impatient. The car was worth €29.000 and had a cramped interior. Although the vendor had stressed that it was not a good car, he really knew nothing about it. When the customer asked about the power windows he, irritated, answered it. So the customer politely left the shop".

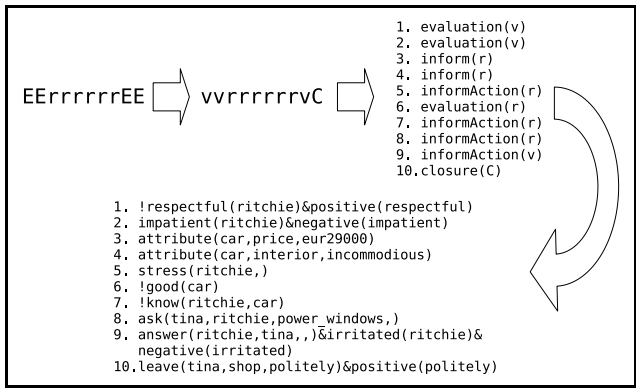

Figure 6: Observer's viewpoint (up to 14 clauses).

However interesting these results, the considerable amount of random decisions made by the system led to a rather high number of incoherent summaries being generated. An analysis of 480 summaries, 160 for each viewpoint, with up to 2, 5, 8, 11, 14, 17, 20 and 23 clauses, randomly generated by the system and manually classified by one of the authors, rendered approximately 68% of the summaries coherent whilst 32% were incoherent (*i.e.*, clauses were not expressed and organised in an effective way [7]).

These figures are, however, very much dependent on the summary size. For those summaries generated under the restricted condition (*i.e.*, those built from templates coming from the experimental condition where summarisers were restricted to 10% of the amount of words in the source dialogue), as many as 94% of the produced summaries were coherent, whereas on the unrestricted condition side that number decreases to as little as 46%. This substantial difference, however, can be directly traced to the high randomness involved in the choice of *Summary Acts* (something that was necessary to allow for different summaries to be generated from the same input data) and the mapping between *Summary Acts* and *Dialogue Acts*, as pointed out in Section 3 and in the beginning of this section.

To sort out this last problem, it would be necessary either to produce a precise semantic description of the dialogues' utterances, or to build a better mapping between the summary

[4]As a matter of fact, NECA is already able to provide dialogues for both English and German, within its *Socialite* scenario. As such, there is nothing preventing the system from generating dialogues in some other language.

clauses (as produced by human summarisers) and the (automatic) dialogue utterances. To accomplish this last task, however, one would need a good deal of data, *i.e.*, handcrafted maps between dialogue utterances and their corresponding summary clauses, so that some learning algorithm could be run on this dataset. By running this algorithm, the mapping could be augmented, for example, with probability values (currently the alternatives in Figure 1 are equally probable), increasing the odds that the system follows the right path from a summary clause to the dialogue utterance from which it originated.

5. CONCLUSION

In this paper, we presented some empirical studies to demonstrate that if a dialogue participant engages in very impolite behaviour, that behaviour will tend to be reported in the dialogue summary. Moreover, this report will be biased by the point of view under which the summary was built, without being affected by constraints on the maximum summary length. The computational applicability of these findings was demonstrated by the construction of an automatic dialogue summariser, capable of generating summaries that take into account a number of the dialogue's emotional and social features, such as the politeness degree of its participants. These features, in turn, are introduced by the system in a way that reflects the bias that different points of view can introduce into a summary.

Although a lot of research on emotion describes it in terms of a combination of valence and arousal (*e.g.* [8]), *i.e.*, a combination of the emotion's polarity – either positive or negative – and the degree of excitement it produces (from calm to excited), the scope of our research was restricted to polarity (or valence) only. Arousal was not dealt with due to the uncertain empirical status of this concept, as demonstrated by the low inter-annotator agreement that we obtained when the data coming from [14] were annotated by nine independent volunteers.

Despite the fact that the current work has made some significant inroads into understanding summarisation of dialogue taking emotion into account, some important questions still remain unanswered and require further research. One such question deals with the choice of the basic unit for annotation that was used in [14], *i.e.*, the clause. Using clauses as the basic unit for annotation had the advantage of dealing with a rather well defined concept and, as a consequence, increasing the reliability of the annotation scheme.

On the other hand, difficulties emerge with sentences such as *"then I <vendor> rudely thanked her <client> for having wasted my time"*. These clauses, if taken separately, might be classified as a negative report about the vendor (*"I rudely thanked her"*) followed by a negative report about the client (*"<client> having wasted my time"*), whereas, if taken together, we might actually have classified the entire set as a negative report about the vendor only.

Also, since the focus of our work was mainly on clauses bearing assessments of emotion or behaviour, we missed a deeper analysis of the other types of clauses and phenomena like, for example, the lack of some information the summariser was expecting to find in the summary (*e.g.*, in some of the dialogues the customer bought the car without asking for its price). Detecting such phenomena, however, strongly depends on previous information about the context into which the interaction is inserted, along with its pragmatic aspects, *i.e.*, something that might indicate, for example, that one of the main characteristics of a business dialogue, where something is being sold or bought, is precisely the negotiated price.

6. ACKNOWLEDGEMENTS

This research was sponsored by CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico – and CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Part of it was also supported by the EC Project NECA IST-2000-28580.

7. REFERENCES

[1] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. An exploration of sentiment summarization. In *AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, USA, March 2004. Technical Report SS-04-07.

[2] M. Bett, R. Gross, H. Yu, X. Zhu, Y. Pan, J. Yang, and A. Waibel. Multimodal meeting tracker. In *Proceedings of RIAO2000*, Paris, France, April 2000.

[3] R. Evans, P. Piwek, and L. J. Cahill. What is NLG? In *Proceedings of International Natural Language Generation Conference INLG02*, New York, USA, 1-3 July 2002.

[4] Y. Hijikata, H. Ohno, Y. Kusumura, and S. Nishida. Social summarization of text feedback for online auctions and interactive presentation of the summary. In *Proceedings of 11th ACM International Conference on Intelligent User Interfaces (ACM IUI 2006)*, pages 242–249, Sydney, Australia, January 2006.

[5] M. Kameyama, G. Kawai, and I. Arima. A real-time system for summarizing human-human spontaneous spoken dialogues. In *Proceedings of the 4th International Conference on Spoken Language (ICSLP 96)*, volume 2, pages 681–684, Philadelphia, USA, 1996.

[6] M. Kearns, C. Isbell, S. Singh, D. Litman, and J. Howe. Cobotds: A spoken dialogue system for chat. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI2002)*, pages 435–430, Edmonton, Canada, 2002.

[7] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-ocurrence statistics. In *Proceedings of the Human Technology Conference (HLT-NAACL-2003)*, Edmonton, Canada, May 27 – June 1 2003.

[8] R. Picard. Affective computing. Technical Report 321, MIT Media Laboratory, Perceptual Computing Section, Cambridge, USA, November 26 1995.

[9] P. Piwek. NECA deliverable D3a: Specification of scene descriptions for the neca domains. Technical report, ITRI – University of Brighton, Brighton, UK, 2002. NECA IST-2000-28580 Deliverable D3a.

[10] P. Piwek, B. Krenn, M. Schröder, M. Grice, S. Baumann, and H. Pirker. RRL: A Rich Representation Language for the description of agent

behaviour in NECA. In *Proceedings of the AAMAS Workshop on Embodied Conversational Agents - Let's Specify and Evaluate them!*, Bologna, Italy, 2002.

[11] E. Reiter and R. Dale. *Building Natural-Language Generation Systems*. Cambridge University Press, 2000.

[12] N. Reithinger, M. Kipp, R. Engel, and J. Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'2000)*, pages 310–317, Hong Kong, China, 2000.

[13] N. T. Roman. *Emoção e a Sumarização Automática de Diálogos*. PhD thesis, Instituto de Computação – Universidade Estadual de Campinas, Campinas, São Paulo, Julho 2007.

[14] N. T. Roman, P. Piwek, and A. M. B. R. Carvalho. *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, chapter Politeness and Bias in Dialogue Summarization: Two Exploratory Studies, pages 171–185. Springer Netherlands, Dordrecht, The Netherlands, January 9 2006. ISBN: 1-4020-4026-1.

[15] J. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

[16] T. Takahashi and Y. Katagiri. Telmea2003: Social summarization in online communities. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 03)*, pages 928–929, Fort Lauderdale, USA, 2003.

[17] K. van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann. Fully generated scripted dialogue for embodied agents. *Artificial Intelligence*, 172(10):1219–1244, June 2008.

[18] K. Zechner and A. Lavie. Increasing the coherence of spoken dialogue summaries by cross-speaker information linking. In *Proceedings of the NAACL-01 Workshop on Automatic Summarization*, Pittsburgh, USA, June 2001.