Chapter X

Politeness and Bias in Dialogue Summarization: Two Exploratory Studies

Norton Trevisan Roman
State University of Campinas
[1]Instituto de Computação,Universidade Estadual de Campinas,
Avenida Albert Einstein 1251
Cx. Postal 6176
CEP 13084-971,Campinas, SP, Brazil
Email: norton@ic.unicamp.br

Paul Piwek
University of Brighton
[2]ITRI, University of Brighton, UK.
Email: Paul.Piwek@itri.brighton.ac.uk

Ariadne Maria Brito Rizzoni Carvalho
State University of Campinas
[1]Computing Institute, State University of Campinas, Brazil.
Email: ariadne@ic.unicamp.br

Abstract

In this chapter, two empirical pilot studies on the role of politeness in dialogue summarization are described. In these studies, a collection of four dialogues was used. Each dialogue was automatically generated by the NECA system and the politeness of the dialogue participants was systematically manipulated. Subjects were divided into groups who had to summarize the dialogues from a particular dialogue participant's point of view or the point of view of an impartial observer. In the first study, there were no other constraints. In the second study, the summarizers were restricted to summaries whose length did not exceed 10% of the number of words in the dialogue that was being summarized.

Amongst other things, it was found that the politeness of the interaction is included more often in summaries of dialogues that deviate from what would be considered normal or unmarked. A comparison of the results of the two studies suggests that the extent to which politeness is reported is not affected by how long a summary is allowed to be. It was also found that the point of view of the summarizer influences which information is included in the summary and how it is presented. This finding did not seem to be affected by the constraint in our second study on the summary length.

Keywords: automatic dialogue summarization, automatic summarization, natural language processing, politeness, bias.

1 Introduction

To build an automatic summarization system capable of producing more "human like" dialogue summaries, one must first understand *how* humans summarize dialogues, *i.e.*, *what* they consider important to report in the summary and *how* they report it. The focus of this paper is on one aspect about which little is known. It concerns the nature of the interaction and, in particular, the politeness of the interlocutors. The question that is addressed is whether politeness is important enough to be mentioned in the summary and, if so, how it is mentioned.

So far, work on both language interpretation and generation has concentrated mainly on interpretation to and generation from truth-conditional representations of content, with some early exceptions such as (Hovy, 1988). The underlying assumption has been that the content of a natural language utterance is captured by the conditions under which it is true. When considering dialogue, the limitations of a strictly truth-conditional approach are apparent. Many types of dialogue acts do not yield to a purely truth-conditional analysis (greetings, acknowledgments, closings, etc.) and the way a dialogue proceeds is often affected by the emotional states of the interlocutors (Craggs and Wood, 2003; Fischer, 1999; Schmitz and Quantz, 1996; Reithinger *et al*., 2000).

Recent research on the influence of emotions on interpretation and generation includes work on recognizing the user's emotional state, e.g., anger (Huber *et al*., 2000), so that a computer dialogue system can adapt its own behaviour to prevent such feeling; the use of humour to facilitate interaction with users (Nijholt, 2003); and strategies to establish a social and personal relationship with the user by means of "small talk" (Bickmore and Cassell, 2000; Bickmore, 2002). An overview of the literature on affect/emotion in natural language generation can be found in (Piwek, 2003). Although there is a body of work on politeness in dialogue generation and interpretation (*e.g.* Ardissono *et al*., 1999; and Walker *et al.*, 1997), as far as we know, there is not yet any systematic work on the role of politeness in dialogue summarization and, more specifically, on how politeness is reported in dialogue summaries.

Work on automated summarization (*e.g*. Mani, 2001; Mani and Maybury, 1999; Marcu, 2000) and more specifically dialogue summarization (*e.g*. Alexandersson and Poller, 2000; Kipp *et al*., 1999; Reithinger *et al*., 2000) has concentrated on summarizing informational content, at the expense of reports on the quality of interaction, such as the politeness of the interlocutors. Approaches so far to politeness in generation and interpretation mainly follow the ideas of Brown and Levinson (1987) about face threatening. These ideas were criticized, amongst others, by Eelen (2001), who pointed out that Brown and Levinson do not provide a balanced account of both polite and

impolite behavior; they are mainly concerned with polite behaviour. Watts (2003) argues that impoliteness is an important feature of interaction. In particular, he has argued that people are more likely to comment on impolite than polite behaviour.

In this chapter, two pilot studies are presented. In both studies, the following hypothesis, based on Watts ' (2003) aforementioned supposition concerning impolite behaviour, is tested:

Hypothesis 1: If a dialogue contains very impolite behaviour by the dialogue participants, this behaviour tends to be reported in the dialogue summary.

This hypothesis was confirmed by the studies, *i.e.*, subjects actually reported politeness for the very impolite dialogues. This is an important result, to the extent that it shows *when* an automatic summarizer should report politeness.

The second tested hypothesis concerns *how* people report politeness. In particular, the way the point of view of a summarizer can bias reports of politeness is investigated. This hypothesis was formulated along the lines of Walton's definition of biased arguments. According to Walton (1999, p.86) "... a biased argument is a one-sided argument, consisting of pure pro-argumentation for one side of an issue in a dialogue, while failing to genuinely interact with the other side on a balanced way. A balanced argument, in contrast, considers all the relevant arguments on both sides and exhibits the characteristics of flexible commitment, empathy, open-mindedness, critical doubt, and evidence sensitivity." The second hypothesis is:

Hypothesis 2: Reporting of politeness in a dialogue summary is biased as a result of the point of view of the summarizer.

Again, the studies confirmed this second hypothesis, *i.e.*, the findings suggest that people bias their summaries according to their point of view.

In the studies, the subjects were instructed to summarize a carefully selected set of automatically generated dialogues. The resulting collection of human-authored summaries constituted the data. The two studies that were carried out differed in only one respect. Whereas in the first study the summarizers were not provided with any instructions concerning the length of their summaries, in the second study, summarizers were given a maximum summary size. The purpose of having two studies was to test the third hypothesis:

Hypothesis 3: Severely restricting the summary length has no influence on Hypotheses 1 and 2.

This third hypothesis, as it will be shown in the sections ahead, was confirmed. No changes in the results pertaining to hypotheses 1 and 2 were found when subjects faced a maximum summary size of 10% of the number of words of the corresponding dialogue.

The eventual aim is to use the insights gained from these studies in the construction of an automatic dialogue summarizer, which will be folded into the NECA system (Krenn *et al.*, 2002). NECA (*Net Environment for Embodied Emotional Conversational Agents*) is a conversational agent platform in which the user can create characters by specifying their roles, personalities and interests. Based on these settings, the system automatically generates dialogues between the characters. The result of the generation process is a script that can be performed by two or more embodied agents (Piwek *et al.*, 2002, Piwek and van Deemter, 2003). Potential applications are for the purpose of, for instance, entertainment, infotainment and education. The system has been

developed for two domains: *eShowRoom* (infotainment), which concerns car sales, and *Socialite* (entertainment), which concerns characters who are inhabitants of a student district in Vienna.

The dialogues for the current studies were taken from the eShowRoom domain (first version of October 2002) where an agent is the vendor while the other is the customer. NECA's scripted dialogues were used as a test corpus because the NECA system allowed us to systematically change, amongst other characteristics, the agents' politeness.

2 First Study: Politeness and Bias in Unconstrained Dialogue Summarization

2.1 Method

The subjects in this study were students and former students from the State University of Campinas, Brazil. Figure 1 shows the educational level and field of study of the subjects. From the total amount of subjects, 18 were men and 12 women. Subjects were given the dialogues, hardcopies or by e-mail, and asked to summarize those dialogues and return the summaries to the researchers, either by e-mail or on paper, typed or handwritten.

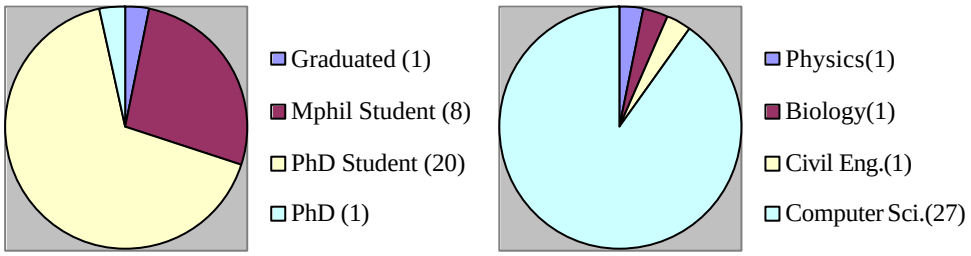

Figure 1. Educational level and field of study of subjects

The study was carried out during July and August 2003 at the State University of Campinas, Brazil. It was conducted as follows. First, four dialogues were generated using the NECA *eShowroom* system. Each dialogue involved a customer and a vendor: a character named Ritchie tries to sell a car to another character named Tina. The dialogues can be found in Appendix I. For the purpose of this study, the politeness of both interlocutors in the dialogues was varied. This was possible because NECA allows us to control some dialogue parameters. These include (a) the interlocutors' interests, which influence the length of the dialogue (if the customer is interested in many features of the car, the dialogue will be longer), (b) their agreeableness, which also influences the dialogue length, for it determines how easily the customer is persuaded to buy the car, or how easily the vendor will give up the sale, and (c) their politeness, which determines how polite the interlocutors will be (currently the system allows only two politeness degrees: polite and impolite).

Transcripts of the automatically generated dialogues were presented to the subjects in the following order:

- Dialogue D_1 showed an interaction in which both vendor and customer were polite and do not use language that is likely to cause offence. The dialogue began with the vendor attending to the customer, and ended with the customer buying the car;

- D_2 showed an interaction in which the vendor is polite and the customer is impolite. The interaction began with the customer not finding the vendor in the showroom (i.e., the vendor was apparently absent for a moment) and ended with the customer not buying the car;
- D_3 showed an interaction in which the vendor is impolite and the customer is polite. Again, the dialogue began with the vendor being apparently absent for a moment, and ended with the customer not buying the car;
- D_4 showed a very short interaction in which both vendor and customer were polite again. Like the first dialogue, it began with the vendor introducing himself to the customer, and ended with the customer buying the car.

The subjects were told to summarize the dialogues as if they had been present during the dialogue and were telling a friend what happened. Subjects were divided into three groups. Firstly, ten subjects were told to summarize the dialogue from the point of view of the customer, i.e., they were asked to pretend that they had been the customer. Secondly, another ten subjects were told to summarize the dialogues as if they were the vendor. Finally, a third group of ten subjects were told to summarize the dialogues as if they had overheard the dialogue without directly participating in it. The subjects were given no other constraint. It was up to the subjects to decide how long the summary should be, as well as whether they should quote parts of the dialogue or rephrase them. Although the dialogues were in English, the summarizers were instructed to summarize them either in Portuguese or English, depending on their preference.

The produced summaries were manually annotated by one of the researchers, sentence by sentence, for whether or not the sentences reported (positive or negative) behaviour. The summaries as a whole were classified based on whether they contained a sentence reporting the interlocutors' behaviour. Thus, a summary was classified as *behav* if it contained at least one sentence concerning the behaviour, such as "I asked a very disgusting vendor..." (negative), "I was extremely attentive..." (positive), "I was in a bad mood..." (negative), and "Ritch, showing lack of respect and politeness..." (negative); and *no behav* otherwise.

For the annotation of bias, Walton's (1999) definition was used. Following Walton, a biased summary was defined as *a summary in which the summarizer strongly defends his/her position in the dialogue, by arguing in favor of it and not genuinely taking the other side in a balanced way*.

To implement the notion of bias based on Walton's work, the reporting of positive/negative behaviour by the vendor or customer was manually annotated on a sentence-by-sentence basis. Examples of sentences that were classified as negative reports were "I lost my patience and rudely said...", "Besides being ill-treated today...", "disgusting vendor", "she was angry and impatient", "what a 'nice' person I received today" (sarcastic), "a loser came...", "annoying customer", "(Ritchie) is an awful vendor", "I was impatient, the vendor came along with small talk...", and "I was not in the mood...she asked silly questions". Similarly, examples of sentences that were classified as positive reports were "I was extremely attentive", "I gently apologized", "Ritchie served her politely", "therefore I preferred to be polite...", "she didn't loose her calm" and "Ritchie showed... with patience".

In order to classify the annotated summaries, two mutually exclusive, but not exhaustive, categories were defined:

Definition 1 (Exclusively Negative Report -- *ENR* -- of an agent A) A summary is classified as an *ENR(A)* if and only if:

1. It has at least one sentence where agent A's behaviour is reported negatively and this behaviour is not blamed on another agent B; and
2. It has no positive report about the agent at all.

Definition 2 (Exclusively Positive Report -- *EPR* – of an agent A) A summary is classified as an *EPR(A)* if and only if:

1. It has at least one sentence reporting the agent A's behaviour positively; and
2. It has no negative report about the agent at all.

Within this framework, a summary can either be classified as *ENR(V)*, when it exclusively reports the vendor's behaviour as negative, or *ENR(C)*, when it exclusively reports the customer's behaviour as negative. Similarly, they can also either be classified as *EPR(V)*, when they exclusively positively report the vendor's behaviour, or *EPR(C)*, when the positively reported agent is the customer. Summaries that do not fit the above categories are left out of consideration, for the moment.

The summary classification mentioned above can now be used to implement this chapter's definition of a biased summary, based on the collective behaviour of the subjects. If one finds that the number of summaries classified as *ENR(V)* and *ENR(C)* varies significantly according to the point of view given the same dialogue, this suggests that subjects did not report the interaction in a balanced way and were biased. The same idea applies to *EPR* summaries.

2.2 Results and Analysis

Figure 2 summarizes the study's results for behaviour reporting[1]. In this figure, the bars differentiate among (a) the users who only reported some party's behaviour (*behav (only)*), ignoring technical information, (b) the users who reported the behaviour, along with technical information (*behav*), and those who did not report any behaviour at all (*no behav*), i.e., those who only included the exchanged technical information in the summary. For each dialogue, behaviour reporting is displayed for each of the three points of view adopted by subjects, i.e., the customer's (**C**), the vendor's (**V**) and an observer's (**O**) point of view.

This Figure shows that although for D_1 and D_4 there was a small percentage of subjects who included the behaviour of some party in the summary (respectively, 7% for D_1 and 17% for D_4), this percentage is remarkably higher for D_2 and D_3 (respectively, 67% for D_2 and 93% for D_3). This percentage consists of the sum of the subjects who have taken the behaviour into account and those who have only taken the behaviour into account.

[1] There are some minor differences between the data in this table and those given in previous technical reports on this research (see Roman *et. al.* (2004a), (2004b) and (2004c)). These differences are due to the correction in this chapter of a typo and a counting mistake. Note that although there are some minor differences between the numbers, the main conclusions have not been affected and are the same for this chapter and the technical reports that predate it.

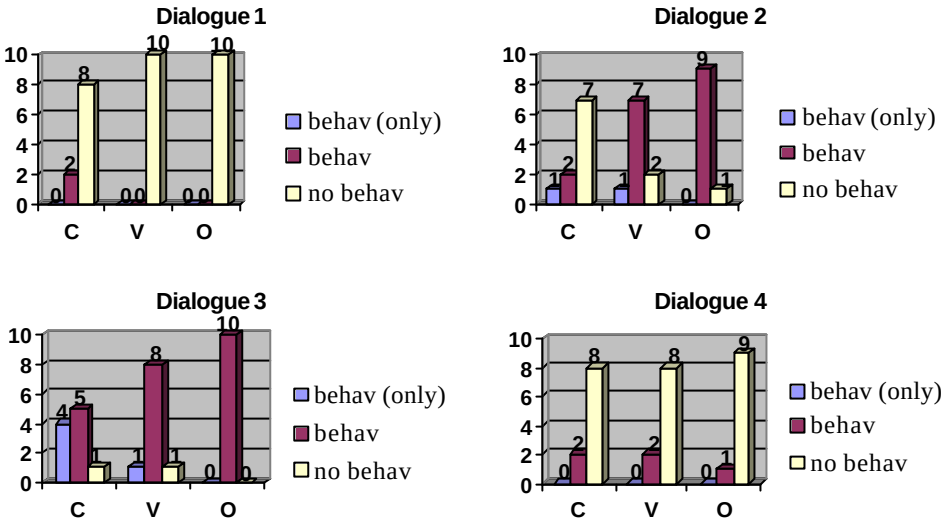

Figure 2. Behaviour reporting data for the first study.

The differences might be explained by the fact that D_2 and D_3 show unusual situations for car sales dialogues. In the first dialogue the customer and in the second the vendor manifested rude behavior. D_1 and D_4, on the other hand, are more usual, or "neutral", dialogues. If one groups D_1 and D_4, and D_2 and D_3 together, one finds approximately 12% of the subjects mentioning the behaviour in D_1 and D_4, and 80% mentioning it in D_2 and D_3. A 2×2 χ^2 analysis reveals that this is a significant difference, χ^2 (1, N=120) = 56.43, at the significance level of p = 0.001. This suggests that whether subjects report some party's behaviour/politeness in a summary depends on the interaction's politeness degree.

Another point that is worth mentioning is the higher number of subjects who reported the behaviour in D_4, when compared to D_1 (from 2 subjects in D_1 to 5 in D_4). This might be because D_2 and D_3 present such an unusual interaction that the subjects notice that in the last dialogue the interaction is again "normal". The difference between D_1 and D_4 in this respect is, however, not statistically significant.

Furthermore, if the total number of summaries mentioning politeness/behaviour and those that do not are counted, and related to the point of view they were summarized under, one has 16 summaries reporting the behaviour (and 24 not reporting it at all) for the customer's viewpoint, 19 reporting (and 21 not reporting) for the vendor's, and 20 reporting (and 20 not reporting) for the observer's. This presents no statistically significant relation between whether politeness/behaviour is reported and the point of view.

Figure 3 shows the bias annotation results. In this figure, the bars differentiate amongst the negatively reported behaviour, for each viewpoint. Applying a χ^2 analysis to the data in Figure 3 results in χ^2 (2, N=40) = 20.39. This indicates that whether subjects report the vendor's or the customer's behaviour exclusively negatively, depends on their point of view at the significance level of p = 0.001. These data lead to the conclusion that there exists a bias: subjects are apparently trying to advocate their side by reporting mostly the counterpart's behaviour. In

addition to the summaries which were either ENR(V) or ENR(C), there were eight other summaries (five vendors and three observers) that were neither ENR(V) or ENR(C). We did not consider these more "balanced" summaries.

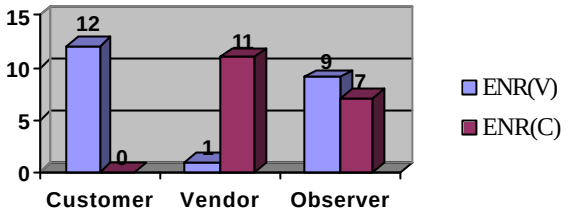

Figure 3. ENR (Exclusively Negative Report) for each agent, according to the point of view.

The results for the positive reports are inconclusive, due to the small number of data points. Figure 4 shows these results.

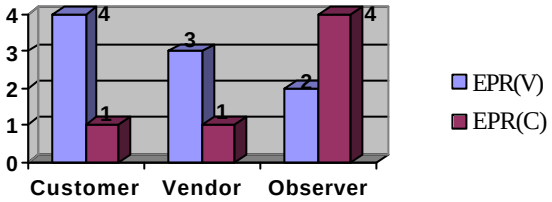

Figure 4. EPR (Exclusively Positive Report) for each agent, according to the point of view.

3 Second Study: Politeness and Bias in Constrained Dialogue Summarization

3.1 Method

This study was also carried out over a period of two months - November and December 2003 - at the State University of Campinas, Brazil. In order to compare results, the same set of dialogues as in the first study was used. The same procedure as for the first study was followed (see Section 2.1 for details). Subjects were asked to produce a summary for each dialogue according to the same point of view they assumed in the first study; the only difference being that the summary length was restricted to 10% of the number of words in the dialogue. For annotation of the summaries, the procedure as described in Section 2.1 was followed.

3.2 Results and Analysis

Figure 5 summarizes the results of the study. In this figure, like in Figure 2, the bars differentiate among the subjects who have only taken some party's behaviour into account (*behav (only)*), the subjects who have taken the behaviour into account as well as technical information (*behav*), and those who have not taken it into account at all (*no behav*).

For this study, the percentage of subjects who included some party's behaviour in the summary, for D_1 and D_4, is 3%. This percentage is 47% and 90% for D_2 and D_3, respectively. Again there is a greater number of subjects who report behaviours when the dialogue is unusual with respect to

the politeness that is displayed. Interestingly, the upper bound on the summary length may have caused summarizers who reported some behaviour to leave out technical information.

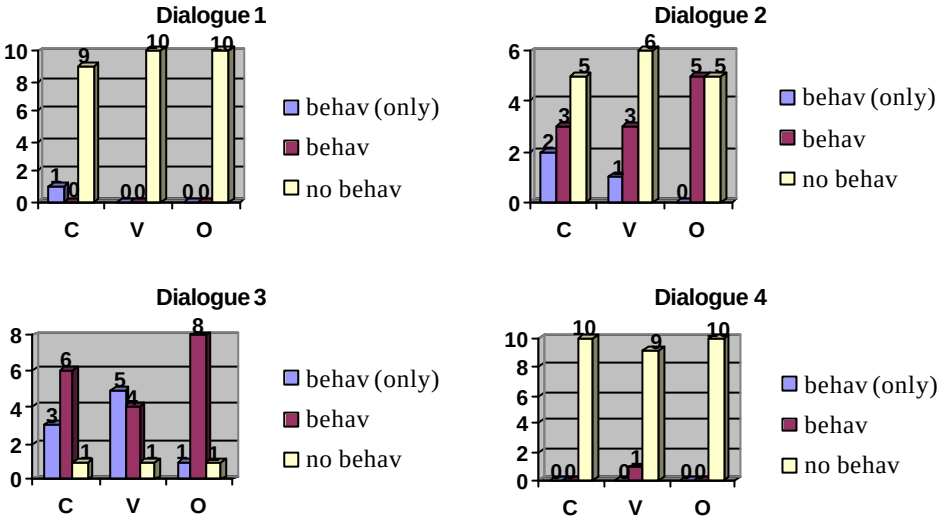

Figure 5. Behaviour data for the second study.

If one groups D_1 and D_4, and D_2 and D_3, one has approximately 3% of the subjects mentioning some behaviour in D_1 and D_4, and 68% mentioning it in D_2 and D_3. A 2×2 χ^2 analysis revealed that this was a significant difference, χ^2 (1, N=120) = 55.13, at the significance level of p = 0.001. Thus, this provides evidence that there may be a relationship between the politeness of the dialogue and subjects reporting dialogue behaviour in the summary, even with the restriction on the summary length.

Notice that the number of subjects who included behaviour in the summaries for D_1 and D_4 is exactly the same, i.e., apparently, the fact that D_4 was presented after two unusual dialogues did not affect the subjects in this second study, where there was a constraint on summary length.

Counting again the total number of summaries mentioning some behaviour and those not mentioning it at all, and relating them to the assumed viewpoint, one has 15 summaries mentioning it and 25 not mentioning it, for the customer's point of view; 14 and 26, respectively, for the vendor's; and 14 and 26 for the observer's. This presents no statistically significant relation between whether summaries report behaviour and the point of view.

Figure 6 summarizes the bias annotation results for this second study. In this figure, like in Figure 3, the bars differentiate amongst the reported behaviour. If one applies a 3×2 χ^2 analysis to the data in Figure 6, one has χ^2 (2, N=39) = 7.91, which is significant at the p = 0.02 level. This result suggests that if subjects report some behaviour, they bias their reports according to their point of view. For this classification, two summaries (one for the vendor's viewpoint and another for the observer's) fit neither in ENR(V) nor in ENR(C).

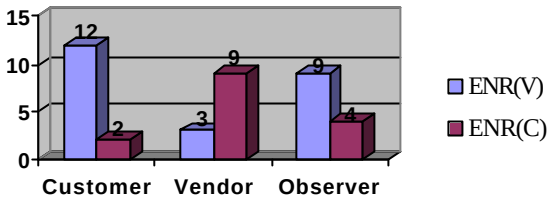

Figure 6. ENR (Exclusively Negative Report) for each agent, according to the point of view.

With respect to the positive reports, the customer's behaviour was positively reported only in one summary (customer), while the vendor was reported in three summaries (one per viewpoint). As in the previous study, no conclusion can be drawn due to the small amount of data. It is worth noticing, however, that for the impolite dialogues, all the positive reports were given as a contrast against the negative behaviour showed by the other party in the dialogue.

4 Comparison

In both studies (Figures 2 and 5) subjects reported the behaviours in the summary mainly for the more impolite dialogues (D_2 and D_3). This did not depend on the summary size. When the summary size was restricted to a value as low as 10% of the number of words present in the dialogue, subjects still included remarks about the emotional state of the interlocutors, related to their politeness, for the dialogues in which one of the interlocutors was impolite/rude.

If, however, one counts the summaries in which subjects reported some behaviour (*behav* and *behave(only)*) and those in which they do not (*no behav*), and groups them according to the study they belong to, then one has no statistically significant evidence ($p = 0.20$) that the overall number of summaries reporting some behaviour depends on the study, i.e., it does not depend on whether subjects are restricted to a 10% summary size or not. Figure 7 summarizes the results.

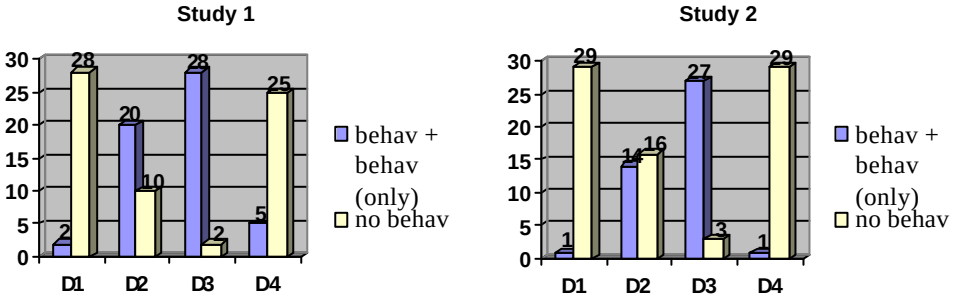

Figure 7. Summaries reporting some behaviour for both studies.

A last point about this part of the studies is that the difference between the number of summaries including behaviour in D_1 and D_4 for the first study is not reproduced in the second study. If the fact that D_4 was presented after two unusual dialogues affected the summaries in the first study, it might be that the restriction on the summary size in the second study suppressed this effect.

Concerning bias, both studies provide evidence that the way people describe the behaviour of some party, and more specifically, which party they describe negatively, depends on the point of view under which they are summarizing, *i.e.*, it is biased. Not only that, but a closer look at Figure 3 and Figure 6 shows that the negatively reported party tends to be the counterpart in the dialogue, as opposed to the summarizer's point of view (vendors negatively reporting customers and *vice versa*). This fact is in accordance with Walton (1999), in the sense that these summaries consist of "pure advocacy for one side of an issue in a dialogue, without genuinely considering the other side in a balanced way".

Comparing data from figures 3 and 6, one has 22 summaries classified as *ENR(V)* and 18 as *ENR(C)*, for the first study, and 24 *ENR(V)* and 15 *ENR(C)*, for the second study. A χ^2 analysis of these data leads to no statistically significant results, meaning that most probably the negative report of behaviours does not depend on the study, i.e., on whether the summary has a fixed length of 10% or no constrain at all.

With respect to the positive reporting of behaviours, the amount of data collected is not enough to lead to any conclusion. Interestingly, the number of positive remarks is 15 in the first study and four in the second study. This could suggest that, as subjects face severe constraints on the summary size, they drop the positive remarks, keeping only the reports on the negative behaviors.

5 Conclusion and Outlook

In this chapter, two exploratory studies were described. These studies were designed to gather information about how people summarize dialogues. More specifically, the studies were intended to shed light on how people deal with politeness/behaviour in dialogues and how their point of view influences their summaries. The eventual goal is to use the results to inform the construction of an automatic dialogue summarization system. The envisaged system should be capable of considering dialogue politeness (through its participants' behaviour) and point of view when summarizing. The underlying rationale is that by basing our work on experiments into the behavior of human summarizers, our system will eventually produce automatically generated summaries that are perceived to be "natural". In practical terms, we intend to evaluate this by finding out to what extent system produced summaries are indistinguishable from summaries produced by humans.

In both studies, the percentage of summaries reporting some behaviour was higher when the dialogues were more impolite. This result was independent of the point of view and summary size. The order in which the dialogues were presented appeared to have little influence on whether people included behaviour information in the summary. No statistically significant differences were found with respect to the reporting of behaviour for the neutral dialogue before the two impolite dialogues, and the neutral dialogue after them.

Regarding the question how people report behaviour, when they do so, it was observed that the point of view adopted by the summarizer influences reporting of behaviours. In particular, whose behaviour is negatively reported (vendor or customer), depends on the point of view of the summarizer rather than the actual dialogue behaviour. This is an indication that people bias their summaries. In our studies, this result did not depend on whether the summary was restricted to 10% of the dialogue size or not.

There is tentative evidence that positive reporting is less subject to bias. The number of data points on this was, however, too limited to apply any statistics. Further studies into positive reporting and bias are required to come to a firm conclusion. Interestingly, the number of subjects positively reporting some of the behaviours is lower in the second than in the first study: it decreases with the shortening of the allowed summary size, indicating that there could be a relationship between them.

The currently used dialogues are automatically generated. In further studies it would be interesting to test the obtained results from these preliminary studies with more complex dialogues, like those produced by human writers for TV, theatre, etc. Currently, the results presented here are being used to build an algorithm for automatic dialogue summarization that takes politeness, behaviours and point of view into account. The algorithm will be tested through an implementation that is embedded in the NECA platform and which takes as input a NECA scripted dialogue. The system will generate the summary according to a point of view defined by the user, and include the politeness of the interaction, if requested by the user.

6 Bibliography

Alexandersson, J. and Poller, P. (2000) Multilingual Summary Generation in a Speech-to-Speech Translation System for Multilingual Dialogues. In *Proc. International Natural Language Generation Conference (INLG-2000)*. Mitzpe Ramon, Israel.

Ardissono, L., Boella, G. and Lesmo, L. (1999) Politeness and Speech Acts. In *Proc. Workshop on Attitude, Personality and Emotions in User-Adapted Interaction.* 41-55. Banff, Canada.

Bickmore, T. and Cassell, J. (2000) How About this Weather? Social Dialogue with Embodied Conversational Agents. In *Proc. AAAI Fall Symposium on Socially Intelligent Agents.* North Falmouth, USA.

Bickmore, T. (2002) Social Dialogue is Serious Business. In *Proc. CHI 2002 Workshop on Socially Adept Technologies.* Minneapolis, USA.

Brown, P. and Levinson, S. (1987) *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Craggs, R. and Wood, M. (2003) Annotating Emotion in Dialogue. In *Proc. 4^{th} SIGdial Workshop on Discourse and Dialogue*. Sapporo, Japan.

Eelen, G. (2001) *A Critique of Politeness Theories.* St. Gerome, Manchester.

Fisher, K. (1999) Annotating Emotional Language Data. *Technical Report 236*. Verbmobil Project.

Hovy, E. (1988) *Generating Natural Language Under Pragmatic Constraints.* Lawrence Erlbaum, New Jersey.

Huber, R., Batliner, A. Buckow, J., Nöth, E., Warnke, V. and Niemann, H. (2000) Recognition of Emotion in a Realistic Dialogue Scenario. In *Proc. International Conference on Spoken Language Processing.* Beijing, China.

Kipp, M., Alexandersson, J. and Reithinger, N. (1999) Understanding Spontaneous Negotiation Dialogue. In *Proc. International Joint Conferences on Artificial Intelligence (IJCAI-1999)*. Stockholm, Sweden.

Krenn, B., Pirker, H., Grice, M., Baumann, S., Piwek, P., van Deemter, K., Schroeder, M., Klesen, M. and Gstrein, E. (2002) Generation of Multimodal Dialogue for Net Environments. In Busemann, S. (Ed.) *KONVENS 2002, Deutsches Forschungszentrum fuer Kuenstliche Intelligenz (DFKI)*. 91-98. Saarbruecken, Germany.

Mani, I. and Maybury, M. (1999) *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.

Mani, I. (2001) *Automatic Summarization*. John Benjamins, Amsterdam.

Marcu, D. (2000) *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Nijholt, A. (2003) Humor and Embodied Conversational Agents. *CTIT Technical Report Series 03-03*. University of Twente.

Piwek, P., Krenn, B., Schröder, M., Grice, M., Baumann, S. and Pirker, H. (2002) RRL: A Rich Representation Language for the Description of Agent Behaviour in NECA, In *Proc. Workshop "Embodied Conversational Agents – Let's Specify and Evaluate Them!"*. Bologna, Italy.

Piwek, P. (2003) An Annotated Bibliography of Affective Natural Language Generation. *Technical Report ITRI-02-02*. ITRI – University of Brighton.

Piwek, P. and van Deemter, K. (2003) Dialogue as Discourse: Controlling Global Properties of Scripted Dialogue. In *Proc. AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*. Menlo Park, CA.

Reithinger, N., Kipp, M., Engel, R. and Alexandersson, J. (2000) Summarizing Multilingual Spoken Negotiation Dialogues. In *Proc. 38th Meeting of the Association for Computational Linguistics (ACL-2000)*. Hong Kong, China.

Roman, N., Piwek, P. and Carvalho, A. (2004a) Politeness and Summarization: an Exploratory Study. In *Proc. AAAI Spring Symposium Exploring Affect and Attitude in Text: Theories and Applications (AAAI-EAAT-2004)*. Palo Alto, USA.

Roman, N., Piwek, P. and Carvalho, A. (2004b) Be Polite or Be Summarized. *Technical Report ITRI-04-05*. ITRI – University of Brighton.

Roman, N., Piwek, P. and Carvalho, A. (2004c) Politeness and Summarization: to Bias or not to Bias. *Technical Report ITRI-04-06*. ITRI – University of Brighton.

Schmitz, B. and Quantz, J. (1996) Dialogue Acts in Automatic Dialogue Interpreting. *Technical Report 173*, Verbmobil Project.

Walker, M., Cahn, J. and Whittaker, S. (1997) Improvising Linguistic Style: Social and Affective Bases for Agent Personality. In *Proc. Conference on Autonomous Agents (AGENTS-97)*. Marina del Rey, USA.

Walton, D. (1999) *One-Sided Arguments: A Dialectical Analysis of Bias.* State University of New York Press.

Watts, R. (2003) *Politeness.* Cambridge University Press.

7 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.

8 Appendix I

The four dia logues used in the experiment, in the same order they were presented to the subjects.

Dialogue 1
Size: 99 words. Summary size: 10 words.

R: Hello! I am Ritchie.
T: Can you tell me something about this car?
R: It is very safe.
R: It has anti-lock brakes.
R: It has airbags.
T: How much does it consume?
R: It consumes 8 litres per 60 miles.
T: Interesting.
T: What kind of luggage compartment does this car have?
R: It has a spacious luggage compartment.
T: Excellent!
T: What kind of interior does this car have?
R: It has a spacious interior.
T: Excellent! (Tina)
T: How much does this car cost?
R: It costs 25 thousand Euros.
T: Well, well.
T: All in all this is a perfect car. It is a deal!
R: Really? I am sure you wont regret it.

Dialogue 2
Size: 182 words. Summary size: 18 words.

T: Anybody here? I haven't got all day.
R: I am sorry to have kept you waiting. I will be with you in a moment.
R: How can I help you?
T: I want you to tell me about this car.
R: It is very friendly for the environment.
R: It is made of recyclable materials.
R: It has a catalytic converter.
T: How much does it consume?
R: It consumes 8 litres per 60 miles.
T: It is not exactly very friendly for the environment.
R: Just consider this.
R: This car is very prestigious.
R: It has leather seats.
R: It has broad tires.
R: It drives 100 and 10 miles per hour.
R: It has 80 horse power.
T: What kind of luggage compartment does it have?
R: It has a spacious luggage compartment.
T: Well, well.
T: What kind of interior does this car have?
R: It has a spacious interior.
T: Well, well.
T: How much does this car cost?
R: It costs 25 thousand Euros.
T: It is not exactly very cheap.
R: Just consider this.
R: This car is very safe.
R: It has anti-lock brakes.
R: It has airbags.
T: Enough is enough! You keep your car!
R: Well, sorry that I cannot help you.

Dialogue 3
Size: 124 words. Summary size: 12 words.

T: Anybody in today?
R: Yeah yeah, keep your calm.
T: Can you tell me something about this car?
R: It is very safe.
R: It has anti-lock brakes.
R: It has airbags.
T: Does it have power windows?
R: What do you think? Of course not!
T: This car is not exactly very prestigious.
R: Just consider this.
R: This car has leather seats.
R: It has broad tires.
R: It drives 100 and 10 miles per hour.
R: It has 80 horse power.
T: What kind of luggage compartment does it have?
R: It has a spacious luggage compartment.
T: Well, well.
T: What kind of interior does this car have?
R: For god's sake, how should I know?
T: No problem.
T: Thank you for your help. I have to think a bit more about this.
R: Great! Thanks for wasting my time.

Dialogue 4
Size: 53 words. Summary size: 5 words.

R: Hello! My name is Ritchie.
R: What can I do for you?
T: Could you tell me something about this car?
R: This is a very sporty car.
R: It can drive 100 miles per hour.
T: How much does it consume?
R: It consumes 8 litres per 60 miles.
T: Not bad.
T: Ok! I'll take it.
R: Congratulations! Exquisite taste.