A Multi-dimensional Annotation Scheme for Behaviour in Dialogues*

Norton Trevisan Roman[1] and Ariadne Maria Brito Rizzoni Carvalho[2]

[1] School of Arts, Sciences and Humanities – University of São Paulo, São Paulo, Brazil
norton@usp.br
[2] Institute of Computing – University of Campinas, Campinas, Brazil
ariadne@ic.unicamp.br

Abstract. In this paper we introduce a multi-dimensional annotation scheme for emotional and behavioural assessment in dialogue summaries. To test the soundness both of the annotation scheme and corresponding guidelines, reliability studies with nine independent annotators were carried out. As an illustration of the utility of our scheme, we have applied it to an already published study and verified whether the same conclusions hold. We hope that, in using our scheme, researchers will be able to save a lot of time and effort that, otherwise, would be spent in planning, developing and testing a scheme of their own.

1 Introduction

Despite the growing interest in emotion and sentiment analysis, recent work on this subject seems to focus mainly on evaluations (*e.g.* [9,3]) and the identification of the semantic orientation of words (*e.g.* [24,10]). With just a few exceptions (*e.g.* [19]), current research seems not to try to identify and classify behaviour and, more specifically, those actions people take that may raise some emotion either on the reader side, as in a text, or on some other conversational party, as in a dialogue.

On this account, empirical evidence for the importance of reporting behaviour and emotion in dialogue summaries has been nevertheless brought up to the community [18]. According to these results, whenever a dialogue presents a very impolite behaviour, as exhibited by any of its parties, human summarisers will tend to report that behaviour in the dialogue's summary, biasing it according to the assumed point of view, *i.e.*, the way some party was reported, which actual party was reported, and even if behaviour should be reported at all, depended to a great extent on the point of view taken by the summariser. Given these results, it becomes clear that automatic dialogue summarisers should deal with this kind of information, were they to build more human-like summaries.

* This research was sponsored by CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico – and CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

A. Kuri-Morales and G. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 386–395, 2010.

Currently, the most feasible alternative available to this end is to apply some corpus linguistics techniques on annotated data sources. To do so, however, it becomes necessary to develop annotation schemes capable of identifying the desired information in a reliable way. As such, in this paper we present a multi-dimensional annotation scheme developed to identify and classify information about emotional behaviour and bias in dialogue summaries. Tests with the scheme were run in a corpus composed of 240 dialogue summaries, produced by 30 independent summarisers [18]. In this corpus, summaries are separated both according to their viewpoint (customer, vendor or observer) and size constraint (unrestricted or limited to 10% of the number of words in the summarised dialogue). Source dialogues were automatic generated by NECA[1] – a platform for automatic generation of dialogues between conversational agents [7] – taking place in a car sales setup where some vendor tries to sell a car to a customer.

In order to demonstrate the usefullness of our annotation scheme, we have applied it to the same corpus used to test it, in an attempt to reproduce the experimental conclusions from [18]. In doing so, we hope to show other researchers the possibilities of our scheme, helping them to carry out their researches about emotion/behaviour reporting and bias without having to spend so much time and effort in developing and testing an annotation scheme of their own. The rest of this paper is organised as follows. Section 2 presents our annotation scheme, describing all dimensions that comprise it. The procedure followed by annotators to apply the scheme in the test corpus, along with its evaluation, is described in Section 3. Our conclusions to this work and venues for future research are presented in Section 5.

2 The Multi-dimensional Annotation Scheme

Following McKeown [14], in this research we take the clause as our basic unit, *i.e.* a unit consisting, as a minimum, of a verb and its complements [15]. As such, by following our scheme, every clause in a summary gets classified according to five distinct dimensions (Figure 1), which try to determine what (or who) was reported (or evaluated) in the clause, by whom and how. These dimensions are:

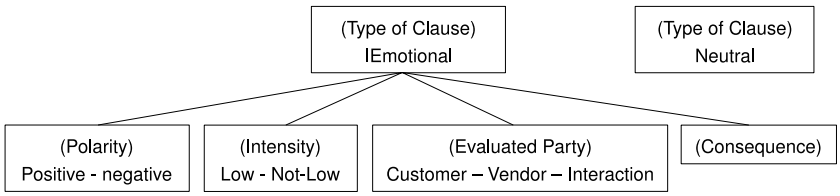

Fig. 1. Dimensions hierarchy in the annotation scheme

[1] Net Environment for Embodied Emotional Conversational Agents.

- *Type of Clause*: following Batliner *et al.* [4] and Fischer [8], this dimension distinguishes between *IEmotional* and *Neutral* clauses. In our research, however, we do not consider a clause as *IEmotional* whenever it presents some emotional feature but, instead, only when it assesses, either positively or negatively, the dialogue participants' politeness degree (more specifically, what Mills [16] called social politeness, or political behaviour), behaviour, humour, feelings and emotions towards some interactional feature (akin to what Keenan *et al.* [11] called the interactional content of the clause), or even when it evaluates the interaction's outcome as a whole.
- *Polarity*: this dimension results both from the theory proposed by Ortony, Clore and Collins [17] and the evaluation proposed by Martin [13], taking only two possible values – *Positive* or *Negative* – that must be applied only to *IEmotional* clauses. Under this dimension, a clause is classified as *Positive* when it describes pleasant actions or feelings, or even when it assesses them so, like in "She was patient" and "The service was fine". On the other hand, it is considered *Negative* when it describes actions or feelings that do not satisfy (and so must be avoided), or when it judges them so, like in "He is rude" or "Tossed all her insatisfaction on me".
- *Intensity*: designed to capture the way perceived emotions or behaviours were described, *i.e.*, towards the lower or higher end in an intensity scale [13], this dimension complements Polarity, in that it tries to determine whether some report was put in a considerably mild way, sounding almost like an euphemism (*Low* intensity) or, rather the opposite, if that demonstration was considered normal, or even above normal, indicating that what was described was taken as something important by the summariser (*Non-low* intensity).
- *Evaluated Party*: this dimension seeks to capture the dialogue participant whose behaviour or emotion was reported either explicitly, like in "the vendor treated me very well", or implicitly, like in "the service was very good", in which case a service necessarily implies a server. Since this dimension was designed primarily for sales dialogues, it takes only one of three possible values: (a) *Vendor*, meaning that the evaluation was made on the vendor's emotions and behaviour; (b) *Client*, when it accounts for the client's; and (c) *Interaction*, which must be used whenever the clause does not assess either of the dialogue participants but, instead, their interaction with each other (directly or indirectly), without particularly focusing in any of them, as in "That was a great sell"[2].
- *Consequence*: must be applied to *IEmotional* clauses describing situations or feeling that were caused by something described in another clause, as in "I was so badly served *that I lost my nerve*", where the later ("I lost my nerve") describes a consequence of the former ("I was so badly served"). More than determining a cause, however, this dimension aims to capture

[2] Clauses like this, however, cannot be taken as *IEmotional* if the reason for the sell being great lies solely on non interactional features, such as a good profit, for example.

the possible ways in which blame is transferred, by having one party justify his/her negative actions on the basis of the other party's behaviour.

Finally, clauses with multiple evaluations, such as "I gently served the rude client", for example, can take multiple classifications too. In this case, the clause may be classified as *IEmotional*, with a positive evaluation about the *Vendor* and a negative evaluation about the *Customer*, both with some *Intensity*.

3 Corpus Annotation and Evaluation

In order to test the soundness both of our annotation scheme and corresponding guidelines [2], we had nine independent annotators apply the scheme to a set of 240 human crafted summaries, comprising 1,773 clauses [18]. As an additional measure to increase reliability [12,2], annotators had to independently go through a training period of about one and a half hour. During this period, they were given a description of the annotation scheme, along with a set of guidelines to help them understand what categories meant (*cf.* [5]). They were then asked to annotate a set of 18 test summaries, with 128 clauses in total, artificially built for this task. Test summaries were written so as to provide a number of specially designed examples, making it easier for annotators to get the grips with the task.

In this research, we assessed our annotation scheme by measuring its reproducibility, *i.e.*, the extent to which annotators will produce the same classifications, working under varying conditions, at different times and locations [12]. To do so, we relied on Krippendorff's alpha as our coefficient of agreement. Ranging from -1 to 1, with 1 meaning total agreement and 0 implying that analysed data cannot be told apart from random events, α can be defined as

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o stands for the number of observed disagreements and D_e represents the number of expected disagreements by chance.

When it comes to subjective assessments, however, as is the case with emotion, behaviour and politeness, which are not so clearly defined, perfect agreement becomes much harder to be achieved [20,23,1]. Not to mention that annotating data depends to a great extent both on the personality and humour of its executors [1], which may vary considerably from time to time. In order to account for the effects of this subjectivity in data annotation, we decided to keep the number of categories minimum, by giving annotators fewer options[3].

Also, and as an attempt to reduce the cognitive load on annotators, we have unified dimensions *Type of Clause*, *Polarity* and *Evaluated Party* into a single dimension, thereby providing annotators with a broader picture of the annotation process, as opposed to having them separately focusing on three different dimensions and their categories. As such, instead of choosing, for example,

[3] This measure was inspired on an experiment carried out by Craggs e Wood [6], in which it was observed that adding a single category to a subject dimension dropped agreement from $\alpha = .52$ to $\alpha = 0.37$.

"IEmotional" for *Type of Clause*, "Positive" for *Polarity* and "Vendor" for *Evaluated Party*, annotators were shown "Positive Report about the Vendor" as an alternative. On this matter, it is worth noticing that this unification was made at the user interface level only, *i.e.*, since from "Positive Report about the Vendor" we can independently acquire the value for the three dimensions underneath, the overall model remains untouched.

Results for our annotation scheme are shown on Table 1. Assuming $\alpha > 0.8$ as good reliability, with $0.67 < \alpha < 0.8$ allowing tentative conclusions [12], only *Polarity* turns out as a reliable dimension, with *Type of Clause* and *Evaluated Party* coming next, at the tentative side. Although such values may seem rather disappointing, they are actually good, when compared to current results on emotion classification, which sometimes deliver around $\alpha = 0.6$ as their highest value (*cf.* [6]). Also, the existence of a "Neutral" category which, however necessary, has been described as a common source of confusion [1], just makes it harder to expect a high agreement on dimensions like *Type of Clause*, for example.

Table 1. Alpha values for the annotation

Dimension	α	*Dimension*	α
Polarity	0.843	Intensity	0.212
Evaluated Party	0.783	Consequence	0.085
Type of clause	0.674		

At the other end of the scale, *Intensity* and *Consequence* were found absolutely unreliable. That *Intensity* would score low was somewhat expected, given its variable definition amongst people [17]. As for *Consequence*, however, results showed people do not agree on the causes for the clause they were analysing. Actually, the figures do not get much better if we rule out the specific clause taken for a cause, *i.e.*, if we only keep information about whether some clause was considered as a consequence of another or not. In that case, agreement only becomes slightly better, at $\alpha = 0,175$, indicating that the very notion of cause/consequence, when it comes to emotional assessment, may be rather obscure.

4 Using the Scheme: A Practical Example

To illustrate a possible use of our annotation scheme, we applied it to the database described in [18], so as to identify reports on emotion and behaviour in dialogue summaries, also determining the existence of bias in these reports. Our utmost goal, with this example, is to demonstrate that, through our scheme, the same conclusions can be drawn, *i.e.*, we can still state that (i) if a dialogue contains very impolite behaviour, this behaviour tends to be reported in the dialogue summary; (ii) these reports are biased towards the point of view taken by the summariser; and (iii) severely restricting the summary length has no influence on the previous hypotheses. To do so, our first step was to assign to each

clause in the dataset the label given by the majority of annotators (*cf.* [21]). By way of caution, however, we found it more appropriate to consider majority only those labels chosen by over 50% of all annotators, having ties solved by one of the researchers, who should opt for one of the competing labels[4] (*cf.* [20,1]).

Notwithstanding, before proceeding with the data analysis we must define the categories used by [18] in terms of those of our scheme. In [18], human generated summaries were classified according to one out of three categories: *behav (excl)*, meaning that the summary solely comprised reports on some party's behaviour, thereby ignoring all technical information; *behav*, to be used with summaries delivering both reports on behaviour and technical information; and *non-behav*, representing those summaries dealing exclusively with the technical information exchanged by the dialogue participants. In that research, summaries were also grouped according to the point of view under which they were built and separated according to the maximum amount of words that summarisers were allowed to use[5].

Given this classification, we can connect our research to that of [18] by taking *behav (excl)* summaries to be those summaries composed uniquely of *IEmotional* clauses (as captured by dimension *Type of Clause*), whereas *non-behav* would deliver only *Neutral* clauses and *behav* would be a mix of both. Figure 2 illustrates the results from applying our annotation scheme to the dataset with no constraint on summary size, with summaries grouped according to the source dialogue (from 1 to 4) and point of view under which they were built (customer, vendor or observer). In this figure, we notice that the amount of *behav* summaries is considerably higher for the dialogues portraying impolite behaviour, *i.e.* dialogues 2 and 3 (83% and 97%, respectively), than for those more neutral (dialogues 1 – 37% – and 4 – 50%). This is a statistically significant difference ($\chi^2(1, N = 120) = 29.40$, at the significance level of $p = 0.001$).

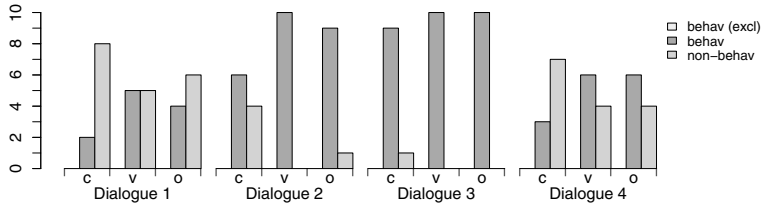

Fig. 2. Behaviour reports for the unrestricted set

When carrying out the same analysis in the restricted set, as illustrated on Figure 3, we once again verify that the amount of *behav (excl)* and *behav* summaries outnumber *non-behav* ones, to a considerable extent, in the second and third dialogues (63% and 90% respectively). By grouping these dialogues (*i.e.* the impolite) at one side and the first and fourth dialogues (neutral ones) at

[4] This problem was nevertheless found in only 69 of the 1,773 clauses, *i.e.* around 3.9% of them.

[5] Either 10% of the number of words in the source dialogue or no restriction at all.

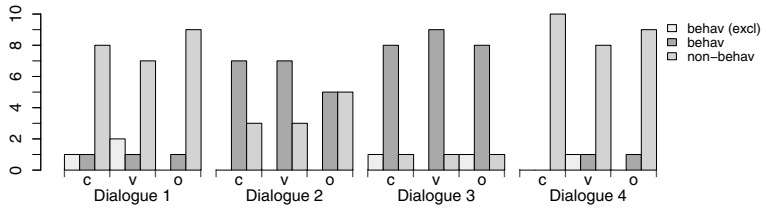

Fig. 3. Behaviour reports for the restricted set

the other, we can see that, even for summaries under the restricted condition, the number of behavioural and non-behavioural summaries depended on the politeness degree of the dialogue. This result was also statistically significant ($\chi^2(1, N = 120) = 45.95$, at the level $p = 0.001$). With these figures, we have confirmed hypothesis (i) and part of Hypothesis (iii) from [18].

Concerning hypothesis (ii), *i.e.* determining the existence of bias in the reports, bias was defined in [18] as 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" [22, pp. 86], *i.e.*, an argument failing to consider all relevant points from both sides. Thus, to determine the existence of bias in summaries, in [18] each summary was first classified either as an Exclusively Positive Report (EPR), *i.e.* a summary containing at least one sentence where some party was positively reported, with no negative reports on that party; or Exclusively Negative Report (ENR), when some party is reported negatively, without transfer of blame to some other agent, there also being no positive report about this party at all.

With our annotation scheme, these categories may be determined by taking together (i) *Type of Clause*, which is responsible for ruling *Neutral* clauses out of further consideration, since we are only interested in determining the presence of bias amongst emotional/behavioural reports; (ii) *Evaluated Party*, allowing us to identify towards whom the report was directed; and (iii) *Polarity*, so we can determine whether it was a positive or negative report. Hence, if, for example, all *IEmotional* clauses (as determined by *Type of Clause*) of some specific summary turn out to be Positive (at the *Polarity* dimension), with Vendor as the value for *Evaluated Party*, then this summary is taken to be an Exclusively Positive Report about the vendor[6].

Figure 4 shows the results for the unrestricted set of summaries, grouped by point of view. In this figure, it is clear the tendency customers have to report on the vendor's behaviour and vice-versa, *i.e.*, whose behaviour was negatively reported depended on the viewpoint assumed by the person reporting it ($\chi^2(2, N = 65) = 14.13$, at the significance level of $p = 0.001$), indicating the presence of a bias. When it comes to positive reports, however, no statistically significant relation between viewpoint and evaluated party was found ($\chi^2(2, N = 29) = 0.96$). Both results match those of [18].

[6] At this point, it is worth noticing that only the reliable and "tentative" dimensions were used in our analysis.

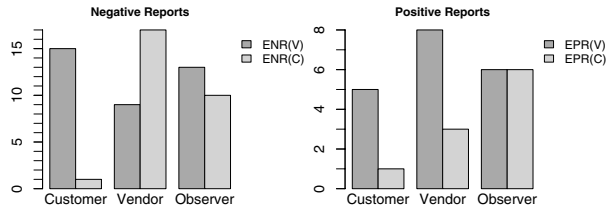

Fig. 4. Exclusively Positive and Negative Reports for the unrestricted set

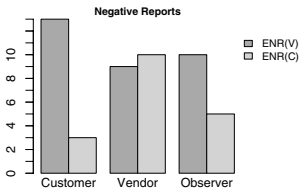

Fig. 5. Exclusively Negative Reports for the restricted set

As for the restricted set, figures change in that, contrary to what was predicted by [18], no relationship can be found between negatively reported behaviour and viewpoint ($\chi^2(2, N = 50) = 4.39$, at $p = 0.1$), as illustrated on Figure 5[7]. If, however, we only account for values under the customer and vendor viewpoints (and therefore put aside the neutral part, *i.e.* the observer), we end up with a statistically significant difference ($\chi^2(1, N = 35) = 4.27$, $p = 0.05$). The reason for this disparity might, in turn, rest on the low reliability of the *Consequence* dimension, which forced us to consider as ENR some summary that otherwise might not fall under this category, due to some undetected blame transfer. The fact that, in [18], the dataset was annotated by a single person, whereas in our research this annotation comes out from nine independent annotators, might have also played a part. Still, if we put all data together, and compare both unrestricted and restricted sets, we end up with 37 ENR(V) and 28 ENR(C) summaries, at the unrestricted side, and 32 ENR(V) and 18 ENR(C) at the restricted one, which is not significant at all ($\chi^2(1, N = 115) = 0.590, p > 0.25$), meaning that, as predicted in [18], considering a summary ENR(V) or ENR(C) did not depend on the summary length.

5 Conclusion

In this paper we presented a multi-dimensional annotation scheme for emotional and behavioural assessment in dialogue summaries. Reliability studies with nine independent annotators were carried out to test the soundness both of the annotation scheme and corresponding guidelines [2]. Results show that, from the

[7] In this figure, negative reports were left out because the dataset was too small to allow for any statistic analysis.

five original dimensions, three were reliable enough to allow for any conclusion to be drawn from the annotation, to wit, *Type of Clause*, *Evaluated Party* and *Polarity*. Unfortunately, *Intensity* and *Consequence* scored too low to deserve any credit.

To illustrate the utility of our scheme to other researchers, we have annotated data from an already published experiment (see [18]), verifying both (i) if it was possible to measure the same phenomena, and (ii) if the conclusions drawn with our annotation scheme would match those from that study. The comparison turned out to be successfull, in that our results confirmed all hypotheses setted up by [18]. We hope that, in using our scheme, researchers will be able to save a lot of time and effort that, otherwise, would be spent in planning, developing and testing a scheme of their own.

As for avenues for future work, we think it would be important determining a way to reliably measure intensity, since this is a feature considered in many theories about emotion (*e.g.* [17]). On this matter, simply adding intensity as a separate dimension definitely did not work for us. Finally, and as a measure to increase our confidence on the reproducibility of the scheme, it would be interesting having other volunteers apply it to a different dataset.

References

1. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: Machine learning for text-based emotion prediction. In: Proceedings of HLT/EMNLP 2005, Vancouver, Canada (2005)
2. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34(4), 555–596 (2008)
3. Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., Martínez-Barco, P.: Summarizing threads in blogs using opinion polarity. In: Proceedings of the Events in Emerging Text Types Workshop of the RANLP, Borovets, Bulgaria (September 18, 2009)
4. Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K.: The Recognition of Emotion. In: Foundations of Speech-to-Speech Translation, pp. 122–130. Springer, Heidelberg (2000)
5. Cohn, T., Callison-Burch, C., Lapata, M.: Constructing corpora for the development and evaluation of paraphrase systems. Computational Linguistics 34(4), 597–914 (2008)
6. Craggs, R., Wood, M.M.: A two dimensional annotation scheme for emotion in dialogue. In: AAAI Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications, Stanford, USA (March 2004); technical Report SS-04-07
7. van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., Baumann, S.: Fully generated scripted dialogue for embodied agents. Artificial Intelligence 172(10) (June 2008)
8. Fischer, K.: Annotating emotional language data. Tech. Rep. 236, Verbmobil Project (December 1999)
9. Hijikata, Y., Ohno, H., Kusumura, Y., Nishida, S.: Social summarization of text feedback for online auctions and interactive presentation of the summary. Knowledge-Based Systems 20(6), 527–541 (2007)

10. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004), San Jose, USA, pp. 755–760 (July 2004)
11. Keenan, J., MacWhinney, B., Mayhew, D.: Pragmatics in memory: A study of natural conversation. Journal of Verbal Learning and Verbal Behavior 16(5), 549–560 (1977)
12. Krippendorff, K.: Content Analysis: An Introduction to its Methodology, 2nd edn. Sage, Thousand Oaks (2004)
13. Martin, J.R.: Beyond Exchange: Appraisal Systems in English. In: Evaluation in Text: Authorial Stance and the Construction of Discourse, pp. 142–175. Oxford University Press, Oxford (1999)
14. McKeown, K.R.: Discourse strategies for generating natural-language text. Artificial Intelligence 27(1), 1–41 (1985)
15. Miller, J.: An Introduction to English Syntax. Edinburgh University Press Ltd., Edinburgh (2002) ISBN 0 7486 1254 8
16. Mills, S.: Gender and Politeness. Cambridge University Press, Cambridge (2003)
17. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
18. Roman, N.T., Piwek, P., Carvalho, A.M.B.R.: Politeness and Bias in Dialogue Summarization: Two Exploratory Studies. In: Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, vol. 20, pp. 171–185. Springer, Netherlands (January 9, 2006)
19. Spertus, E.: Smokey: Automatic recognition of hostile messages. In: Innovative Applications of Artificial Intelligence (IAAI 1997), pp. 1058–1065 (1997)
20. Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4), 315–346 (2003)
21. Vieira, R., Poesio, M.: An empirically based system for processing definite descriptions. Computational Linguistics 26(4), 539–593 (2000)
22. Walton, D.: One-Sided Arguments: A Dialectical Analysis of Bias. State University of New York Press (1999)
23. Watts, R.: Politeness. Cambridge University Press, Cambridge (2003)
24. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics 35(3), 399–433 (2009)