Dialogue Segmentation with Large Numbers of Volunteer Internet Annotators

T. Daniel Midgley
Discipline of Linguistics, School of Computer Science and Software Engineering
University of Western Australia
Perth, Australia
Daniel.Midgley@uwa.edu.au

Abstract

This paper shows the results of an experiment in dialogue segmentation. In this experiment, segmentation was done on a level of analysis similar to *adjacency pairs*. The method of annotation was somewhat novel: volunteers were invited to participate over the Web, and their responses were aggregated using a simple voting method. Though volunteers received a minimum of training, the aggregated responses of the group showed very high agreement with expert opinion. The group, as a unit, performed at the top of the list of annotators, and in many cases performed as well as or better than the best annotator.

1 Introduction

Aggregated human behaviour is a valuable source of information. The Internet shows us many examples of collaboration as a means of resource creation. Wikipedia, Amazon.com reviews, and Yahoo! Answers are just some examples of large repositories of information powered by individuals who voluntarily contribute their time and talents. Some NLP projects are now using this idea, notably the 'ESP Game' (von Ahn 2004), a data collection effort presented as a game in which players label images from the Web. This paper presents an extension of this collaborative volunteer ethic in the area of dialogue annotation.

For dialogue researchers, the prospect of using volunteer annotators from the Web can be an attractive option. The task of training annotators can be time-consuming, expensive, and (if inter-annotator agreement turns out to be poor) risky.

Getting Internet volunteers for annotation has its own pitfalls. Dialogue annotation is often not very interesting, so it can be difficult to attract willing participants. Experimenters will have little control over the conditions of the annotation and the skill of the annotators. Training will be minimal, limited to whatever an average Web surfer is willing to read. There may also be perverse or uncomprehending users whose answers may skew the data.

This project began as an exploratory study about the intuitions of language users with regard to dialogue segmentation. We wanted information about how language users perceive dialogue segments, and we wanted to be able to use this information as a kind of gold standard against which we could compare the performance of an automatic dialogue segmenter. For our experiment, the advantages of Internet annotation were compelling. We could get free data from as many language users as we could attract, instead of just two or three well-trained experts. Having more respondents meant that our results could be more readily generalised to language users as a whole.

We expected that multiple users would converge upon some kind of uniform result. What we found (for this task at least) was that large numbers of volunteers show very strong tendencies that correspond well to expert opinion, and that these patterns of agreement are surprisingly resilient in the face of noisy input from some users. We also gained some insights into the way that people perceived dialogue segments.

2 Segmentation

While much work in dialogue segmentation centers around topic (e.g. Galley et al. 2003, Hsueh et al. 2006, Purver et al. 2006), we decided to examine dialogue at a more fine-grained level. The level of analysis that we have chosen corresponds most closely to *adjacency pairs* (after Sacks, Schegloff and Jefferson 1974), where a segment is made of matched sets of utterances from different speakers (e.g. *question/answer* or *suggest/accept*). We chose to segment dialogues this way in order to improve dialogue act tagging, and we think that examining the back-and-forth detail of the mechanics of dialogue will be the most helpful level of analysis for this task.

The back-and-forth nature of dialogue also appears in Clark and Schaefer's (1989) influential work on *contributions* in dialogue. In this view, two-party dialogue is seen as a set of cooperative acts used to add information to the common ground for the purpose of accomplishing some joint action. Clark and Schaefer map these speech acts onto *contribution trees*. Each utterance within a contribution tree serves either to present some proposition or to acknowledge a previous one. Accordingly, each contribution tree has a *presentation phase* and an *acceptance phase*. Participants in dialogue assume that items they present will be added to the common ground unless there is evidence to the contrary. However, participants do not always show acceptance of these items explicitly. Speaker B may repeat Speaker's A's information verbatim to show understanding (as one does with a phone number), but for other kinds of information a simple 'uh-huh' will constitute adequate evidence of understanding. In general, less and less evidence will be required the farther on in the segment one goes.

In practice, then, segments have a tailing-off quality that we can see in many dialogues. Table 1 shows one example from Verbmobil-2, a corpus of appointment scheduling dialogues. (A description of this corpus appears in Alexandersson 1997.)

A segment begins when WJH brings a question to the table (utterances 1 and 2 in our example), AHS answers it (utterance 3), and WJH acknowledges the response (utterance 4). At this point, the question is considered to be resolved, and a new contribution can be issued. WJH starts a new segment in utterance 5, and this utterance shows features that will be familiar to dialogue researchers: the number of words increases, as does the incidence of new words. By the end of this segment (utterance 8), AHS only needs to offer a simple 'okay' to show acceptance of the foregoing.

1	WJH	<uhm> basically we have to be in Hanover for a day and a half
2	WJH	correct
3	AHS	right
4	WJH	okay
5	WJH	<uh> I am looking through my schedule for the next three months
6	WJH	and I just noticed I am working all of Christmas week
7	WJH	so I am going to do it in Germany if at all possible
8	AHS	okay

Table 1. A sample of the corpus. Two segments are represented here.

Our work is not intended to be a strict implementation of Clark and Schaefer's contribution trees. The segments represented by these units is what we were asking our volunteer annotators to find. Other researchers have also used a level of analysis similar to our own. Jönsson's (1991) initiative-response units is one example.

Taking a cue from Mann (1987), we decided to describe the behaviour in these segments using an atomic metaphor: dialogue segments have *nuclei*, where someone says something, and someone says something back (roughly corresponding to adjacency pairs), and *satellites*, usually shorter utterances that give feedback on whatever the nucleus is about.

For our annotators, the process was simply to find the nuclei, with both speakers taking part, and then attach any nearby satellites that pertained to the segment.

We did not attempt to distinguish nested adjacency pairs. These would be placed within the same segment. Eventually we plan to modify our system to recognise these nested pairs.

3 Experimental Design

3.1 Corpus

In the pilot phase of the experiment, volunteers could choose to segment up to four randomly-chosen dialogues from the Verbmobil-2 corpus. (One longer dialogue was separated into two.) We later ran a replication of the experiment with eleven dialogues. For this latter phase, each volunteer started on a randomly chosen dialogue to ensure evenness of responses.

The dialogues contained between 44 and 109 utterances. The average segment was 3.59 utterances in length, by our annotation.

Two dialogues have not been examined because they will be used as held-out data for the next phase of our research. Results from the other thirteen dialogues appear in part 4 of this paper.

3.2 Annotators

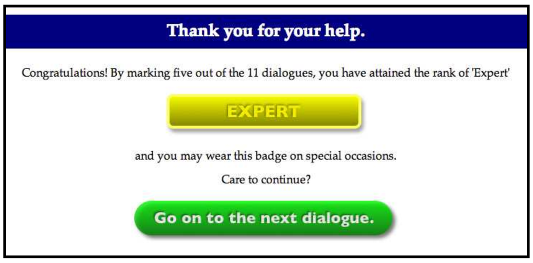
Volunteers were recruited via postings on various email lists and websites. This included a posting on the university events mailing list, sent to people associated with the university, but with no particular linguistic training. Linguistics first-year students and Computer Science students and staff were also informed of the project. We sent advertisements to a variety of international mailing lists pertaining to language, computation, and cognition, since these lists were most likely to have a readership that was interested in language. These included Linguist List, Corpora, CogLing-L, and HCSNet. An invitation also appeared on the personal blog of the first author.

At the experimental website, volunteers were asked to read a brief description of how to annotate, including the descriptions of nuclei and satellites. The instruction page showed some examples of segments. Volunteers were requested not to return to the instruction page once they had started the experiment.

The annotator guide with examples can be seen at the following URL:

`http://tinyurl.com/ynwmx9`

A scheme that relies on volunteer annotation will need to address the issue of motivation. People have a desire to be entertained, but dialogue annotation can often be tedious and difficult. We attempted humor as a way of keeping annotators amused and annotating for as long as possible. After submitting a dialogue, annotators would see an encouraging page, sometimes with pretend 'badges' like the one pictured in Figure 1. This was intended as a way of keeping annotators interested to see what comments would come next. Figure 2 shows statistics on how many dialogues were marked by any one IP address. While over half of the volunteers marked only one dialogue, many volunteers marked all four (or in the replication, all eleven) dialogues. Sometimes more than eleven dialogues were submitted from the same location, most likely due to multiple users sharing a computer.



Thank you for your help.

Congratulations! By marking five out of the 11 dialogues, you have attained the rank of 'Expert'

and you may wear this badge on special occasions.

Care to continue?

Figure 1. One of the screens that appears after an annotator submits a marked form.

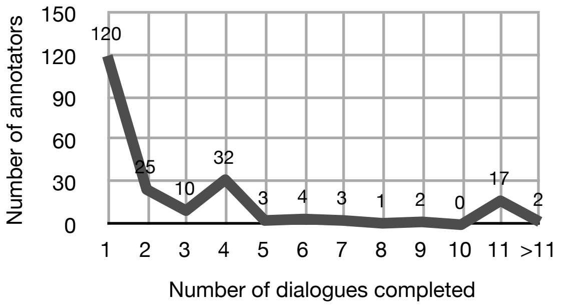


Figure 2. Number of dialogues annotated by single IP addresses

In all, we received 626 responses from about 231 volunteers (though this is difficult to determine from only the volunteers' IP numbers). We collected between 32 and 73 responses for each of the 15 dialogues.

3.3 Method of Evaluation

We used the WindowDiff (WD) metric (Pevzner and Hearst 2002) to evaluate the responses of our volunteers against expert opinion (our responses). The WD algorithm calculates agreement between a reference copy of the corpus and a volunteer's hypothesis by moving a window over the utterances in the two corpora. The window has a size equal to half the average segment length. Within the window, the algorithm examines the number of segment boundaries in the reference and in the hypothesis, and a counter is augmented by one if they disagree. The WD score between the reference and the hypothesis is equal to the number of discrepancies divided by the number of measurements taken. A score of 0 would be given to two annotators who agree perfectly, and 1 would signify perfect disagreement.

Figure 3 shows the WD scores for the volunteers. Most volunteers achieved a WD score between .15 and .2, with an average of .245.

Cohen's Kappa (κ) (Carletta 1996) is another method of comparing inter-annotator agreement

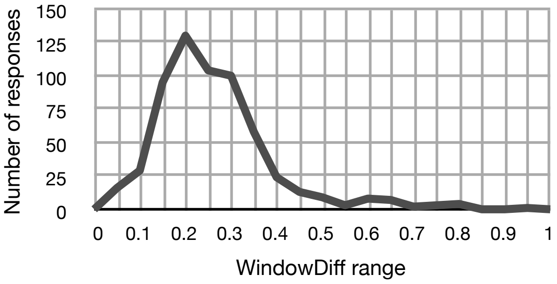


Figure 3. WD scores for individual responses. A score of 0 indicates perfect agreement.

in segmentation that is widely used in computational language tasks. It measures the observed agreement (A_O) against the agreement we should expect by chance (A_E), as follows:

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

For segmentation tasks, *κ* is a more stringent method than WindowDiff, as it does not consider near-misses. Even so, *κ* scores are reported in Section 4.

About a third of the data came from volunteers who chose to complete all eleven of the dialogues. Since they contributed so much of the data, we wanted to find out whether they were performing better than the other volunteers. This group had an average WD score of .199, better than the rest of the group at .268. However, skill does not appear to increase smoothly as more dialogues are completed. The highest performance came from the group that completed 5 dialogues (average WD = .187), the lowest from those that completed 8 dialogues (. 299).

3.4 Aggregation

We wanted to determine, insofar as was possible, whether there was a group consensus as to where the segment boundaries should go. We decided to try overlaying the results from all respondents on top of each other, so that each click from each respondent acted as a sort of vote. Figure 4 shows the result of aggregating annotator responses from one dialogue in this way. There are broad patterns of agreement; high ‘peaks’ where many annotators agreed that an utterance was a segment boundary, areas of uncertainty where opinion was split between two adjacent utterances, and some background noise from near-random respondents.

Group opinion is manifested in these peaks. Figure 5 shows a hypothetical example to illustrate how we defined this notion. A *peak* is any local maximum (any utterance *u* where *u - 1* < *u* > *u + 1*) above background noise, which we define as any utterance with a number of votes below the arithmetic mean. Utterance 5, being a local maximum, is a peak. Utterance 2, though a local maximum, is not a peak as it is below the mean. Utterance 4 has a comparatively large number of votes, but it is not considered a peak because its neighbour, utterance 5, is higher. Defining peaks this way allows us to focus on the points of highest agreement, while ignoring not only the relatively low-scoring utterances,

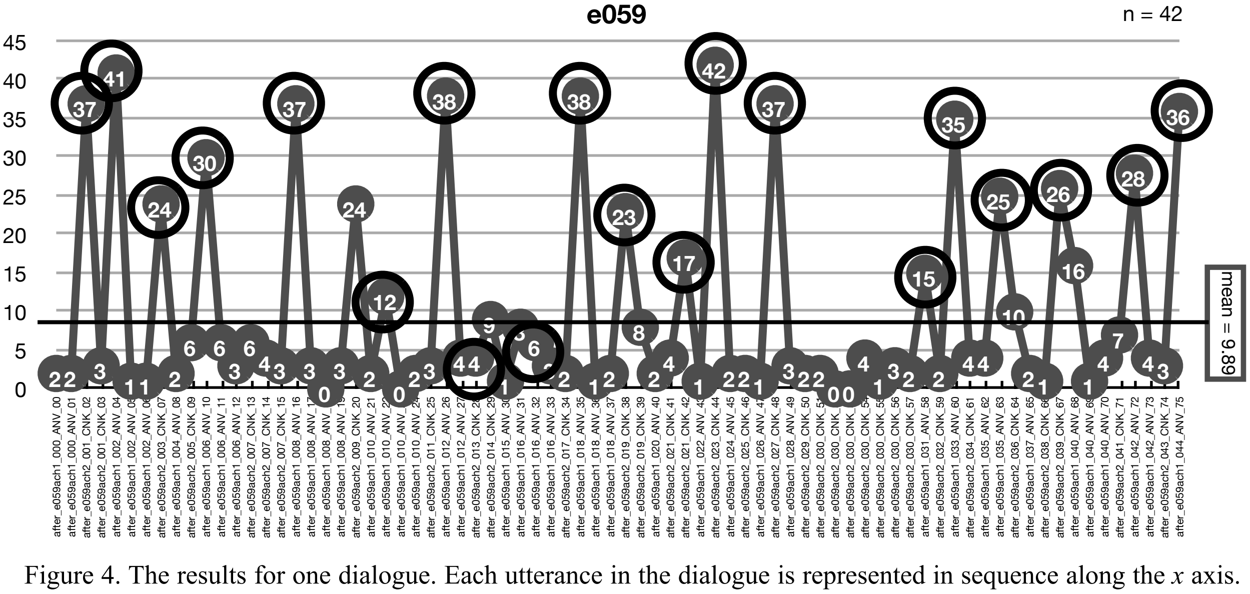


Figure 4. The results for one dialogue. Each utterance in the dialogue is represented in sequence along the *x* axis. Numbers in dots represent the number of respondents that ‘voted’ for that utterance as a segment boundary. Peaks appear where agreement is strongest. A circle around a data point indicates our choices for segment boundary.

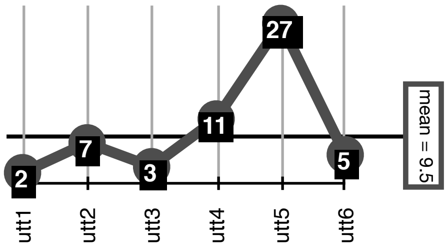


Figure 5. Defining the notion of 'peak'. Numbers in circles indicate number of 'votes' for that utterance as a boundary.

but also the potentially misleading utterances near a peak.

There are three disagreements in the dialogue presented in Figure 4. For the first, annotators saw a break where we saw a continuation. The other two disagreements show the reverse: annotators saw a continuation of topic as a continuation of segment.

4 Results

Table 2 shows the agreement of the aggregated group votes with regard to expert opinion. The aggregated responses from the volunteer annotators agree extremely well with expert opinion. Acting as a unit, the group's WindowDiff scores always perform better than the individual annotators on average. While the individual annotators attained an average WD score of .245, the annotators-as-group scored WD = .108.

On five of the thirteen dialogues, the group performed as well as or better than the best individual annotator. On the other eight dialogues, the group performance was toward the top of the group, bested by one annotator (three times), two annotators (once), four annotators (three times), or six annotators (once), out of a field of 32–73 individuals. This suggests that aggregating the scores in this way causes a 'majority rule' effect that brings out the best answers of the group.

One drawback of the WD statistic (as opposed to κ) is that there is no clear consensus for what constitutes 'good agreement'. For computational linguistics, $\kappa \geq .67$ is generally considered strong agreement. We found that κ for the aggregated group ranged from .71 to .94. Over all the dialogues, $\kappa = .84$. This is surprisingly high agreement for a dialogue-level task, especially considering the stringency of the κ statistic, and that the data comes from untrained volunteers, none of whom were dropped from the sample.

5 Comparison to Trivial Baselines

We used a number of trivial baselines to see if our results could be bested by simple means. These were random placement of boundaries, majority class, marking the last utterance in each turn as a boundary, and a set of hand-built rules we called 'the Trigger'. The results of these trials can be seen in Figure 6.

Dialogue name	WD average as marked by volunteers	WD single annotator best	WD single annotator worst	WD for group opinion	How many annotators did better?	Number of annotators
e041a	0.210	0.094	0.766	0.094	0	39
e041b	0.276	0.127	0.794	0.095	0	39
e059	0.236	0.080	0.920	0.107	1	42
e081a	0.244	0.037	0.611	0.148	4	36
e081b	0.267	0.093	0.537	0.148	4	32
e096a	0.219	0.083	0.604	-	-	32
e096b	0.160	0.000	0.689	0.044	1	36
e115	0.214	0.079	0.750	0.079	0	34
e119	0.241	0.102	0.610	-	-	32
e123a	0.259	0.043	1.000	0.174	6	34
e123b	0.193	0.093	0.581	0.047	0	33
e030	0.298	0.110	0.807	0.147	2	55
e066	0.288	0.063	0.921	0.063	0	69
e076a	0.235	0.026	0.868	0.053	1	73
e076b	0.270	0.125	0.700	0.175	4	40
ALL	**0.245**	**0.000**	**1.000**	**0.108**	**60**	**626**

Table 2. Summary of WD results for dialogues. Data has not been aggregated for two dialogues because they are being held out for future work.

5.1 Majority Class

This baseline consisted of marking every utterance with the most common classification, which was 'not a boundary'. (About one in four utterances was marked as the end of a segment in the reference dialogues.) This was one of the worst case baselines, and gave WD = .551 over all dialogues.

5.2 Random Boundary Placement

We used a random number generator to randomly place as many boundaries in each dialogue as we had in our reference dialogues. This method gave about the same accuracy as the 'majority class' method with WD = .544.

5.3 Last Utterance in Turn

In these dialogues, a speaker's turn could consist of more than one utterance. For this baseline, every final utterance in a turn was marked as the beginning of a segment, except when lone utterances would have created a segment with only one speaker.

This method was suggested by work from Sacks, Schegloff, and Jefferson (1974) who observed that the last utterance in a turn tends to be the first pair part for another adjacency pair. Wright, Poesio, and Isard (1999) used a variant of this idea in a dialogue act tagger, including not only the previous utterance as a feature, but also the previous speaker's last speech act type.

This method gave a WD score of .392.

5.4 The Trigger

This method of segmentation was a set of hand-built rules created by the author. In this method, two conditions have to exist in order to start a new segment.

- Both speakers have to have spoken.
- One utterance must contain four words or less.

The 'four words' requirement was determined empirically during the feature selection phase of an earlier experiment.

Once both these conditions have been met, the 'trigger' is set. The next utterance to have more than four words is the start of a new segment.

This method performed comparatively well, with WD = .210, very close to the average individual annotator score of .245.

As mentioned, the aggregated annotator score was WD = .108.

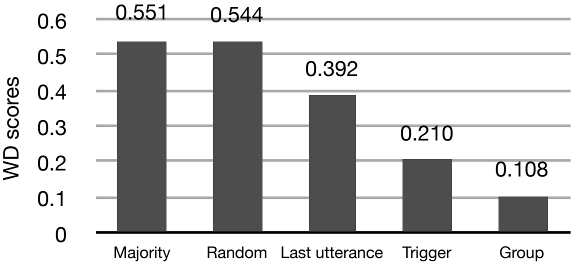


Figure 6. Comparison of the group's aggregated responses to trivial baselines.

5.5 Comparison to Other Work

Comparing these results to other work is difficult because very little research focuses on dialogue segmentation at this level of analysis. Jönsson (1991) uses initiative-response pairs as a part of a dialogue manager, but does not attempt to recognise these segments explicitly.

Comparable statistics exist for a different task, that of multiparty topic segmentation. WD scores for this task fall consistently into the .25 range, with Galley et al. (2003) at .254, Hsueh et al. (2006) at .283, and Purver et al. (2006) at .284. We can only draw tenuous conclusions between this task and our own, however this does show the kind of scores we should be expecting to see for a dialogue-level task. A more similar project would help us to make a more valid comparison.

6 Discussion

The discussion of results will follow the two foci of the project: first, some comments about the aggregation of the volunteer data, and then some comments about the segmentation itself.

6.1 Discussion of Aggregation

A combination of factors appear to have contributed to the success of this method, some involving the nature of the task itself, and some involving the nature of aggregated group opinion, which has been called 'the wisdom of crowds' (for an informal introduction, see Surowiecki 2004).

The fact that annotator responses were aggregated means that no one annotator had to perform particularly well. We noticed a range of styles among our annotators. Some annotators agreed very well with the expert opinion. A few annotators seemed to mark utterances in near-random ways. Some 'casual annotators' seemed to drop in, click only a few of the most obvious boundaries in the dialogue, and then submit the form. This kind of behaviour would give that annotator a disastrous individual score, but when aggregated, the work of the casual annotator actually contributes to the overall picture provided by the group. As long as the wrong responses are randomly wrong, they do not detract from the overall pattern and no volunteers need to be dropped from the sample.

It may not be surprising that people with language experience tend to arrive at more or less the same judgments on this kind of task, or that the aggregation of the group data would normalise out the individual errors. What is surprising is that the judgments of the group, aggregated in this way, correspond more closely to expert opinion than (in many cases) the best individual annotators.

6.2 Discussion of Segmentation

The concept of segmentation as described here, including the description of nuclei and satellites, appears to be one that annotators can grasp even with minimal training.

The task of segmentation here is somewhat different from other classification tasks. Annotators were asked to find segment boundaries, making this essentially a two-class classification task where each utterance was marked as either a boundary or not a boundary. It may be easier for volunteers to cope with fewer labels than with many, as is more common in dialogue tasks. The comparatively low perplexity would also help to ensure that volunteers would see the annotation through.

One of the outcomes of seeing annotator opinion was that we could examine and learn from cases where the annotators voted overwhelmingly contrary to expert opinion. This gave us a chance to learn from what the human annotators thought about language. Even though these results do not literally come from one person, it is still interesting to look at the general patterns suggested by these results.

1	MGT	so what time should we meet
2	ADB	<uh> well it doesn't matter as long as we both checked in I mean whenever we meet is kind of irrelevant
3	ADB	so maybe about try to
4	ADB	you want to get some lunch at the airport before we go
5	MGT	that is a good idea

Table 3. Example from a dialogue.

'let's see': This utterance usually appears near boundaries, but does it mark the end of a segment, or the beginning of a new one? We tended to place it at the end of the previous segment, but human annotators showed a very strong tendency to group it with the next segment. This was despite an example on the training page that suggested joining these utterances with the previous segment.

Topic: The segments under study here are different from topic. The segments tend to be smaller, and they focus on the mechanics of the exchanges rather than centering around one topic to its conclusion. Even though the annotators were asked to mark for adjacency pairs, there was a distinct tendency to mark longer units more closely pertaining to topic. Table 3 shows one example. We had marked the space between utterances 2 and 3 as a boundary; volunteers ignored it. It was slightly more common for annotators to omit our boundaries than to suggest new ones. The average segment length was 3.64 utterances for our volunteers, compared with 3.59 utterances for experts.

Areas of uncertainty: At certain points on the chart, opinion seemed to be split as one or more potential boundaries presented themselves. This seemed to happen most often when two or more of the same speech act appeared sequentially, e.g. two or more questions, information-giving statements, or the like.

7 Conclusions and Future Work

We drew a number of conclusions from this study, both about the viability of our method, and about the outcomes of the study itself.

First, it appears that for this task, aggregating the responses from a large number of anonymous volunteers is a valid method of annotation. We would like to see if this pattern holds for other kinds of classification tasks. If it does, it could have tremendous implications for dialogue-level annotation. Reliable results could be obtained quickly and cheaply from large numbers of volunteers over the Internet, without the time, the expense, and the logistical complexity of training. At present, however, it is unclear whether this volunteer annotation technique could be extended to other classification tasks. It is possible that the strong agreement seen here would also be seen on *any* two-class annotation problem. A retest is underway with annotation for a different two-class annotation set and for a multi-class task.

Second, it appears that the concept of segmentation on the adjacency pair level, with this description of nuclei and satellites, is one that annotators can grasp even with minimal training. We found very strong agreement between the aggregated group answers and the expert opinion.

We now have a sizable amount of information from language users as to how they perceive dialogue segmentation. Our next step is to use these results as the corpus for a machine learning task that can duplicate human performance. We are considering the Transformation-Based Learning algorithm, which has been used successfully in NLP tasks such as part of speech tagging (Brill 1995) and dialogue act classification (Samuel 1998). TBL is attractive because it allows one to start from a marked up corpus (perhaps the Trigger, as the best-performing trivial baseline), and improves performance from there.

We also plan to use the information from the segmentation to examine the structure of segments, especially the sequences of dialogue acts within them, with a view to improving a dialogue act tagger.

Acknowledgements

Thanks to Alan Dench and to T. Mark Ellison for reviewing an early draft of this paper. We especially wish to thank the individual volunteers who contributed the data for this research.

References

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 319–326.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997. *Dialogue acts in VERBMOBIL-2*. Verbmobil Report 204, DFKI, University of Saarbruecken.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4): 543–565.

Jean C. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2): 249–254.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Michael Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 562–569.

Pei-Yun Hsueh, Johanna Moore, and Steve Renals. 2006. Automatic segmentation of multiparty dialogue. In *Proceedings of the EACL 2006*, pp. 273–280.

Arne Jönsson. 1991. A dialogue manager using initiative-response units and distributed control. In *Proceedings of the Fifth Conference of the European Association for Computational Linguistics*, pp. 233–238.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. In *IPRA Papers in Pragmatics* 1: 1-21.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1): 19–36.

Matthew Purver, Konrad P. Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 17–24.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of COLING/ACL'98*, pp. 1150–1156.

James Surowiecki. 2004. *The wisdom of crowds: Why the many are smarter than the few*. Abacus: London, UK.

Helen Wright, Massimo Poesio, and Stephen Isard. 1999. Using high level dialogue information for dialogue act recognition using prosodic features. In *DIAPRO-1999*, pp. 139–143.