Empirical Verification of Adjacency Pairs Using Dialogue Segmentation

T. Daniel Midgley **Shelly Harrison**
Discipline of Linguistics,
University of Western Australia
(dmidgley, shelley)@cyllene.uwa.edu.au

Cara MacNish
School of Computer Science and
Software Engineering,
University of Western Australia
cara@csse.uwa.edu.au

Abstract

A problem in dialogue research is that of finding and managing expectations. Adjacency pair theory has widespread acceptance, but traditional classification features (in particular, 'previous-tag' type features) do not exploit this information optimally. We suggest a method of dialogue segmentation that verifies adjacency pairs and allows us to use dialogue-level information within the entire segment and not just the previous utterance. We also use the χ^2 test for statistical significance as 'noise reduction' to refine a list of pairs. Together, these methods can be used to extend expectation beyond the traditional classification features.

1 Introduction

Adjacency pairs have had a long history in dialogue research. The pairs of question/answer, inform/backchannel, and others have been well-known ever since they were proposed by Sacks and Schegloff in 1973. They have been used by dialogue researchers to assist in knowing 'what comes next' in dialogue.

Unfortunately, this dialogue information has been difficult to leverage. Most dialogue act (DA) classification research uses some kind of dialogue history, but this usually takes the form of some kind of 'previous tag' feature, perhaps even 'two-previous tag'. Dialogue information from three or more utterances previous is not normally used because, in the words of one researcher, "[n]o benefit was found from using higher-order dialog grammars" (Venkataraman et al. 2002). This could be due to the sparse data problem; more permutations means fewer repetitions.

Part of the problem, then, may lie in the way the 'previous tag' feature is used. Consider the following example from the Verbmobil-2 corpus (Verbmobil 2006)[1]:

A:	how does does November fourteenth and fifteenth look	SUGGEST
B:	no	REJECT

Here, the second pair part occurs directly after the first pair part that occasioned it. But sometimes performance factors intervene as in the following example, where B is engaging in floor-holding using a dialogue act annotated here as DELIBERATE:

A:	so that maybe I if I need to if I need to order like a limo or something	SUGGEST
B:	<hes> let us see	DELIBERATE
B:	the this is the <hes> wrong month	DELIBERATE
B:	the third	DELIBERATE
B:	let us see	DELIBERATE
B:	I don't have anything scheduled that morning and we are leaving at one	INFORM

The response (INFORM) finally comes, but the forgetful 'previous tag' feature is now looking for what comes after DELIBERATE.

What is needed is a way to not only determine what is likely to happen next, but to retain that expectation over longer distances when unfulfilled, until that expectation is no longer needed. Such information would conform more closely to this description of a conversational game (but which could be applied to any communicative subgoal):

[1]For a full description of the Verbmobil speech acts, see Alexandersson 1997.

Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pages 104–108,
Sydney, July 2006.

> A conversational game is a sequence of moves starting with an initiation and encompassing all moves up *until that initiation's purpose is either fulfilled or abandoned*. (Carletta 1997, italics mine.)

2 Dialogue segmentation

This work grew out of related research into finding expectations in dialogue, but we were also interested in dialogue segmentation. Dialogues taken as a whole are very different from each other, so segmentation is necessary to derive meaningful information about their parts. The question is, then, how best to segment dialogues so as to reveal dialogue information or to facilitate some language task, such as DA classification?

Various schemes for dialogue segmentation have been tried, including segmentation based on fulfilment of expectation (Ludwig et al. 1998), and segmenting by propositionality (Midgley 2003).

One answer to the question of how to segment dialogue came from the pioneering work of Sacks and Schegloff (1973) article.

> A basic rule of adjacency pair operation is: given the recognizable production of a *first pair part*, on its first possible completion its speaker should stop and a next speaker should start and produce a *second pair part* from the same pair type of which the first is recognizably a member. (p. 296, italics mine.)

Thus, if a speaker stops speaking, it is likely that such a handover has just taken place. The last utterance of a speaker's turn, then, will be the point at which the first speaker has issued a first pair part, and is now expecting a second pair part from the other speaker. This suggests a natural boundary.

This approach was also suggested by Wright (1998), who used a "most recent utterance by previous speaker" feature in her work on DA tagging. This feature alone has boosted classification accuracy by about 2% in our preliminary research, faring better than the traditional 'previous tag' feature used in much DA tagging work.

We collected a training corpus of 40 English-speaking dialogues from the Verbmobil-2 corpus, totalling 5,170 utterances. We then segmented the dialogues into *chunks*, where a chunk included everything from the last utterance of one speaker's turn to the last-but-one utterance of the next speaker.

3 Results of segmentation

This segmentation revealed some interesting patterns. When ranked by frequency, the most common chunks bear a striking resemblance to the adjacency pairs posited by Schegloff and Sacks.

Here are the 25 most common chunks in our training corpus, with the number of times they appeared. The full list can be found at http://www.csse.uwa.edu.au/~fontor/research/chi/fullseg.txt

```
SUGGEST:ACCEPT                          176
INFORM:FEEDBACK_POSITIVE                166
FEEDBACK_POSITIVE:FEEDBACK_POSITIVE
                                        104
FEEDBACK_POSITIVE:INFORM                 97
ACCEPT:FEEDBACK_POSITIVE                 65
FEEDBACK_POSITIVE:SUGGEST                60
INFORM:INFORM                            57
REQUEST:INFORM                           46
INFORM:BACKCHANNEL                       41
INFORM:SUGGEST                           40
REQUEST_COMMENT:FEEDBACK_POSITIVE        40
INIT:FEEDBACK_POSITIVE                   35
BYE:NONE                                 34
ACCEPT:INFORM                            32
BYE:BYE                                  31
REQUEST:FEEDBACK_POSITIVE                30
POLITENESS_FORMULA:FEEDBACK_POSITIVE
                                         29
REQUEST_CLARIFY:FEEDBACK_POSITIVE        28
BACKCHANNEL:INFORM                       28
NOT_CLASSIFIABLE:INFORM                  28
REQUEST_SUGGEST:SUGGEST                  28
NONE:GREET                               27
SUGGEST:SUGGEST                          27
ACCEPT:SUGGEST                           26
SUGGEST:REQUEST_CLARIFY                  26
```

The data suggest a wide variety of language behaviour, including traditional adjacency pairs (e.g. `SUGGEST: ACCEPT`), acknowledgement (`INFORM: BACKCHANNEL`), formalised exchanges (`POLITENESS_FORMULA: FEEDBACK_POSITIVE`) offers and counter-offers (`SUGGEST: SUGGEST`), and it even hints at negotiation subdialogues (`SUGGEST: REQUEST_CLARIFY`).

However, there are some drawbacks to this list. Some of the items are not good examples of adjacency pairs because the presence of the first does not create an expectation for the second half (e.g. `NOT_CLASSIFIABLE: INFORM`). In

some cases they appear backwards (ACCEPT: SUGGEST). Legitimate pairs appear further down the list than more-common bogus ones. For example, SUGGEST: REJECT is a well-known adjacency pair, but it does not appear on the list until after several less-worthy-seeming pairs. Keeping the less-intuitive chunks may help us with classification, but it falls short of providing empirical verification for pairs.

What we need, then, is some kind of noise reduction that will strain out spurious pairs and bring legitimate pairs closer to the top of the list.

We use the well-known χ^2 test for statistical significance.

4 The χ^2 test

The χ^2 test tells how the observed frequency of an event compares with the expected frequency. For our purposes, it tells whether the observed frequency of an event (in this case, one kind of speech act following a certain other act) can be attributed to random chance. The test has been used for such tasks as feature selection (Spitters 2000) and translation pair identification (Church and Gale 1991).

The χ^2 value for any two speech acts *A* and *B* can be calculated by counting the times that an utterance marked as tag A (or not) is followed by an utterance marked as tag B (or not), as in Table 1.

	$U_i = A$	$U_i \neq A$
$U_{i+1} = B$	AB	$\neg AB$
$U_{i+1} \neq B$	$A\neg B$	$\neg A\neg B$

Table 1. Obtaining counts for χ^2.

These counts (as well as *N*, the total number of utterances) are plugged into a variant of the χ^2 equation used for 2x2 tables, as in Schütze et al. (1995).

$$\chi^2 = \frac{N(AB \cdot \neg A\neg B - A\neg B \cdot \neg AB)}{(AB + A\neg B)(AB + \neg AB)(A\neg B + \neg A\neg B)(\neg AB + \neg A\neg B)}$$

We trained the χ^2 method on the aforementioned chunks. Rather than restrict our focus to only adjacent utterances, we allowed a match for pair A:B if B occurred *anywhere* within the chunk started by A. By doing so, we hoped to reduce any acts that may have been interfering with the adjacency pairs, especially hesitation noises (usually classed as DELIBERATE) and abandoned utterances (NOT_CLASSIFIABLE).

5 Results for χ^2

Here are the 25 pairs with the highest χ^2 scores. With tail probability $p = .0001$, a χ^2 value > 10.83 is statistically significant. The full list can be found at http://www.csse.uwa.edu.au/~fontor/research/chi/fullchi.txt.

```
NONE:GREET                          1576.87
BYE:NONE                             949.89
SUGGEST:ACCEPT                       671.81
BYE:BYE                              488.60
NONE:POLITENESS_FORMULA              300.46
POLITENESS_FORMULA:
POLITENESS_FORMULA                   272.95
GREET:GREET                          260.69
REQUEST_CLARIFY:CLARIFY              176.63
CLARIFY:CLARIFY                      165.76
DEVIATE_SCENARIO: DEVIATE_SCENARIO
                                     159.45
SUGGEST:FEEDBACK_POSITIVE            158.12
COMMIT:COMMIT                        154.46
GREET:POLITENESS_FORMULA             111.19
INFORM:FEEDBACK_POSITIVE              84.82
REQUEST_SUGGEST:SUGGEST               83.17
SUGGEST:REJECT                        83.11
THANK:THANK                           76.25
SUGGEST:EXPLAINED_REJECT              69.31
POLITENESS_FORMULA:INIT               67.76
NONE:INIT                             59.97
FEEDBACK_POSITIVE:ACCEPT              59.41
DEFER:ACCEPT                          56.07
THANK:BYE                             51.82
POLITENESS_FORMULA:THANK              50.21
POLITENESS_FORMULA:GREET              45.17
```

Using χ^2 normalises the list; low-frequency acts like REJECT and EXPLAINED_REJECT now appear as a part of their respective pairs.

These results give empirical justification for Sacks and Schegloff's adjacency pairs, and reveals more not mentioned elsewhere in the literature, such as DEFER:ACCEPT. As such, it gives a good idea of what kinds of speech acts are expected within a chunk.

In addition, these results can be plotted into a directed acyclic graph (seen in Figure 1). This graph can be used as a sort of conversational map.

6 Conclusions and Future Work

We can draw some tentative conclusions from this work. First of all, the dialogue segmentation combined with the χ^2 test for significance yields information about what is likely to happen, not just for the next utterance, but somewhere in the next chunk. This will help to overcome the limitations imposed by the traditional 'previous

tag' feature. We are working to implement this information into a model where the expectations inherent in a first pair part are retained when not immediately fulfilled. The expectations will also decay with time.

Second, this approach provides empirical evidence for adjacency pairs mentioned in the literature on conversation analysis. The noise reduction feature of the χ^2 test gives more weight to legitimate adjacency pairs where they appear in the data.

An intriguing possibility for the chunked data is that of *chunk matching*. Nearest-neighbour algorithms are already used for classification tasks (including DA tagging for individual utterances), but once segmented, the dialogue chunks could be compared against each other as a classification tool as in a nearest-neighbour algorithm.

References

J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. 1997. *Dialogue acts in Verbmobil-2*. Verbmobil Report 204.

J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13--31.

K. W. Church and W. A. Gale. 1991. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, Oxford.

D. Midgley. 2003. Discourse chunking: a tool for dialogue act tagging. In *ACL-03 Companion Volume to the Proceedings of the Conference*, pages 58–63, Sapporo, Japan.

E. A. Schegloff. and H. Sacks. 1973. Opening up closings. *Semiotica*, 8(4):289–327.

H. Schütze, D. Hull, and J. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR '95*, pages 229–237.

M. Spitters. 2000. "Comparing feature sets for learning text categorization." In *Proceedings of RIAO 2000*, April, 2000.

A. Venkataraman, A. Stolcke, E. Shriberg. Automatic dialog act labeling with minimal supervision. In *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, Melbourne, Australia, 2002.

Verbmobil. 2006. "Verbmobil" [online]. Available <http://verbmobil.dfki.de/>.

H. Wright. 1998. Automatic utterance type detection using suprasegmental features. In *ICSLP (International Conference on Spoken Language Processing) '98*. Sydney, Australia.

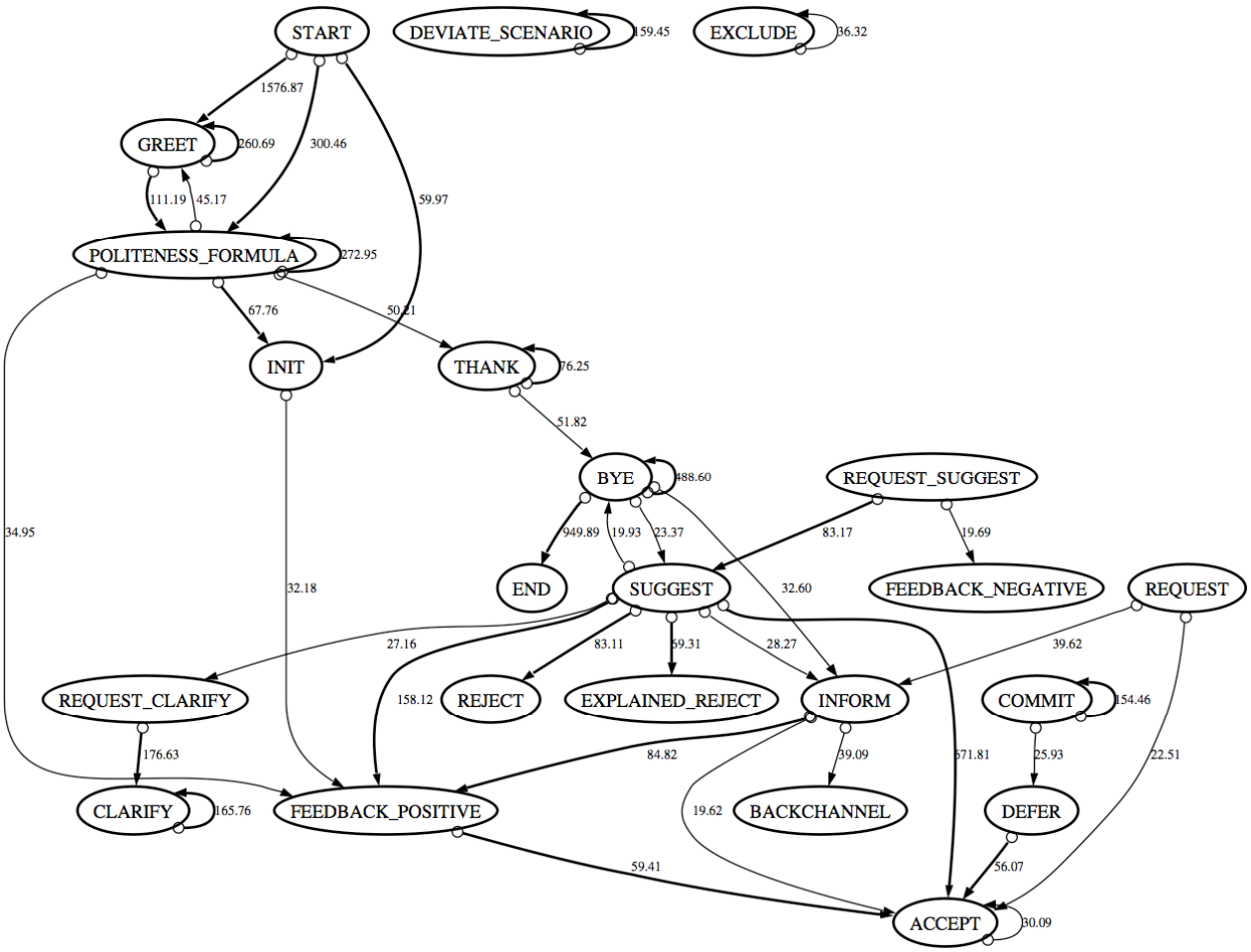

Figure 1. A directed acyclic graph using the χ^2 data for the 40 highest pairs. For any pair of connected nodes, the first node represents the last utterance in a speaker's turn, and the second could be any utterance in the other speaker's turn. The numbers are χ^2 scores. For illustrative purposes, higher χ^2 values are shown by bold lines. The complete graph can be found at http://www.csse.uwa.edu.au/~fontor/research/chi/fullchart.jpg