Discourse chunking: a tool in dialogue act tagging

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

Abstract

Discourse chunking is a simple way to segment dialogues according to how dialogue participants raise topics and negotiate them. This paper explains a method for arranging dialogues into chunks, and also shows how discourse chunking can be used to improve performance for a dialogue act tagger that uses a case-based reasoning approach.

1 Dialogue act tagging

A dialogue act (hereafter DA) is an encapsulation of the speaker's intentions in dialogue—what the speaker is trying to accomplish by saying something. In DA tagging (similar to part-of-speech tagging), utterances in a dialogue are tagged with the most appropriate speech act from a tagset. DA tagging has application in NLP work, including speech recognition and language understanding.

The Verbmobil-2 corpus was used for this study, with its accompanying tagset, shown in Table 1.1.

Much of the work in DA tagging (Reithinger, 1997; Samuel, 2000; Stolcke et al. 2000; Wright, 1998) uses lexical information (the words or *n*-grams in an utterance), and to a lesser extent syntactic and phonological information (as with prosody). However, there has traditionally been a lack of true discourse-level information in tasks involving dialogue acts. Discourse information is typically limited to looking at surrounding DA tags (Reithinger, 1997; Samuel, 2000). Unfortunately, knowledge of prior DA tags does not always translate to an accurate guess of what's coming next, especially when this information is imperfect.

Theories about the structure of dialogue (for example, centering [Grosz, Joshi, & Weinstein 1995], and more recently Dialogue Macrogame Theory [Mann 2002]) have not generally been applied to the DA tagging task. Their use amounts to a separate tagging task of its own, with the concomitant time-consuming corpus annotation.

In this work, I present the results from a DA tagging project that uses a case-based reasoning system (after Kolodner 1993). I show how the results from this DA tagger are improved by the use of a concept I call "discourse chunking." Discourse chunking gives information about the patterns of topic raising and negotiation in dia-

Tag	Example
ACCEPT	sounds good to me
BACKCHANNEL	mhm
BYE	see you
CLARIFY	I said the third
CLOSE	okay <uhm> so I guess that is it
COMMIT	I will get that arranged then
CONFIRM	well I will see you <uhm> at the airport on the third
DEFER	and I will get back to you on that
DELIBERATE	so let us see
DEVIATE_SCENARIO	oh I have tickets for the opera on Friday
EXCLUDE	January is basically shot for me
EXPLAINED_REJECT	I am on vacation then
FEEDBACK	gosh
FEEDBACK_NEGATIVE	not really
FEEDBACK_POSITIVE	okay
GIVE_REASON	because that is when the express flights are
GREET	hello Miriam
INFORM	<uhm> I I have a list of hotels here
INIT	so we need to schedule a trip to Hanover
INTRODUCE	Natalie this is Scott
NOT_CLASSIFIABLE	and <uh>
OFFER	<uhm> would you like me to call
POLITENESS_FORMULA	good of you to stop by
REFER_TO_SETTING	want to step into your office since we are standing right outside of it
REJECT	no that is bad for me unfortunately
REQUEST	you think so?
REQUEST_CLARIFY	I thought we had said twelve noon
REQUEST_COMMENT	is that alright with you
REQUEST_COMMIT	can you take care of <uhm> arranging those reservations
REQUEST_SUGGEST	do you have any preference
SUGGEST	we could travel on a Monday
THANK	okay thanks John

Table 1.1. The tagset for the Verbmobil-2 corpus. (Verbmobil 2003)

logue, and where an utterance fits within these patterns. It is also able to use existing DA tag information within the corpus, without the need for separate annotation.

2 Discourse chunking

In order to accomplish a mutual goal (for example, two people trying to find a suitable appointment time), dialogue participants engage in predictable kinds of activity, structuring the conversation in a coherent way in order to accomplish their goals.

Alexandersson et al. (1997) have noted that these conversations tend to follow certain patterns, particularly with regard to the way that topics get raised and dealt with:

Hello The dialogue participants greet each other. They introduce themselves, unveil their affiliation, or the institution or location they are from.
Opening The topic to be negotiated is introduced.
Negotiation The actual negotiation, between opening and closing.
Closing The negotiation is finished (all participants have agreed), and the agreed-upon topic is (sometimes) recapitulated.
Good Bye The dialogue participants say good bye to each other.

Within a conversation, the opening-negotiation-closing steps are often repeated in a cyclical pattern.

This work on discourse chunking combines the opening, negotiation, and closing sections into a single chunk. One reason for this is that these parts of the conversation tend to act as a single chunk; when they appear, they regularly appear together and in the same order. Also, some of these parts may be missing; a topic of negotiation is frequently brought up and resolved without an explicit opening or closing. Very often, the act of beginning a topic of negotiation defines the opening by itself, and the act of beginning a new negotiation entails the closing of the previous one.

A slightly simplified model of conversation, then, appears in Figure 2.1.

In this model, participants greet each other, engage in a series of negotiations, and finish the conversation when the goals of the dialogue are satisfied.

These three parts of the conversation are "dialogue chunks". These chunks are relevant from a DA tagging perspective. For example, the DA tags used in one of these chunks are often not used in

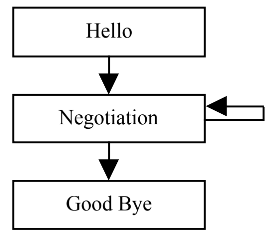

Figure 2.1. A slightly simplified model of conversation.

other chunks. For an obvious example, it would be almost unheard of for the `GREET` tag to appear in the "Good Bye" chunk. Other DA's (such as `FEEDBACK_POSITIVE`) can occur in any of the three chunks. Knowing which chunk we are in, and where we are within a chunk, can facilitate the tagging task.

Within chunks, some patterns emerge. Note that in the example from the Verbmobil-2 corpus (shown in Table 2.1), a negotiation topic is raised, and dealt with (by an `ACCEPT` speech act). Then there follows a sequence of `FEEDBACK_POSITIVE`s as the negotiation topic winds down. This "winding down" activity is common at the end of a negotiation chunk. Then a new topic is raised, and the process continues.

One-word utterances such as "okay" or "yeah" are particularly problematic in this kind of task because they have rather general semantic content and they are commonly used in a wide range of contexts. The word "yeah" on its own, for example, can indicate acceptance of a proposition, mere

Speaker ID	Words	DA Tag
KNT	some other time oh actually I see that I have got some free time in like the fifth sixth and seventh of January	SUGGEST
KNT	how does that	NOT_CLASSIFIABLE
LMT	yeah that is fine	ACCEPT
KNT	great so let us do that then	FEEDBACK_POSITIVE
LMT	okay	FEEDBACK_POSITIVE
KNT	okay	FEEDBACK_POSITIVE
LMT	okay good	FEEDBACK_POSITIVE

Table 2.1 An example of tagged conversation from the Verbmobil-2 corpus.

acknowledgement of a proposition, feedback, deliberation, or a few of these at once (Core & Allen 1997). In Verbmobil-2, these utterances can be labeled either `ACCEPT`, `FEEDBACK_POSITIVE`, `BACK-CHANNEL`, or `REQUEST_COMMENT`. Without knowing where the utterance appears within the structure of the dialogue, these utterances are very difficult to classify.

Some previous work has used prosody to solve this kind of problem (as with Stolcke 2000). I propose discourse chunks as an alternative method. It can pull information from the text alone, without the computational overhead that prosody can entail.

3 Chunk segmentation

Just where do the discourse chunk boundaries lie? For this exercise, I have constructed a very simple set of rules to determine chunk boundaries. These rules come from my observations; future work will involve automatic chunk segmentation. However, these rules do arise from a principled assumption: the raising of a new topic shows the beginning of a discourse chunk. Therefore, a speech act that (according to the definitions in Alexandersson 1997) contains a topic or proposition represents the beginning of a discourse chunk.

By definition, only four DA's contain or may contain a topic or proposition. These are `INIT`, `EXCLUDE`, `REQUEST_SUGGEST`, and `SUGGEST`.

Spkr ID	Words	Discourse Chunk	DA Tag
KNT	some other time oh actually I see that I have got some free time in like the fifth sixth and seventh of January	1	SUGGEST
KNT	how does that	17.5	NOT_CLASSIFIABLE
LMT	yeah that is fine	34	ACCEPT
KNT	great so let us do that then	50.5	FEEDBACK_POSITIVE
LMT	okay	67	FEEDBACK_POSITIVE
KNT	okay	83.5	FEEDBACK_POSITIVE
LMT	okay good	100	FEEDBACK_POSITIVE

Table 3.1 An example from the corpus, now tagged with discourse chunks.

3.1 Chunking rules

The chunking rules are as follows:

1. The first utterance in a dialogue is always the start of chunk 1 (hello).

2. The first `INIT` or `SUGGEST` or `REQUEST_SUGGEST` or `EXCLUDE` in a dialogue is the start of chunk 2 (*negotiation*).

3. `INIT`, `SUGGEST`, `REQUEST_SUGGEST`, or `EXCLUDE` marks the start of a subchunk within chunk 2.

4. If the previous utterance is also the start of a chunk, and if it is spoken by the same person, then this utterance is considered to be a continuation of the chunk, and is not marked.

5. The first `BYE` is the start of chunk 3 (*good bye*).

Items within a chunk are numbered evenly from 1 (the first utterance in a chunk) to 100 (the last), as shown in Table 3.1. This normalizes the chunk distances to facilitate comparison between utterances.

4 The case-based reasoning (CBR) tagger

A thorough discussion of this CBR tagger goes beyond the scope of this paper, but a few comments are in order.

Case-based reasoning (Kolodner 1993) is a form of machine learning that uses examples. In general, classification using a case-based reasoner involves comparing new instances (in this case, utterances) against a database of correctly-tagged instances. Each new instance is marked with the same tag of its "nearest neighbour" (that is, the closest match) from the database. A *k*-nearest neighbour approach selects the closest *k* matches from the database to be committee members, and the committee members "vote" on the correct classification. In this implementation, each committee member gets a vote equal to its similarity to the test utterance. Different values of *k* performed better in different aspects of the test, but this work uses $k = 7$ to facilitate comparison of results.

The choice of features largely follows those of Samuel 2000, and are as follows:

- Speaker change
- Word number
- Word similarity
- *n*-gram similarity
- Previous DA tag

and the following two features not included in that study,

- 2-previous DA tag

Inclusion of this feature enables more complete analysis of previous DA tags. Both 'previous DA tag' and '2-previous DA tag' features use the "best guess" for previous utterances rather than the "right answer", so this run allows us to test performance even with incomplete information.

- Discourse chunk tag

Distances for this tag were computed by dividing the larger discourse chunk number from the smaller. Comparing two "chunk starter" utterances would give the highest similarity of 1, and comparing a chunk starter (1) to a chunk-ender (100) would give a lower similarity (.01).

Not all features are equally important, and so an Evolutionary Programming algorithm (adapted from Fogel 1994) was used to weight the features. Weightings were initially chosen randomly for each member of a population of 100, and the 10 best performers were allowed to "survive" and "mutate" their weightings by a Gaussian random number. This was repeated for 10 generations, and the weightings from the highest performer were used for the CBR tagging runs.

A total of ten stopwords were used (the, of, and, a, an, in, to, it, is, was), the ten most common words from the BNC (Leech, Rayson, & Wilson 2001). These stopwords were removed when considering word similarity, but not *n*-gram similarity, since these low-content words are useful for distinguishing sequences of words that would otherwise be very similar.

The database consisted of 59 hand-tagged dialogues (8398 utterances) from the Verbmobil-2 corpus. This database was also automatically tagged with discourse chunks according to the rules above. The test corpus consisted of 20 dialogues (2604 utterances) from Verbmobil-2. This corpus was tagged with correct information on discourse chunks; however, no information was given on the DA tags themselves.

5 Discussion and future work

Table 5.1 shows the results from two DA tagging runs using the case-based reasoning tagger: one run without discourse chunks, and one with.

Without discourse chunks	With discourse chunks
53.68% (1385/2604 utterances)	65.44% (1704/2604 utterances)

Table 5.1: Overall accuracy for the CBR tagger

To put these results in perspective, human performance has been estimated at about 84% (Stolcke 2000), since human taggers sometimes disagree about intentions, especially when speakers perform more than one dialogue act in the same utterance. Much of the recent DA tagging work (using 18-25 tags) scores around the mid-fifty to mid-sixty percentiles in accuracy (see Stolcke 2000 for a review of similar work). This work uses the Verbmobil-2 tagset of 32 tags.

It could be argued that the discourse chunk information, being based on tags, gives the DA tagger extra information about the tags themselves, and thus gives an unfair 'boost' to the performance. At present it is difficult to say if this is the only reason for the performance gains. If this were the case, we would expect to see improvement in recognition for the four tags that are "chunk starters", and less of a gain in those that are not.

In the test run with discourse chunks, however, we see across-the-board gains in almost all categories, regardless of whether they begin a chunk or not. Table 5.2 shows performance measured in terms of the well-known standards of precision, recall, and f-measure.

One notable exception to the upward trend is `EXCLUDE`, a beginning-of-chunk marker, which performed slightly worse with discourse chunks. This would suggest that chunk information alone is not enough to account for the overall gain. Both `ACCEPT` and `FEEDBACK_POSITIVE` improved slightly, suggesting that discourse chunks were able to help disambiguate these two very similar tags.

Table 5.3 shows the improvement in tagging scores for one-word utterances, often difficult to tag because of their general use and low information. These words are more likely to be tagged ACCEPT when they appear near the beginning of a chunk, and FEEDBACK_POSITIVE when they appear nearer the end. Discourse chunks help their classification by showing their place in the dialogue cycle.

One weakness of this project is that it assumes knowledge of the correct chunk tag. The test corpus was tagged with the "right answers" for the chunks. Under normal circumstances, the corpus would be tagged with the "best guess," based on the DA tags from an earlier run. However, the goal for this project was to see if, given perfect information, discourse chunking would aid DA tagging performance. The performance gains are persuasive evidence that it does. Ongoing work involves seeing how accurately a new corpus can be tagged with discourse chunks, even when the DA tags are unknown.

6 Acknowledgements

This work was supported by an Australian Postgraduate Award. Thanks to Cara MacNish and Shelly Harrison for supervision and advice. Many thanks to Verbmobil for generously allowing use of the corpus which formed the basis of this project.

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.

M. G. Core, and J. F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*. Cambridge, MA.

D. Fogel. 1994. An introduction to evolutionary computation. *Australian Journal of Intelligent Information Processing Systems*, 2:34–42.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225

J. Kolodner. 1993. *Case-Based Reasoning*. Academic Press/Morgan Kaufmann.

G. Leech, P. Rayson, and A. Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman.

W. Mann. 2002. Dialogue Macrogame Theory. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, pages 129–141, Philadelphia PA.

N. Reithinger and M. Klesen. 1997. Dialogue act classification using language models. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proceedings of the 5th European Conference on Speech Communication and Technology*, volume 4, pages 2235-2238, Rhodes, Greece.

K. Samuel. 2000. *Discourse learning: An investigation of Dialogue Act tagging using transformation-based learning*. Ph.D. thesis, University of Delaware.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Verbmobil. 2003. "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.

Tag	Without discourse chunks			With discourse chunks		
	precision	recall	f-measure	precision	recall	f-measure
INIT	0.590	0.411	**0.484**	0.735	0.446	**0.556**
SUGGEST	0.446	0.399	**0.421**	0.778	0.912	**0.839**
REQUEST_SUGGEST	0.308	0.078	**0.125**	0.550	0.216	**0.310**
EXCLUDE	0.500	0.063	**0.111**	0.143	0.031	**0.051**
GREET	0.926	0.926	**0.926**	0.926	0.926	**0.926**
BACKCHANNEL	0.824	0.875	**0.848**	0.824	0.875	**0.848**
BYE	0.719	0.976	**0.828**	0.816	0.952	**0.879**
POLITENESS_FORMULA	0.821	0.742	**0.780**	0.889	0.774	**0.828**
THANK	0.875	0.636	**0.737**	0.875	0.636	**0.737**
FEEDBACK_POSITIVE	0.567	0.843	**0.678**	0.615	0.839	**0.710**
COMMIT	0.778	0.500	**0.609**	0.733	0.393	**0.512**
DELIBERATE	0.568	0.582	**0.575**	0.600	0.570	**0.584**
INFORM	0.493	0.682	**0.572**	0.655	0.812	**0.725**
FEEDBACK_NEGATIVE	0.700	0.304	**0.424**	0.667	0.348	**0.457**
REQUEST_COMMENT	0.425	0.327	**0.370**	0.500	0.288	**0.366**
REJECT	0.500	0.278	**0.357**	0.316	0.333	**0.324**
NOT_CLASSIFIABLE	0.534	0.265	**0.354**	0.696	0.274	**0.393**
DEFER	0.750	0.214	**0.333**	0.800	0.286	**0.421**
ACCEPT	0.392	0.290	**0.333**	0.476	0.429	**0.451**
REQUEST	0.351	0.191	**0.248**	0.525	0.456	**0.488**
REQUEST_CLARIFY	0.400	0.130	**0.197**	0.600	0.196	**0.295**
EXPLAINED_REJECT	0.333	0.133	**0.190**	0.600	0.600	**0.600**
GIVE_REASON	0.200	0.077	**0.111**	0.182	0.077	**0.108**
CLOSE	0.333	0.063	**0.105**	0.500	0.063	**0.111**
CLARIFY	0.400	0.056	**0.098**	0.000	0.000	**0.000**
CONFIRM	0.000	0.000	**0.000**	0.500	0.074	**0.129**
DEVIATE_SCENARIO	0.000	0.000	**0.000**	0.000	0.000	**0.000**

Table 5.2: Results for all DA types that appeared more than ten times in the corpus. The first group of four DA's represents those that signal the beginning of a discourse chunk; the second group shows those that do not.

	Percent classified correctly without discourse chunk information	Percent classified correctly with discourse chunk information
okay	71.90 (151/210)	75.24 (158/210)
yeah	69.90 (72/103)	74.76 (77/103)
right	62.16 (23/37)	72.97 (27/37)
mhm	88.23 (60/68)	88.23 (60/68)
bye	93.33 (14/15)	93.33 (14/15)

Table 5.3: Some examples of one-word utterances in the corpus, before and after discourse chunking.