Dialog Act Classification Using N-Gram Algorithms

没有具体的操作方法和代码，
只是在特定的地图任务上的行为识别
上看起来是有用的。

Max Louwerse and Scott Crossley

Institute for Intelligent Systems
University of Memphis
{max, scrossley } @ mail.psyc.memphis.edu

Abstract

Speech act classification remains one of the challenges in natural language processing. This paper evaluates a classification system that assigns one of twelve dialog acts to an utterance from the Map Task Corpus. The dialog act classification system chooses a dialog act based on n-grams from a training set. The system's performance is comparable to other classification systems, like those using support vector machines. Performance is high given the fact that the system only considers an utterance out of context and from written input only. Moreover, the system's performance is on par with human performance.

1. Introduction

One of the challenges in natural language processing is the relation between utterances and context. Consider somebody expressing the utterance *I'll be there tonight*. We know that *I* is a deictic personal pronoun, which most likely refers to the speaker. There is also a deictic tonight, referring to the night following the time the utterance was expressed. Furthermore, we know that in the near future, after the utterance has been expressed, there is going to be a location in which an *I* will be present. But the expression conveys much more. How should the utterance be interpreted? How should it be responded to? For instance, the speaker might have made a promise about being there tonight, or the speaker might have predicted to be there tonight. It might also be the case, however, that the speaker is threatening to be there tonight. Furthermore, it is not only the illocutionary force of the speaker that is of importance, but also the perlocutionary effect of the expression (Grice, 1975; Searle, 1975). As such, addressees have certain felicity expectations based on previous utterances (Austin, 1962). For instance, in the case of a promise, we assume that the promise is kept. In the case of a threat, we hope the threat will not become reality. Recognition of the speaker's intention based on context links meaning and expression and helps create speech acts. It is also this recognition of a speaker's intention that makes speech act classification a challenge for NLP.

Our focus on speech acts comes from interest in multimodal communication (Louwerse, Bard, Steedman, Hu and Graesser, 2004). Speech acts for instance determine the structure of a dialog, which in turn can predict intonational patterns. For instance, Taylor, King, Isard, and Wright (1998) and Hastie-Wright, Poesio, and Isard (2002) have shown that speech recognition can be improved by taking into account the sequence of dialogue acts and the association between such moves and observed intonation contours, while Flecha-Garcia (2002) even showed that dialog acts can partly predict eye brow movements. Speech acts thus plays a fundamental role in multimodal communication.

2. The Challenge of Speech Act Recognition

Language understanding requires recognition of the speaker's intentions. As Allen, Byron, Dzikovska, Ferguson, Galescu and Stent (2001) point out intention recognition forms one of the main challenges in the development of dialog management systems. There are three related reasons why intention recognition is so difficult. First of all, there are various distinct ways of formulating an intention. First, as we have shown earlier, the same utterance can contain different speaker's intentions. The utterance 'okay' could check the attention or agreement of the dialog partner (S1: Okay? Ready?), could be an acknowledgement showing that the speaker heard, understood and accepted the previous utterance (S1: "Just head southwards"; S2: "Okay"), a yes-answer (S1: "Do you want me to run by that again"; S2: "Okay") or could initiate a new conversation ("Okay, now go straight down").

Secondly, speech acts often remain linguistically unmarked. We do not necessarily have to use speech act verbs (Wierzbicka, 1987) or illocutionary force devices (Levinson, 1983) to promise, ask a question or threaten. Without intentions being linguistically marked, dialog management identification of the intention is difficult.

Thirdly, classifying a speaker's intention seems to require some underlying framework. Existing speech act classifications are very different from one another. On the one side of the spectrum there are classifications that cover a very small number of speech acts (five in Searle, 1975); on the other side are those that cover a very large number of speech acts (150 in Verschueren, 1980). Furthermore, the construction of these classifications is often based on

different approaches, for instance sentence modality (Vanderverken, 1990) or verb meanings (Wierzbicka, 1987).

These three issues (ambiguity, specificity and source) make speech act classification highly problematic (see also Clark, 1996, Levinson, 1983). We acknowledge that one utterance can have different illocutionary forces, and that one illocutionary force can be represented by various utterances. We also acknowledge that speech act classification cannot be based purely on language input, but should be considered in context, ideally in combination with other linguistic and paralinguistic modalities like intonation, facial expressions, head and body posture and gesture. The question to be addressed in this paper however is to what extent speech acts can be classified solely on the basis of language input and how the findings compare to human performance in an identical task. The purpose of this research is to identify an utterance into one of more than 12 speech act categories that have been proposed for the Map Task Corpus (Anderson et al., 1991) based solely on the available linguistic input.

2. Map Task

The Map Task is a restricted-domain route-communication task which makes clear to experimenters exactly what each participant knows at any given time and which permits the creation of corpora of spontaneous speech under controlled conditions.

In each Map Task dialogue, an Instruction Giver and an Instruction Follower collaborate to reproduce on the follower's map a route which is pre-printed on the giver's map.

By way of instructions, participants are told that they and their partners have maps of the same location but drawn by different explorers and so potentially different in detail. They are not told where or how the maps differ and neither can they see the other's map. The maps are of fictional locations and players have only three sources of knowledge in their initial encounter with a map: 1) the instructions, 2) what appears on the visible map (cartoon landmarks, their labels, and in the case of the giver, the location of the route) and 3) what has been said during the dialogue.

Dialog Act	%	Description	Example Utterance	Common n-grams
INSTRUCTION	14.38	Commands partner to carry out action	*Go round, ehm, until you get to just above them*	*you get to, you go, well*
EXPLANATION	6.56	States information not directly elicited by partner	*Yeah, that's what I thought you were talking about*	*I have a, I think, right I'm*
CHECK	6.81	Requests partner to confirm information that speaker has reason to believe but is not sure about	*So going down to Indian Country?*	*so I'm going, am I, so I*
ALIGN	8.46	Checks attention and agreement of partner, or his readiness for next dialog act	*This is the left-hand edge of the page, yeah?*	*what I mean, see the, are you*
QUERY-YN	6.5	Question that takes a 'yes' or 'no' answer and does not count as a check or align	*I've mucked this up completely have I?*	*have you got, do you, you have*
QUERY-W	3.09	Any query not covered by the other categories	*Left of the bottom or left of the top of the chestnut tree?*	*where do I, where are, what*
ACKNOWLEDGE	24.18	Verbal response which minimally shows that the speaker has heard (and often understood and accepted) the move to which it responds	*Mmhmm*	*that's okay, right, okay*
REPLY-Y	11.33	Reply to any query with a yes-no surface form which means 'yes' however that is expressed	*Uh-huh*	*yeah I do, uh-huh, yes*
REPLY-N	4.75	Reply to any query with a yes-no surface form which means 'no' however that is expressed	*No, no at the moment*	*I've not got, no I, nope*
REPLY-W	2.52	Reply to any type of query which doesn't simply means 'yes or 'no'	*Because I say*	*to the right, to the left, hand side of*
CLARIFY	3.84	Reply to some kind of question in which the speaker tells the partner something over and above what was strictly asked	*.. Mm, no you are still on land*	*you should be, and then, sort of*
READY	7.57	dialog act that occurs after end of a dialog game and prepares conversation for a new game to be initiated	*Okay. Now go straight down.*	*wait a minute, well right, okay now*

Table 1. The 12 move types used in the Map Task, their frequency in percentages, a description and an (out-of-context) example

The Map Task dialogs have been manually coded according to three levels of the dialog structure (Carletta, et al. 1996; 1997): 1) transactions, subdialogs that accomplish a major step in the participants' plan to achieve the map task; 2) conversational games, transactions between the participants that form coherent units, as in the case of question-answer pairs; 3) conversational moves, different categories of utterances (initiations and responses).

This lowest level of dialog structure, the conversational moves, is the focus of this paper. A total of 12 different categories of conversational moves are distinguished in Map Task (Carletta, et al. 1996; 1997). These are presented in Table 1.

3. Map Task Dialog Act Classifier

3.1 Algorithm

Map Task dialogs differ from other dialogs in their content. Crossley and Louwerse (under review) conducted a bigram analysis comparing ten different registers of spoken corpora (the TRAINS Corpus (Allen & Heeman, 1995), the Santa Barbara Corpus (Du Bois, Chafe, Meyer & Thompson, 1997), the Switchboard Corpus (Godfrey and Holliman, 1997), the Map Task Corpus (Human Communication Research Centre, 1997) and the six spoken corpora used in the London Lund Corpus (broadcast speeches, face to face conversations, telephone conversations, interview, spontaneous speeches, and prepared speeches)) (Svartvik, 1990). Frequencies of the bigrams in each register were entered in a factor analysis using the methodology described in Biber (1988). This analysis returned four dimensions, with the third dimension being identified as 'spatial.' Compared to the other corpora, the Map Task corpus loaded significantly higher on this third dimension, which allowed Crossley and Louwerse to conclude that bigram frequency alone might allow for the identification of a register like the Map Task corpus.

Based on this previous work, we now address the question to what extent n-grams can be used to identify dialog acts? The culmination of this research is a dialog act classifier based on the lexical construction of the Map Task corpus. Prior to the design of the speech act classifier, the Map Task corpus was divided into its coded dialog acts and then further subdivided into participant's roles (giver and follower). These individual sections were then analyzed for unique speech patterns based on n-gram occurrences.[1] For the purposes of the speech act classifier, only uni-grams, bi-grams, and tri-grams were evaluated. Those n-gram occurrences that were unique to individual speech acts were then written into a program that tracked and labeled each utterance in the Map Task corpus. The program was designed to separate giver utterances from follower utterances and search each separately for individual speech acts based on their rate of occurrence within the subdivided participant corpus. The program was written to first search each Map Task utterance for specific tri-grams and, if a match were made, label that utterance and remove it from the corpus. The program then repeated the same process first for bi-grams and then lastly for uni-grams. The program was trained on a corpus that represented about half of the entire Map Task corpus (73,074 words) and tested on the remaining half (71,308 words). Additionally, in an effort to examine the effects of discourse integration, all moves in the corpus were also coded according to the labeling of the move that preceded them. Discourse integration only proved valuable for the moves REPLY-Y and ACKNOWLEDGE (which had similar n-gram occurrences, but differed in their preceding moves) and REPLY-W (which was indistinguishable based on n-gram occurrence, but generally followed the QUERY-W move).

[1] Because of notational issues in the transcription, regular expressions were used to collapse utterances like 'mmmhhhh' and 'mmhh'.

3.2 Results Compared to Optimal Performance

Performance of the dialog act classification was compared with the gold standard of codes in the Map Task Corpus as discussed in Carletta et al. (1996). Results are presented in Table 2.

Dialog Act	Precision	Recall	F-measure
INSTRUCTION	66	54	59
EXPLANATION	65	18	28
CHECK	58	43	49
ALIGN	85	13	23
QUERY-YN	66	75	70
QUERY-W	54	61	57
ACKNOWLEDGE	59	95	73
REPLY-Y	84	61	71
REPLY-N	63	93	75
REPLY-W	29	74	42
CLARIFY	18	16	17
READY	50	7	12

Table 2. Precision, Recall, and F-measure for each dialog act

Average accuracy of the system was 58.08%. This is obviously considerably higher than the accuracy of randomly selecting a dialog act (7%) and assigning each utterance to the most frequent dialog act (20%) (Poesio & Mikkheev, 1998). This latter observation is important as our results may just be following the frequency patterns of dialog acts in the Map Task corpus. To falsify this possibility we conducted non-parametric Mann-Whitney tests that uses rank ordering of the data to determine differences. A significant difference between the frequency and classification findings falsifies the

possibility that our data set is a sheer reflection of the frequency of dialog acts. Tests comparing the frequency of the dialog acts with precision ($U = 1$, $z = -4.10$, $p < .001$, $N = 12$), recall ($U = 9$, $z = -3.64$, $p < .001$, $N = 12$) and F-measure ($U = 4$, $z = -3.93$, $p < .001$, $N = 12$) indeed confirmed this.

Recently, Surendran and Levow (2005) used support vector machine (SVM) algorithms to classify the Map Task dialogs into their twelve dialog acts using 1) acoustic features only (duration, intensity, pitch, speaking rate and speaker identity), 2) text only, and 3) a combination of text and acoustic features. Because the detailed report of precision and recall data is available, a statistical comparison can be made between the two methods. Not surprisingly, our data – like Surendran and Levow's text data – outperformed the acoustic condition on precision (χ^2 (1) = 90.63, $p < .001$) and recall (χ^2 (1) = 77.39, $p < .001$). Our findings were comparable to Surendran and Levow's data in the text condition for precision (χ^2 (1) = .104, $p = .78$) and recall (χ^2 (1) = 1.33, $p = .25$). When Surendran and Levow used both text and acoustic features in their SVM algorithm their findings outperformed our findings in recall (χ^2 (1) = 9.26, $p = .002$) but not precision (χ^2 (1) = 1.87, $p = .17$). These findings show that the performance of the dialog act classification is comparable with other approaches that are similar in their (text-based) task. However, the findings also support the claim that linguistic and paralinguistic information, as well as context should be considered in the identification of dialog acts (see also Taylor, King, Isard and Wright, 1998; Stolcke, et al., 2000). This is not surprising, particularly because an ACKNOWLEDGE 'okay', a CHECK 'okay', a READY 'okay' and a REPLY-Y 'okay' are identical in their isolated written input, but will be different when intonation and context are considered. Because of this, the performance results given here are unfairly biased. In precision and recall scores our classification findings are compared with the gold standard in the Map Task Corpus (Carletta, et al., 1996). This gold standard consists of the codes assigned to utterances by raters who had access to both context and speech.

3.3 Results Compared to Equivalent Performance

In the previous results the performance of the dialog act classification system was compared to a gold standard of human raters who did have access to previous and following dialog acts as well as to prosody and other speech cues in the dialog. To allow for a fair comparison of a system that uses written utterances out of context, our findings need to be compared with human ratings of the same materials.

To that end, a survey was designed in order to test the accuracy of human evaluators against the overall accuracy of the dialog act classifier.

The human raters in this survey first read through the HCRC dialogue structure coding manual (Carletta, 1996) and were then given the opportunity to clarify any questions they might have had. They were next given a survey in which they were asked to evaluate Map Task dialog acts and label them.

While the dialog act classifier can assess and label thousands of utterances within minutes, human raters cannot make such quick judgments. For this reason, the human raters used in this survey were only given a small sample of utterances taken from the original Map Task corpus.

Initially 120 moves were chosen for the survey with 10 moves from each category being included. In order to make the sampling more representative of the actual corpus, an additional 19 dialog acts were added. These dialog acts were based on occurrence rates in the Map Task corpus so that more common moves received more prominence than less common moves. When finished, the surveys were analyzed for accuracy and then compared to the findings of the dialog act classifier.

	Human raters			Dialog Act Classifciation System		
Dialog Act	precision	recall	F-measure	precision	Recall	F-measure
INSTRUCTION	49	77	60	58	54	56
EXPLANATION	37	5	43	75	27	40
CHECK	32	41	36	75	27	40
ALIGN	21	13	16	1	25	40
QUERY-YN	48	45	46	1	55	71
QUERY-W	88	68	77	1	82	90
ACKNOWLEDGE	36	61	45	41	1	58
REPLY-Y	61	58	59	77	83	80
REPLY-N	95	81	87	67	91	77
REPLY-W	5	18	26	5	36	42
CLARIFY	27	23	25	4	18	25
READY	25	14	18	75	27	40

Table 3. Precision, Recall, and F-measure for each dialog act (dialog act classification system and human raters)

4. Discussion and Conclusion

This paper investigated dialog acts in the Map Task corpus by developing and evaluating a dialog act classification system that classified utterances from the Map Task corpus into one of twelve dialog acts. The system used out-of-context transcribed utterances and applied an n-gram algorithm in its classification of these utterances. Such an algorithm seemed warranted by corpus linguistic work that showed that the Map Task scenario has unique content features.

The question can of course be raised what the contribution is of a speech-act classification system that uses n-grams of utterances out-of-context, while we know that performance is better when utterances are considered within their dialog context and when other modalities, such as intonation, are taken into account. The answers to this question are practical. First, to compare our n-gram algorithm to Surendran and Levow's (2005) linear support vector machines and hidden Markov models the same input had to be used, out-of-context textual input. Secondly, to compare the speech act classification with additional linguistic and paralinguistic channels, one needs to have at least a comparison group. With most intelligent systems currently not being able to recognize channels other than textual input (Graesser, McNamara, & VanLehn, 2005), speech act classification necessarily starts there,

We have shown that the performance of the out-of-context textual dialog act classification system was not different from other systems like those using state vector machine algorithms and was on par with human performance. From two performance studies two conclusions were drawn.

First, identification of dialog acts benefits from discourse context and other linguistic and paralinguistic cues, like intonation. Though this conclusion may be obvious, it is nevertheless noteworthy to see that the performance of a system that does not use context and intonation still has an acceptable performance when compared to a gold standard for which these context and intonation were in fact used.

The second conclusion relates to the evaluation of systems. In most system evaluations the system's output is compared to ideal human data. What the system's output should really be compared with is human performance in an identical task. When we conducted such a study for our dialog act classification system, the results between system and human raters did not significantly differ from the results between human raters themselves.

We have argued that dialog acts and discourse structure are important because they help to understand the speaker's intentions. Furthermore, they are important in multimodal communication tasks. They help intelligent systems to become more intelligent in their feedback (e.g. REPLY-Y or REPLY-N to a QUERY-YN). More importantly, other modalities like intonation, eye gaze, facial expressions and gesture correlate with discourse structure. That is, by identifying dialog acts we can predict the expression of certain modalities. At the same time, by using cues from these modalities the performance of speech act classification system becomes better. With more areas of research studying multimodal communication the need of speech act classification systems becomes larger, while the NLP challenge of speech act classification thereby becomes smaller.

Acknowledgements

This research was supported by grant NSF-IIS-0416128 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution. We would like to thank Gwyneth Lewis for her help in the data analysis and her comments on previous drafts of this paper. Correspondence concerning this article should be addressed to Max M. Louwerse, Institute for Intelligent Systems / Department of Psychology, University of Memphis, 202 Psychology Building, Memphis, Tennessee 38152-3230.

References

Allen, J. and Heeman, P.A. 1995. *TRAINS spoken dialog corpus* (CD ROM). Philadelphia, PA: Linguistic Data Consortium.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. 1991. The HCRC Map Task Corpus. *Language and Speech* 34: 351-366.

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A. 2000. Towards a Generic Dialogue Shell. *Natural Language Engineering* 6: 1-16.

Austin, J. L. 1962. How to Do Things with Words. Oxford: Oxford University Press.

Biber, D. 1988. Linguistic Features: Algorithms and Functions in Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Clark, H. H. 1996. Using Language, Cambridge: Cambridge University Press.

Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. 1996. HCRC Dialogue Structure Coding Manual (HCRC/TR-82). Edinburgh, Scotland: Human Communication Research Centre, Univ. of Edinburgh.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics* 23: 13-31.

Crossley, S. A. and Louwerse, M. M. (under review). *Register Classification Using Bigram Analysis*.

Du Bois, J.W., Chafe, W.L., Meyer, C. and Thompson, S.A. 1997. *Santa Barbara Corpus of Spoken American English Part-I* (CD ROM). Philadelphia, PA: Linguistic Data Consortium.

Flecha-Garcia, M. L. 2002. Eyebrow Raising and Communication in Map Task Dialogues. Proceedings of the 1st Congress of the International Society for Gesture Studies, Univ. of Texas at Austin.

Godfrey, J.J. and Holliman, E. 1997. *SWITCHBOARD-1 Transcripts* (CD ROM). Philadelphia, PA: Linguistic Data Consortium.

Graesser, A.C., McNamara, D.S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist, 40*, 225-234.

Grice, H. P. 1975. Logic and Conversation. In *Syntax and Semantics* 3: Speech Acts, eds. P. Cole, P. and J. Morgan, 41-58. New York: Academic Press.

Hastie-Wright, H., Poesio, M. and Isard, S. 2002. Automatically Predicting Dialogue Structure Using Prosodic Features. *Speech-Communication* 36: 63-79.

Human Communication Research Centre 1997. *The HCRC Map Task Corpus* (CD ROM). Philadelphia, PA: Linguistic Data Consortium.

International Computer Archive of Modern and Medieval English (2000). *The London-Lund Corpus of Spoken English* (CD-ROM).

Landis, J. R., and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorial Data. *Biometrics* 36:159-174

Levinson, S. C. 1983. Pragmatics. Cambridge: Cambridge University Press.

Louwerse, M. M., Bard, E. G., Steedman, M., Hu, X., and Graesser, A. C. 2004. Tracking Multimodal Communication in Humans and Agents. Technical Report, Inst. for Intelligent Systems, Univ. of Memphis, Memphis, TN.

Poesio, M., and Mikheev, A. The Predictive Power of Game Structure in Dialogue Act Recognition: Experimental Results Using Maximum Entropy Estimation. In Proceedings of ICSLP-98, November, 1998.

Searle, J. 1975. A Taxonomy of Illocutionary Acts. In Minnesota Studies in the Philosophy of Language, ed. K. Gunderson, 334-369. Minnesota: Univ. of Minnesota Press.

Stolcke, A. Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3): 339-373.

Surendran, D. and Levow, G. Dialog Act Tagging with Support Vector Machines and Hidden Markov Models, Proceedings of the 9th IEEE Automatic Speech Recognition and Understanding Workshop, November.

Jan Svartvik (ed) (1990), *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82. Lund University Press.

Taylor, P., King, S., Isard, S. and Wright, H. 1998. Intonation and Dialogue Context as Constraints for Speech Recognition. *Language and Speech* 41: 493-512.

Vanderverken, D. 1990. On the Unification of Speech Act Theory and Formal Semantics. In Intentions in Communication, eds. P. R. Cohen, J. Morgan and M. E. Pollack, 195-220. Cambridge, MA: MIT Press.

Verschueren, J. 1980. On Speech Act Verbs. Amsterdam: John Benjamins.

Wierzbicka, A. 1987. English Speech Act Verbs. A Semantic Dictionary. Sydney: Academic Press.