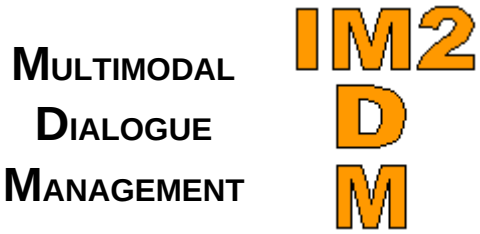
http://www.im2.ch



http://www.issco.unige.ch/projects/im2/mdm/

ISSCO/TIM/ETI, Université de Genève

Machine Learning Approaches to Shallow Discourse Parsing: A Literature Review

Alexander Clark

IM2.MDM/ Discourse Parsing WP Project Deliverable–
March 2003

Literature review of Machine Learning Approaches to Shallow Discourse Parsing

Alexander Clark

March 25, 2003

Abstract

This document reviews the literature on shallow discourse parsing, in particular the use of machine learning techniques. This is deliverable Y1.M6 of the Discourse Parsing White Paper which is part of the MDM IP of the IM2 project.

1 Introduction

This literature review is intended to summarise research into a rather narrow field: Machine learning approaches to shallow discourse parsing (SDP). Even this sub-field turns out to have rather too much literature for any survey to be truly exhaustive. We will not do a paper-by-paper analysis since as is so often the case there is quite a lot of overlap between different approaches. We shall rather try to draw out some particular themes and issues that we think are particularly crucial. Discourse parsing here refers to the task of attaching to each utterance in a discourse, whether it is a multi-party dialogue or a single person discourse, a label drawn from a small finite set of Discourse tags, that indicates the function of the utterance or the type of the speech act (Searle & Vanderveken, 1985) associated with that utterance. There are a number of different annotation schemes that have been used in the past: we shall not discuss them here (see (Midgley, 2001) for a useful summary. Here we shall focus on algorithms for assigning them that are based on a Machine learning methodology, by which we mean a methodology based not on manually crafting an algorithm to assign the classification (such as (Hinkelman & Allen, 1989; Core, 1998)) but on using an algorithm to extract a classifier from some labelled or unlabelled data.

We shall start with a few general and inevitably inaccurate remarks. In general there are two techniques which have been used: either specific sequences of words have been manually or automatically identified as possible features – so-called discourse markers or cue-phrase, or alternatively language models have been used. There are a number of advantages and disadvantages to both approaches, which we shall discuss later. These two techniques are used to define a classifier for each individual utterance considered in isolation. There

is of course some further information that can be incorporated and that is the context of the utterance, which can clearly help disambiguation in certain circumstances.

One further information source is that of prosody – intonation, stress, speaking rate, volume and other acoustic features that can be important, depending of course on the language in use.

In line with our restriction to *shallow* discourse processing, we shall not consider work which attempts the more difficult and interesting task of constructing deep discourse trees such as for eaxmple (Forbes, Miltsakaki, Prasad, Sarkar, Joshi, & Webber, 2001; Marcu, 2000). In any event, these approaches have generally been applied to very clean written language, and it is not clear that they would generalise well to spontaneous dialogue.

2 Discourse Marker Based Algorithms

The earliest work in SDP used simple algorithms based on identifying discourse markers as discussed in (Schiffrin, 1987). An obvious problem with this approach, that was recognised early on, is that discourse markers, such as "now" or "say", are often ambiguous, and can be used either in a discourse function, or with a sentential function. Thus much work focuses, on the problem of identifying and disambiguating these cue phrase (Hirschberg & Litman, 1993; Litman, 1994; Siegel & McKeown, 1994; Litman, 1996; Heeman, Byron, & Allen, 1998), using standard techniques. Researchers have either used existing lists of cue words or have automatically derived them (Samuel, Carberry, & Vijay-Shanker, 199). The discourse markers then serve as features for some classifier. Classifiers that have been used include neural networks (Kipp, 1998; Ries, 1999), decision trees (Litman, 1996; Cettolo & Corazza, 1998), and LVQ (nearest neighbour) (Jokinen, Hurtig, Hynnä, Kanto, Kaipainen, & Antti, 2001). An interesting comparison is presented in (Marineau, Wiemer-Hastings, Harter, Olde, Chipman, Karnavat, Pomeroy, S., & A., 2000).

3 Language Model based Algorithms

Language model based techniques are those which classify by explicitly modelling the sequence of words for each utterance type. These techniques have been explored by a number of researchers, especially in the VERBMOBIL project, as part of the shallow translation component (Mast, Nöth, Niemann, & Schukat-Talamazzini, 1995; Mast, Kompe, Harbeck, Kiessling, Niemann, Nöth, Schukat-Talamazzini, & Warnke., 1996; Reithinger & Klesen, 1997). A particularly thorough and clear example of this technique by another group is (Stolcke, Shriberg, Bates, Coccaro, Jurafsky, Martin, Meteer, Ries, Taylor, & Van Ess-Dykema, 1998; Stolcke, Ries, Coccaro, Bates, Jurafsky, Taylor, Martin, Meteer, & Van Ess-Dykema, 2000). A particular merit of these approaches is that they can be incorporated naturally into an Automatic Speech Recognition (ASR)

system, by summing over the most likely outputs from the recogniser. Thus, the resulting classification can be surprisingly resistant to the noise introduced by the use of an ASR system rather than a near-perfect human transcription. In addition to incorporating lexical cues, language models that use low-order n-grams can also capture the fact that individual discourse acts have an identifiable 'micro-syntax' Moreover because of their principled probabilistic basis, it is straightforward to incorporate them into larger systems, and to combine them with other sources of information.

4 Use of context

It is clear that the context, i.e. the preceding or following sequence of dialog acts provides an important source of information. Thus in addition to a language model that models the sequence of words in an individual sentence, researchers have also used similar techniques to model the sequence of dialog acts. These have generally been finite-state models such as ngrams (Stolcke et al., 2000) (deterministic given the correct tags, non-deterministic given the words) though stochastic context free grammars were used in (Alexandersson & Reithinger, 1997).

Some early papers used rather ad hoc methods to combine the predictions derived from the utterance and its context (Alexandersson & Reithinger, 1997), but rapidly research seems to have converged on using Bayes rule combined with standard HMM decoding techniques (Rabiner, 1989) Many researchers have noticed the similarity between dialogue act tagging and POS tagging, and have used similar methodologies. Thus the HMM tagging methodology (Church, 1988) has been used extensively (Stolcke et al., 2000), in addition to attempts to import other techniques such as Brill's Transformation-Based Learning (Samuel, Carberry, & Vijay-Shanker, 1998). As noted by (Stolcke et al., 2000) there are a number of different decoding algorithms that can be used, depending on whether one has access to later utterances – in this case one can use forward-backward decoding. Additionally Bayes' nets have been used to combine several different knowledge sources (Keizer, 2001; Keizer, op den Akker, & Hijholt, 2002).

The exact model that is used to model the sequence of dialog acts varies a certain amount. Ngram models have been used for obvious technical reasons, but there may be longer range structure to be detected (Poesio & Mikheev, 1998) though this must depend on the exact data under investigation. Thus the Switchboard data studied by (Stolcke et al., 2000) consist of spontaneous conversations between two people on a pre-determined topic, so the sequence of dialog acts seems not to have much global structure. More task-oriented dialog, or indeed of recordings of meetings that must for administrative reasons follow a specific pattern, would exhibit, we expect, a noticeable global structure. Context Free Grammars have also been used in the Verbmobil project (Alexandersson & Reithinger, 1997). The only other work using more complex models that we are familiar with is the use of ALERGIA, a grammatical inference technique, used by (Kita, Fukui, Nagata, & Morimoto, 1996).

5 Prosody

Prosody is clearly very important in the task of segmenting utterances, and appears also to be helpful in classification of utterance type (Jurafsky, Shriberg, Fox, & Curl, 1998; Shriberg, Bates, Stolcke, Taylor, Jurafsky, Ries, Coccaro, Martin, Meteer, & Van Ess-Dykema, 1998; Stolcke et al., 1998, 2000), though this depends to a certain extent on the dialogue act concerned. Particular examples where prosody helps include utterances like "yeah" which can be very ambiguous, even in context without prosody. It is also possible to examine the contribution of prosody on its own (Fernandez & Picard, 2002). This paper is also interesting for the use of support vector machines. A variety of different acoustic and prosodic features have been studied – the exact choice is language dependent.

6 Unsupervised techniques

Up to now we have only discussed supervised classifications where the set of labels has been manually constructed by discourse theoreticians, either a priori derived from a philosophical analysis of speech act theory or by examination of naturally occurring data (Hinkelman, 1990).

A radical alternative is to derive the set of labels themselves using an unsupervised technique from the data. Only a few researchers have experimented with this: (Andernach, 1996; Andernach, Poel, & Salomons, 1997) use two different sorts of clustering algorithm, one a Bayesian soft-clustering package and the other Kohonen self-organizing maps. (Möller, 1997, 1998) presents an integrated system that uses a custom hierarchical clustering algorithm. In both of these lines of research, the empirical evaluation is incomplete, and perhaps the arguments for using completely unsupervised algorithms are not completely convincing.

Another very interesting paper is (Marcu & Echihabi, 2002). Here they use an unsupervised technique to extract pairs of lexical cue words from a very large (billion-word) corpus. They show dramatic improvements as a result of this. Here, the use of unsupervised techniques is well motivated as an adjunct to a labelled classification task.

7 Discussion

It appears to be clear that a lot of dialog acts cannot be classified purely on the basis of discourse markers, because discourse markers are not always present. Therefore, one has to use some sort of more general lexical information to get above a threshold of performance. The obvious way to do this is through using language models, though it may also be possible to extract particular words, or pairs of words, that identify particular dialog acts.

Some researchers argue for a close integration of utterance segmentation and identification (Ries, 1999; Warnke, Kompe, Niemann, & Nöth, 1997) – indeed

it does appear that in spoken dialog, the two tasks are cloesly related.

Existing techniques make some radical independence assumptions: the standard HMM model assumes that the words in an utterance are conditionally independent of the words in preceding utterances, given the current state of the discourse model. Since the models of the sequence of dialog acts do not incorporate topic information, this is clearly false. This will become more important as the shift to multi-party dialogue continues.

References

Alexandersson, J., & Reithinger, N. (1997). Learning dialogue structure from a corpus. In *Proceedings of EuroSpeech-97*, pp. 2231–2235 Rhodes.

Andernach, T. (1996). A machine learning approach to the classification of dialogue utterances. In *Proceedings of NeMLaP-2* Ankara, Turkey.

Andernach, T., Poel, M., & Salomons, E. (1997). Finding classes of dialogue utterances with kohonen networks. In Daelemans, W., Van den Bosch, A., & Weijters, A. (Eds.), *Workshop Notes of the ECML / MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pp. 85–94 Prague, Czech Republic.

Cettolo, M., & Corazza, A. (1998). History integration into semantic classification. In Ponting, K. (Ed.), *Computational Models of Speech Pattern Processing*, Vol. 169 of *NATO ASI Series F*, pp. 356–361. Springer Verlag. Rather misleading title.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on applied natural language processing*, pp. 136–143.

Core, Mark, G. (1998). Analyzing and predicting patterns of DAMSL utterance tags. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing* Stanford, CA.

Fernandez, R., & Picard, R. W. (2002). Dialog act classification from prosodic features using support vector machines. In *Proceedings of Speech Prosody 2002* Aix-en-Provence, France.

Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A., & Webber, B. (2001). D-LTAG system – discourse parsing with a lexicalized tree adjoining grammar. In *Information Stucture, Discourse Structure and Discourse Semantics Workshop held at ESSLLI 2001.*

Heeman, P. A., Byron, D., & Allen, J. F. (1998). Identifying discourse markers in spoken dialog. In *AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 44–51 Stanford.

Hinkelman, E. (1990). *Linguistic and Pragmatic Constraints on Utterance Interpretation*. Ph.D. thesis, University of Rochester.

Hinkelman, E., & Allen, J. (1989). Two constraints on speech act ambiguity. In *Proceedings of ACL*, pp. 212–219.

Hirschberg, J., & Litman, D. (1993). Empirical studies in the disambiguation of cue phrases. *Computational Linguistics*, *19*(3), 501–530.

Jokinen, K., Hurtig, T., Hynnä, K., Kanto, K., Kaipainen, M., & Antti, K. (2001). Self-organizing dialogue management. In *Proceedings of 2nd Workshop on Neural Networks and Natural Language Processing MLPRS* Tokyo, Japan.

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, prosodic and syntactic cues for dialog acts. In *AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 114–120.

Keizer, S. (2001). Dialogue act modelling with bayesian networks. In Striegnitz, K. (Ed.), *Proceedings of the 6th ESSLLI Student Session*, pp. 143–153.

Keizer, S., op den Akker, R., & Hijholt, A. (2002). Dialogue act recognition with bayesian networks for dutch dialogues. In *Proceedings of 3rd SIGdial Workshop on Discourse and Dialogue* Philadelhia, PA.

Kipp, M. (1998). The neural pathway to dialogue acts. In *Proceedings of the 13th ECAI*, pp. 175–179.

Kita, K., Fukui, Y., Nagata, M., & Morimoto, T. (1996). Automatic acquisition of probabilistic dialogue models. In *Proceedings of the 4th International Conference on Spoken Dialogue Processing*, pp. 196–199.

Litman, D. J. (1994). Classifying cue phrases in text and speech using machine learning. In *Proc. Annual Meeting of the American Association for Artificial Intelligence*, pp. 806–813 Seattle.

Litman, D. (1996). Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, *5*, 53–94.

Marcu, D., & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL* Philadelphia.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., S., R., & A., G. (2000). Classification of speech acts in tutorial dialogue. In *Proc. of Intelligent Tutoring Systems ITS2000*.

Mast, M., Kompe, R., Harbeck, S., Kiessling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E. G., & Warnke., V. (1996). Dialog act classification with the help of prosody. In *ICSLP 96* Philadelphia.

Mast, M., Nöth, E., Niemann, H., & Schukat-Talamazzini, E. G. (1995). Automatic classification of speech acts with semantic classification trees and polygrams. In *IJCAI-95 Workshop: New Approaches to Learning for Natural Language Processing*, pp. 71–79 Montreal.

Midgley, D. (2001). Literature review. www.cs.uwa.edu.au/ fontor/research/index.html. University of Western Australia.

Möller, J.-U. (1997). Classitall: Incremental and unsupervised learning in the dia-mole framework. In Daelemans, W., Van den Bosch, A., & Weijters, A. (Eds.), *Workshop Notes of the ECML / MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pp. 95–104 Prague, Czech Republic.

Möller, J.-U. (1998). Using unsupervised learning for engineering of spoken dialogues. In *AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing*.

Poesio, M., & Mikheev, A. (1998). The predictive power of game structure in dialogue act recognition: Experimental results using maximum entropy estimation. In *ICSLP 98*.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–285.

Reithinger, N., & Klesen, M. (1997). Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pp. 2235–2238 Rhodes.

Ries, K. (1999). HMM and neural network based speech act detection. In *Proceedings of ICASSP 99* Phoenix, Arizona.

Samuel, K., Carberry, S., & Vijay-Shanker, K. (199). Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics* Waterloo, Ontario, Canada.

Samuel, K., Carberry, S., & Vijay-Shanker, K. (1998). An investigation of transformation-based learning in discourse. In *In Machine Learning: Proceedings of the Fifteenth International Conference*, pp. 497–505 Madison, Wisconsin.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge University Press.

Searle, J. R., & Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge University Press.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech. *Language and Speech*, *41*(3-4), 439–487.

Siegel, E. V., & McKeown, K. R. (1994). Emergent linguistic rules from inducing decision trees: disambiguating discourse clue words. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* Seattle, WA.

Stolcke, A., Ries, K., Coccaro, N., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., & Van Ess-Dykema, C. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–371.

Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., Meteer, M., Ries, K., Taylor, P., & Van Ess-Dykema, C. (1998). Dialog act modeling for conversational speech. In *AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing* Stanford.

Warnke, V., Kompe, R., Niemann, H., & Nöth, E. (1997). Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Proc. European Conf. on Speech Communication and Technology*, Vol. 1, pp. 207–210.