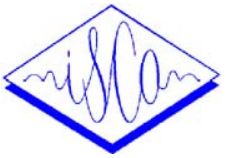
TOWARDS THE DETECTION AND DESCRIPTION OF TEXTUAL MEANING INDICATORS IN SPONTANEOUS CONVERSATIONS

Klaus Ries
riescs.cmu.edu
Interactive System Labs

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999



Carnegie Mellon University, Pittsburgh, PA, USA
University of Karlsruhe, Karlsruhe, Germany

10.21437/Eurospeech.1999-365

ABSTRACT

The description of textual and stylistic features has so far been largely neglected in the empirical study of conversational speech. In this paper we want to make a couple of strong initial points towards the use textual meaning and stylistic features in language engineering: First of all we want to show that there are other besides the traditional features in spontaneous speech that are worth studying and that might reveal good information: These are related to the interactive nature of the language and to the distribution of the most frequent (non-topical) words. Secondly we want to present two tasks that we have chosen as our benchmark and present detection results. Finally we want to motivate how this can be used in information access applications.

1. INTRODUCTION

The study of textual meaning has often be confined to the discussion of thematic or topical meaning. [9, p.334] suggests that the textual grammar of English is composed of the components

structural *thematic structure* and *information structure and focus*

cohesive reference, ellipsis and substitution, conjunction and lexical cohesion

While this reflects the a very specific view however there are also a number of publications that argue that the way we talk about things is also interesting. I want to point out three larger strands of research here [1]: The first type is concerned with the creation of a set of "genres" that following certain patterning rules like narratives [5, 11]:

`Abstract? Orientation? Complication Evaluation Resolution Coda?`

[1] This overview does not claim to contain all references but rather the most important ones that should allow to follow-up on the respective literature.

Different genres are defined as belonging to different patterns, typically expressed in a regular expression syntax. This definition is problematic for a number of reasons: The definition of pattern rules seems to be arbitrary and annotating them reliably seems to be complicated. However marking genre boundaries often easier and better to define than to mark thematic segment boundaries which has to be done by tracking the cohesion of the text. A second type of research tries to identify to what extent grammatical and other choices vary along "register" or within a discourse typology [18, 7, 14]. This idea has been used in corpus based linguistic work [3, 4] and in information retrieval [19, 10] recently. [13] is using word categories derived from a semantic network to determine stylistic features in conversations and assumes that they are related to social distance. The third type of research [16, 12] describes local stylistic variation in the context of emphasizing. One could also subsume this feature as special markers for the peak of a story or for the display of empathy – in some sense this seems to be a special combination of the first and second type. Additionally work in socio- and psycholinguistic and ethnomethodology has been looking into the use of style in communication, which is not as relevant to this work as the abovementioned research.

Our own research is geared towards information access for spontaneous human to human interactions [20] and we want to combine the notion of topical, genre and stylistic information. While the notion of topical information has been popularized by internet search engines stylistic information is still rarely exploited [10]. Generic information is in some sense the pendant to information extraction in human to human interactions: We assume e.g. that we know that we have a conversation between close acquaintances. Knowing that close acquaintances often tell each other stories about their friends and relatives and what happened to them recently we may focus on extracting the information (a) who is being talked about (b) what was the story (c) how was the story judged by the participants.

A field even less covered in traditional information access is the style of the conversation. Some of these stylistic parameters are relatively easy to access and can be easily used by a human: A seminar talk e.g. is usually more or less a monologue, a telephone conversation typically has two participants, a meeting has many people taking turns etc. However it is not clear a priori how more fine grained distinctions can be made and whether there are stylistic features in human to human discourse that have not been studied in the largely text oriented research reviewed above.

This paper will explore the question of features for these stylistic differences. The question will be addressed by defining a stylistic discrimination task: Given a text tell me which style it has, where style is assumed to be correlated with the corpus used or properties of the speaker. The feature sets and data sets used will be introduced in section 2, whereas section 3 is reserved for the neural networks used and the experimental results. In section 4 we will present our conclusion and an outlook.

2. TASKS AND FEATURES

We decided to pick two prototypical situations in which we would assume that the style is different. The first task is a cross-corpus discrimination task in which we included four different spoken language corpora. The second task is a within-corpus discrimination task where we try to distinguish speakers with different features (gender). The idea is to train a classifier on those tasks and see which features are important for classification performance. However we may run into problems with both task-definitions: We might be measuring correlation between the types and features that we would not call style otherwise. We will therefore carefully analyze suspicious feature groups in the experimental section and have indeed found unexpected correlations.

2.1. Task definition

The first task is to discriminate between four spontaneous speech corpora: CallHome English, CallHome Spanish, Broadcast News and Switchboard, which are all published by [1]. All corpora but Broadcast news are telephone conversations and all corpora but CallHome Spanish are in English. Unless we want a trivial 100% result we have to exclude a couple of potential features: We cannot rely directly on word identities (since we have multiple languages) and on features that would immediately uncover the fact that Broadcast-News contains music, a variety of speakers and speaking situations etc. The same argument could me made for the fact that CallHome speakers are relatives.

This radical approach is not necessary if we draw the stylisticly different sources from one corpus. We have decided to use the Switchboard corpus and to distinguish between male and female speakers since we assume that unacquainted male and female speakers will show stereotypical gender specific discourse behaviors. In Switchboard the topic of the conversation is also given in advance by the system that connects "random" people. However, as can be seen from Tab. 1 the participants did not stick to their assigned topics and drifted off to private discussions which results in some of the best gender discriminating keywords.

husband watch children seen family she T uh-huh
her usually Texas care wasn't either Yes kids feel
um haven't fact Do We nice ago find

Table 1: **Salient words for gender discrimination**: Keywords for talking about family life are very salient for gender discrimination.

2.2. Feature definition

Originally we have only been looking at features that can be extracted either from the signal directly (pause information) or that require almost no processing. Features that would be relying on any kind of processing from the word level would not be suitable for the first task and were therefore omitted from our first set of experiments. As an initial preprocessing step we also calculated the "active segments" which are segments on each channel where one speaker is talking with pauses shorter than 0.3sec (except when the words surrounding that pause are very short themselves, e.g. individual backchannels or human noises).

In the within-corpus discrimination task we have been using a histogram of the most frequent 300 words as well. To see whether we can restrict ourselves to a smaller number of keywords a salience analysis was conducted by ranking the words according their contribution to a Naive Bayes classifier [8] (Tab. 1)

Additionally we have been using WordNet [6] to derive verb and noun classes. We used the 45 lexicographers classes (see Tab. 2) to describe a small set of distinctive classes and chose for each word / parts-of-speech pair the most frequent class occured in. In a sense we are approximating the more developed classes as they are present in the MCCA toolkit and have been used for the description of social features of speakers [13].

We have implemented those grammatical features of [2] that could be easily derived automatically using a part-of-speech tagger, the regular expressions as defined in [2], a simple morphological analysis and various word classes [2, 14]. We have selected 46 of the original 67 features and interestingly a lot of those that account for high factor loads in [2] belong to that

noun classes
act animal artifact attribute body cognition communication event feeling food group location motive object person phenomenon plant possession process quantity relation shape state substance time
verb classes
body change cognition communication competition consumption contact creation emotion motion perception possession social stative weather

Table 2: **Verb and noun classes in WordNet**

category. All features were normalized for text-length such that these differences should not affect the classification problem.

3. EXPERIMENTS

The first set of experiments was carried out with a reduced set of features such that we could experiment with both the cross-corpus task as well as with the with within-corpus task. All experiments were using a neural network using shortcut connections and 5 hidden units. The output function was softmax, the corresponding error function the Kullbach-Leibler divergence. The error function was regularized using a momentum term and the training method used was RPROP [15]. Numeric features were represented using a histogram based bucketing. The first interesting result from Tab. 3 is that the cross-corpus task is extremely simple. If we actually look at the means and variances of these features it becomes immediately obvious that the discrimination should be trivial.

The results for the within-corpus task are a little harder to interpret. Using just the fact that a word was said on one channel versus. the other ("only word" in Tab. 3) resulted in a good baseline of 66.3%. If we compare this to the results of the pause and length measurements most of them seem to do worse than that, with the exception of the active segment length and interestingly enough especially the word length. We first suspected that the speaking rate is different between male and female speakers but in our database [17] we have found no significant difference.

The next observation is that while the most salient words, from which a lot seem to be keying towards family affairs (Tab. 1), can give us a good performance. However we can see that (a) this does not explain all information in the unigram distribution and (b) if we edit out the obvious "family affair" words the performance goes down but is still very high. We may therefore conclude that the discrimination is done mostly by the non-topical words and that – as verified by further experiments – all of these words contribute to the classification.

We have had only limited success in using the grammatical features [2]. Interestingly the best result we obtained used only the parts of speech histogram, which has been entered in Tab. 3. All other features could not rival with any of the word level features which we will investigate further.

Feature	cross corpus	within task
pick most likely category	25.0%	50.0%
length of active segments	100.0%	68.8%
length of overlap of active segments	100.0%	61.3%
length of words	95.5%	71.3%
all of the above	97.5%	67.5%
+ most freq 300 word		83.0%
most freq 300 word		83.0%
no "family affair" words		80.0%
40 most salient words		75.0%
20 most salient words		72.5%
10 most salient words		68.8%
only "word"		66.3%
WordNet		
noun and verb classes		65.0%
Stylistic features		
parts-of-speech histogram		72.5%

Table 3: **Style detection results**

4. CONCLUSIONS

The current experiments indicate that stylistic features between different corpora vary so enormously that any detection technique using even the simplest features should work. Simple features such as word lengths and overlap between speakers carry a long way and also help in the within-corpus discrimination. It is however much more difficult to discriminate in one single corpus. Word level information from the most frequent words seems to be very useful but it is not clear what exactly makes the difference or whether it is just the combination of all these features. We have tested a number of simple and traditional explanations but they all fell short so far.

In our future work we want to try to find more reasons for these results and apply our results to a stylistic characterization of human to human conversations. We hope that this stylistic characterization, combined with topical as well as genre information will provide a rich textual representation of spontaneous conversations that can be applied to information access applications. The research in stylistic variation should also help to improve the detection of genre and/or topical segments since we may assume that stylistic features do not change too much or in specific ways during these segments.

5. ACKNOWLEDGEMENTS

The work reported in this paper was conducted as part of project Clarity and was funded by the US Department of Defense. The author is grateful for their support. The author is indebted for the discussions with the project leaders Alex Waibel, Alon Lavie and Lori Levin. This work would have been impossible with the help and ressources of many other people: Forced alignments for Broadcast News from the Sphinx speech group at CMU, a part of speech tagger for Switchboard from Klaus Zechner, the initial implementation of the neural networks from Jürgen Fritsch, the prosodic database for Switchboard from the John Hopkins Summer Workshop and all the people that I met there, discussions with Doug Biber and Dong Hoon van Uytsel, the WordNet database and probably many more that I forgot to mention.

6. REFERENCES

[1] Linguistic Data Consortium (LDC) catalogue. http://www.ldc.upenn.edu/.

[2] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1988.

[3] Douglas Biber. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(219-241), 1993.

[4] Douglas Biber, Susan Conrad, and Randi Reppen. *Corpus Lingusitics: Investigating Language Structure and Use*. Cambridge University Press, 1998.

[5] Suzanne Eggins and Diana Slade. *Analysing Casual Conversation*. Cassell, 1998.

[6] Christiane Fellbaum, editor. *WordNet – An Electronic Lexical Database*. MIT press, 1998.

[7] Gerd Fritz and Franz Hundsnurscher. *Handbuch der Dialoganalyse*. Niemeyer, Tuebingen, 1994.

[8] Allen Gorin. On automated language acquistition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.

[9] M.A.K. Halliday. *An introduction to functional grammar*. Oxford University Press, 1994.

[10] Brett Kessler, Geoff Nunberg, and Hinrich Schütze. Automatic detection of genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Morgan Kaufmann Publishers, San Francisco CA, 1997.

[11] W. Labov and J. Waletzky. Narrative analysis: oral versions of personal experience. In J. Helm, editor, *Essays of the Verbal and Visual Arts*, pages 12–14, Washington DC, 1967. American Ethnological Society, proceedings of Spring Meeting 1967, University of Washington Press.

[12] Robert E. Longacre. *The grammar of discourse*. Plenum Press, 2nd edition, 1996.

[13] McTavish, Litkowski, and Schrader. A computer content analysis approach to measuring social distance in residential organizations for older people. In *Conference of the Society for Conceptual and Content Analysis by Computer*, Mannheim, Germany, September 1995. http://www.clres.com/index.html#papers.

[14] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A comprehensive grammar of the English language*. Longman, 1985.

[15] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proc. of the IEEE Int. Conf. on Neural Networks*, pages 586–591, 1993.

[16] Margret Selting. *Stilfragen*, volume Jahrbuch 1994 of *Institut fuer deutsche Sprache*, chapter Sprechstile als Kontextualisierungshinweise (speaking styles as contextualization hints). de Gruyter, 1995.

[17] Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):439–487, 1998.

[18] Deborah Tannen. *Converational Style*. Language and Learning for Human Service Professions. Ablex Publishing Corporation, 1984.

[19] van Bretan, Johan Dewe, Anders Hallberg, Jussi Karlgren, and Niklas Wolkert. Genres defined for a purpose, fast clustering, and an iterative information retrieval interface. In *Eighth DELOS Workshop on User Interfaces in Digital Libraries Lngholmen*, pages 60–66, October 1998.

[20] Alex Waibel, Michael Bett, and Michael Finke. Meeting browser: Tracking and summarising meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.