Socio-Situational Setting Classification based on Language Use

Yangyang Shi, Pascal Wiggers, Catholijn M. Jonker

Man-Machine interaction group, Mediamatics department, EWI
Delft University of Technology, the Netherlands
shiyang1983@gmail.com

***Abstract*—We present a method for automatic classification of the socio-situational setting of a conversation based on the language used. The socio-situational setting depicts the social background of a conversation which involves the communicative goals, number of speakers, number of listeners and the relationship among the speakers and the listeners. Knowledge of the socio-situational setting can be used to search for content recorded in a particular setting or to select context-dependent models for example for speech recognition. We investigated the performance of different feature sets of conversation level features and word level features and their combinations on this task. Our final system, that classifies the conversations in the Spoken Dutch Corpus in one of 14 socio-situational settings, achieves an accuracy of 89.55%.**

I. Introduction

Language is situated. Conversations take place in a particular social context and documents are written with, among other things, a particular purpose and audience in mind. Knowledge of this socio-situational setting can greatly benefit language processing applications. For example, a search engine could only return those documents or videos that match a particular conversation style. In automatic speech processing, the socio-situational setting can be used to select dedicated language models and acoustic models for that context. In this paper we explore methods for the classification of the socio-situational setting of conversations, based on the language used.

The socio-situational setting can be characterized by situational features such as communicative goals, number of speakers participating, and the relationship between speakers and listeners. It influences the way people speak. In different settings we use a different speaking style and use different words. For example, a professor lecturing on a particular topic may place emphasis on important terms by repeating them and pronouncing them clearly. In a spontaneous conversation with one of his students about the same topic, the professor may articulate less carefully and use more informal speech and when explaining the topic to a family member he may avoid technical terms altogether.

As becomes clear in this example, the socio-situational setting of a conversation is independent of the topic of that conversation. The socio-situational setting can be seen as an aspect of genre. However, whereas a genre often denotes a particular set of stylistic and rhetoric elements as well as some content related aspects to classify a text, for example as fiction or mystery, the socio-situational setting as we define it here relates to broad categories of spoken language use such as spontaneous face-to-face conversations, debates or reading. Depending on the setting people may display differences in the acoustic and prosodic aspects of their speech as well as in the word use [1], [2]. In this work, we specifically look at language use.

The paper is organized as follows. In the next section, we give a brief overview of related work. In section 3, we describe the CGN corpus. Section 4 discusses the features that we extracted for socio-situational classification. Section 5 presents the classification experiments we performed with different feature sets. Finally, based on the results, conclusions are drawn.

II. Related work

Genre classification is a classical text classification problem. Kessler *et al.* [3] point out that by taking genre into account, parsing accuracy, part-of-speech (POS) tagging accuracy and word-sense disambiguation can be enhanced. In automatic speech recognition, language models are quite sensitive to genre changes, even if the changes are subtle [4]. For example, the perplexity of a language model trained on Dow-Jones newswire text will be doubled when it is applied to the very similar Associated Press newswire [4]. Genre dependent language models demonstrate perplexity reductions compared to global n-gram language models [5].

The fundamental problem of automatic genre classification is how to define genre. As noted by Kessler [3] and used in some studies [6], [7], [8], the genre is the way a text is created, the way it is distributed, the register of language it uses and the kind of audience it is addressed to, such as Editorial, Reportage and Research articles. Recently some studies [9], [10] focus on internet-based document genre classification, in which the genre includes different types of homepages, linklists and blogs.

In studies on automatic genre classification, various features have been proposed. Karlgren and Cutting [11] use some structural cues (such as adverb count, character count, sentence count), lexical cues ("Me" count, "Therefore" count, etc) and token cues (characters per sentence average, character

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

per word average, etc) with discriminant analysis. Kessler *et al.* [3] classify cues in four categories: structural cues (passives, topicalized sentences and counts of part-of-speech tags, etc), lexical cues (words in expressing date, title, etc), character-level cues (punctuations, separators, delimiters, etc) and derivative cues (ratios and variation measures derived from measures of lexical and character level features). They do not use structural cues since these require high computation in parsing and tagging text. Stamatatos *et al.* [8] propose an automatic text genre detection method of restricted text, using the frequencies of occurrence of the common words of an entire written language instead of a certain training corpus. Argamon *et al.* [2] exploits syntactic features in ten different genres in the British national corpus. More recently, Feldman *et al.* [12] propose the use of POS histograms instead of POS n-grams in naive Bayes models. In this paper, in addition to words and POS-tags, we propose some simple and low computation cost features such as sentence length, single occurrence word ratio and function word ratio.

These features are in part inspired by the work of van Gijsel *et al.* [13], who analyzed lexical richness of conversation from a socio-situational setting perspective. They showed that the lexical richness of texts is influenced by topic dependence as well as socio-situational effects. Conversations containing more informal, dialogic and/or spontaneous speech typically have lower type-token ratios than formal, monologic and/or prepared conversations.

III. The Spoken Dutch Corpus

Most prior studies focus on written text. Moreover, the corpora used are not designed according to genre categories. For example, the Brown corpus needs to be manually preprocessed to eliminate some texts that do not fall unequivocally into one of the predefined genre categories [3]. In contrast, the Corpus Spoken Dutch (Corpus Gesproken Nederlands, CGN) [14] we use in our experiments contain several socio-situational settings by design.

The CGN contains audio recordings of standard Dutch spoken by adults in Netherlands and Flanders. As shown in table I, it contains nearly 9 million words divided into 14 components that correspond to different socio-situational settings. Components comp-a to comp-h contain dialogues or multilogues and the components comp-i to comp-o contain monologues. For this research we combined components comp-c and comp-d as both contain spontaneous telephone dialogues. These components only differ in the recording platform used, which is not relevant for our work as we look at linguistic features. We performed all analyses and experiments described below on the transcripts of the recordings in the CGN. As these are transcripts of spoken language they do contain ungrammaticalities, incomplete sentences, hesitations and broken-off words. To limit the size of our vocabulary and guarantee reliable statistics, we only selected words that appeared at least three times in the whole data set. This resulted in a vocabulary of 44368 words. All other words were replaced by an out-of-vocabulary token.

TABLE I
Overview of the CGN

components	socio-situational setting	words
comp-a	Spontaneous conversations ('face-to-face')	2,626,172
comp-b	Interviews with teachers of Dutch	565,433
comp-c&d	Spontaneous telephone dialogues	2,062,004
comp-e	Simulated business negotiations	136,461
comp-f	Interviews/ discussions/debates	790,269
comp-g	(political) Discussions/debates/ meetings	360,328
comp-h	Lessons recorded in the classroom	405,409
comp-i	Live (eg sports) commentaries (broadcast)	208,399
comp-j	Newsreports/reportages (broadcast)	186,072
comp-k	News (broadcast)	368,153
comp-l	Commentaries/columns/reviews (broadcast)	145,553
comp-m	Ceremonious speeches/sermons	18,075
comp-n	Lectures/seminars	140,901
comp-o	Read speech	903,043

IV. Language socio-situational setting classification features

We extracted features at both the conversation level and the word level. The conversation level features are sentence length, single occurrence word ratio and function word ratio. The word level features are POS tags and words.

A. Sentence length

Wiggers *et al.* [15] show that the sentence length (SL) distribution varies for different socio-situational settings. For example, in spontaneous speech (comp-a, comp-c&d) the average sentence length is below 7. In spontaneous face-to-face conversations almost 25% of the sentences contain only one word such as yes or no answers and interjections. In contrast, the sentence length means in political discussion/debates/meetings (comp-f) and ceremonious speeches/sermons (comp-m) are 15 and 20 respectively.

In Fig. 1 shows the sentence length distribution of 6 components. The sentence length distribution in components comp-a, comp-e and comp-f are similar with each other, however, the mean and standard deviation (std) values are different.

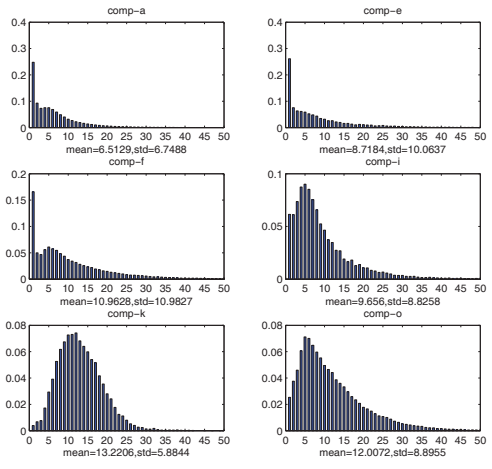

Fig. 1. Sentence length distribution of components a, e, f, i, k, o

B. Single occurrence word ratio

A word in the vocabulary which only appears once in a conversation is treated as a single occurrence word (SW). We calculate the single occurrence word ratio (SWR) of a conversation as the number of single occurrence words divided by the total number of words in the conversation. We find that the SWR distribution is different for different socio-situational settings. Fig. 2 shows some examples. In spontaneous speech (comp-a, comp-e), the SWR is less than for broadcasted speech such as interviews, discussion, debates (comp-f) and live commentaries and news report (comp-i, comp-k). Compared with other components, news broadcasts (comp-k) uses the most single occurrence words. The average SWR for news broadcasts is 0.627, while for example the SWR in business negotiations is below 0.1. Based on this analysis, we believe that the single occurrence word feature plays an important role.

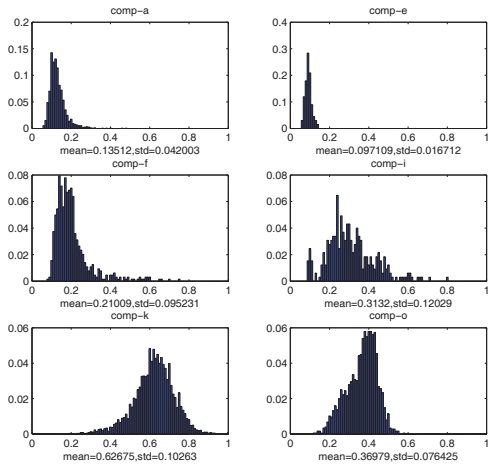

Fig. 2. Single occurrence word ratio distribution of components a, e, f, i, k, o

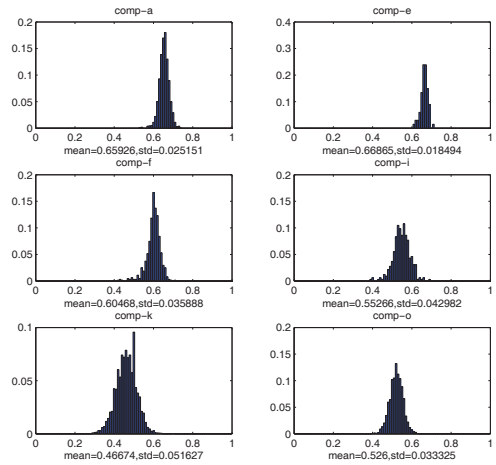

Fig. 3. Function word ratio distribution of components a, e, f, i, k, o

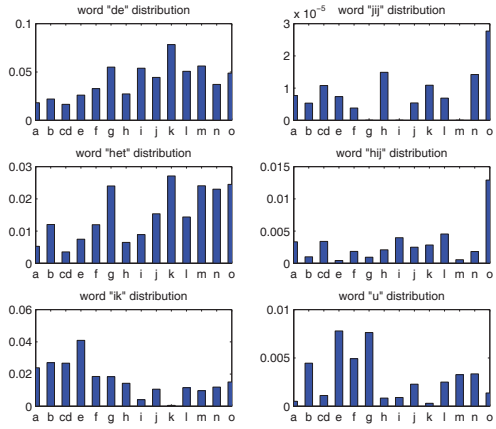

Fig. 4. The distribution of function words “de” (the), “het” (the), “ik” (I), “jij” (you), “u” (you, formal), “hij” (he)

C. Function words

While for topic classification function words are usually removed, function words can serve as important cues in socio-situational setting classification. Typically, more function words are used in spontaneous speech than in more formal speech [15]. For every conversation we calculate the function word ratio as the number of function words divided by the total number of words in that conversation. Fig. 3 shows that the CGN news broadcast has the smallest function word ratio, while business negotiations (comp-e) have the highest average function word ratio.

Not only does the function word ratio vary for different socio-situational settings, but the distributions of specific function words differ for different socio-situational settings. Fig.4 depicts the frequency distribution of 6 common function words over all components.

D. Words and POS-tags

The choice of words is context dependent. We can capture this by using the word frequencies of all words in the vocabulary as features as is done for many text classification tasks [3], [7], [6], [8], [12]. Part-of-speech tag frequencies also give useful information. For example, in spontaneous speech more adjectives are used on average than in formal speech, while in more formal setting more nouns are used on average [15].

Rather than using the direct frequency counts we apply a modified version of the term frequency inverse document frequency (tf-idf) metric, which is widely used in information retrieval [16], to calculate the weights of POS-tag and word features. The term frequency $\mathrm{tf}_{i,j}$ is the number of times term i appears in document j. The document frequency df_i is the number of documents that contain term i. Inverse document

frequency $\mathrm{idf}(i)$ can be calculated by:

$$\mathrm{idf}_i = \log(\frac{N}{\mathrm{df}_i}),$$

where N is the total number of documents. The tf-idf weight is the combination of $\mathrm{tf}_{i,j}$ and idf_i.

$$\mathrm{weight}(i,j) = \begin{cases} (1+\log(\mathrm{tf}_{i,j}))\mathrm{idf}_i & \mathrm{tf}_{i,j} > 0, \\ 0 & \mathrm{tf}_{i,j} = 0, \end{cases} \quad (1)$$

$\mathrm{weight}(i,j)$ indicates the importance of term i in discriminating document j from other documents. To emphasize terms that are discriminative for socio-situational setting, we modify the inverse document frequency as

$$\mathrm{idf}_i = \log(\sqrt{\frac{N}{\mathrm{df}_i}\frac{S}{\mathrm{sf}_i}}),$$

where S is the total number of socio-situational settings in the CGN, sf_i represents the number of socio-situational settings that contain term i.

V. Experiments

A. Data

We used the CGN corpus with the 14 different socio-situational settings shown in table I. The corpus comprises 12,767 manual transcripts of conversations. Of these, we randomly selected a test set of 2000 transcripts; the remaining transcripts were used as training data.

We represented each conversation as a feature vector. The dimension of the vector is determined by the features used to represent the data. We experimented with several subsets of the seven features discussed above: sentence length (SL), function word ratio (FWR), function word (FW), single occurrence word ratio (SWR), POS tags, POS-trigrams and words. Table II shows each of the subsets and the dimensions of the corresponding feature vectors.

B. Method

For classification we chose Support Vector Machines (SVMs) as these have shown good performance for high dimensional features spaces [17] and have successfully been applied in several text classification tasks [18], [19]. The basic idea of SVMs is to use a kernel function to map the features to a high dimensional space in which the different classes can be separated by a hyperplane. A hyperplane is found that separates the classes with a maximal margin. The kernel function that we used depends on the size of the feature vector. For small feature vectors, such as feature set 1, feature set 5 and feature set 9, we adopted the popular radial basis function (RBF)(2) as our kernel function:

$$K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2), \gamma > 0. \quad (2)$$

For large size document vector, we don't need to map data to a higher dimensional space, so the linear function (3) is applied as our kernel function:

$$K(x_i, x_j) = x_i^T x_j. \quad (3)$$

The classifiers using feature set 1, 3, 5 were trained with Libsvm [20] using C-SVM, the others were trained by Liblinear [21] using the L2-regularized L2-loss SVM. The scale parameters and regularization weights are calculated by grid search algorithm.

C. Result

The results of these classifiers on the test set are shown in table II. 'training method' shows the training method used and the last column is the prediction accuracy of each classifier. The lowest prediction accuracy was obtained by only using SL, FWR and SWR features; however, these features have the lowest computational cost. The highest prediction accuracy of 89.55% is achieved by combining SL, FWR, SWR, POS and word features.

Table III shows the confusion matrix of the best classifier in our experiments. Each column except the last one represents the label predicted by our classifier, each row stands for the correct label. The last column depicts the prediction accuracy of the classifier on every component. For example, row 'a' shows that 209 conversations in comp-a are correctly classified, and 6, 5, 1, and 2 conversations are wrongly classified to comp-c&d, comp-f, comp-h and comp-j, respectively. The third row shows that 32 of the conversations in comp-c&d are incorrectly classified as comp-a (while all others are classified correctly). The confusion between comp-a and comp-c&d is not surprising, as both contain spontaneous conversations. The only difference is that comp-a is face-to-face, while comp-c&d is by telephone. We can also see in table III that comp-b and comp-e are 100% correctly classified by our classifier. Component comp-l has the lowest accuracy. It is confused most often with components comp-j and comp-k – which are also confused with each other several times. All three of these components contain news related broadcasts. The low accuracy of comp-m most likely indicates that this component contains too little data to train a reliable classifier.

VI. Conclusion

We proposed a classifier that predicts the socio-situational setting of a conversation. We extracted the average sentence length, the single occurrence word ratio and the function word ratio as features at the conversation level and TF-IDF counts of words, POS tags, POS-trigrams and function words as features on the word level. Our experiments showed that a combination of conversation level features and word level features performs best with a classification accuracy of 89.55%. This is a 7% improvement over a unigram based classifier. The conversations that were classified incorrect were typically classified as a similar socio-situational setting. In the future, we plan to use this classifier to automatically select context specific language models.

References

[1] W. Labov, *Sociolinguistic patterns*. University of Pennsylvania Press, 1972.

[2] S. Argamon, M. Koppel, and G. Avneri, "Routing documents according to style," in *In Proceedings of First International Workshop on Innovative Information Systems*, 1998.

TABLE II
SELECTED FEATURE SETS, TRAINING METHOD AND RESULTS

feature set	features	dimension	training method	prediction accuracy
1	POS	326	libsvm	87.20%
2	words	44,368	liblinear	82.45%
3	FW	2,026	liblinear	83.65%
4	POS-trigrams	8,466	liblinear	80.80%
5	SL, FWR, SWR	4	libsvm	74.05%
6	SL, FWR, SWR and FW	2,030	liblinear	87.15%
7	POS and FW	2,352	liblinear	86.15%
8	POS and words	44,694	liblinear	88.85%
9	SL, FWR, SWR and POS	330	libsvm	87.85%
10	SL, FWR, SWR, FW and POS	2,356	liblinear	85.00%
11	SL, FWR, SWR and word	44,372	liblinear	87.40%
12	SL, FWR, SWR, FW and POS-trigrams	10,496	liblinear	85.40%
13	SL, FWR, SWR, and POS-trigrams	8,470	liblinear	84.45%
14	SL, FWR, SWR, POS-trigrams and words	52,838	liblinear	86.15%
15	FW and POS-trigram	10,492	liblinear	83.10%
16	POS-trigrams and words	52,834	liblinear	86.25%
17	POS and POS-trigrams	8,792	liblinear	82.70%
18	SL, FWR, SWR, POS and words	44,700	liblinear	89.55%

TABLE III
CONFUSION MATRIX OF TYPE 18 CLASSIFIER

A\P	a	b	c&d	e	f	g	h	i	j	k	l	m	n	o	accuracy
a	209	0	6	0	5	0	1	0	2	0	0	0	0	0	93.72%
b	0	32	0	0	0	0	0	0	0	0	0	0	0	0	100.00%
c&d	32	0	166	0	0	0	0	0	0	0	0	0	0	0	83.84%
e	0	0	0	16	0	0	0	0	0	0	0	0	0	0	100.00%
f	3	0	0	0	78	1	1	0	10	2	0	0	0	2	80.41%
g	1	0	0	0	1	35	0	0	0	1	0	0	0	1	89.74%
h	2	0	1	0	8	0	37	0	1	0	0	0	0	1	74.00%
i	0	0	0	0	1	0	0	46	3	4	1	0	0	0	83.64%
j	1	0	0	0	12	1	0	2	43	23	9	0	0	2	46.24%
k	0	0	0	0	1	0	0	0	6	874	1	0	0	2	98.87%
l	0	0	0	0	4	2	0	4	15	15	6	0	0	8	11.11%
m	0	0	0	0	0	1	0	0	0	0	1	1	0	0	33.33%
n	1	0	0	0	1	0	0	0	0	0	0	0	8	0	80.00%
o	0	0	0	0	0	0	0	0	0	4	2	0	0	240	97.56%

[3] B. Kessler, G. Numberg, and H. Schütze, "Automatic detection of text genre," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, ser. EACL '97. Stroudsburg, PA, USA: Association for Computational Linguistics, 1997, pp. 32–38.

[4] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the Ieee*, vol. 88, no. 8, pp. 1270–1278, 2000.

[5] Y. Shi, P. Wiggers, and C. M. Jonker, "Language modelling with dynamic bayesian networks using conversation types and part of speech information," in *The 22nd Benelux Conference on Artificial Intelligence (BNAIC)*, 2010.

[6] M. Santini, "A shallow approach to syntactic feature extraction for genre classification," in *7th Annual CLUK Research Colloquium*, 2004.

[7] Y.-B. Lee and S. H. Myaeng, "Text genre classification with genre-revealing and subject-revealing features," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '02. New York, NY, USA: ACM, 2002, pp. 145–150.

[8] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in *Proceedings of the 18th conference on Computational linguistics - Volume 2*, ser. COLING '00, 2000, pp. 808–814.

[9] G. Chen and B. Choi, "Web page genre classification," in *Proceedings of the 2008 ACM symposium on Applied computing*, ser. SAC '08. New York, NY, USA: ACM, 2008, pp. 2353–2357.

[10] M. Santini, "Some issues in automatic genre classification of web pages," in *In JADT 2006 - 8mes Journes*, 2006.

[11] J. Karlgren and D. Cutting, "Recognizing text genres with simple metrics using discriminant analysis," 1994, pp. 1071–1075.

[12] S. Feldman, M. Marin, M. Ostendorf, and M. Gupta, "Part-of-speech histograms for genre classification of text," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 4781 –4784.

[13] S. V. Gijsel, D. Speelman, and D. Geeraerts, "Locating lexical richness : a corpus linguistic, sociovariational analysis," *Les journes internationales danalyse des donnes textuelles JaDT Proceedings of the 8th International Conferene on the statistical analysis of textual data JADT*, vol. 2, pp. 961–972, 2006.

[14] N. Oostdijk, W. Goedertier, F. V. Eynde, L. Boves, J. pierre Martens, M. Moortgat, and H. Baayen, "Experiences from the spoken dutch corpus project," in *Araujo (eds), Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002, pp. 340–347.

[15] P. Wiggers and L. J. M. Rothkrantz, "Exploratory analysis of word use and sentence length in the spoken dutch corpus," in *Temporal Logic in Specification*, 2007, pp. 366–373.

[16] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, May 1999.

[17] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2009.

[18] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the European Conference on Machine Learning*, 1998.

[19] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, March 2002.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[21] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, Aug. 2008, software available at http://www.csie.ntu.edu.tw/ cjlin/liblinear.