Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 988–1002

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Classifying the socio-situational settings of transcripts of spoken discourses

Yangyang Shi [a,*], Pascal Wiggers [b], Catholijn M. Jonker [a]

[a] Intelligent Systems Department, Delft University of Technology, The Netherlands
[b] CREATE-IT Applied Research, Amsterdam University of Applied Sciences (HvA), The Netherlands

Received 29 May 2012; received in revised form 18 June 2013; accepted 19 June 2013
Available online 29 June 2013

Abstract

In this paper, we investigate automatic classification of the socio-situational settings of transcripts of a spoken discourse. 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 in speech recognition. The subjective experiment we report on in this paper shows that people correctly classify 68% the socio-situational settings. Based on the cues that participants mentioned in the experiment, we developed two types of automatic socio-situational setting classification methods; a static socio-situational setting classification method using support vector machines (S3C-SVM), and a dynamic socio-situational classification method applying dynamic Bayesian networks (S3C-DBN). Using these two methods, we developed classifiers applying various features and combinations of features. The S3C-SVM method with sentence length, function word ratio, single occurrence word ratio, part of speech (POS) and words as features results in a classification accuracy of almost 90%. Using a bigram S3C-DBN with POS tag and word features results in a dynamic classifier which can obtain nearly 89% classification accuracy. The dynamic classifiers not only can achieve similar results as the static classifiers, but also can track the socio-situational setting while processing a transcript or conversation. On discourses with a static social situational setting, the dynamic classifiers only need the initial 25% of data to achieve a classification accuracy close to the accuracy achieved when all data of a transcript is used.
© 2013 Elsevier B.V. All rights reserved.

Keywords: Socio-situational setting; Support vector machine; Dynamic Bayesian networks; Genre classification; Part of speech

1. Introduction

"You shall know a word by the company it keeps" (Firth, 1957). We also shall know a conversation by the situation which it is used. 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 may only return those documents or videos that match a particular speech style. In automatic speech processing, the socio-situational setting can be used to select dedicated language models and acoustic models for that context.

The socio-situational setting can be characterized by situational features such as: communicative goals, the number of speakers participating, and the relationship between speakers and listeners. It influences the way people speak. In different settings people use different speaking styles and different words. Socio-situational setting is a concept that is related to, but different from, the concepts of topic and genre that are well-known in the literature on natural language processing.

The socio-situational setting of a spoken discourse is independent of the topic of the discourse. 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

* Corresponding author. Address: HB12.290, Mekelweg 4, 2628CD Delft, The Netherlands. Tel.: +31 0681861586.
E-mail address: shiyang1983@gmail.com (Y. Shi).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.06.011

students about the same topic, the professor may articulate less carefully and use more informal language and when explaining the topic to a family member the technical terms might be missing altogether. Different types of spoken discourses can relate to the same topics. For example, in web search one might be looking for a lecture on Western civilization, rather than a political debate which refers to Western civilization.

The socio-situational setting is related to but different from the genre. It can be seen as an aspect of genre. However, 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 (Kessler et al., 1997). Depending on the setting people may display differences in the acoustic and prosodic aspects of their conversations as well as in the word use (Labov, 1972; Argamon et al., 1998). 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.

In this paper, we address socio-situational setting classification by automatic classifiers as well as humans. Two types of automatic classification methods are developed based on the features from the literature on automated document classification, see e.g., Kessler et al. (1997), and features that are based on the way humans do this classification task.

To obtain information about the performance of humans in this task and about the clues in the text they use to do the classification we set up and performed an experiment. In addition to the features, the subjective experiment of socio-situational setting classification provides a baseline for the automatic classification of socio-situational settings.

Two types of socio-situational setting classification methods are presented in this paper, which are a static socio-situational setting classification method using Support Vector Machines (SVM) (which we call S3C-SVM) and a dynamic socio-situational setting classification method using Dynamic Bayesian Networks (DBN) (which we call S3C-DBN). A set of static classifiers is constructed by the S3C-SVM method using different features as well as combinations of these features. In dynamic classifiers which are developed using the S3C-DBN method, we investigate the impact to the performance of classification not only from the perspective of different features but also from the perspective of dependencies among these features.

In the static classification method, we investigate the effect of sentence length, single occurrence word ratio, function word ratio, word, POS tags, POS trigrams and their combinations on the classification results. Static classifiers need to observe the complete discourse to do classification. When the context information is applied to language modeling (Iyer et al., 1994; Shi et al., 2010), the static classification method usually separates the usage of context information in language model testing into two sequential phases: one phase for context classification, the other phase for combining the context information into language modeling. Static classifiers are unsuitable for online classification as they need the complete text to make classifications. Therefore, in addition to the static classification methods, we propose a dynamic classification method of the socio-situational settings of a spoken discourse.

The dynamic classification method developed in this paper is an online classification method that sequentially processes text of a transcript. It reevaluates classification each time a word in the transcript is observed. We investigate how much of the text information is needed to achieve an acceptable classification accuracy. For example, guessing that the conversation beginning with "In this class, we will discuss something" is a lecture can be done with confidence, and the prediction that a conversation beginning with "Hello, this is Mike speaking.", is a spontaneous conversation by phone can also be made with confidence. Therefore, classifying the socio-situational setting of a spoken discourse is a task for which dynamic classification is highly appropriate. In fact, the results of dynamic classifiers in our experiments, reach their final classification results having processed about 25% of the text.

The dynamic classification method can benefit context based adaptive language modeling. Knowing the socio-situational setting of a spoken discourse would benefits context based adaptive language modeling, as the socio-situational setting is a form of context information. In addition, the dynamic socio-situational setting classification method introduced in this paper can make classification of socio-situational settings on the fly. Therefore, the dynamic classification method can be directly integrated into context based adaptive language models in online prediction for next word.

The paper is organized as follows. In the next section, we give an overview of related work. In Section 3, we describe the Spoken Dutch Corpus which we used as the test data set. Section 4 discusses the possible differences of discourses in terms of their socio-situational settings. Our subjective experiment is described in Section 5. Section 6 discusses the features that we extracted from the discourse transcripts. Section 7 presents the S3C-SVM classification methods and their classification results. Section 8 discusses the S3C-DBN classification method, the structure of different models and the classification results. Finally, we compare the results from the human experiment with the results of our automatic socio-situational stetting classifiers and draw conclusions.

2. Related work

In this section, we discuss related work in socio-situation setting classification, genre classification and the features used in genre classification. We also present some related work on Support Vector Machines (SVMS) and probabilistic classifiers.

The socio-situational setting classification is related to traditional genre classification. The fundamental problem

of automatic genre classification is how to define genre. As noted by Kessler et al. (1997) and used in some studies (Rosenfeld, 2000; Lee and Myaeng, 2002; Stamatatos et al., 2000), 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, Research articles etc. Some research (Chen and Choi, 2008; Santini, 2006) focus on internet-based document genre classification, in which the genre includes different types of homepages, linklists, blogs etc. In (Levinson, 1979; Ries et al., 2000), they use the terminology activities rather than the genre. They suppose that the choice of individual discourse is restricted by different goals. They categorize the dialogues into story-telling, planning, discussion, etc. In this paper, we propose the term of socio-situational setting which defines the social restriction of a speaker's utterances.

The genre classification can benefit practical applications. It is pointed out by Kessler et al. (1997) 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 sensitive to genre changes, even if the changes are subtle (Rosenfeld, 2000). For example, the performance of a language model trained on Dow–Jones newswire text will be seriously degraded when it is applied to the Associated Press newswire (Rosenfeld, 2000).

In studies on automatic genre classification of discourse, various features have been proposed. Some structural cues (such as adverb count, character count, sentence count), lexical cues ("Me" count, "Therefore" count, etc) and token cues (chars per sentence average, character per word average, etc) have been used with discriminant analysis by Karlgren and Cutting (1994). Their work has achieved a classification accuracy of 65% on a data set with 15 genres. In the work done by Kessler et al. (1997), the cues have been classified in four categories: structural cues (passive, topicalized sentences and counts of part-of-speech tags, etc), lexical cues (words in expressing date, title, etc), character-level cues (punctuation, separators, delimiters, etc) and derivative cues (ratios and variation measures derived from measures of lexical and character level features). Using the same data set as Karlgren and Cutting (1994), around 78% classification accuracy has been reported by Kessler et al. (1997). Using the frequencies of occurrence of the common words and punctuation markers of an entire written language instead of a certain training corpus, an automatic text genre detection method for restricted text has been proposed by Stamatatos et al. (2000). In the work by Stamatatos et al. (2000), more than 97% classification accuracy has been reported on the Wall Street Journal corpus of 1989 with 4 genres. The syntactic features in ten different genres in the British national corpus have been exploited by Argamon et al. (1998). More recently, the use of POS histograms instead of POS *n*-grams in naive Bayes models has been proposed by Feldman et al. (2009). However, all of this previous work is based on edited text rather than on spoken discourses.

In additional to words and POS-tags, we use some simple and low computational cost features such as: sentence length, single occurrence word ratio and function word ratio in socio-situational setting classification in this paper. These features are in part inspired by the work of Tong et al. (2002), who analyzed the type-token ratio of a speaker's utterances from the socio-situational setting perspective. The type-token ratio is the ratio of the number of different words to the number of total words in a text or speech. They show that the type token ratio 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.

Support Vector Machines (SVMS) (Cortes and Vapnik, 1995) are well suited for text classification (Joachims, 1998a). SVMS separate the data with a functional margin, which is not dependent on the number of features. In this paper, we apply the SVMS for the static classification of socio-situational settings.

Probabilistic classifiers offer alternative approaches to classification. One important probabilistic classifier in document classification is the naive Bayesian classifier described by Langley et al. (1992). The naive Bayesian classifier is extended by Pearl (1988) to a chain augmented naive Bayesian classifier, which can be viewed as a combination of a naive Bayesian classifier and an *n*-gram language model. In this paper, we present a dynamic Bayesian (DB) approach to socio-situational setting classification, and compare it with the static approach.

The performance of humans in a genre classification task is investigated before. In the work done by Obin et al. (2010), they investigate that whether participants use prosodic features in discourse genre identification. However, they did not propose to take advantage of this feature in an automatic classification methods. In this paper, we investigate people's performance in socio-situational setting classification as well as the performance of the automatic classifiers using the features mentioned by the participants in their classification task.

3. The Spoken Dutch Corpus

Previous genre classification 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 (Kessler et al., 1997).

In contrast, in the overall design of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) (Oostdijk et al., 2002; Oostdijk, 1999) which we use in our experiments, the principal parameter is taken to be the socio-situational setting. The recordings were collected along with

Table 1
Overview of the Spoken Dutch Corpus (CGN).

Components	Socio-situational setting	Words	Discourse
Comp-SC	Spontaneous conversations ('face-to-face')	2,626,172	1537
Comp-IT	Interviews with teachers of Dutch	565,433	160
Comp-ST	Spontaneous telephone dialogues	2,062,004	1230
Comp-BN	Simulated business negotiations	136,461	67
Comp-DD	Interviews/ discussions/debates	790,269	642
Comp-PD	(political) Discussions/debates/ meetings	360,328	248
Comp-LR	Lessons recorded in the classroom	405,409	265
Comp-LS	Live (eg sports) commentaries (broadcast)	208,399	325
Comp-NR	Newsreports/reportages (broadcast)	186,072	506
Comp-NB	News (broadcast)	368,153	5581
Comp-CC	Commentaries/columns/reviews (broadcast)	145,553	364
Comp-CS	Ceremonious speeches/sermons	18,075	16
Comp-LE	Lectures/seminars	140,901	78
Comp-RS	Read speech	903,043	1761

the socio-situational settings. Details about the construction of the CGN can be found in Oostdijk (1999).

The CGN contains audio recordings of standard Dutch spoken by adults in Netherlands and Flanders. As shown in Table 1, it contains nearly 9 million words divided into 14 components that correspond to different socio-situational settings. Components from comp-SC to comp-LR contain dialogues or multilogues and the components comp-LS to comp-RS contain monologues.

We performed all experiments and analyses described below on the correct 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 make statistics reliable, we only selected words that appeared at least three times in the whole data set. This resulted in a vocabulary of 44,368 words. All other words were replaced by an out-of-vocabulary token.

4. Differences among discourses from varied socio-situational settings

Socio-situational settings depict the social restrictions for spoken discourses. In this section, we analyze the differences among spoken discourses with different socio-situational settings from the following aspects: the social roles and the social goal of the participants in the discourses, and the social function of the discourses.

The social role of the participants in the discourse varies for the different socio-situational settings listed in Table 1. From "Spontaneous conversations ('face-to-face')" to "Lessons recorded in the classroom", the spoken discourses are dialogues or multilogues which need the participation from at least two speakers. In the rest of the socio-situational settings, usually there is only one speaker. In dialogue or mulitlogue situations, the participation of each speaker varies according to his or her social role. For example, in "Lessons recorded in the classroom" and "Interviews with teachers of Dutch", there usually is one dominant speaker, who speaks most of time. The others respond to the dominant speaker. However, in "Simulated business negotiations" and "(political) Discussions/debates/meetings", usually the dominant speaker is not easy to spot. In monologues, the differences in the social roles of the participants can be reflected by their different immersion and involvement. For example, in "News (broadcast)", the speakers usually depict the News from third-person perspective, in which the speakers have less immersion than the speakers in "Ceremonious speeches/sermons".

The social function of the discourse can also serve as a feature to characterize different socio-situational settings. Public formal discourse is different from the private informal discourse. For example, in "News (broadcast)", the speakers hesitate less and there are less incomplete sentences than in "Spontaneous conversations ('face-to-face')" and "Spontaneous telephone dialogues". For some special social events, discourses even have their own distinguishable syntactic structures and terminologies. This is for example the case in "Ceremonious speeches/sermons". The discourses bearing the function to disseminate knowledge usually have more repetitions than others, for example, "Lessons recorded in the classroom" and "Lectures/seminars".

The social goal of the participants also distinguishes some socio-situational settings from others. For example, in "Interviews with teachers of Dutch", the goal determines that in most cases, there is one speaker asking questions and the other one answering the questions. In "Spontaneous conversations ('face-to-face')" and "Spontaneous telephone dialogues", the social goal requires the involvement from participants. So there are many interruptions.

5. Socio-situational setting classification by humans

To get a feeling for the difficulty of the task and for possible features for classification, we set up a small experiment to answer the following questions:

1. What is the average accuracy people obtain in socio-situational setting classification?
2. How do humans do socio-situational setting classification and what kind of cues do people mention in socio-situational setting classification?

After reading a conversation thoroughly, a participant chose one of the 14 socio-situational settings listed in Table 1. In addition, the participant had to answer an open question on the kind of features which could help in socio-situational setting classification. Ten participants with a

Table 2
Overview of the experiment samples.

Comp	Sample	Socio-situational settings	Sentences	Words
SC	2	Spontaneous conversations ('face-to-face')	67	574
IT	2	Interviews with teachers of Dutch	89	812
ST	2	Spontaneous telephone dialogues	65	622
BN	1	Simulated business negotiations	36	398
DD	2	Interviews/ discussions/debates	90	765
PD	2	(political) Discussions/debates/ meetings	117	1271
LR	1	Lessons recorded in the classroom	43	485
LS	1	Live (eg sports) commentaries (broadcast)	5	49
NR	1	Newsreports/reportages (broadcast)	11	114
NB	2	News (broadcast)	35	324
CC	1	Commentaries/columns/reviews (broadcast)	15	171
CS	1	Ceremonious speeches/sermons	31	314
LE	1	Lectures/seminars	38	405
RS	1	Read speech	42	315

Master degree or higher, were invited to do the experiment. The age of the participants ranged from 27 to 44. Seven of them were male, the rest female.

As shown in Table 2, in total twenty samples were selected from the CGN. In comp-SC, comp-IT, comp-ST, comp-DD, comp-PD and comp-NB, two transcripts were randomly sampled. In the rest of the components, only one sample was randomly selected. Each participant was asked to label exactly the same twenty pieces of transcripts in different order. All the selected samples are directly used without length normalization. In this way, the classification made by participants is based on the same information as the automatic classifiers.

Table 3 shows the confusion matrix of human performance in socio-situational setting classification. The human prediction accuracy ranges from 30% to 75%. The average prediction accuracy of the participants is 68%. The standard deviation is 13.75%. People identified "(political) Discussions/debates/meetings", "Ceremonious speeches/sermons" and "Read speech" with 100% accuracy. People achieved low classification accuracy on "Spontaneous telephone dialogues", "Lessons recorded in the classroom", "Live (e.g. sports) commentaries (broadcast)", "Commentaries/columns/reviews" and "Lectures/seminars". Half of the participants misclassified spontaneous telephone dialogues as spontaneous face to face dialogue. Seventy percent of the participants classified the "Lessons recorded in classroom", as an "Interview with a teachers of Dutch". In Table 3, we find that people could tell news related broadcasting apart from other categories (eg. spontaneous conversation), but they made mistakes in telling apart "Live (e.g. sports) commentaries (broadcast)", "News reports/ reportages (broadcast)", "News (broadcast)" and "Commentaries/columns/reviews (broadcast)".

Based on the answers of the participants to the second question, we compiled a list of cues that were repeatedly mentioned. The detailed statistics of these cues are listed in Table 4.

SN gives the number of speakers involved in the conversation. For example, the speaker number of a spontaneous conversation is two, while it is one in read speech. In 51 out of 200 answers, the number of speakers is mentioned as an important cue.

Table 3
Confusion matrix of human classification on socio-situational setting.

Comp	SC	IT	ST	BN	DD	PD	LR	LS	NR	NB	CC	CS	LE	RS	Sum	ac(%)
SC	15		5												20	75
IT	2	15	1		2										20	75
ST	10		10												20	50
BN				8	2										10	80
DD					16	1			2		1				20	80
PD						20									20	100
LR		7			1		1						1		10	10
LS								5			5				10	50
NR									6	3	1				10	60
NB									7	12	1				20	60
CC					1				4		5	1			10	50
CS												10			10	100
LE						2						5	2	1	10	20
RS														10	10	100

Table 4
Features human reported in classification. "-" means this feature is not mentioned by participant in that classification. √ indicates the feature is mentioned. The symbol "m" in the first row means the number of the speakers in the conversation is bigger than 2. In the row of "SL", "S" and "L" stand for short and long average sentence length, respectively. In the row "IS", the "C" stands for complete structure; "I" for incomplete structure. "II" means informal and interruptive, "QA" refers to the question-answer style conversation like interview, "FF" means formal and fluent. "time" column list the times people mention these features in classification.

F\C	SC	IT	ST	BN	DD	PD	LR	LS	NR	NB	CC	CS	LE	RS	Time
SN	2	2	2	m	2	m	2	–	1	1	-	1	1	1	51
SL	S	–	S	-	L	–	–	–	–	–	–	–	–	L	11
IS	I	C	–	I	C	–	–	–	–	C	–	C	–	C	11
SW	√	√	√	√	√	√	–	–	–	√	–	–	√	–	20
SS	–	√	–	–	√	√	–	–	–	√	√	√	√	–	24
CT	√	√	√	√	√	√	√	√	√	√	√	√	√	√	47
FM	II	QA	II	II	QA	FF	II	–	FF	FF	–	FF	II	–	31

SL stands for the average sentence length in a conversation. The average sentence length is shorter in spontaneous conversation than in formal lectures or read speech. A third of the answers related to spontaneous conversations mentioned this cue.

IS depicts whether a spoken discourse has disfluency, hesitations and incomplete structures or not. For example, in News reports or ceremonies, the discourse is well prepared and contains less hesitation, disfluency and incomplete structure than spontaneous conversations.

SW Special words or lexicons are also reported by participants in their classification. For example, some participants identify a conversation as spontaneous because it contains many short words like "ja", "nee" , "uh" and "mm".

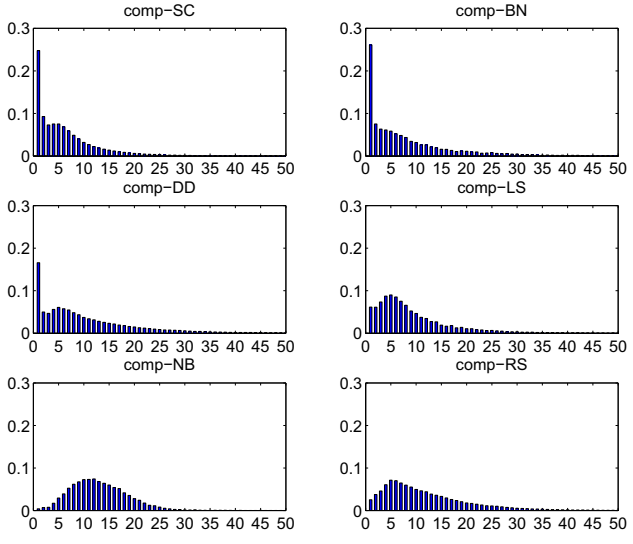

Fig. 1. Sentence length distribution of components SC, BN, DD, LS, NB, and RS. The distribution varies among all components. Here we use components SC, BN, DD, LS, NB, and RS as examples. Horizontal direction stands for sentence length. Vertical direction stands for the probability of the sentence length in one component. Each bar represents the ratio of the number of sentences with the given length to the total number of sentences in that component.

SS Special sentences clearly characterize some socio-situational setting. For example, all the participants correctly identify a discourse as an "Interview/discussion/debate (broadcast)", because they noticed a special sentence: "welkom in de studio" (welcome to the studio). In fact, 25 out 200 answers mentioned the special sentences in classifying socio-situational settings.

CT Content and topic take 23.5% of all cues mentioned by participants in our experiments. For example, in spontaneous conversation, some content reflects that speakers have visual connection with each other. In a sermon, the content is religion related.

FM Formality characterizes the way a discourse is structured. It reflects the social status of each speaker in the conversations. For example, spontaneous conversations are informal and involve many interruptions. In interviews, the conversation generally could be in a question-answer style. According to the questionnaire results, we categorize "FM" into the following 3 types: informal and interruptive (II), question-answer style (QA), formal and fluent (FF).

6. Language socio-situational setting classification features

Based on the results described in the previous section and on the literature mentioned in Section 2, we extracted features at both the discourse level and the word level. The discourse level features are sentence length, single occurrence word ratio and function word ratio. The word level features are POS tags and words.

6.1. Sentence length

Wiggers and Rothkrantz (2006) show that the sentence length (SL) distribution varies for different socio-situational settings. For example, in the CGN, for spontaneous speech (comp-SC, comp-ST) 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

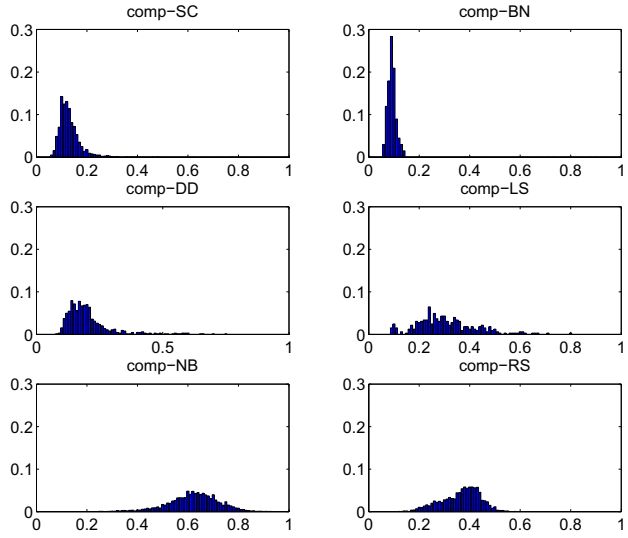

Fig. 2. Single occurrence word ratio distribution of components SC, BN, DD, LS, NB, and RS. The distribution varies among all components. Here we use components SC, BN, DD, LS, NB, and RS as examples. Horizontal direction stands for single occurrence word ratio. Vertical direction stands for the probability of the single occurrence word ratio. Each bar represents the ratio of the number of transcripts that have the given single occurrence word ratio to the total number of transcripts in that component.

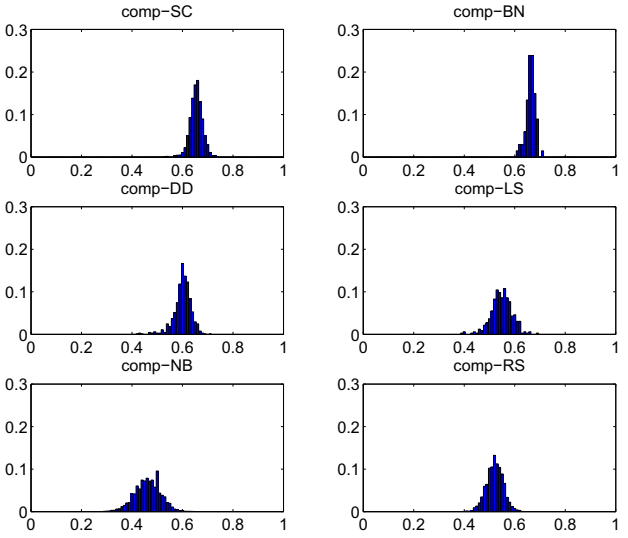

Fig. 3. Function word ratio distribution of components SC, BN, DD, LS, NB, and RS. The distribution varies among all components. Here we use components SC, BN, DD, LS, NB, and RS as examples. Horizontal direction stands for function word ratio. Vertical direction stands for the probability of function word ratio. Each bar represents the ratio of the transcripts that have the given function word ratio to total number of transcripts in that component.

means of sentence length in "(political) Discussion/debates/meetings" (comp-PD) and "Ceremonious speeches/sermons" (comp-CS), are 15 and 20, respectively. Fig. 1 shows the sentence length distribution of 6 CGN components.

6.2. Single occurrence word ratio

A word in the vocabulary that only appears once in a conversation is treated as a single occurrence word (SW). We calculate the single occurrence word ratio (SWR) of a discourse 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-SC, comp-BN), the SWR is less than for example broadcasted speech such as "(political) Discussion/debates/meetings" (comp-PD) and live commentaries and news report (comp-LS, comp-NB). Compared with other components, "News(broadcasts)" (comp-NB) uses 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 in socio-situational setting classification.

6.3. Function words

While for topic classification function words are usually removed, function words can serve as important cues in socio-situational setting classification.

Typically, the relative number of function words is higher in spontaneous speech than in more formal speech (Wiggers and Rothkrantz, 2006). For every discourse we calculate the function word ratio as the number of function words divided by the total number of words in that discourse. Fig. 3 shows that the CGN news broadcast component (comp-NB) has the smallest function word ratio, while business negotiations (comp-BN) have the highest average function word ratio.

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

6.4. 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 (Kessler et al., 1997; Lee and Myaeng, 2002; Rosenfeld, 2000; Stamatatos et al., 2000; Feldman et al., 2009).

In addition to words, 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 speech more nouns are used on average (Wiggers and Rothkrantz, 2006).

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 (Baeza-Yates and Ribeiro-Neto, 1999), to

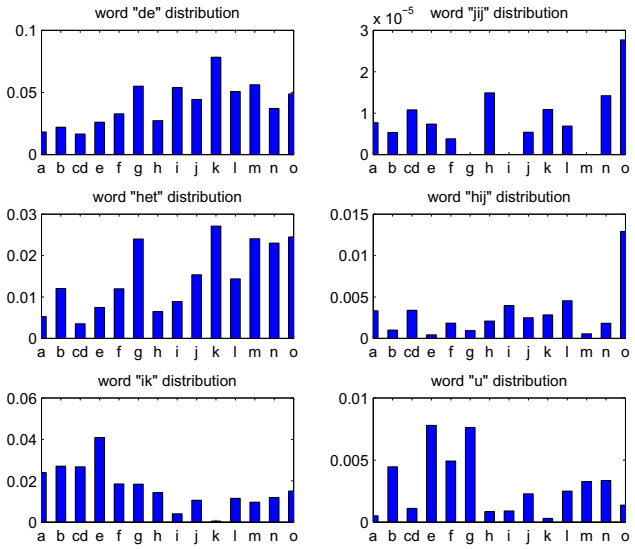

Fig. 4. The distribution of function words "de" (the), "het" (the), "ik" (I), "jij" (you), "u" (you, formal), "hij" (he). Horizontal direction stands for components. Vertical direction stands for the frequency of the special function word in each component. In order to illustrate the distribution shape clearly, different vertical direction scales were chosen.

calculate the weights of POS-tag and word features. The (tf-idf) helps to reduce the weight of common POS-tag and word features which have little discriminative power and to increase the weight of rare features which have much discriminative power. 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:

$$\mathrm{idf}_i = \log\left(\frac{N}{\mathrm{df}_i}\right), \tag{1}$$

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

$$\mathrm{weight}(i,j) = \begin{cases} (1+\log(\mathrm{tf}_{i,j}))\mathrm{idf}_i & \mathrm{tf}_{i,j} > 0, \\ 0 & \mathrm{tf}_{i,j} = 0. \end{cases} \tag{2}$$

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

$$\mathrm{idf}_i = \log\left(\sqrt{\frac{N}{\mathrm{df}_i}\frac{S}{\mathrm{sf}_i}}\right), \tag{3}$$

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. In fact, this modification is intend to average the inverse document frequency with inverse socio-situational setting frequency in terms of logarithm value.Based on the extracted features such as sentence length, single occurrence word ratio, function words, words and POS-tags discussed in this section, we will show two socio-situational setting classification methods in the following sections.

7. Static socio-situational setting classification

For static socio-situational classification, we chose Support Vector Machines (SVMS) as these have shown good performance for high dimensional features spaces (Theodoridis and Koutroumbas, 2009) and have successfully been applied in several text classification tasks (Joachims, 1998b; Tong and Koller, 2009).

We represented each spoken discourse 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 5 shows each of the subsets and the dimensions of the corresponding feature vectors.

Depending on the size of document vectors, different kernel functions are used in our experiment. For small feature vectors, such as feature set 1, feature set 4 and feature set 8, we adopted the radial basis function (RBF)(4) as our kernel function:

$$K(x_i,x_j) = exp(-\gamma\|x_i - x_j\|^2), \quad \gamma > 0. \tag{4}$$

For large size document vectors, we do not need to map data to a higher dimensional space, so the linear function (5) is applied as our kernel function:

$$K(x_i,x_j) = x_i^T x_j. \tag{5}$$

The classifiers using feature set 1, 3, 5 were trained with Libsvm (Chang and Lin, 2011) using C-SVM, the others were trained by Liblinear (Fan et al., 2008) using the L2-regularized L2-loss SVM. For small data sets such as set 1, set 5, and set 9, the grid parameter search algorithm (Chang and Lin, 2011) is directly applied to calculate the scale parameters and regularization weights. When dealing with large data sets, a small subset is randomly selected to calculate the parameters by the grid parameter search algorithm. The results of these 18 SVM classifiers using different feature sets, are shown in Table 5. 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. This classifier also gets the best accuracy of 88.62%, when 10 fold cross validation is used.

Table 6 shows the confusion matrix of the best classifier in our experiments. Each column except the last one represents the label obtained from the automatic classification, each row stands for the correct label. The last column depicts the classification accuracy of the classifier on every component.

The third row in Table 6 shows that 32 of the conversations in comp-ST are incorrectly classified as comp-SC (while all others are classified correctly). The misclassification between comp-SC and comp-ST is not surprising, as both contain spontaneous conversations. The only difference is that comp-SC is face-to-face, while comp-ST is

Table 5
Selected feature sets and their related classifiers prediction accuracy.'C/γ' refers to the penalty parameter C and the kernel parameter γ. Both C and γ can be represented as exponentiation 2^n. In the table, we show the power n of C and γ. The 'dim' stands for dimension and 'ac' column for the prediction accuracy of the SVM classifiers on the test data. The 'cv ac ' gives the 10 folders cross validation accuracy of the SVM classifiers.

Set	Features	Dim	C/γ	ac (%)	cv ac (%)
1	POS	326	5/−5	87.20	86.74
2	Words	44,368	−3/0	82.45	81.08
3	FW	2026	−2/0	83.65	85.25
4	POS-trigrams	8,466	−3/0	80.80	83.74
5	SL, FWR, SWR	4	7/3	74.05	74.48
6	SL, FWR, SWR and FW	2,030	−4/0	87.15	86.83
7	POS and FW	2352	−1/0	86.15	87.58
8	POS and words	44,694	−1/0	88.85	88.56
9	SL, FWR, SWR and POS	330	1/−3	87.85	87.11
10	SL, FWR, SWR, FW and POS	2356	−3/0	85.00	84.91
11	SL, FWR, SWR and word	44,372	−3/0	87.40	88.02
12	SL, FWR, SWR, FW and POS-trigrams	10,496	−5/0	85.40	86.72
13	SL, FWR, SWR, and POS-trigrams	8470	−3/0	84.45	85.35
14	SL, FWR, SWR, POS-trigrams and words	52,838	−4/0	86.15	87.04
15	FW and POS-trigram	10,492	−1/0	83.10	86.87
16	POS-trigrams and words	52,834	−2/0	86.25	85.82
17	POS and POS-trigrams	8,792	−3/0	82.70	85.83
18	SL, FWR, SWR, POS and words	44,700	−3/0	89.55	88.62

Table 6
Confusion matrix of type 11 classifier.

comp	SC	IT	ST	BN	DD	PD	LR	LS	NR	NB	CC	CS	LE	RS	ac(%)
SC	209		6		5		1		2						93.72
IT		32													100.00
ST	32		166												83.84
BN				16											100.00
DD	3				78	1	1		10	2				2	80.41
PD	1				1	35				1				1	89.74
LR	2		1		8		37		1					1	74.00
LS					1			46	3	4	1				83.64
NR	1				12	1		2	43	23	9			2	46.24
NB					1				6	874	1			2	98.87
CC					4	2		4	15	15	6			8	11.11
CS						1					1	1			33.33
LE	1				1								8		80.00
RS										4	2			240	97.56

by telephone. As is discussed earlier, we found the same confusion for human classification.

We can also see in Table 6 that comp-IT and comp-BN are 100% correctly classified by our classifier. Component comp-CC has the lowest accuracy. It is confused most often with comp-NR and comp-NB – which are also confused with each other several times. All these three components contain news related broadcasts. The low accuracy of comp-CS most likely indicates that this component contains too little data to train a reliable classifier.

Fundamentally the performance of the automatic classification is jointly determined by the training data size as well as the distinguishability of corresponding components. In general, the classifier can get better accuracy with more data to train on. A Large training data set can improve the classification accuracy. For the distinguishable components, our results seem to show that even a small data set is sufficient for training a good classifier. For example, on the relatively small components such as comp-IT and comp-BN, the classifier actually obtain 100% accuracy. However, when the training data becomes too small, the distinguishability of the specific component is easy to be ignored by the automatic classification. For example, all the automatic classification methods get low classification accuracy on comp-CS, even though this component is obviously different from other components from a human's perspective.

8. Dynamic Bayesian document classification

The static classification method treats each document as a whole. For applications such as adaptive language modeling, this is not desirable. Therefore, we also investigate a dynamic classification method of socio-situational setting using dynamic Bayesian Networks (S3C-DBN). This method updates the classification result for each word that is

observed. Before introducing the classifier, we briefly discuss DBNS.

8.1. Dynamic Bayesian networks

Bayesian networks are methods for reasoning with uncertainty based on Bayes rule of probability theory (Pearl, 1988). A Bayesian network represents the joint probability distribution over a set of random variables $\mathbf{X} = X^1, X^2 \ldots X^N$. It consists of two parts:

1. A directed acyclic graph (DAG) G. The variables X^i in the domain X are mapped one-to-one to the nodes v^i of G. The directed arcs in the network represent the direct dependencies between variables. The absence of an arc between two nodes means that the variables corresponding to the nodes do not directly depend on each other.
2. A set of conditional probability distributions (CPD). With each variable X^i a conditional probability distribution $P(X^i|Pa(X^i))$ is associated, which quantifies how X^i depends on $Pa(X^i)$, the set of variables represented by the parents of the node v^i in G representing X^i.

Dynamic Bayesian networks (DBNS) (Dean and Kanazawa, 1989; Murphy, 2002) are an extension of Bayesian networks. They can model probability distributions of semi-infinite sequences of variables that evolve over time. A DBN can be represented by two Bayesian networks: an a priori model $P(X_1)$ and a two slice temporal Bayesian network which defines the dependence between the variables at a particular step and the variables at the previous time step:

$$P(X_t|X_{t-1}) = \prod_{i=1}^{N} P(X_t^i|Pa(X_t^i)), \tag{6}$$

where $\mathbf{X}_t$ is the set of random variables at time t and X_t^i is the ith variable at time step t. $Pa(X_t^i)$ are the parents of X_t^i.

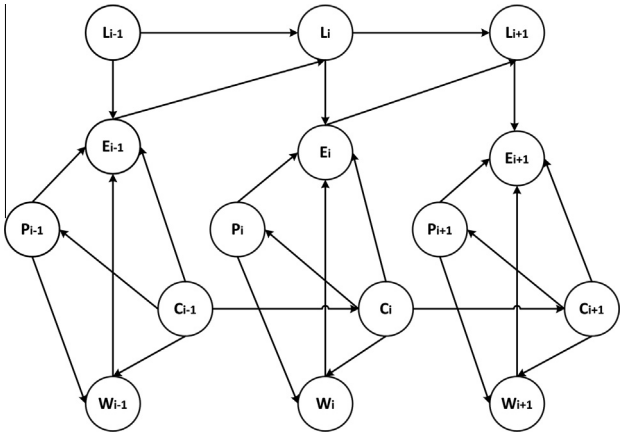

Fig. 5. Unigram DB classification model with words, POS-tags and sentence length. W_i, P_i, C_i and L_i stand for current word, POS-tags, classification label and sentence length, respectively. E_i indicates that whether current word is the end of a sentence.

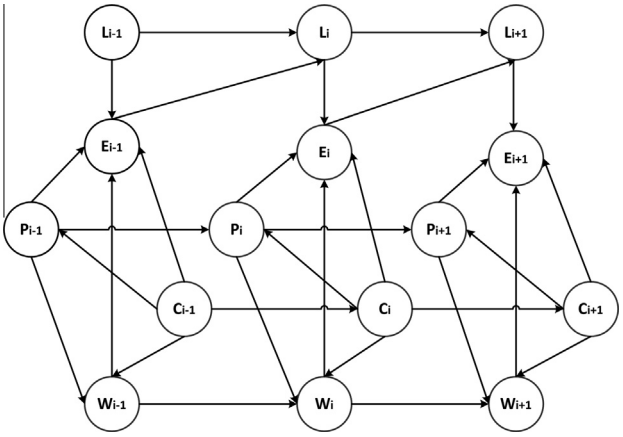

Fig. 6. Bigram DB classification model with words, POS-tags and sentence length. W_i, P_i, C_i and L_i stand for current word, POS-tags, classification label and sentence length, respectively. E_i indicates that whether current word is the end of a sentence.

8.2. Dynamic Bayesian document classifier

As before, we classify discourses based on their lexical transcripts. This can be seen as document classification, which maps a document d to one of a set of predefined classes $\mathbf{C} = \{c^1, c^2, \ldots, c^n\}$. In this paper, 18 different DB classification models are implemented. Words, POS-tags and sentence length and their combinations are used as features.

8.2.1. Unigram DB classification

Fig. 5 shows the graphical structure of the interpolated unigram DB classification model. The interpolated conditional probability of words in the unigram DB classification method is:

$$P_{int}(w_t|c_t) = \lambda_1 P(w_t) + \lambda_2 P(w_t|c_t). \tag{7}$$

In case of using the combination of words and POS-tags, the interpolated probability is:

$$P_{int}(w_t|pos_t, c_t) = \lambda_1 P(w_t) + \lambda_2 P(w_t|c_t) + \lambda_3 P(w_t|pos_t) + \lambda_4 P(w_t|, pos_t, c_t). \tag{8}$$

L in Fig. 5 represents the current word position in a sentence. Together with the end of sentence node E, it reflects the sentence length. The relation between C, L and E indicates that different socio-situational settings have different sentence length distributions. The interpolation method is also applied in computation of the conditional probabilities for L_t and E_t:

$$P_{int}(l_t|l_{t-1}, e_{t-1}) = \alpha_1 P(l_t|l_{t-1}) + \alpha_2 P(l_t|e_{t-1}) + \alpha_3 P(l_t), \tag{9}$$

$$P_{int}(e_t|l_t, p_t, c_t, w_t) = \beta_1 P(e_t|l_t) + \beta_2 P(e_t|p_t) + \beta_3 P(e_t|c_t) + \beta_4 P(e_t|w_t) + \beta_5 P(e_t). \tag{10}$$

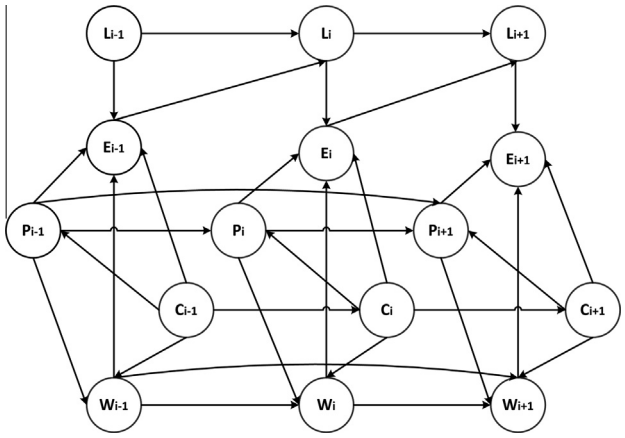

Fig. 7. trigram DB classification model with words, POS-tags and sentence length. W_i, P_i, C_i and L_i stand for current word, POS-tags, classification label and sentence length, respectively. E_i indicates that whether current word is the end of a sentence.

8.2.2. Bigram DB classification

The bigram DB classification using the combination of word, POS and sentence length, is depicted in Fig. 6. These models assume a 1-order Markov chain. The models which only use some of these features are sub-graphs of Fig 6.

For bigram DB classification only using words or POS tags, the features at a particular time step t only depend on the features at $t-1$ and the current hidden variable c_t. For example, the following Eq. (11) gives the interpolated conditional probability of w in the bigram DB classification model which only considers the word feature:

$$P_{int}(w_t|w_{t-1},c_t) = \lambda_1 P(w_t|w_{t-1},c_t) + \lambda_2 P(w_t|w_{t-1},c_t) + \lambda_3 P(w_t). \quad (11)$$

For bigram DB classification models using both word and POS features, the current word w_t depends on the previous word w_{t-1} as well as the current p_t and socio-situational setting c_t. The interpolated conditional probability of current word w_t is calculated by:

$$\begin{aligned} P_{int}(w_t|w_{t-1},p_t,c_t) = {} & \lambda_1 P(w_t) + \lambda_2 P(w_t|c_t) \\ & + \lambda_3 P(w_t|w_{t-1},c_t) + \lambda_4 P(w_t|p_t) \\ & + \lambda_5 P(w_t|w_{t-1},p_t) \\ & + \lambda_6 P(w_t|w_{t-1},p_t,c_t). \end{aligned} \quad (12)$$

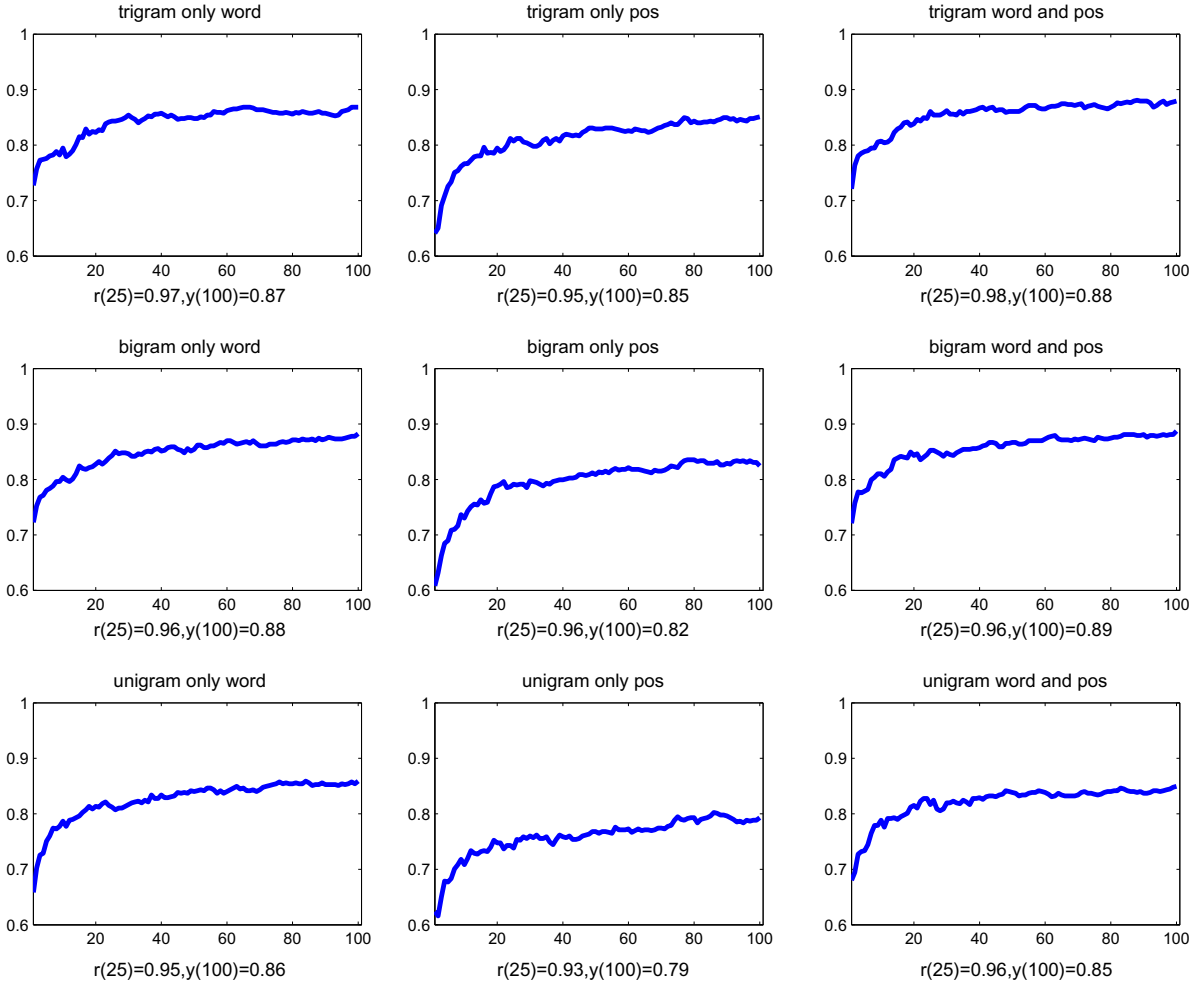

Fig. 8. Classification accuracy trend over percent of each conversation, x, y axis represent the percentage of a conversation and prediction accuracy, respectively. $y(100)$ represent prediction accuracy using 100% information, $r(25) = y(25)/y(100)$.

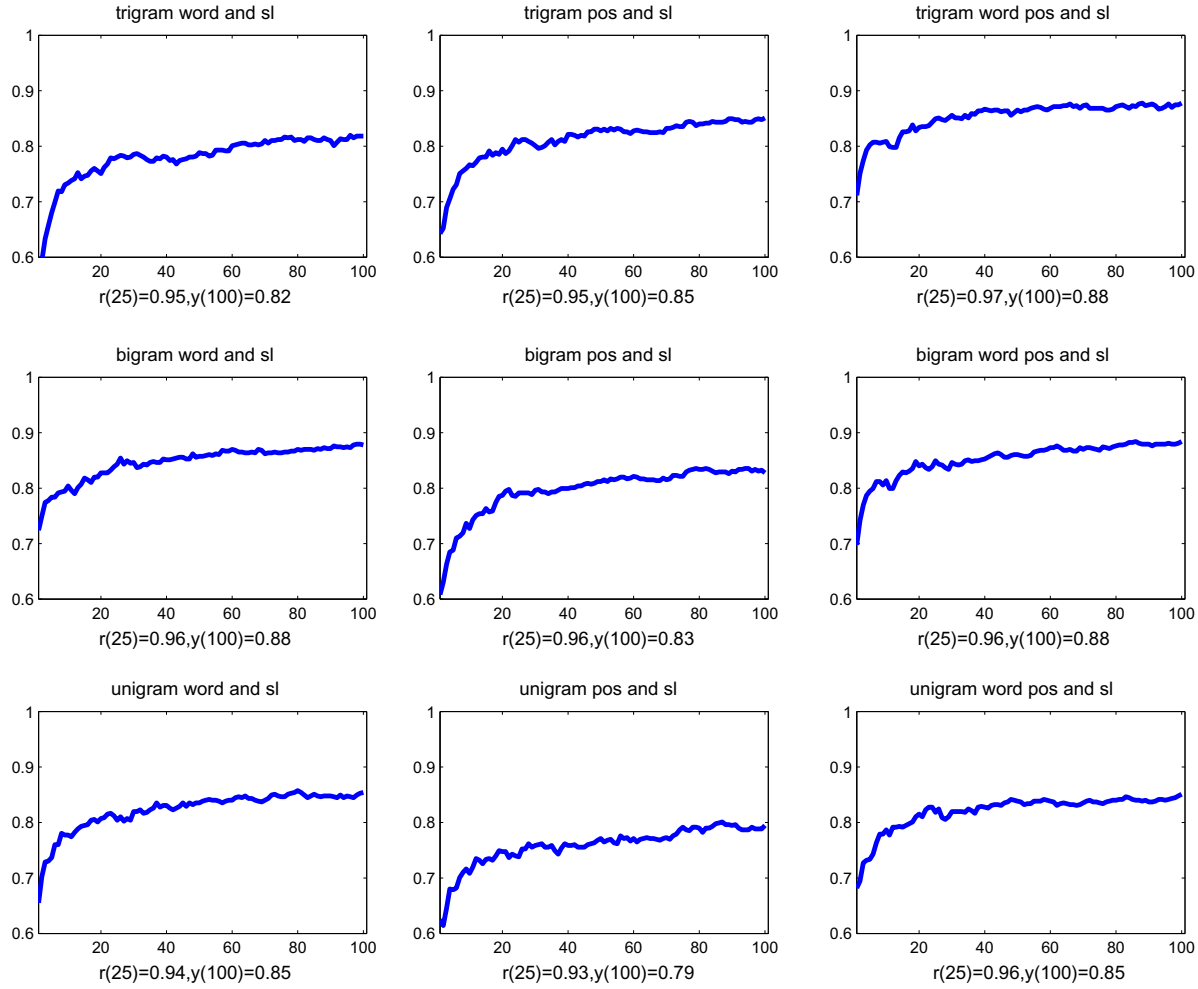

Fig. 9. Classification accuracy trend over percent of each conversation, x, y axis represent the percentage of a conversation and prediction accuracy, respectively. $y(100)$ represent prediction accuracy using 100% information, $r(25) = y(25)/y(100)$.

8.2.3. Trigram DB classification

A 2nd order Markov chain is applied in these models. Fig 7 shows the trigram DB classification model using word, POS and sentence length features. The word and POS features in this case depend on the features of the previous two time slices.

The conditional probability of the current word w_t given w_{t-1}, w_{t-2} and c_t is given by (13).

$$P_{int}(w_t|w_{t-1},w_{t-2},c_t) = \lambda_1 P(w_t|w_{t-1},w_{t-2},c_t) + \lambda_2 P(w_t|w_{t-1},c_t) + \lambda_3 P(w_t|c_t) + \lambda_4 P(w_t). \quad (13)$$

In the trigram DB model combining word and POS features, the POS-tags are conditioned on the previous two POS-tags and the current socio-situational setting. The conditional probability of POS can be calculated using Eq. (13). The w_t in this case, depends on w_{t-1}, w_{t-2}, as well as on the current pos_t and socio-situational setting class label c_t. The following Eq. (14) gives the interpolation of the conditional probability of w_t:

$$\begin{aligned} P_{int}(w_t|w_{t-1},w_{t-2},pos_t,c_t) = {} & \lambda_1 P(w_t) + \lambda_2 P(w_t|c_t) \\ & + \lambda_3 P(w_t|w_{t-1},c_t) \\ & + \lambda_4 P(w_t|pos_t) \\ & + \lambda_5 P(w_t|w_{t-1},w_{t-2},c_t) \\ & + \lambda_6 P(w_t|w_{t-1},pos_t) \\ & + \lambda_7 P(w_t|w_{t-1},w_{t-2},pos_t) \\ & + \lambda_8 P(w_t|w_{t-1},w_{t-2},pos_t,c_t). \end{aligned} \quad (14)$$

In Eqs. (9)–(14), all interpolated parameters are treated as hidden variables in DB models. These parameters are trained on the held-out development set.

8.3. Experiment

To test the DB classifiers we once again used the CGN data set. Of the data set 80% was randomly selected as training data, 10% was selected as developing data, and the remaining 10% was treated as testing data.

Table 7
The prediction accuracy of 18 dynamic classifiers.

Models	Information	Prediction accuracy		
		25% Data (%)	50% Data (%)	100% Data (%)
Trigram	word	84.33	84.80	86.83
	POS	80.72	82.92	85.11
	word, POS	86.05	86.05	87.93
	word, sl	78.06	78.84	81.82
	POS, sl	80.72	82.76	85.11
	word, POS, sl	84.95	86.52	87.77
Bigram	word	84.33	85.42	88.24
	POS	79.15	80.88	82.45
	word, POS	85.27	86.68	88.71
	word, sl	84.33	85.74	87.77
	POS, sl	79.15	81.19	82.76
	word, POS, sl	84.95	86.05	88.40
Unigram	word	81.19	84.01	85.89
	POS	73.82	76.80	79.31
	word, POS	81.66	83.86	84.95
	word, sl	80.41	83.54	85.42
	POS, sl	73.82	77.12	79.47
	word, POS, sl	81.82	83.86	85.11

Figs. 8 and 9 show the prediction accuracy of the 18 classifiers as a function of the percentage of test data observed. The exact classification accuracies of the 18 classifiers with 25, 50 and 100 percent data are listed in Table 7. In terms of overall performance, the DB classifier using POS-tag and word bigrams, which achieves a classification accuracy of 88.71%, performs best among the 18 classifiers. Its confusion matrix is shown in Table 8. As is indicated in Fig. 8 and Fig. 9, the classification accuracy increases rapidly for the first 20% of the data, then flattens. The DB classifiers using only words are more stable and precise than the classifiers that use only POS-tags. Based on 1% of the information, both trigram and bigram DB classifiers using words can correctly classify 70% the discourses, while systems that use only POS-tags achieve less than 65% accuracy.

In this section, we show and compare the 18 dynamic socio-situational setting classifiers. In the following section, we discuss the relationship among the static classification, dynamic classification and human classification.

9. Discussion

Comparing the confusion matrices of the static, dynamic, and human classification we find three similarities and three differences. The similarities are:

1. The confusion between spontaneous face to face dialogue (comp-SC) and spontaneous telephone dialogue (comp-ST) is the main cause of misclassification in both components. In both components, the spoken discourses have many short ungrammatical sentences, repetitions and repairs. People usually use fewer determiners in spontaneous conversations than in read speech.
2. In all experiments, "Read speech" (comp-RS) is classified with high accuracy. Both human and dynamic classifiers can correctly classify all the "Read speech" (comp-RS). The static classification method can correctly classify more than 97% "Read speech" (comp-RS).
3. The sub-matrix of comp-LS, comp-NR, comp-NB and comp-CC of each confusion matrix has relative high density. There are non-zero values on non-diagonal elements. In particular, in the human based experiment and static classification method, the misclassification of these four news related components is caused by the confusion with each other.

The differences are:

1. In classifying lectures in the classroom (comp-LR), humans performed much worse than the static classifiers and the dynamic classifiers. In the confusion matrix in Table 3, seven out of ten people mistook the lectures in the classroom to be the interview with a Dutch teacher. Even though participants knew that the content of the conversations was about teaching, most of them were still misled by the question/answer style between teacher and one student.

Table 8
Confusion matrix of bigram DB classifier with word and POS.

comp	SC	IT	ST	BN	DD	PD	LR	LS	NR	NB	CC	CS	LE	RS	ac (%)
SC	72	2	11												84.71
IT		8													100.00
ST	8		55		1									0	85.94
BN				1										0	100.00
DD	3		1		29	1									85.29
PD	1					13									92.86
LR	1						11								91.67
LS	1							19							95.00
NR	5				6	1			5	4				1	22.73
NB					1			1		264				3	98.14
CC	1		2		7	1				2	5			5	21.74
CS						1									0.00
LE					1								4	2	57.14
RS														80	100.00

2. In classifying interviews with teachers (comp-IT), both static and dynamic classification methods got 100% classification accuracy. But in the subjective experiment, participants only achieve a 75% classification accuracy. Table 3 shows that some participants categorized the interview as a spontaneous conversation.
3. The participants got 100% accuracy in classifying ceremonious speech/sermons, but both the static and dynamic classification method do not perform well in these cases. The reason probably is that there is limited training data in these components of the CGN.

10. Conclusion

This paper studies the classification of socio-situational setting of a spoken discourse based on word level transcripts by humans and by automatic classification methods. The differences among socio-situational settings of discourses are discussed from the following perspectives: the social role and the social goals of the participants in the discourse, and the social function of the discourse.

In order to get a baseline for socio-situational setting classification, a subjective experiment was performed in which participants were asked to classify the socio-situational settings of discourses. The experimental results show that people can correctly classify 68% of the socio-situational settings. Inspired by the features mentioned by the participants, we extracted the average sentence length, the single occurrence word ratio and the function word ratio as features on the discourse level and TF-IDF counts of words, POS tags, POS-trigrams and function words as features on the word level.

A static S3C-SVM classification method was constructed with these features. The experiments on the static classifiers show that a combination of discourse level features and word level features performed best with a classification accuracy of almost 90%.

In addition to the static S3C-SVM classification method, a S3C-DBN method was proposed, which can achieve similar classification accuracy as the S3C-SVM method, but also can make socio-situational setting classification on a word-by-word basis. We experimented with 18 different S3C-DBN classifiers. In particular, the best S3C-DBN classifier we developed was the bigram DB classifier using word and POS tags which obtained a classification accuracy of almost 89%.

Both the static and the dynamic classifiers can be applied to provide the context information for language models. When the static classification methods are applied, the usage of the socio-situational setting information in the language models needs two phases, one phase for obtaining the socio-situational setting information by classifying the discourses, the other phase for applying the information in language modeling. The advantage of the dynamic classifiers is that they can provide online classification results to word level language models. The experimental results show that all the S3C-DBN classifiers using the initial 25% of the text in the transcripts can get at least 93% of the accuracy which they achieved on the complete transcripts.

In comparison, both the static and the dynamic classifiers outperform the human participants. Our experiments show that some socio-situational settings, such as "read speech", are easy to identify, as both humans and all automated classifiers we developed scored 100% accuracy on these discourses.

Acknowledgments

The authors would like to thank the anonymous reviewers for their excellent comments.

References

Argamon, S., Koppel, M., Avneri, G., 1998. Routing documents according to style. In: Proceedings of First International Workshop on Innovative Information Systems.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison Wesley.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27.

Chen, G., Choi, B., 2008. Web page genre classification. In: Proceedings of the 2008 ACM Symposium on Applied, Computing, pp. 2353–2357.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

Dean, T., Kanazawa, K., 1989. A model for reasoning about persistence and causation. Computational Intelligence 5 (3), 142–150.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. Liblinear: a library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.

Feldman, S., Marin, M., Ostendorf, M., Gupta, M., 2009. Part-of-speech histograms for genre classification of text. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4781–4784.

Firth, J.R., 1957. A synopsis of linguistic theory. In: Firth, J.R. (Ed.), Studies in Linguistic Analysis. Basil Blackwell, Oxford, pp. 1930–1955.

Iyer, R., Ostendorf, M., Rohlicek, J.R., 1994. Language modeling with sentence-level mixtures. In: Proceedings of the Workshop on Human Language Technology, pp. 82–87.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of 10th European Conference on, Machine Learning, pp. 137–142.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of 10th European Conference on, Machine Learning, pp. 137–142.

Karlgren, J., Cutting, D., 1994. Recognizing text genres with simple metrics using discriminant analysis. In: Proceedings of the 15th Conference on Computational linguistics, vol. 2, pp. 1071–1075.

Kessler, B., Numberg, G., Schütze, H., 1997. Automatic detection of text genre. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 32–38.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press.

Langley, P., Iba, W., Thompson, K., 1992. An analysis of Bayesian classifiers. In: Proceedings of the Tenth National Conference on, Artificial Intelligence, pp. 223–228.

Lee, Y.-B., Myaeng, S.H., 2002. 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, pp. 145–150.

Levinson, S.C., 1979. Activity types and language. Linguistics 17 (5-6), 365–400.

Murphy, K., 2002. Dynamic Bayesian networks: representation, inference and learning. Ph.D. Thesis, University of California, Berkeley.

Obin, N., Dellwo, V., Lacheret, A., Rodet, X., 2010. Expectations for discourse genre identification: a prosodic study. In: Proceedings of Interspeech, pp. 3070–3073.

Oostdijk, N., 1999. Building a corpus of spoken dutch. URL <http://lands.let.kun.nl/cgn/>.

Oostdijk, N., Goedertier, W., Eynde, F.V., Boves, L., Pierre Martens, J., Moortgat, M., Baayen, H., 2002. Experiences from the spoken dutch corpus project. In: Araujo (Eds.), Proceedings of the Third International Conference on Language Resources and Evaluation, pp. 340–347.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems – Networks of Plausible Inference. Morgan Kaufmann Publishers, Inc..

Peng, F., Schuurmans, D., 2003. Combining naive Bayes and n-gram language models for text classification. In: 25th European Conference on Information Retrieval, Research, pp. 335–350.

Ries, K., Levin, L., Valle, L., Lavie, A., Waibel, A., 2000. Shallow discourse genre annotation in callhome spanish. In: Proceecdings of the International Conference on Language Ressources and Evaluation.

Rosenfeld, R., 2000. Two decades of statistical language modeling: where do we go from here?. Proceedings of the IEEE 88 (8) 1270–1278.

Santini, M., 2004. A shallow approach to syntactic feature extraction for genre classification. In: Seventh Annual CLUK Research Colloquium.

Santini, M., 2006. Some issues in automatic genre classification of web pages. In: JADT 2006–8mes Journes.

Shi, Y., Wiggers, P., Jonker, C.M., 2010. Language modelling with dynamic Bayesian networks using conversation types and part of speech information. In: The 22nd Benelux Conference on, Artificial Intelligence, pp. 154–161.

Stamatatos, E., Fakotakis, N., Kokkinakis, G., 2000. Text genre detection using common word frequencies. In: Proceedings of the 18th Conference on Computational Linguistics, vol. 2, pp. 808–814.

Theodoridis, S., Koutroumbas, K., 2009. Pattern Recognition, fourth ed. Academic Press.

Tong, S., Koller, D., 2009. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66.

Van Gijsel, S., Speelman, D., Geeraerts, D., 2006. Locating lexical richness: a corpus linguistic, sociovariational analysis. Les journes internationales danalyse des donnes textuelles JaDT. In: Proceedings of the Eighth International Conference on the Statistical Analysis of Textual Data JADT 2, pp. 961–972.

P. Wiggers, L.J.M. Rothkrantz, Exploratory analysis of word use and sentence length in the spoken dutch corpus. In: Proceedings of the International Conference on Text, Speech and Dialogue, pp. 366–373.