Features for modelling characteristics of conversations

Notebook for PAN at CLEF 2012

Gunnar Eriksson and Jussi Karlgren

Gavagai AB
Skånegatan 97, 116 35 Stockholm
www.gavagai.se
{ guer | jussi }@gavagai.se

Abstract In this experiment, we find that features which model interaction and conversational behaviour contribute well to identifying sexual grooming behaviour in chat and forum text. Together with the obviously useful lexical features — which we find are more valuable if separated by who generates them — we achieve very successful results in identifying behavioural patterns which may characterise sexual grooming. We conjecture that the general framework can be used for other purposes than this specific case if the lexical features are exchanged for other topical models, the conversational features characterise interaction and behaviour rather than topical choice.

1 Introduction

1.1 Identifying sexual grooming behaviour

The goal of this experiment is to identify one specific class of authors in written interactive on-line chat forum logs, namely authors who attempt to convince other participants to provide sexual favours. The task involves both identifying specific users and identifying those conversational turns which are typical of of sexual grooming behaviour. The details of the task are given in the introduction to the "Uncovering Plagiarism, Authorship and Social Software Misuse" Lab of the 2012 CLEF conference.

1.2 Experimental material

The material for the two tasks were chat site conversation logs. The training material consisted of 66 928 conversations in which 97 689 different people participated. 142 of them were considered to exhibit sexual grooming behaviour in the conversations, and they participated in 2026 of the conversations. The corresponding figures for the test collection were 155 129 conversations with 218 702 different authors. The number of people pre-assessed by human judges to exhibit sexual grooming behaviour was 250, and they participated in 3 740 conversations. We will here adhere to the terminology of the workshop and refer to users who exhibit sexual grooming behaviour as "predators", the conversations they partake in as "marked conversations" and other users in those conversations as "counterparts".

2 Exploring the uncharted conversational landscape

To model the chat site behaviour of sexual predators and the conversations they participate we initially inspected a number of marked conversations visually. The observations made gave rise to a number of hypotheses:

Hypothesis: Sexual predators frequently participate in two-part conversations – dialogues – and in short conversations without a counterpart – monologues. The participation in group conversations with more than two participants is more uncommon and presumably less conducive to negotiating sexual favours from conversational counterparts.

Hypothesis: The length of marked conversations is usually very short (mostly in the case of monologues) or rather long with a greater number of turns.

Hypothesis: Sexual predators and their counterparts often have had previous on-line conversations and a marked conversation is often a continuation of a previous one.

Hypothesis: Introduction of conversational topics is frequently done by the counterpart.

Hypothesis: Both parts seemed often to be willing to elaborate on a topic chosen by the other conversator.

Based on these observations and hypotheses it seemed reasonable to model sexual predator behaviour on different feature sets: sets which capture the various topics chosen by conversational participants and their elaboration of them, as well as sets which capture conversational turn-taking, conversation length and number of conversational participants.

3 Topical modelling – lexical features

3.1 Lexical tokens

The bulk of lexical features were word token unigrams and bigrams, extracted from each utterance after normalization and tokenization.

This process was straightforward and simplistic: each string of non-whitespace characters delimited by whitespace was considered a proto-token. In each such token, each leading or trailing sequence of punctuation characters was then considered a token of its own. This gave a sequence of unigram tokens consisting of maximal strings of alphanumeric and non-alphanumeric characters. Finally, all alphanumeric characters were translated into lower case.

This unigram sequence of tokens was then input to bigram extraction, and also fed into a process where special unigram tokens were identified and in some cases modified to tailor the lexical features towards the task at hand and to the content of chat site conversations: References to the names of conversation interlocutors was generalised to OTHER and mentions of the participant's own name were generalized to SELF. These features were combined with information on conversation type, cf. below in Sec. 4 , to produce participant "names" such as SELF or OTHER–GROUP. Furthermore, URLs in the utterance's unigram sequence were separated so that each URL component was represented as a unigram of its own, alongside with its full URL representation.

Feature	# of utterances
FAIL	0-1
HANDSHAKE	2-7
PRELUDE	8-25
BRIEF	26-50
DISCOURSE	51-100
LDISCOURSE	101-160
VLDISCOURSE	161-

Table 1. Conversation length

3.2 Vocabulary of self and others

All unigrams and bigrams obtained through this process constituted a person's primary vocabulary to model the topical range of a person: what they are talking *about* across a (set of) conversation(s). Below, this feature set is referred to as SLEX.

But since not only the things that you yourself utter to another person determine the content of a conversation, we also, for each participant, let the joint vocabulary of the primary lexical features of each other conversation participant become an additional set of lexical features for the participant at hand. This secondary vocabulary was meant to model the things that other people say *to* a participant across a (set of) conversation(s). This feature set is referred to as OLEX.

As an alternative to the above disjoint sets of SLEX and OLEX features, we also tested the combination of these sets, to facilitate comparison between the above split of the vocabulary of a conversation with a simpler model which would use the joint vocabulary of all participants as the topical feature set for the conversation. This all-in conversation feature type is referred to as CLEX.

4 Conversational modeling

The initial observations of the training corpus led us to believe that the number of participants in a conversation, the length of it, and how people behaved towards the other interlocutor(s) in terms of turn-taking was significant and useful for the identification task.

First, type of discourse was determined by the number of participants. We implement a simple categorisation into three different categories: MONOLOGUE, DIALOGUE, or GROUP for conversations with 1, 2, and >2 participants, respectively.

Secondly, each conversation was assigned to one of seven length categories based on the number of utterances in the conversation. The categories are defined after manual inspection of the trial corpus and are given in Tab. 1.

These two dimensions are included in each person's feature list both separately and jointly. The participation in a BRIEF group conversation would thus increment a person's features BRIEF, and GROUP, but also the joint feature BRIEF–GROUP. The rationale behind these cross-product features was the observation that group conversations tend to

Feature type	Comment	Examples
SLEX	lexical features of SELF:	'darling'–SL, 'i am'–SL, 'bored'–SL
OLEX	lexical features of OTHERS:	'darling'–OL, 'i am'–OL
CLEX	comb. of SLEX and OLEX:	'darling', 'i am', 'bored'
CTYPE	# of conversation participants:	DIALOGUE, MONOLOGUE
LTYPE	conversation length:	DISCOURSE, BRIEF–GROUP
TTAKE	interlocutor turn-taking:	SELF − OTHER-IN-GROUP − SELF, OTHER − END

Table 2. Features types

be longer than dialogues but consist of fewer utterances per participant and that thus the separation of brief dialogues from brief group conversations would seem motivated.

To model aspects of interlocutor behaviour, we formulate features to model turn-taking, opening, and closing of each conversation a person participates in. By the same rationale which separated conversation length categorizations, each utterance in a conversation is annotated with who generated it (SELF vs OTHER-IN-DIALOGUE vs OTHER-IN-GROUP). This annotation is used to generate trigram features such as SELF – OTHER-IN-DIALOGUE –SELF (normal turn-taking) or SELF – SELF –SELF (monologic turn-taking). Openings and closings are marked with START and END in these trigram sequences to yield features such as START – SELF – OTHER-IN-GROUP.

5 The full feature set

The features described above, lexical and conversational, were collated to a conversation and topical profile for each participant. The frequency information of the features for a participant was removed, so that the mere presence of a feature in the profile of a participant was used for classification. Table 2 gives an overview of the feature types with examples.

6 Training and classification

The bag-of-features profiles of each person in the set were used for the classification of predators. We used a freely available off-the-self maximum entropy classifier[1] with standard settings. Features with occurrences < 2 in the data, i.e. features that occurred with only one person, were discarded in the resulting classifier models.

For the predator identification subtask, different combinations of the feature types were tested in a 10-fold cross validation setup in which one tenth of the authors was used for test and the remaining nine tenths of authors for training. A model with a stable good performance was selected to classify the participants of the test corpus.

For lexically oriented subtask of identifying specifically distinctive utterances of most characteristic of grooming behaviour, only the SLEX feature set was used. In the training material, the features were extracted separately for each utterance of persons

[1] http://search.cpan.org/∼laye/AI-MaxEntropy-0.20/lib/AI/MaxEntropy.pm

Conversation type features	Predator	Non-predator	P %	Non-p %
MONOLOGUE	1840	27722	49.20	18.31
DIALOGUE	1898	103964	50.75	68.67
GROUP	2	19703	0.05	13.01

Table 3. Conversation type features

Conversation length features	Predator	Non-predator	P %	Non-p %
FAIL	1123	15590	30.03	10.30
HANDSHAKE	882	99582	23.58	65.78
PRELUDE	335	14262	8.96	9.42
BRIEF	319	11919	8.53	7.87
DISCOURSE	623	7574	16.66	5.00
LDISCOURSE	458	2449	12.25	1.62
VLDISCOURSE	0	13	0.00	0.01

Table 4. Conversation length features

marked as predators in the first sub-task. These SLEX profiles of utterances were fed to the resulting classifier from the first task to obtain a "predator score", a probability score for this feature set being a viable component for a predator profile. All utterances where then ranked by this score, and after visual inspection of the ranking, a threshold was set for categorisation. When repeating this procedure for the test corpus, all utterances above the set threshold were selected as examples of grooming behaviour.

7 Results

Conversation type Table 3 confirms the hypothesis given in Section 2, based on initial visual observations of the training data: predators do indeed tend to participate in monologues and dialogues, and do not participate in group conversations.

Conversation length Table 4 again confirms the hypothesis given in Section 2, based on initial visual observations of the training data: FAIL, DISCOURSE, and LDISCOURSE are overrepresented for the predator profiles, meaning that they participate disproportionately often in extremely short or very long conversations. This provides a partial explanation to the number of monologues — they tend to be test shots where the person attempts to initiate a conversation without receiving any response, here marked as FAIL.

Turntaking Given the above results, it will not be surprising to find that turn-taking features also reflect the tendency of predators to initiate brief monologous conversations: the feature START-SELF-END is overrpresented among predators as are other related features; the fact that non-predators engage in group conversations whereas predators

do not give the predators an attendand underrepresentation for features such as OTHER-OTHER-OTHER. Also, the features for vanilla conversational turntaking such as OTHER-SELF-OTHER and its converse SELF-OTHER-SELF have a clear overrepresentation in predator profiles than in non-predators', reflecting the predator willingness to engage in lengthy and attentive grooming conversations.

7.1 Feature class usefulness for identifying users who exhibit sexual grooming behaviour

Using a classifier on the complete feature set (OLEX+SLEX+CTYPE+LTYPE+TTAKE) gives a first indication of the strength of the respective features. Table 6 demonstrates the relative strength of the various features, with a large number of lexical features together with some highly indicative conversational features.

Table 5 goes on to show the relative contribution of each class of feature, measured by performance on the test set. The first line of the table demonstrates the quality of the complete feature set, which was the one submitted as our official experimental run. The second line demonstrates the effect of combining the SLEX and OLEX features into the CLEX set. We can conclude that separating the two sets is worth while.

The third group of lines shows the effect of either SLEX and OLEX. We see here that the contribution of OLEX is more valuable than that of SLEX.

The fourth group of lines shows the effect of either SLEX and OLEX without any conversational features. We see here that the contribution of the conversational features is considerable.

The fifth group of lines shows, somewhat unsurprisingly, that on their own, conversational features fail to deliver. Lexical features are necessary for identifying predators: without a topical model, the conversational features cannot pick out the specific class of users under consideration here, i.e. those attempting to negotiate sexual favours from their counterparts.

The sixth group of lines first shows how adding one conversational feature set at a time to the lexical features yields similar results. They each have a similar contribution to the end result. Second, it shows how adding two of the conversational feature sets at a time gives the best results on the training set, better than using the full feature set. The reason for this interference has not been established yet: the absolutely best results were from the combination where the LTYPE feature set was dropped.

8 Results

In the official results of the workshop we submitted one run for each subtask.

For the first subtask, that of identifying users who exhibit sexual grooming behaviour, our submission using the full feature set was the best submission if measured by the original F (β=1) scoring scheme and third best if measured by the later F (β=0.5) precision-oriented scoring scheme. It identified 265 users as potential predators, whereof 223 out of the 250 known predators were assessed to be correctly identified, yielding a precision of 0.8415 and a recall of 0.8920.

Features types used	Precision	Recall	F (β=1)
OLEX+SLEX+CTYPE+LTYPE+TTAKE	0.84	0.89	0.87
CLEX+CTYPE+LTYPE+TTAKE	0.56	0.55	0.55
OLEX+CTYPE+LTYPE+TTAKE	0.90	0.79	0.84
SLEX+CTYPE+LTYPE+TTAKE	0.64	0.86	0.73
SLEX	0.21	0.84	0.33
OLEX	0.27	0.80	0.35
OLEX+SLEX	0.43	0.85	0.57
CLEX	0.48	0.72	0.58
CTYPE	0	0	0
LTYPE	0	0	0
TTAKE	0	0	0
CTYPE+LTYPE+TTAKE	0.68	0.04	0.06
OLEX+SLEX+LTYPE	0.86	0.85	0.86
OLEX+SLEX+CTYPE	0.88	0.78	0.83
OLEX+SLEX+TTAKE	0.89	0.90	0.90
OLEX+SLEX+CTYPE+LTYPE	0.90	0.90	0.90
OLEX+SLEX+LTYPE+TTAKE	0.95	0.93	0.94
OLEX+SLEX+CTYPE+TTAKE	0.97	0.97	0.97

Table 5. Model performance

For the second subtask, the lexically oriented subtask of identifying specifically distinctive utterances of most characteristic of grooming behaviour, our SLEX features set was used. It turned out overgenerate, hurting recall, and our feature set retrieved 10 416 lines whereof 1 116 were assessed to be correctly identified out of some 6 000 possible hits, yielding a precision of 0.1071 and a recall of 0.1743.

9 Analysis and findings

We found in this experiment that conversational features strengthened precision in identifying sexual grooming behaviour in chat and forum text. On their own, without lexical features, as dicussed in Section 7.1, conversational features cannot predict sexual grooming behaviour, but we can conjecture that the general framework can be used for other purposes than this specific case. If the topical model — i.e. the lexical features — are exchanged for other topical models, other categories of user might be identified using conversational features, since they characterise conversational behaviour rather than topical choice.

Features type	Predator class	Non-predator class
CTYPE	1	2
LTYPE	1	6
TTAKE	0	26
tot. conversation feat.	2	34
OLEX	264	186
SLEX	734	780
tot. lexical feat.	998	966

Table 6. Feature types among top 1000 strongest features for each class in model with all feature types used.