Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification
Notebook for PAN at CLEF 2012

José María Gómez Hidalgo[1] and Andrés Alfonso Caurcel Díaz[2]

[1] Optenet, Research and Development Department, Las Rozas, Spain
[2] Universidad Politécnica de Madrid, Dept. of Applied Intelligent Systems, Madrid, Spain
{jgomez,andres.caurcel}@optenet.com

Abstract In this paper we present a system for sexual predator detection which combines two different approaches: a knowledge-based system that makes use of pattern matching according to hand-coded patterns that represent typical predator behaviors, and a learning-based system which employs surface linguistic features like capitalization and chat-like expressions. These approaches are combined in a chained fashion, being the learner applied to the suspicious predators as reported by the knowledge-based system. While the results of the system evaluation on the training collection are nice, the test run results for the official Sexual Predator Identification sub-task have shown much room for improvement.

1 Introduction

This paper describes the system designed and used by Optenet in the PAN Author Identification task, Sexual Predator Identification sub-task. The system is intended to validate our previous (unreported) research on predator detection in Spanish language, plus our current research on age detection in Social Networks [1,9][3].

In order to achieve this goal, we have adapted a previously existing knowledge-based system that makes use of hand-coded predation patterns in Spanish language. Additionally, we have augmented this system with learning capabilities based on linguistic features, trying to separate adults from children, and from adults posing as children. The overall results of our system when evaluated on the training collection are nice, but the system has demonstrated very poor performance on the test runs. We believe this is mainly due to three reasons: first, the adaptation of the knowledge-based system has been done by the automatic translation of patterns from Spanish to English; second, we do not make use of learning on any kind of word-level features (words, word n-grams, character n-grams, etc.); finally, the previously existing knowledge-based system makes use of some additional language-dependent mechanisms that we have not adapted to English due to time restrictions.

In the next sections, we present the general architecture and processing approach of the system. We also describe the modules included in it, along with some examples and conclusions.

[3] These references are available in Spanish language only.

2 General Architecture

We have built a system that is composed of two modules:

1. A knowledge-based conversation filter (KBF), which processes single-speaker conversations, and retains predator utterances while discarding most neutral and victim conversations. This module analyzes each sentence and applies pattern matching to detect suspicious ones, based on a set of hand-crafted patterns which correspond to typical predator behavior. This sub-system largely reuses a previously existing knowledge-based system for predator detection instant messaging conversations written in Spanish.
2. A learning-based detection sub-system for Chat Language (CHL), which makes use of chat-like and linguistic features, and it is trained on the retained conversations. This sub-system highlights the candidate predators according to their language/writing style. This approach is being tested on age detection in Social Networks.

We adopt a chaining approach. In the training phase, we have applied the following process to the training data:

- All conversations are processed by the KBF, which reports a subset of users as potential predators.
- Those conversations in which a user has been reported as a potential predator[4], are used as a training collection for the CHL. This module learns a classifier which is stored for the classification phase.

We roughly follow the same approach in the classification phase. The only difference is that, instead of using the CHL sub-system to learn a classifier, we use it to classify those users marked as suspicious by the KBF module.

The Sexual Predator Identification sub-task requires not only to spot candidate predators, but to highlight suspicious sentences in their conversations. During the classification phase, our system reports the matched sentences by the KBF module in those speakers reported by the whole system as predators.

3 The Knowledge-Based Conversation Filter

The KBF builds on previously unreported research by Optenet on sexual predation detection in chats, in the Spanish language. This module is based on the Spanish NGO Protegeles [8] predator behavior characterization, which is quite similar to that reported in [5], and latter used by Kontontathis *et al.* to build Chatcoder [3].

It must be noted that chats used in the PAN task do not correspond to real grooming cases, as the victim is a volunteer acting as a hook. Moreover, many of the cases are quite fast[5], while according to our experience in the real world, sex predators spend

[4] Here we mean full conversations, including the victim or any neutral user in a false positive example.

[5] The harassment happens even in the first conversation between the predator and the volunteer.

several months when approaching and seducing a child before they meet on the real world. Our previously existing system was designed to specifically target those slow but more difficult cases.

The KBF module features 51 hand-coded patterns in Spanish, which have been automatically translated to English using the Google Translator in order to evaluate the language portability of the system. The obtained patterns have not been corrected in the case of translation mistakes. These patterns represent twelve typical predation behaviors (BH) that correspond to four main predation phases:

1. Hooking: trying to locate the child (BH11), avoiding direct questions (BH12).
2. Fidelization: questions about family settings (BH21).
3. Seduction: personal questions (BH31), reducing sexual inhibition (BH32), sending pictures (BH33), flattery (BH34), generating debt perception (BH35).
4. Harassment: nude and sexual pictures (BH41), virtual sex (BH42), trying to meet (BH43), concept manipulation (e.g. child sex was accepted in ancient Greece – BH44).

The KBF includes some techniques for detecting other predator behaviors, but they are not easy to port from Spanish to English. Due to the time limit of the competition, we have deactivated those techniques in the English version of the system.

In the table 1, we show several examples of patterns used to identify some of the previous behaviors. The pattern-matching algorithm requires 2 or more unsorted words in common with a sentence to fire, except for BH12 and BH44, in which a full ordered match is required. We consider only the blank space as a word separator both in speaker sentences and in patterns. A speaker is reported as suspicious if at least 10% of his/her sentences match any behavior, or 3 or more in case of conversations with less than ten sentences. These thresholds have been retained from the original KBF module in Spanish.

Behavior	Pattern	Behavior	Pattern
BH11	"how old age"	BH21	"where are your brothers mother father"
BH32	"consolation console touch"	BH34	"handsome love you like you better"
BH41	"naked picture of you"	BH44	"everyone does"

Table 1. Examples of patterns for detecting particular predation behaviors

We present an example of a speaker conversation tagged as suspicious by the KBF sub-system in the table 2, along with the codes of the behaviors identified by the pattern-matching algorithm. We only report the lines corresponding to the suspicious speaker. It must be noted in this example that there are some false positive lines (for instance, several lines are recognized as BH31 while not being – personal questions), while other lines that should be matched, do not (for instance, line 22 corresponds to BH34 – flattery). We interpret this as a consequence of directly translating Spanish language patterns to English without reviewing nor enriching them.

Line	Text	Behavior
1	hii....jayne, how r ya	BH11, BH31
2	ya, why?	–
3	i answered	BH31, BH35, BH42
4	i said, hi, how r ya	BH11, BH42
5	actually	–
6	hi jayne...how r ya	BH11
7	lol	–
8	wait	–
9	iam_j4yn3 : hey T	–
10	hi	–
11	ugh, r u there?	–
12	lol.....i said hi, u didnt answer	–
13	ok np	–
14	aww ty	–
15	im back	–
16	all ready to go?	BH42
17	oops	–
18	it was ok	–
19	didnt mean to put a question mark there lol	–
20	ya	–
21	how bout if i call u when im close and you tell me where to meet you	BH31, BH35, BH42
22	awww ty.....ur a sweetie	–
23	we can argue bout that later, lol........im ready to go, so c u	–
24	>:D<	–
25	:D	–
26	bye 4 now	–

Table 2. Example of speech acts from a suspicious speaker according to the KBF module.

Among the original training conversations provided in the task, which correspond to 139,573 speakers, this module reports 14,236 as potential predators, retaining 100% of reported predators ($recall = 1$), and filtering out 89,89% of neutral or victim speakers.

4 The Learning-Based Detection sub-system

Over the years, some authors have been relatively successful in detecting the age of speakers in online chats [7,10]. Based on these works, we have formulated the hypothesis that linguistic and chat-like language properties can help to cluster speakers in three groups: real kids speaking as digital natives, adults simulating kid language in chats (as potential predators), and actual adults. Thus, we have selected six surface linguistic properties as features for a learning-based module to be trained on the output of the KBF sub-system:

- Number of uppercased letters.
- Number of words (guessing that chat-language tries to minimize communication symbols).

- Number of SMS/Chat words (according to the dictionary available at [4]).
- Number of emoticons (according to the dictionary available at the same site).
- Number of typos (word that do not occur in the Freeling [6] English dictionary nor in our SMS/Chat dictionary).
- Number of punctuation symbols (.,;:).

The values for these features are taken as absolute numbers and as relative numbers by dividing them among the number of lines in the conversation, and rounding them. A vector of values is computed for each conversation speaker. In consequence, each vector has twelve values. For instance, the values for the previous speaker conversation example are the following ones:

$$70aca6a54d7d6b260273282143a685e0 \Rightarrow [279, 255, 3, 24, 28, 266, 10, 9, 0, 0, 1, 10]$$

We have used two classes (predator, neutral), and trained a WEKA [2] Naïve Bayes Multinomial classifier with default parameters on the retained speakers from the KBF sub-system. The evaluation of this classifier on the output of the KBF module taken as training collection, and using 4-fold cross validation, leads to a recall of 0.986, a precision of 0.639, and a F_1 value of 0.775.

5 Conclusions

We have designed and evaluated an English-language sexual predator detection system that largely reuses a previously existing one for the Spanish language. Our system combines a knowledge-based predation detection system that makes use of hand-crafted patterns, and a learning-based classifier that employs simple linguistic features.

While the results of the evaluation on the training collection are nice, the performance on the test run of the PAN Author Identification task, Sexual Predator Identification sub-task, has been very poor. We believe this is mainly due to two reasons:

- First, the adaptation of the knowledge-based system has been done by the automatic translation of patterns from Spanish to English. We believe that the obtained patterns are not correct, so they must be reviewed by a native speaker with experience on Internet chats. Moreover, they should be augmented with specific patterns for the English language, and the numbers of matched patterns may be optimized by a learning based classifier. It must be noted that the mistakes made by this module may impact on the linguistic learning-based sub-system as well, making the effectiveness of the learned classifier much worse than expected.
- Second, and according to the ground truth provided by the PAN organizers, it is clear for us that word-level features (specifically, word n-grams and character n-grams) should be used as well. Given the architecture of our system, we believe the most straightforward way to do this, is training an additional text-based classifier on the suspicious users reported by the KBF, or, most likely, on all the users, and combine it with the linguistic classifier using stacking [11].
- Third, we have not used some of the techniques currently implemented in the knowledge-based system because they are highly language-dependent and due to the time limit for submitting the test run results.

We plan to do these improvements in future versions of our system, and to port them to the Spanish language sexual predator detector as well.

6 Acknowledgements

The work described here has been partly funded by the Spanish Ministry of Economy and by the Spanish Centre for Industrial Technological Development (CDTI), under the research project "WENDY: WEb-access coNfidence for chilDren and Young" (TSI-020100-2010-452).

References

1. Gómez Hidalgo, J., Caurcel Díaz, A.: Avances tecnológicos en la protección del menor en redes sociales. Novática, Revista de la ATI (218) (2012)
2. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
3. Kontostathis, A., Edwards, L., Bayzick, J., McGhee, I., Leatherman, A., , Moore, K.: Comparison of rule-based to human analysis of chat logs. In: Proceedings of the First International Workshop on Mining Social Media (MSM09) (2009)
4. Lingo2Word Home Page: http://www.lingo2word.com (June 2012)
5. Olson, L.N., Daggs, J.L., Ellevold, B.L., Rogers, T.K.K.: Entrapping the innocent: Toward a theory of child sexual predators' luring communication. Communication Theory 17(3), 231–251 (2007)
6. Padró, L., Collado, M., Reese, S., Lloberes, M., Castellón, I.: Freeling 2.1: Five years of open-source language processing tools. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta. European Language Resources Association (2010)
7. Pendar, N.: Toward spotting the pedophile telling victim from predator in text chats. In: Proceedings of the International Conference on Semantic Computing. pp. 235–241. IEEE Computer Society, Washington, DC, USA (2007)
8. Protegeles Home Page: http://www.protegeles.com (June 2012)
9. Santos Sierra, A., Sánchez Ávila, C., Carmonet Bravo, M., Guerra Casanova, J., Santos Sierra, D.: Control de edad en redes sociales mediante biometría facial. In: XII Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2012) (2012)
10. Tam, J., Martell, C.H.: Age detection in chat. In: Proceedings of the International Conference on Semantic Computing. pp. 33–39. IEEE Computer Society, Washington, DC, USA (2009)
11. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241–259 (1992)