Approaches of anonymisation of an SMS corpus

Namrata Patel, Pierre Accorsi, Diana Inkpen, Cédric Lopez, Mathieu Roche

Approaches of anonymisation of an SMS corpus

Namrata Patel[1], Pierre Accorsi[1], Diana Inkpen[2],
Cédric Lopez[3], Mathieu Roche[1]

[1] LIRMM – CNRS, Univ. Montpellier 2, France
[2] Univ. of Ottawa, Canada
[3] Objet Direct – VISEO, France

Abstract. This paper presents two anonymisation methods to process an SMS corpus. The first one is based on an unsupervised approach called *Seek&Hide*. The implemented system uses several dictionaries and rules in order to predict if a SMS needs anonymisation process. The second method is based on a supervised approach using machine learning techniques. We evaluate the two approaches and we propose a way to use them together. Only when the two methods do not agree on their prediction, will the SMS be checked by a human expert. This greatly reduces the cost of anonymising the corpus.

1 Introduction

In the past few years, SMS (Short Message Service) communication has become a veritable social phenomenon. Although numerous scientific studies (namely in the fields of linguistics, sociology, psychology, mass communication, etc.) have been conducted on this recent form of communication, there remains a general gap in our accumulated knowledge of the subject. This is mainly due to the fact that researchers have limited access to suitable data for their studies. Typically, they require large volumes of authentic data for their work to be significant.

The international project *sms4science* (http://www.sms4science.org/) aims at building and studying precisely such a body of data by collecting authentic text messages from different parts of the world. In the context of the sud4science project (http://www.sud4science.org/), over 90,000 authentic text messages in French have been collected. But the publication of these resources requires to meticulously remove all traces of identification from each SMS. In order to perform this anonymisation task, we have developed the *Seek&Hide*[4] software [1]. After the summarisation of the principle of this system, this paper focuses on machine learning approaches in order to predict SMS to anonymise.

In this paper, we begin by introducing the distinctive aspects of our work by looking at pre-existing anonymisation techniques (section 2). We present two solutions: the first one based on rules (section 3) and the second one based on machine learning (section 4). A combined approach is finally proposed in section 5. To conclude our study, we present and discuss the obtained results (section 6).

[4] Not "Hide and Seek", but "Seek and Hide": with this tool, we seek to hide words that are to be anonymised.

2 Related work

Anonymisation is indispensable when one seeks to mask an individual's identity. For example, this is essential before the distribution of court orders pertaining to children, juvenile delinquents, victims of sexual harassment, etc. [2], or when one needs to put together a medical corpus [3,4]. In the medical field, it is customary to resort to automatic anonymisation techniques using rules and medical dictionaries in order to process the most common cases [2,5,6,7,8]. These systems primarily aim at the automatic recognition of names, dates, places, and other elements which could lead to the identification of people covered by publication restrictions. Generaly the used methods to recognize Named Entity are based on specific rules and dictionaries. Moreover, supervised methods can be applied. For instance [9] have trained several classifiers, and they have combined decision functions for an anonymisation task.

We agree with [10] that the process of anonymisation cannot be entirely automated. Their work focuses on the creation of an interface by which the researcher can identify personal data and decide whether or not to render it anonymous. Given the size of our *sud4science LR* corpus (over 90,000 SMS), an automated procedure considerably benefits the annotator, as shall be evidenced in our paper. The distinguishing feature in our approach, as put forth by [11], is that we take into consideration the numerous linguistic particularities of the forms used in SMS writing.

[12] present the first freely available corpus of Dutch SMS, where anonymisation was performed automatically by replacing sensitive data, including dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, street numbers, etc.), e-mail addresses, URLs, and IP addresses. [13] have collected about 60,000 SMS, focusing on English and Mandarin Chinese. Previous works consider the same sensistive data: It seems that no names (or nicknames) were automatically anonymised. We call attention to the fact that our paper focuses on the most complex part of the anonymisation process: that of the processing of first names.

Sometimes, the identity of the markers that need to be anonymised are trivial names made up by the senders themselves, are subject to syntactic variations (often significant) and become cultural footprints (nicknames, diminutives, repetition of letters into the name) [14], i.e., in the following text message, "cece" requires anonymisation.

Coucou mon cece *! J'espere [. . .]*

Other anonymisation tasks are accomplished using regular expressions to identify the appropriate words; for example, e-mail addresses, telephone numbers, and URLs.

In the following section we propose two anonymisation techniques which are adapted to the demands of text messages.

3 Seek&Hide and anonymisation of SMS data

3.1 Principle

As stated in the previous section, our approach to anonymise/de-identify corpus of French text messages is to adopt a two-phase procedure.

This two-way procedure ensures the dependability of the system: the combined use of Natural Language Processing (NLP) techniques and human evaluation helps minimise computer as well as human errors, greatly improving the overall result.

Let us now take a deeper look into the workings of the system by considering each of its processes individually.

The main purpose of the system is to process a corpus using (a) dictionaries as reference material and (b) word-processing techniques so as to identify and eventually hide words that have to be anonymised. This constitutes the preliminary treatment of the corpus. Text messages that are processed by this phase undergo the following transformation, the details of which are given in the next paragraph:

"Coucou Patrice, ça va?" $\mapsto$ *"Coucou <PRE_7_17316>, ça va?"*

When a word is anonymised, it is replaced by a code conforming to the following format: <[Tag]_[#characters]_[cross-reference]> where "Tag" indicates the type of the word (e.g. First name, Last name). Thus, "Patrice" is replaced by $< PRE\_7\_17316 >$ where:

- *PRE* $\rightarrow$ First name (prénom)
- *7* $\rightarrow$ number of characters in "Patrice"
- *17316* $\rightarrow$ "Patrice"'s ID in the dictionary of first names

3.2 Global Process

Having seen what *Seek&Hide* does in its automatic phase, let us now find out how this is done. *Seek&Hide* operates in a three-part procedure:

(1) Pre-processing. Each SMS in the corpus, in its raw state, is basically just a string of characters. In order for *Seek&Hide* to make any sense out of this data, it needs to break the string of characters into words. This is the pre-processing phase, called "Tokenisation". Once tokenised, the SMS becomes a coherent sentence: a series of identifiable words. The SMS tokenisation is a complex processing [15]. For an anonymisation task that does not need a precise analysis of message content, such as ours, we consider that the simple use of a "space" as separator for tokenisation is satisfactory.

(2) Identification. In this phase of the automatic process, *Seek&Hide* uses a technique of identification which uses specific kinds of dictionaries to analyse each word of an SMS. The idea behind is simple: each word of an SMS can either be classified as "To anonymise" or as "Nothing to anonymise". We thus use two kinds of dictionaries corresponding to this classification, distinguishing them by "Dictionary" and "Anti-dictionary" on the basis of their content (pertaining to the task of anonymisation):

- The "Dictionary" contains words that need to be anonymised.
- The "Anti-dictionary" contains words that do not require anonymisation.

The following list shows the different resources used as reference material to identify the words in the text messages:

Dictionary: Dictionary of first names (21,921 first names)
Anti-dictionaries:

- Dictionary of inflected forms of the French language (LExique des Formes Fléchies du Français, LEFFF)[5] (105,595 lemmas),
- Dictionary of some forms used in SMS writing (739 words),
- Dictionary of places (9,463 cities and 194 countries).

Each word is then labelled according to its presence, or the lack of it, in the dictionaries used by *Seek&Hide* (*cf.* Table 1).

(3) Treatment. The words of the corpus are processed according to their labels and are thus (a) anonymised, (b) ignored, (c) highlighted. Words that could not be identified in a dictionary, and words that were identified in both types of dictionaries, are highlighted. These will be processed by users via a web-interface of the system [1].

Table 1 summarises this treatment by giving the range of possible cases encountered. As can be seen, "Cédric" is anonymised because it is identified only in the dictionary (of first and last names). Similarly, "crayon" is ignored as it is identified only in the anti-dictionary (LEFFF). "Pierre" and "Namrata" are problematic: "Pierre" is ambiguous as it belongs to both, the dictionary and the anti-dictionary. "Namrata" is unknown as it belongs to neither of the dictionaries. These two words are consequently highlighted for further processing. This one is based on a semi-automatic system based on human-machine interactions.

Moreover, in order to take into account the specificities of the SMS data, we added different heuristics to solve these cases:

Misspelled words: ex.: *surment* (instead of *sûrement*)
Words written without their accents: ex.: *desole* (instead of *désolé*)
Words with misplaced accents: ex.: *dèsolè* (instead of *désolé*)
Letter repetitions: ex.: *nicooooolllaassss* (instead of *nicolas*)
Onomatopoeias: ex.: *mouhahaha*
Omission of the apostrophe: ex.: *jexplique* (instead of *j'explique*)

[5] http://www.labri.fr/perso/clement/lefff/

Concatenation: ex.: *jtaime* (instead of *je t'aime*)

These heuristics work particularly well for the SMS data because there are a number of cases in text messages in which words are not always written using their correct spellings. Performing an "accent-insensitive" word-search in such cases, for example, is one of the heuristics employed by *Seek&Hide*. This and other heuristic solutions are further discussed in [1].

Word	**Dico**	**Anti-dico**	**Label**	**Treatment**
Cédric	yes	no	Dictionary	Automatically anoymised
crayon	no	yes	Anti-dictionary	Ignored (not to be anonymised)
Pierre	yes	yes	Ambiguous	Candidate to anonymise
Namrata	no	no	Unknown	Candidate to anonymise

Table 1. Different cases to take into account.

In order to improve the system and to reduce the workload in the manual validation phase, we decided to implement a second technique, based on supervised machine learning. These methods learn from annotated training data, and are able to make predictions on new test data.

4 Machine learning approach to SMS anonymisation

We wanted to see if it is possible to train a classifier to decide which messages need to be anonymised and which do not need. The classifier works at the SMS level not at word level. That is, if there are no words to be anonymise, the classifier will signal that there is nothing to anonymise; but if there is one or more words to anonymise, the classifier will signal that this SMS needs anonymisation. The reason we trained the classifier this way is that the data available for training a classifier were labeled at message level (not at word level).

The features used by the classifier are inspired from the linguistics analysis from the previous section. The values of the features are calculated by using the lexical resources mentioned above.

Here is the list of features extracted from each SMS:

- The number of words from the SMS that are in the dictionary of abbreviated forms specific to SMS texts (anti-dictionary).
- The number of words that are in the LEFFF dictionary of French (anti-dictionary).
- The number of words that are in the dictionary of first names. We expect this to be particularly useful for the class of messages to be anonymised.

– The number of words that are in the dictionary of country names.
– The length of the SMS.
– The number of words in upper case in the text.
– The average words length in the SMS.
– The number of pronouns in the text.
– The number of numbers.
– The number of punctuation tokens.
– Elongation: the number of words with elongated / repeated vowels.

For the features that count numbers of various elements, we experimented with the counts and we also normalized by the length of each SMS, but the results were similar, because most of the messages have similar lengths (usually short texts).

The algorithm that we selected for the classification is the Decision Trees algorithm (DT). In fact we tested several algorithms from Weka [16], but the DT algorithm worked better than other classifiers that we tried and it has the advantage that we can see what the classifier learnt by examining the learnt decision tree. Other classifiers, such as SVM and Naive Bayes, learn separation planes or probabilities, and these numbers are not understandable for a human examiner.

We also experimented with several meta-classifiers, and the Bagging algorithm based on Decision Trees was successful in improving the results with another 2 percentage points (as it will be shown in Section 6.3). Bagging is a form of voting with several classifiers trained on various parts of the training data, in order to obtain a more generic classifier.

The following section describes how we can combine the learnt model (section 4) with *Seek&Hide* system based on the use of rules (section 3).

5 Discussion: How to combine both approaches?

Seek&Hide system predicts if it is necessary

– to anonymise the SMS (TA)
– to anonymise nothing (NTA)
– to give the SMS to the expert because the prediction is impossible (Untagged).

The learnt model obtained by machine learning methods proposes two classes (TA and NTA). So, we can combine both approaches to propose a general prediction. This general prediction is based on the

– agreement/disagreement between *Seek&Hide* and the learnt model,
– class found with the learnt model if *Seek&Hide* can not predict.

Then the general prediction is based on the situation presented in Table 2. The principle is to minimize the intervention of the expert. With the use of both methods (i.e. *Seek&Hide* and the learnt model), the manual analysis by an expert is useful only if there exists a disagreement between the two automatic methods.

Seek&Hide	**Learnt model** (Machine leaning)	**Action**
TA	TA	**TA**
TA	NTA	**expert**
NTA	TA	**expert**
NTA	NTA	**NTA**
Untagged	TA	**TA**
Untagged	NTA	**NTA**

Table 2. Different possible actions regarding the predictions of both systems (*Seek&Hide* and learnt models).

6 Experiments

6.1 Experimental Protocol

Seek&Hide was tested on a sample of our SMS corpus containing 23,055 SMS that were manually tagged as "To anonymise (TA)" or "Nothing to anonymise (NTA)" by a student-annotator. During the acquisition of the corpus, a fourth year student was employed for a three-month internship, in order to read the incoming messages and make sure they respected certain rules and regulations. He thus labelled those text messages that needed to be anonymised as "To anonymise" and those that were to be left as-is as "Nothing to anonymise (NTA)". Out of the 23,055 SMS in our sample, 90.7% (i.e. 20,913 SMS) were noted by him as NTA and 9.3% (i.e. 2,142 SMS) as TA.

In the following section, we propose a method of evaluation whereby *Seek&Hide*'s results on the sample are compared with those of the student-annotator.

6.2 Global Analysis of *Seek&Hide* Results

Table 3 presents the SMS distribution in the sample according to 3 categories: Those tagged TA, those tagged NTA, and those left untagged. The untagged label corresponds to the text messages that our automatic system cannot tag (TA or NTA) because they contain ambiguous and/or unknown words. These will be processed via the web-interface of the semi-automatic phase of our tool.

We note that our system returns results for 65.3% of the sample corpus (i.e. 15,052 SMS with TA and NTA tags). The other part of the corpus (34.7 % of the

corpus left untagged) has to be processed by the semi-automatic system. In this section we focus our analysis on the evaluation of the 15,052 SMS automatically processed by *Seek&Hide* as NTA and TA. In this context, the confusion matrix (see Table 4) shows a more detailed analysis of the results obtained.

The top left box indicates the number of true positives - TP (i.e. the 13,904 SMS correctly classified by the application as 'NTA'), the top right, the false negatives - FN (i.e. the 59 SMS classified as 'NTA' by *Seek&Hide* and 'TA' by the student-annotator), the bottom left, the false positives - FP (i.e. the 413 SMS classified as 'TA' by Seek&Hide and 'NTA' by the student-annotator), and the bottom right, the true negatives TN (i.e. the 676 SMS correctly classified as 'TA'). Note that *correct* and *incorrect* terms are based on the tags specified by the student-annotator. A deeper analysis of Table 4 shows that *Seek&Hide* predicts 13,963 SMS (i.e. first line of Table 4: $13,904 + 59$) that do not need anonymisation (tagged NTA). Of these, only 59 SMS are irrelevant, when compared with the manual evaluation done by the student-annotator. This shows that the NTA-tagging performed by *Seek&Hide* is very efficient. However, the *Seek&Hide* prediction of text messages that require anonymisation (i.e. second line of Table 4) is not as good, as only 676 of 1,089 SMS are relevant.

In our case we obtain a value of accuracy at 0.96. This score validates the relevance of our methods in predicting text messages that may or may not require anonymisation.

Sample Processed	Tag: TA	Tag: NTA	Untagged	Total
By Seek&Hide	1,089	13,963	8,003	23,055
	4.72%	**60.57%**	**34.71%**	

Table 3. Results of the *Seek&Hide* System on the Sample Analysed by the Student-Annotator.

Confusion Matrix	Student-Annotator: NTA	Student-Annotator: TA
Seek&Hide: NTA	13,904	59
Seek&Hide: TA	413	676

Table 4. Confusion Matrix

The following section presents results obained with other methods based on machine learning approaches.

6.3 Global Analysis of Machine Learning Results

The first 4000 SMS texts from our corpus were selected as training data for the classifier. The labels (to anonymise or not) were available from the student anno-

tator, as mentioned above. The reason to limit the size of the training data is that we do not want to burden the human expert with a lot of manual annotation, since the goal of the system is to save expert's time for the SMS anonymisation task. In fact, when we experimented with half the amount of the training data, we obtained similar results.

In the 4000 training messages, there were 529 that were positive examples (labelled with the class TA, to anonymise) and the rest of 3471 were negative examples. With this high imbalance, the classifier learns mostly the characteristics of the negative class. Such a classifier is not useful, since it would anonymise only very few examples.

To deal with the high data imbalance, we use a simple undersampling technique. We balanced the training data by keeping only 529 of the negative examples. We successfully trained a classifier on the 1058 examples (529 positive examples, 529 negative examples).

An example of learnt decision tree is presented in figure 1. Only a part of the tree is presented, because it has 127 nodes, from which 64 are leaf nodes. The leaf nodes contain the predicted class, followed in brackets by the number of correct / incorrect instances from the training data classified into the node. By examining the learnt decision tree, we note that the best feature, used in the root of the tree is the number of words that start with capital letters. The length of the words also seems to be important. We expected the occurrences in the dictionary of First names to be higher up in the tree. Table 5 shows the InfoGain values for each feature. The table ranks the features by their ability to discriminate between the two classes, since the InfoGain measures the entropy of classification when one feature at a time is used. We can see that the FristName feature is the third most important.

Rank	Feature	InfoGain Value
1	UpperCase	0.2754
2	SMSLength	0.2121
3	FirstName	0.1366
4	Countries	0.0986
5	Cities	0.0986
6	WordLenght	0.0776
7	Numbers	0.0367
8	Elongations	0.0353
9	Abbreviations	0.034
10	Punctuation	0.0269
11	LEFFF	0.0262
12	Pronouns	0.000

Table 5. Importance of the features in the classification.

```
UpperCase <= 1
|   Numbers <= 2: NTA (273.0/8.0)
|   Numbers > 2
|   |   WordLength <= 5: NTA (5.0/1.0)
|   |   WordLength > 5: TA (18.0/3.0)
UpperCase > 1
|   Numbers <= 3
|   |   UpperCase <= 2
|   |   |   Punctuation <= 1
|   |   |   |   SMSLength <= 27: NTA (5.0)
|   |   |   |   SMSLenght > 27
|   |   |   |   |   WordLength <= 7: TA (48.0/6.0)
|   |   |   |   |   WordLength > 7: NTA (2.0)
......
|   Numbers > 3
|   |   WordLength <= 5
|   |   |   FirstName <= 7
|   |   |   |   Pronouns <= 0
|   |   |   |   |   Elongations <= 0: TA (5.0)
|   |   |   |   |   Elongations > 0
|   |   |   |   |   |   SMSLength <= 78: NTA (2.0)
|   |   |   |   |   |   SMSLength > 78: TA (4.0)
|   |   |   |   Pronouns > 0
|   |   |   |   |   Punctuation <= 6: NTA (6.0)
|   |   |   |   |   Punctuation > 6: TA (2.0)
|   |   |   FirstName > 7: TA (35.0/3.0)
|   |   WordLength > 5: TA (94.0/4.0)
```

Fig. 1. Part of the learnt decision tree.

Table 6 presents some results of the Bagging DT classifier, by 10-fold cross-validation. This standard evaluation technique in machine learning uses 9 parts of the data for training and tests on the remaining part, then it repeats this for other 9 parts. At the end, it averages over the 10 folds. The baseline in the table is 50%, for a random classifier that would choose any of the two classes. The DT algorithm achieves an accuracy of 77.8%, therefore it is much better than the baseline. The meta-classifier (Bagging DT) reaches 79.4%. The performance for each of the two classes is presented in terms of Precision (how many examples classified into that class are correct), Recall (how many correct examples of that class are retrieved) and F-measure (the harmonic mean of Precision and Recall). The values in Table 6 show that the classifier is doing equally well for the two classes.

We did additional testing for the machine learning method. We trained the classifier on the entire training data, and we tested it on the next 2000 messages from the corpus (from the 4000th SMS to 5999th). This is a realistic test set, because the test data comes later in the time line, and it might have differences compared to the training data. For this test, the accuracy of the Decision Tree was 74.6% DT, while Bagging DT achieved an accuracy of 76.9%. The accuracy is slightly lower than the one obtained by 10-fold cross validation. This shows that the classifier is general enough to obtain similar results on new test data.

Classes	Real: NTA	Real: TA
Model: NTA	383	72
Model: TA	146	457

Class	Precision	Recall	F-Measure
NTA	0.842	0.724	0.778
TA	0.758	0.864	0.807

Table 6. Machine learning results for each class (i.e., confusion matrix and precision/recall/F-measure).

7 Conclusion and Future Work

The system proposed in this article performs the anonymisation/de-identification of a corpus. To this end, it uses (a) a dictionary of first names and (b) anti-dictionaries (of ordinary language and of some forms of SMS writing) to identify the words that require anonymisation. Note that the adopted principle is sufficiently generic for it to be adapted to various types of corpus, irrespective of their language.

In its automatic phase, the system processes over 70% of the corpus. This corresponds to the unambiguous text messages present in it: Those that contain words that are neither unknown, nor ambiguous (found in both the dictionary as well as the anti-dictionaries). A comparative analysis of its performance, based on the manual evaluation of a significant albeit small portion of the corpus (i.e. 23,055 SMS), yielded positive results on 96% of the text messages processed (whether considered to be anonymised or not to be anonymised).

As future work, two students will perform the task of anonymisation. A thorough analysis on their part will allow us to improve our techniques and enrich our dictionaries. Another immediate direction of future work is to design more features for the supervised machine learning algorithms, in order to increase their accuracy for this task.

We would also like to apply the tool and its associated algorithms to other types of data (e.g., medical data) that require anonymisation/de-identification.

Acknowledgements

We thank **Rachel Panckhurst**, leader of the sud4science LR project. This project is part of a vast international SMS data collection project, entitled sms4science (http://www.sms4science.org), and started at the CENTAL (Centre for Natural Language Processing, Université Catholique de Louvain, Belgium) in 2004. Our work is supported by the MSH-M (Maison des Sciences de l'Homme de Montpellier – France) and the DGLFLF (Délégation générale à la langue française et aux langues de France).

References

1. Accorsi, P., Patel, N., Lopez, C., Panckhurst, R., Roche, M.: Seek&Hide: Anonymising a French SMS Corpus Using Natural Language Processing Techniques. Lingvisticæ Investigationes **35**(2) (2012)

2. Plamondon, L., Lapalme, G., Pelletier, F.: Anonymisation de décisions de justice. TALN'04 (2004)
3. Grouin, C., Rosier, A., Dameron, O., Zweigenbaum, P.: Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. Risques, technologies de l'information pour les pratiques médicales (2009) 23–34
4. Tamersoy, A., Loukides, G., Nergiz, M., Saygin, Y., Malin, B.: Anonymization of longitudinal electronic medical records. Information Technology in Biomedicine, IEEE Transactions on **16**(3) (2012) 413 –423
5. Sweeney, L.: Replacing personally-identifying information in medical records, the scrub system. In: Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association (1996) 333
6. Aramaki, E., Imai, T., Miyo, K., Ohe, K.: Automatic deidentification by using sentence features and label consistency. In: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. (2006) 10–11
7. Gardner, J., Xiong, L., Wang, F., Post, A., Saltz, J., Grandison, T.: An evaluation of feature sets and sampling techniques for de-identification of medical records. In: Proceedings of the 1st ACM International Health Informatics Symposium, ACM (2010) 183–190
8. Gicquel, Q., Proux, D., Marchal, P., Hagége, C., Berrouane, Y., Darmoni, S., Pereira, S., Segond, F., Metzger, M.: Évaluation dun outil daide á lanonymisation des documents médicaux basé sur le traitement automatique du langage naturel. Systèmes dinformation pour lamélioration de la qualité en santé (2012) 165–176
9. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. JAMIA **14**(5) (2007) 574–580
10. Reffay, C., Blondel, F., Giguet, E., et al.: Stratégies pour l'anonymisation systématique d'un corpus d'interactions plurilingues. In: Proceedings of IC2012. (2012) 1–21
11. Fairon, C., Klein, J.: Les écritures et graphies inventives des sms face aux graphies normées. Le Français aujourd'hui (3) (2010) 113–122
12. Treurniet, M., De Clercq, O., van den Heuvel, H., Oostdijk, N.: Collection of a corpus of dutch sms. (2012)
13. Chen, T., Kan, M.: Creating a live, public short message service corpus: The nus sms corpus. Language Resources and Evaluation 1–37
14. Reffay, C., Teutsch, P.: Anonymisation de corpus réutilisables Masquer l'identité sans altérer l'analyse des interactions. Proceedings of EIAH'2007 : Environnements Informatiques pour l'Apprentissage Humain (2007)
15. Beaufort, R., Roekhaut, S., Cougnon, L.A., Fairon, C.: A hybrid rule/model-based finite-state framework for normalizing sms messages. In: Proceedings of ACL. (2010) 770–779
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)