A Machine-Translation Method for Normalization of SMS

Darnes Vilariño, David Pinto, Beatriz Beltrán, Saul León, Esteban Castillo, and Mireya Tovar

Faculty of Computer Science
Benemérita Universidad Autónoma de Puebla, Mexico
{darnes,dpinto,bbeltran,mtovar}@cs.buap.mx,
saul.ls@live.com, ecjbuap@gmail.com

Abstract. Normalization of SMS is a very important task that must be addressed by the computational community because of the tremendous growth of services based on mobile devices, which make use of this kind of messages. There exist many limitations on the automatic treatment of SMS texts derived from the particular writing style used. Even if there are suficient problems dealing with this kind of texts, we are also interested in some tasks requiring to understand the meaning of documents in different languages, therefore, increasing the complexity of such tasks. Our approach proposes to normalize SMS texts employing machine translation techniques. For this purpose, we use a statistical bilingual dictionary calculated on the basis of the IBM-4 model for determining the best translation for a given SMS term. We have compared the presented approach with a traditional probabilistic method of information retrieval, observing that the normalization model proposed here highly improves the performance of the probabilistic one.

1 Introduction

The growing use of mobile devices has lead to novel/fashion patterns of communication which must be studied because of their lexical, syntactic and semantic particularities. The cost for sending SMS and the easy access to the purchase of mobile devices have made instant messaging emerge as the preferred communication medium just after spoken media and e-mails. An SMS is basically a short text message containing no more than 160 characters. Mobile device users are using this kind of messages for accessing different computational services that years ago were only accesed by means of a personal computer.

SMS messages contain a high number of terms that do not appear in regular dictionaries. In SMS written context, it is common to introduce new terminology that very often is associated with contractions of the original words or with an ortographical representation of their corresponding phonetic. It is worth noting that such terminology changes according to the different age ranges of the users, since young people (teenagers) used to employ phonetic representations which may be derived from the fact that they do not domain enough vocabulary of

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2012, LNCS 7329, pp. 293–302, 2012.

Table 1. Examples of terminology used in SMS

SMS message	Interpretation
10q	Thank you
tna	temporarily not available
afk	away from keyboard
L8	late
LOL	Laugh Out Loud
26Y4U	Too sexy for you
@WRK	at work
⋮	⋮

their native language. Some examples of the terms used and their corresponding interpretation are shown in Table 1.

Here the importance of proposing techniques for the automatic treatment of SMS. From the point of view of information retrieval systems it is important to exploit the potential number of users using such kind of devices and communication services. In this paper we aim, as a study case, to face the problem of searching answers of FAQs (Frequently Asked Questions) when one SMS is used as query for the information retrieval system. It is well known that millions of Instant Messaging (IM) users, including SMS, generate e-content in a language that adheres to conventional grammar, or punctuation rules and usual pronunciation. The words are intentionally compressed, using abbreviations, acronyms, phonetic transliterations and even neologisms. Additionally, it is important to note that the reduced space of the mobile device screen and the size of the keyboard leads very often messages to contain inadvertent typographical errors and spelling mistakes. This fact, emphasizes the complexity of constructing an information retrieval system which considers SMS queries, even when the corpus is conformed by questions whose answers are well documented (FAQs).

For the purpose of analysing the quality of the normalized SMS texts, in this research work we present one information retrieval system which considers the monolingual, crosslingual and multilingual approaches for three different languages: English, Hindi and Malayalam. The dataset was obtained from FIRE, a well-known evaluation forum promoted by IBM Research India for information retrieval systems[1]. The aim of using this corpus is to evaluate two different approaches by using bilingual statistical dictionaries in the task of SMS-based FAQ retrieval. In the first case, we have proposed to normalize the queries (SMS) by using the most frequent translation calculated from a training corpus, whereas the second approach considers a straightforward probabilistic search engine which integrates in a single step the process of searching and translating the query. It is important to note, that even in the case of the monolingual task, we are considering to solve the problem by means of machine translation techniques, assuming that both, the queries and the target documents are written in

[1] http://www.isical.ac.in/~clia/

a different language. Therefore, the statistical bilingual dictionaries are in fact, a kind of association thesaurus which probabilistically determines which sentence is the "translation" (correct way of writing) for a given SMS word.

There are some works reported in literature dealing with the particular task of FAQ retrieval and SMS-based FAQ retrieval which are presented as follows. Harksoo Kim has studied the problem of FAQ retrieval as it may be seen in [1], where he presented a trustly way of recovering FAQs by using a clustering of previous query logs. In fact, he also improved this first approach in [2] by employing latent semantic analysis, and also by using latent term weights [3]. In [4] it is proposed the use of machine translation techniques for the alignment of questions and answers of a FAQ corpus, with the aim of constructing a bilingual statistical dictionary which is further used for expanding the queries introduced in an information retrieval system. The experiments were performed with the English and Chinese languages, and we consider interesting the idea of using machine translation techniques for generating a terminological association model between the question and answer vocabulary, which may be further used in order to estimate the probability of an answer given a query (set of terms). We have extended this approach by considering the same type of alignment, but in this case between the SMS and the FAQ questions. In [5] it is presented an approach for domain specific FAQ retrieval based on a concept named "independent aspects". This concept basically consists of extracting terms and relationships by employing WordNet and Hownet which are then used in a mixture-based probabilistic model with the aim of analyzing queries and query-answer pairs by means of independent aspects. It is worth noting that any of the presented approaches use SMS as an input query, but they use queries written in normal text, which is a very important difference with respect to the experiments carried out in this research paper.

An excelent work for SMS normalization may be found in [6]. They prepared a training corpus of 5000 SMS aligned with reference messages manually prepared by two persons which are then introduced to a phrase-based statistical method to normalize SMS messages. The obtained results are very interesting despite they do not apply their method to any particular task such as information retrieval. Their findings include the observation of a great difference between SMS and normal written texts due to the particular written style of SMS writers and the high frequency of non-standarized terms which very often occur in short versions, shortened, truncated or phonetically transliterated.

Even so, there exist some works facing the problem of FAQ retrieval based on SMS queries. In [7] and [8], for instance, a web service for retrieving Hindi FAQs considering SMS as queries. This propossal consists on formulate the similarity criterion of the search process as a combinatorial problem in which the search space is conformed of all the different combinations for the vocabulary of the query terms and their N best translations. Unfortunately, the corpus used in these experiments is not available and, therefore, it is not possible to use it for comparison with our approach.

The obtained results in this paper show that we may significantly improve the retrieval results by normalizing the queries (SMS) before applying the typical information retrieval procedures. We consider that given the great success of instant messaging in the world, these findings would be of high benefit for all users of mobile devices, in particular in India because we have studied three languages widely used by close of 400 millions of indian mobile device users.

The rest of this paper is structured as follows. In Section 2 we introduce the approach proposed here for normalizing SMS. Section 3 describes the probabilistic model used as a reference for comparison with the normalization approach. In Section 4 the experimental results are presented and discussed. Finally, in Section 5 the conclusions are given together with some ideas of future work.

2 Normalization of SMS

We are proposing a SMS normalization approach for being applied in the framework of the SMS-based FAQ retrieval task (monolingual, crosslingual and multilingual). This problem is very relevant due to the particular terminology used in the SMS which necessarily requires of a normalization process. For the experiments carried out in the monolingual task, we have considered that both, the SMS and the FAQs are written in the same language (English, Hindi or Malayalam). In the case of the crosslingual task, the SMS is written in English, whereas the FAQs are written in Hindi. Finally, in the case of the multilingual task, both the SMS and the FAQs may be written in any of the three different languages.

Formally, the problem faced in this paper may be formulated as follows: Let ζ be the set of questions in the FAQ corpus, and $S = s_1 s_2 \dots s_n$ be an SMS. Both, the SMS and each question $q \in \zeta$, are seen as a sequence of terms. The aim is to find the question q^* belonging to the corpus ζ that obtains the maximum degree of similarity with respect to the SMS S. For this purpose, we have used a single model of information retrieval based on set theory, in particular, by using intersection of term sets. We have proposed a normalization model for queries (SMS) which is described in the following paragraphs.

For the problem we are interested in, we must consider that SMS may contain not only those terms that occur very frequently, but also those that are not so frequent. Let us take for example the following phrase taken from the test corpus: "whr can i find info abt pesticide estb reg and rep", which may be interpreted as "where can i find information about pesticide establishment registration and reporting". Therefore, the idea of mantaining a generic dictionary of frequently used terms on the SMS context may be unuseful on narrow domains.

With the aim of determining the correct "meaning" of terms appearing in an SMS query, we propose to substitute each query term for the closest translation offered by a bilingual statistical dictionary. In order to construct this dictionary, we have used the Giza++ tool[2] which allowed us to calculate the IBM-4 model

[2] `http://code.google.com/p/giza-pp/`

Table 2. Distribution of the different statistical bilingual dictionaries

Task	Source language (SMS)	Target language (Question-FAQ)
Monolingual	English	English
	Hindi	Hindi
	Malayalam	Malayalam
Crosslingual	English	Hindi
Multilingual	English	English
	English	Hindi
	English	Malayalam
	Hindi	Hindi
	Hindi	English
	Hindi	Malayalam
	Malayalam	Malayalam
	Malayalam	English
	Malayalam	Hindi

by using a training corpus conformed of a set of aligned phrases (one SMS with its corresponding FAQ). In total, we have constructed 13 different statistical bilingual dictionaries as it is shown in Table 2.

The similarity among the SMS terms and each one of the FAQ questions (P_{FAQ}) were calculated using the Jaccard similarity coefficient, as it is shown in Eq. (1).

$$Similarity(SMS, P_{FAQ}) = \frac{| \quad SMS \quad \cap \quad P_{FAQ} \quad |}{| \quad SMS \quad \cup \quad P_{FAQ} \quad |} \tag{1}$$

The pseudocode associated to the FAQ retrieval is shown in the Algorithm 1. This algorithm receives as input the set of SMS (topics o search queries), the target dataset or FAQs (ζ) and the statistical bilingual dictionary (ϕ) used in the SMS normalization process. Each topic (SMS) is normalized according to the criterion explained in Section 2. The normalized message is then compared with each FAQ question by means of the Jaccard similarity measure. All those values greater than the minimum of the N best similarity values (which we will known as *threshold*), are returned in the answer set (P_{FAQ}).

3 A Probabilistic Model for SMS-Based FAQ Retrieval

We have used a probabilistic model which considers both, the translation and the search process in a single step. It basically uses a statistical bilingual dictionary for calculating the probability of each topic (search query) to be associated to a target document. The training phase is done by applying the IBM-1 model to a set of pairs of query vs. relevant documents. The obtained statistical dictionary is used in conjunction with the set of target documents in order to show the

Algorithm 1. SMS-based FAQ retrieval

Input: Topics: $SMS = \{sms_1, sms_2, \ldots, sms_n\}$
Input: FAQs: $\zeta = \{q_1, \cdots, q_n\}$
Input: Statistical bilingual dictionary: $\phi = p(t_{SMS}, t_q)$
Output: N best answers for each SMS: Q

```
foreach sms_i ∈ SMS do
    smsN_i = Normalize(sms_i, φ);
    Q[i] ← {∅};
    foreach P_FAQ ∈ ζ do
        if Similarity(smsN_i, P_FAQ) > Threshold then
            Q[i] = Q[i] ∪ {P_FAQ, Similarity(smsN_i, P_FAQ)};
        end
    end
    if |Q[i]| > N then
        Q[i] = NBestValues(Q[i]);
    end
end
return Q
```

most relevant ones given a query which is written in a different language of that of the target documents (see [9]).

Formally, let x be a query text in a certain (source) language, and let $y_1, y_2, \ldots, y_n$ be a collection n of documents written in a different (target) language. Given a number $k < n$, we are interested in finding the k most relevant documents with respect to the source query x. For this purpose, it is employed a probabilistic approach in which the k most relevant documents are computed as those most probable given x, i.e.,

$$\hat{S}_k(x) = \underset{\substack{S \subset \{y_1,\ldots,y_n\} \\ |S|=k}}{\arg\max} \min_{y \in S} p(y \mid x) \tag{2}$$

Actually, $p(y \mid x)$ is modelled by using the IBM-1 model, which assumes that the order of the words in the query is not important and, therefore, each position in a document is equally likely to be connected to each position in the query. Although this assumption is unrealistic in machine translation, it is considered that the IBM-1 model is particularly well-suited for crosslingual information retrieval. In our case, we have observed that the best behaviour for the training dataset was obtained when the IBM-4 was used.

4 Experimental Results

In this section we present and discuss the results obtained for each task evaluated. First, we present a general description of the training and test corpora used in the experiments. Secondly, we provide a discussion attempting to find evidence of a relationship between the values obtained in the evaluation and the peculiarity of each language considered in the experiments.

4.1 Training Dataset

The training corpora are conformed by aligned sentences (SMS with its corresponding FAQ). The different FAQ corpora is divided according to the language. In Table 3 we observe the number of FAQs associated to each language in the training corpora.

Table 3. Distribution of FAQs in the training corpora

Idioma	# de FAQs
Inglés	7,251
Hindi	1,994
Malayalam	681

The SMS used as topics are distributed according to their language and task, as presented in Table 4.

Table 4. Distribution of SMS in the training corpora

Task	Language	# of SMS
Monolingual	English	1,071
	Hindi	230
	Malayalam	140
Crosslingual	English	472
Multilingual	English	460
	Hindi	230
	Malayalam	80

4.2 Test Dataset

In the test dataset, we use the same number of FAQs (see Table 3), however, the number of SMS changes. The number of SMS used in the test phase (evaluation) are distributed according to their language and task (see Table 5).

Table 5. Distribution of SMS in the test corpora

Task	Language	# of SMS
Monolingual	English	3,405
	Hindi	324
	Malayalam	50
Crosslingual	English	3,405
Multilingual	English	3,405
	Hindi	324
	Malayalam	50

4.3 Evaluation Results

In Table 6 we show the Mean Reciprocal Rank (MRR) obtained for each run submitted to the SMS-based FAQ retrieval task. The normalization model is identified by the "NORM" tag, whereas the probabilistic one is identified by the "PROB" tag. As it can be seen, in the multilingual task, we did not send a probabilistic run because we observed a very low performance when we used the training dataset.

The most interesting result is that we have greatly outperformed the probabilistic model by using the normalization one. After a simple analysis of the obtained results we conclude that the use of the maximum probability of translation instead of the product of the probable translation is the best solution for this particular task.

A quite interesting finding is that the best MRR obtained is when the Malayalam language is used. We consider that this result comes from the fact that the people is using a very low number of terms out of the vocabulary, even when they are writing SMS-type messages. It seems that people talking in Malayalam consistently use standard vocabulary of that language. This fact should be an expected result, because this language can be considered as a variation of the Tamil which is used in education and administration. Actually, the Malayalam has borrowed thousands of nouns, hundreds of verbs and some indeclinable part of speechs from sanscrit, which is language assumed to be used by aristocracy and academics, as it happened with latin in the European countries.

From our particular point of view, the obtained results are competitive in the case monolingual and multilingual, but they are not in the case of the crosslingual evaluation, given the very low performance obtained. We conclude that there are not sufficient statistical evidence on the training corpus for obtaining the best translation/association of the SMS terms for different languages.

As future work, we plan to propose a mechanism for improving the translation association by means of bootstraping and corpus enrichment based on snippets extracted from the web.

Table 6. Results obtained at the SMS-based FAQ retrieval task

Task	Run	In Domain correct	Out of Domain correct	Mean Reciprocal Rank (MRR)
Crosslingual	NORM	1/37 (0.027)	163/3368 (0.048)	0.0398
Crosslingual	PROB	0/37 (0.000)	170/3368 (0.050)	0
Monolingual-English	NORM	385/704 (0.546)	75/2701 (0.027)	0.6006
Monolingual-English	PROB	1/704 (0.001)	140/2701 (0.051)	0.0025
Monolingual-Hindi	NORM	153/200 (0.765)	0/124 (0.000)	0.8070
Monolingual-Hindi	PROB	0/200 (0.000)	5/124 (0.040)	0.0051
Monolingual-Malayalam	NORM	39/50 (0.780)	0/0 (NaN)	0.8304
Monolingual-Malayalam	PROB	1/50 (0.020)	0/0 (NaN)	0.0714
Multlingual-English	NORM	353/704 (0.501)	25/2701 (0.009)	0.5360
Multilingual-Hindi	NORM	113/200 (0.565)	0/124 (0.000)	0.6163
Multilingual-Malayalam	NORM	32/50 (0.640)	0/0 (NaN)	0.7037

5 Conclusions and Further Work

The aim of this paper was to evaluate a normalization model which may be used in the SMS-based FAQ retrieval task of FIRE 2011. Unfortunately, we can not compare this results with others because at the moment of publishing this working note, the evaluation results for the rest of the teams that participated at the task were not available. We can, however, say that the normalization model proposed in this paper has obtained an aceptable performance in the task.

The proposed model is based on the use of statistical bilingual dictionaries and, therefore, as far as we have a training corpus, we may extend the same procedure to different languages and similar tasks. As future work we would like to improve the information retrieval model used. We should replace the Jaccard similarity measure by other one that take into account the frequency of the terms among the documents to be compared.We are also interesting in proposing a mechanism for improving the translation association by means of bootstraping and corpus enrichment based on snippets extracted from the web.

Acknowledgments. This project has been partially supported by projects CONACYT #106625, VIEP #VIAD-ING11-II and #PIAD-ING11-II.

References

1. Kim, H., Seo, J.: High-performance faq retrieval using an automatic clustering method of query logs. Inf. Process. Manage. 42, 650–661 (2006)
2. Kim, H., Lee, H., Seo, J.: A reliable faq retrieval system using a query log classification technique based on latent semantic analysis. Inf. Process. Manage. 43, 420–430 (2007)
3. Kim, H., Seo, J.: Cluster-based faq retrieval using latent term weights. IEEE Intelligent Systems 23, 58–65 (2008)
4. Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., Liu, Y.: Statistical machine translation for query expansion in answer retrieval. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 464–471. Association for Computational Linguistics, Prague (2007)
5. Wu, C.H., Yeh, J.F., Chen, M.J.: Domain-specific faq retrieval using independent aspects. ACM Transactions on Asian Language Information Processing (TALIP) 4, 1–17 (2005)
6. Aw, A., Zhang, M., Xiao, J., Su, J.: A phrase-based statistical model for sms text normalization. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions, COLING-ACL 2006, pp. 33–40. Association for Computational Linguistics, Stroudsburg (2006)
7. Kothari, G., Negi, S., Faruquie, T.A., Chakaravarthy, V.T., Subramaniam, L.V.: SMS based interface for FAQ retrieval. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL-IJCNLP 2009, vol. 2, pp. 852–860. Association for Computational Linguistics, Morristown (2009)

8. Contractor, D., Kothari, G., Faruquie, T.A., Subramaniam, L.V., Negi, S.: Handling noisy queries in cross language faq retrieval. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 87–96. Association for Computational Linguistics, Stroudsburg (2010)
9. Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., Rosso, P.: A statistical approach to crosslingual natural language tasks. J. Algorithms 64, 51–60 (2009)