The Soundex Phonetic Algorithm Revisited for SMS Text Representation*

David Pinto[1], Darnes Vilariño[1], Yuridiana Alemán[1],
Helena Gómez[1], Nahun Loya[1], and Héctor Jiménez-Salazar[2]

[1] Faculty of Computer Science
Benemérita Universidad Autónoma de Puebla, Mexico
{dpinto,darnes}@cs.buap.mx,
{yuridiana.aleman,helena.adorno}@gmail.com,
israel_loya@hotmail.com
[2] Information Technologies Department
Universidad Autónoma Metropolitana, DF, Mexico
hgimenezs@gmail.com

Abstract. The growing use of information technologies such as mobile devices has had a major social and technological impact such as the growing use of Short Message Services (SMS), a communication system broadly used by cellular phone users. In 2011, it was estimated over 5.6 billion of mobile phones sending between 30 and 40 SMS at month. Hence the great importance of analyzing representation and normalization techniques for this kind of texts. In this paper we show an adaptation of the Soundex phonetic algorithm for representing SMS texts. We use the modified version of the Soundex algorithm for codifying SMS, and we evaluate the presented algorithm by measuring the similarity degree between two codified texts: one originally written in natural language, and the other one originally written in SMS "sub-language". Our main contribution is basically an improvement of the Soundex algorithm which allows to raise the level of similarity between the texts in SMS and their corresponding text in English or Spanish language.

1 Introduction

SMS is a very popular short message-based communication service among mobile phone users. However, SMS is also synonym of the short message itself which can contain up to 160 characters. The length limitation of an SMS has lead to create a sort of "sub-language" which includes a vocabulary of words, phonetically similar to that of the original natural language, but that regularly omit grammatical forms, punctuation marks and vowels.

In this paper we present a study based on lexical similarity, for different adaptations to the Soundex phonetic algorithm in order to represent SMS texts. The input is an SMS text and the output is a code or a set of codes. The aim is to have a family of words that matches with the same code, and to use the codified text version in

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

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 47–55, 2012.

some natural language tasks, such as information extraction or question answering. The presented algorithms have been evaluated in three different datasets (two languages, English and Spanish) by comparing the lexical similarity between pairs of texts (SMS, natural language) already codified with the purpose of having an overview of the level of generality obtained with the different codifications.

The remainder of the paper is organized as follows. In section 2 we summarize different works reported in the literature that are related to the one presented in this paper. In Section 3 we discuss the Soundex phonetic algorithm. In Section 4 we present different modifications we have done to the Soundex algorithm in order to have a proper codification for SMS texts, together with a preliminary study that indicates the degree of similarity between pairs of texts which express exactly the same meaning, having different textual representation, in one case they are written using the standard vocabulary of the language (Spanish or English) and in the other case they are written in the SMS sub-language. Finally we conclude this paper by resuming the strengths of our contributions and sketching future research issues.

2 Related Work

The Soundex code has often been applied in the information retrieval task, particularly when it is based on transcriptions of spoken language, because it is known that speech recognition produce transcription errors that are phonetically similar but ortographically dissimilar. In [1], however, it is claimed that the use of this codification does not improve regular string-matching based IR. The purpose of this paper is to study phonetic-based representations for SMS messages. Therefore, we are interested in those works reported in litereature dealing with the task of SMS analysis, basically by considering normalization of SMS. In [2], for instance, the authors provide a brief description on their input pre-processing work for an English to Chinese SMS translation system using a word group model. The same authors provide an excellent work for SMS normalization in [3]. They prepared a training corpus of 5,000 SMS aligned with reference messages manually prepared by two persons which are then introduced to a phrase-based statistical method to normalize SMS messages. In the context of SMS-based FAQ retrieval, the most salient works are the ones presented in [4] and [5], where authors formulate a 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 in our experiments. To the best of our knowledge, in the literature there is not a particular phonetic algorithm particularly adapted for representing SMS and, therefore, we consider that the approach presented in this paper would be of high benefit.

3 Phonetic Representation

The phonetic representation has several applications. It allows to search concepts based on pronunciation rather than on the spelling, as it is traditionally done. There exist different algorithms for codifying text according to its phonetic pronunciation. Some of the

Table 1. Soundex phonetic codes for the English language

Numeric code	*Letter*
0	a,e,i,o,u,y,h,w
1	b,p,f,v
2	c,g,j,k,q,s,x,z
3	d,t
4	l
5	m,n
6	r

most known and used phonetic algorithms are: Soundex [6,7], NYSIIS [8], Metaphone and Double Metaphone [9]. For the purposes of these preliminar experiments, we have started by considering the Soundex algorithm, which is better described as follows:

The Soundex phonetic algorithm was mainly used in applications involving searching of people's names like air reservation systems, censuses, and other tasks presenting typing errors due to phonetic similarity [10]. As shown in [11], the Soundex algorithm evaluates each letter in the input word and assigns a numeric value. The main function of this algorithm is to convert each word into a code made up of four elements.

Soundex uses numeric codes (see Table 1) for each letter of the string to be codified.

The Soundex algorithm can be depicted as follows:

1. Replace all but the first letter of the string by its phonetic code
2. Eliminate any adjacent reptitions of codes
3. Eliminate all occurrences of code 0 (that is, eliminate vowels)
4. Return the first four characters of the resulting string

The Soundex algorithm has the following features:

- It is intuitive in terms of operation.
- The simplicity of the code allows to implement changes according to the objective.
- The processing time is relatively short.
- It has a high tolerance for variations in words that sound very similar or are exactly the same.

3.1 Different Adaptations to the Soundex Algorithm

We have observed the SMS representation in some languages like Spanish and English. That is why we propose some improvements to the basic Soundex algorithm with the purpose of obtaining the same code for a word in the SMS representation and its corresponding normal way of writing (standard vocabulary). We have done different adaptations of the two languages studied in this paper: Spanish and English. Below we resume the changes made to the Soundex algorithm for the Spanish language:

1. To keep the four digits of the Soundex code, but replace the first letter with its numeric representation.

2. If the letter "X" appears aside to a consonant, then change it for the letters "PR", and immediately code this two letters to its numeric representation. The rationale of this proposal is that the "x" letter is often used for expressing the word "por (because)".
3. To replace all symbols for its corresponding name, i.e. $ = pesos (Mexican pesos), % = "por ciento (percent)", hs = horas (hours), among others using the SMS dictionaries latterly introduced.
4. To replace all numbers in the original text for its textual representation, and then compute its Soundex code.

For the experiments carried out in this paper, we have tested four different approaches. The first one is when no phonetic codification is applied, which we have named *Uncodified* version. The second approach is named *Soundex* because it uses de basic Soundex algorithm. The following two approaches: *NumericSoundex* and *SoundexMod* are summarized as follows:

1. *NumericSoundex:* To keep the four digits of the Soundex code, replacing the first letter with its numeric representation.
2. *SoundexMod:* Before applying *NumericSoundex*, we use a dictionary of common SMS acronyms and phrases abbreviations for normalizing the SMS texts. A freely SMS dictionary available online[1] was used.

Similar changes to the Soundex algorithm was done in order to deal with the English language, obtaining four approches to be compared in the experiments carried out (*Uncodified*, *Soundex*, *NumericSoundex* and *SoundexMod*). For the English *SoundexMod* version, the freely SMS dictionary available online[2] was used.

4 Evaluating the Different Soundex Adaptations for SMS Text Representation

In this section we study the behavior of the different phonetic representations proposed when considering lexical similarity. The description of the corpus characteristics is done in the following SubSection. The metric used for determining the performance of the different phonetic algorithms together with the lexical similarity values found are shown in SubSection 4.2.

4.1 Datasets

In order to evaluate the adaptations made to the Soundex Algorithm for the Spanish language, we have constructed a parallel SMS corpus on the basis of the book named "*En Patera y haciendo agua*" [12] which is freely available online[3]. The salient features of this parallel corpus are shown in Table 2.

[1] http://www.diccionariosms.com

[2] http://smsdictionary.co.uk/

[3] http://www.adiccionesdigitales.es/libro

In order to evaluate the phonetic modification for the English language, we have used two different corpora. The first one is a parallel SMS corpus of 5,000 SMS aligned with reference messages manually prepared by two persons prepared by [3] as a training dataset. The salient features of this corpus are shown in Table 3.

Table 2. A Spanish parallel corpus of SMS

Feature	SMS	Original text
Number of messages	316	316
Number of tokens	30,195	30,270
Vocabulary size	7,061	6,448
Average message length in words	95.55	95.79
Average message length in characters	454.11	552.70
Average characters per word	4.75	5.77

Table 3. An English parallel corpus of SMS [3]

Feature	SMS	Original text
Number of messages	5,000	5,000
Number of tokens	68,666	69,521
Vocabulary size	6,814	5,746
Average message length in words	13.73	13.9
Average message length in characters	57.95	62.44
Average characters per word	4.22	4.49

The second corpus used in this experiment was the corpus of the SMS-based FAQ Retrieval task of the FIRE 2011 competition (Forum for Information Retrieval Evaluation)[4]. The salient features of this comparable corpus were already shown in Table 4.

4.2 Evaluation Based on Lexical Similarity

We are using a similarity measure in order to observe how well the different approaches codify correctly the SMS. We basically, applied a similarity measure for verifying in which percentage the Soundex-like codes in Spanish or English original texts are equal/similar to the Soundex-like codes for the text written in SMS for both languages. In particular, we use the Jaccard coefficient [13] to measure the similarity between the sample sets. Let SMS' be the SMS codified and T' be the original text codified, both with one of the above presented Soundex-like phonetic representation, then in Eq. 1 it is shown the Jaccard coefficient between SMS' and T'.

$$Jaccard(SMS', T') = \frac{|SMS' \bigcap T'|}{|SMS' \bigcup T'|} \quad (1)$$

[4] http://www.isical.ac.in/%7Efire/

Table 4. An English comparable corpus of SMS

Feature	SMS	Original text
Number of messages	721	721
Number of tokens	5,573	7,337
Vocabulary size	2,121	2,034
Average message length in words	7.74	10.14
Average message length in characters	37.28	56.62
Average characters per word	4.82	5.58

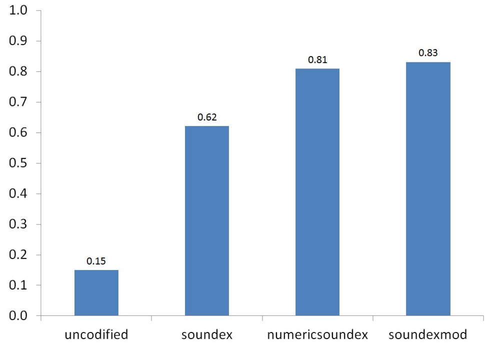

Fig. 1. Average Jaccard similarity for the different SMS text representations with the Spanish SMS parallel corpus

The intersection represents the number of matches between the SMS codified and the original text in Spanish or English also codified with some of the proposed Soundex-like algorithm. The union represents the total number of words in the data set. We apply the Jaccard coefficient between the SMS and the correct translation (or associated text in the case of the comparable corpus). The greater the value of the Jaccard coefficient, the better, the matching between the pair of codified texts. In Figure 1 we show the average Jaccard coefficient values obtained after comparing each pair (text,SMS) for the Spanish SMS parallel corpus. A 0.15 of average similarity in the *Uncodified* corpus show the great difference that exist between SMS and the corresponding translation. By applying the *Soundex* algorithm we obtain a similarity average value of 0.62. However, by just modifying the Soundex algorithm considering the first element of the code to be a numeric value (*NumericSoundex*), we improve the Soundex representation obtaining a similarity value of 0.81. Finally, the *SoundexMod* approach obtains the best value of similarity (0.83).

In Figure 2 we show the average Jaccard coefficient values obtained after comparing each pair (text,SMS) for the English SMS parallel corpus. A 0.652 of average similarity in the *Uncodified* corpus indicates that there exist a more or less stable way of writing SMS in this corpus, using a small number of acronyms, contractions and elimination of vowels. By applying the *Soundex* algorithm we obtain a similarity average value

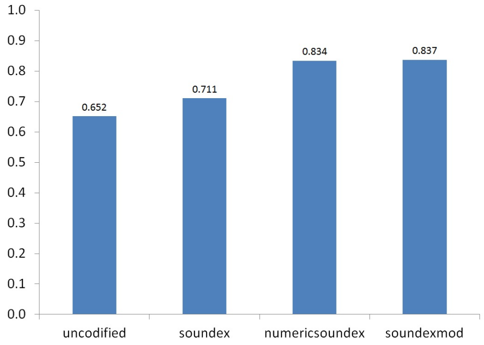

Fig. 2. Average Jaccard similarity for the different SMS text representations with the English SMS parallel corpus

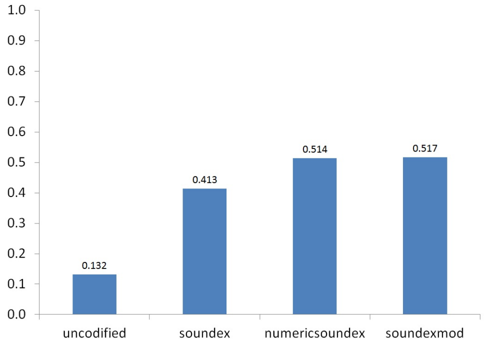

Fig. 3. Average Jaccard similarity for the different SMS text representations with the English SMS comparable corpus (FIRE)

of 0.711. However, by just modifying the Soundex algorithm considering the first element of the code to be a numeric value (*NumericSoundex*), we improve the Soundex representation obtaining a similarity value of 0.834. Finally, the *SoundexMod* approach obtains the best value of similarity (0.837).

In Figure 3 we show the average Jaccard coefficient values obtained after comparing each pair (text,SMS) for the English SMS comparable corpus. A 0.131 of average similarity in the *Uncodified* corpus show the great difference that exist between the query written in SMS format and the corresponding FAQ question associated. By applying the *Soundex* algorithm we obtain a similarity average value of 0.413. However, by just modifying the Soundex algorithm considering the first element of the code to be a numeric value (*NumericSoundex*), we improve the Soundex representation obtaining a similarity value of 0.514. Finally, the *SoundexMod* approach obtains the best value of similarity (0.517).

In summary, we have observed that the Soundex algorithm is useful for codifying SMS, but the simple modification *NumericSoundex* greatly improves the similarity between SMS codified and original texts codified. Ad-hoc modifications to the Soundex method for a particular language slightly improves the results, but generates a language dependent algorithm.

5 Conclusions and Future Work

We have proposed adaptations to the Soundex phonetic algorithm in order to provide a suitable algorithm for codifying SMS which may further be used in other natural language tasks such as text extraction, question answering, information retrieval, etc., which use SMS as part of the written texts.

We have used two parallel corpora and one comparable corpus with the purpose of evaluating the performance of the proposed algorithms in a more challenging environment. The Soundex method greatly improves the matching between SMS and their corresponding associated words, but the modifications proposed in this paper improve also the Soundex method in all the cases.

As future work, we would like to study the familiy of words clustered around a single SMS word in order to determine the existence of semantic groups. Also we would like to evaluate the performance of the presented phonetic codification in real natural language tasks.

References

1. Reyes-Barragán, A., Villaseñor Pineda, L., Montes-y-Gómez, M.: INAOE at QAst 2009: Evaluating the usefulness of a phonetic codification of transcriptions. In: Proceedings of CLEF 2009 Workshop. Springer (2009)
2. Aiti, A., Min, Z., Pohkhim, Y., Zhenzhen, F., Jian, S.: Input normalization for an english-to-chinese sms translation system. In: MT Summit 2005 (2005)
3. 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)
4. 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)
5. 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)
6. Hall, P.A.V., Dowling, G.R.: Approximate string matching. ACM Comput. Surv. 12, 381–402 (1980)
7. Rajkovic, P., Jankovic, D.: Adaptation and application of daitch-mokotoff soundex algorithm on Serbian names. In: XVII Conference on Applied Mathematics (2007)
8. Taft, R.: Name search techniques. Special report. Bureau of Systems Development, New York State Identification and Intelligence System (1970)
9. Philips, L.: Hanging on the metaphone. Computer Language Magazine 7, 38–44 (1990)

10. Knuth, D.E.: The art of computer programming, vol. 3: sorting and searching, 2nd edn. Addison Wesley Longman Publishing Co., Inc., Redwood City (1998)
11. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
12. Romero, J., et al.: En patera y haciendo agua. Adicciones Digitales (2011), `http://www.adiccionesdigitales.es/libro`
13. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de Société vaudoise des Sciences Naturelles 37, 547–579 (1901)