Automatic normalization of short texts by combining statistical and rule-based techniques

Marta R. Costa-jussà[†] **and Rafael E. Banchs**[‡]

[†]Barcelona Media Innovation Center
Av. Diagonal 177, 08018, Barcelona
marta.ruiz@barcelonamedia.org
[‡] Institute for Infocomm Research
1 Fusionopolis Way, 138632, Singapore
rembanchs@i2r.a-star.edu.sg

Abstract

Short texts are typically composed of small number of words, most of which are abbreviations, typos and other kinds of noise. This makes the noise to signal ratio relatively high for this specific category of text. A high proportion of noise in the data is undesirable for analysis procedures as well as machine learning applications. Text normalization techniques are used to reduce the noise and improve the quality of text for processing and analysis purposes.

In this work, we propose a combination of statistical and rule-based techniques to normalize short texts. More specifically, we focus our attention on SMS messages. We base our normalization approach on a statistical machine translation system which translates from noisy data to clean data. This system is trained on a small manually annotated set. Then, we study several automatic methods to extract more general rules from the normalizations generated with the statistical machine translation system.

We illustrate the proposed methodology by conducting some experiments with a SMS Haitian-Créole data collection. In order to evaluate the performance of our methodology we use several Haitian-Créole dictionaries, the well-known perplexity criteria and the achieved reduction of vocabulary.

Keywords: Normalization of short texts, Statistical Machine Translation, Automatic extraction of rules, Perplexity

1. Introduction

Recently, new Information and Communication Technologies (ICTs) have transferred the authorship of contents from institutions to the people. Therefore, the new information and communication channels are not a place to publish institutional information as traditional channels used to be, but they have provide the means for users to exchange, explain and write about their lives, opinion and interests, as well as to upload media, photos and files. The main characteristic of this new way of sharing information is the casual way in which it is performed. The users, young and not so young, use their personal computers and personal devices to communicate informally by using short text messages or chats, without minding about spelling and deliberately shortening the words and using contextual slang. As a consequence, the new challenges researchers must deal with when analyzing the content in this new information society context are related to the fact that quite often they are not following linguistic rules.

During years, there has been a big effort to produce natural language processing tools that try to understand well written documents and sentences, but these tools cannot be applied out of the box to analyze the contents of the new information society context. Not even basic morpho-syntactic tools like stemmers can bring to common stems words that have been shortened (like Xmas or Christmas). Figure 1 shows an example of how natural language expressions can be re-coded in the context of chat or SMS communications in English.

Because of the specific nature of short texts, as well as the specific challenges they pose to traditional natural language processing and computational linguistic methods, there have been a recent increment in the research events related to these problems. For instance, several tracks have been organized in the context of the different evaluation frameworks at TREC (blog and Web tracks), CLEF (Web people search laboratory), NTCIR (opinion analysis pilot task), INEX (ad-hoc passage retrieval task), ROMIP (track on news clustering), and FIRE (ad-hoc task on retrieval from technical forums and mailing lists). More specifically, the CAW2 workshop [1] organized

[1] http://caw2.barcelonamedia.org/

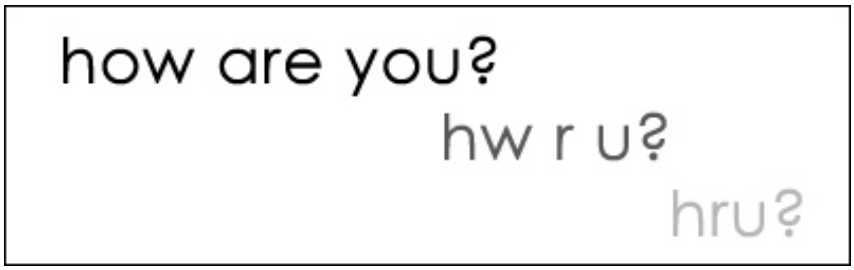

Figure 1: *Chat or SMS English language.*

a text normalization task where the objective was to correct sample texts extracted from the web 2.0. The main objective of the task was to normalize chat-style sentences in order to produce sentences that were at least syntactically correct, and that could be suitable for further treatment by means of existing analysis tools and information retrieval systems. The task was in English, which was quite an advantage because there are many resources available in the web for the English language. In the evaluation, participants were required to deal with textual contents from sites where users behave in different ways: twitter, chats, comments on products, forums, news and comments on news. There was no parallel corpus of misspelled/corrected text for this evaluation. What was given was a train set of some hundred thousand sentences with a high proportion of misspelled words, produced from different users in different contexts. Therefore, participants crawled the web to find dictionaries and other resources that could help in the conversion from chat-style language into standard language (Henriquez and Hernández, 2009).

In this paper, we propose a much more challenging scenario. We work with the Haitian-Créole language which is a language with very low resources. Our objective is to normalize a relatively large corpus of short messages given very scarce resources. We only count with a small set of parallel corpus, at the sentence level, of SMS texts and their corresponding standard language normalization, plus a dictionary of standard Haitian-Créole words. This dataset corresponds to a subset of SMS communications in the aftermaths of Haiti's earthquake, when many people sent SOS SMS to inform and request for help. This dataset has been recently released for research purposes in the context of the statistical machine translation competition organized in WMT 2011 (Callison-Burch et al., 2011). The main challenge of the WMT 2011 task was to translate SMS texts into English. We instead, focus the use of the data collection here on the problem of normalization of short texts in the specific scenario of a language with scarce resources.

The normalization task presented here is different from the one presented in CAW2 because of: (1) the availability of a small parallel corpora instead of monolingual raw corpora; and (2) the low resource language that is considered does not allow crawling the web in order to find dictionaries for chat-style idioms and expressions. Therefore, we cannot directly compare the methodologies from CAW2 with our methodology.

The normalization procedure described here is implemented with the few resources available and, accordingly, it is extensible to any other scenario of languages with scarce resources. Additionally, it can also leverage over existent resources for the cases of languages for which there is a lot of available resources. The proposed method is based on a statistical normalization methodology that combines statistical machine translation techniques with some rule-based strategies. In a first stage, similarly to previous work (Aw et al., 2006), we train a phrase-based statistical machine translation system. Then, differently from previous works, we use a translation dictionary to extract automatic rules that can be applied to normalize the text. Several experiments are performed using different combinations of these two techniques.

The rest of the paper is organized as follows. Section 2. describes some details about the scarce resource language under consideration. Section 3. presents a description of the phrase-based statistical machine translation system that we are using for implementing the first step of the normalization method. Then, section 4. describes the methodology we use to automatically extract rules from the resulting phrase-based translation table. These automatically extracted rules are the ones to be used in the second step of the normalization process. Additionally, we present how we can further exploit the combination of both methodologies. Section 5. presents the experiments related to the normalization of Haitian-Créole SMS messages. Details on the data used for normalization and for training the statistical machine translation system are presented. The performance of our methodology is

evaluated using a standard language model technique, based on the information-theoretic measure of perplexity. We show and discuss the quality of improvement resulting from using both systems (statistical and rule-based) as well as the concatenation of both. Finally, section 6. presents the most relevant conclusions of this work and the future plans for continuing this research.

2. Haitian-Créole

Haitian-Créole is a creolized language derived mainly from 17th to 18th century French, and with some minor influences from other linguistic groups such as African, Arabic, Spanish, Tano, and English. Table 5 shows some examples in Haitian-Créole and its corresponding translation into French and English.

A demen	À demain	See you tomorrow
A pi ta	À plus tard	See you later
Adye	Au revoir	Good bye
Anchante	Enchanté	Enchanted
Bon apre-midi	Bon après-midi	Good afternoon
Bnn nui	Bon nuit	Good night
Bonjou	Bonjour	Good morning
Bonswa	Bon soir	Good evening
Dezole	Desolé	Sorry
Eskize m	Excuse-moi	Excuse me
Ki jan ou rele?	Comment vous appelez-vous?	What is your name?
Ki jan ou ye?	Comment allez-vous?	How are you?
Ki laj ou?	Quelle age avez-vous?	What is your age?

Table 1: *Haitian-Créole examples, including the corresponding translations in French and English*

Haitian-Créole is one of the two official languages of Haiti, besides French. It is spoken by a population of nearly twelve million people in Haiti, and by about two to three million emigrants residing in the Bahamas, Canada, United States, France, Cayman Islands, Cuba, French Guiana, Martinique, Guadeloupe, Belize, Puerto Rico, Dominican Republic, and other Caribbean countries.
Two theories are currently considered about the origin of Haitian-Créole. The first of these theories states that a rudimentary type of Creole was actually originated among captive slave communities in Africa and was brought to the Caribbean from Africa and evolved into current Haitian-Créole. The second one states that Haitian-Créole was directly originated and developed in the Caribbean by slaves that were native speakers of languages from the Fon family, who included French vocabulary into their native Fon grammars.
Haitian-Créole's grammar is actually different from modern French, as it is closer to 17th century colonial French that was spoken by low class white people who emigrate from Europe to the colonies in the Caribbean; more specifically, the isle of Sainte-Domingue where nowadays both Haiti and Dominican Republic are located. Among some of the relevant characteristics of Haitian-Créole, we can mention the following:

- The primary word order is Subject-Verb-Object (SVO).
- Adjective and Articles are not inflected for gender.
- Verbs are not inflected neither for person nor for tenses.
- Uses suffixes for indicating possession and pluralization.
- Lexicon is principally composed of French derived words.
- Sentence structures are related to Africa's Fon language family.

Haitian-Créole is considered a living language, as it constantly creates and borrows new words from different languages and cultures to describe new or old concepts and realities [2].

[2]http://en.wikipedia.org/wiki/Haitian_Creole_language

3. Phrase-based Statistical Machine Translation System

Statistical Machine Translation (SMT) uses statistical algorithms to decide the most likely translation of a word. Thus, given a source string $s_1^J = s_1 \dots s_j \dots s_J$ to be translated into a target string $t_1^I = t_1 \dots t_i \dots t_I$, the aim is to choose, among all possible target strings, the string with the highest probability:

$$\tilde{t_1^I} = \underset{t_1^I}{argmax}\, P(t_1^I|s_1^J)$$

where I and J are the number of words of the target and source sentence, respectively.
The first SMT systems were reformulated using Bayes' rule. Recent systems have expanded such an approach to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och, 2003). This approach leads to maximizing the corresponding weights (λ_m) of a linear combination of feature functions (h_m):

$$\tilde{t} = \underset{t}{argmax} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t, s) \right\}.$$

This log-linear combination of feature functions can be represented as shown in Figure 2.

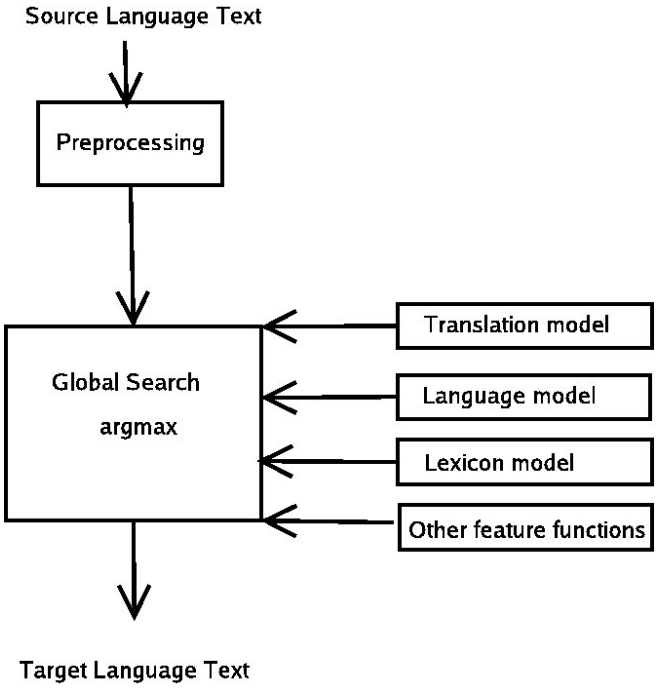

Figure 2: *SMT log-linear schema*

Given some parallel corpora at the level of sentence in a target and source language, the translation model tries to assign a probability that the target sentence t_1^I generates the source sentence s_1^J. While these probabilities can be estimated by thinking about how each individual word is translated, modern statistical MT is based on the intuition that a better way to compute these probabilities is by considering the behavior of phrases, i.e. sequences of words (Koehn and Knight, 2003). The basic idea of phrase-based translation is to segment the given source sentence into units (hereinafter called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.
Basically, a bilingual phrase is a pair of m source words and n target words. For extraction from a bilingual word aligned training corpus, two additional constraints are considered:

1. words are consecutive, and,
2. they are consistent with the word alignment matrix.

Given the collected phrase pairs, the phrase translation probability distribution is commonly estimated by relative frequency in both directions. One of the most common baseline implementations of the phrase-based system combines the relative frequencies together with the following additional feature models:

- Target language model, which is formulated as a probability distribution over strings that attempts to reflect how likely a string occurs inside a language (Stolcke, 2002). This model takes into account the quality of the target language in the generated hypothesis, regardless of the appropriateness of the translation.
- Word bonus and phrase bonus, which aim at compensating the natural tendency of statistical machine translation systems to generate short translation outputs. These models favour the generation of wordier translation hypotheses.
- Source-to-target and target-to-source lexical models, also known as IBM1 models (Brown et al., 1993). These models assign probabilities to each translation unit based on the probability of word per word translations within the unit. The probability estimated by lexical models tends to be in some situations less sparse than the probability given directly by the relative frequency models.
- Reordering model, which provides a measure of the plausibility of word movements during decoding; these models have been widely investigated in SMT (Costa-jussà and Fonollosa, 2009). Nowadays, one of the most used algorithms is the lexicalized reordering that was originally proposed in (Tillmann, 2004). This reordering model learns local orientations with probabilities for each bilingual phrase from training data. Therefore, for each bilingual unit, the basic idea is to learn the likelihood that the particular unit directly follows a previous bilingual unit (monotone), it is swapped with a previous bilingual unit (swap), or it is determined to be disconnected from the previous unit (discontinuous). The model learns local orientations (monotone vs. non-monotone) with probabilities for each bilingual phrase from the training material. During decoding, the model attempts to find a Viterbi local orientation sequence.

Finally, the log-linear combination of these models is optimized for decoding by following the minimum error rate procedure (Och, 2003). This method performs a series of consecutive adjustments to the feature weight vector, taking advantage of special properties of the mapping from sets of feature weights to the resulting translation quality measurement, with the objective of minimizing the translation errors over a development dataset.
This phrase-based methodology is used to build an SMT system from "raw" Haitian-Créole to "clean" Haitian-Créole. The training corpus, which was manually built, is detailed in section 5.

4. Generating automatic rules given a translation table

In the previous section we described how the SMT system is built from raw Haitian-Créole to clean Haitian-Créole. Following the phrase-based schema, the SMT system learns a translation table given the available training parallel corpora at the level of sentence. Therefore, from the specific raw-to-clean Haitian-Créole SMT system we described in the previous section, the resulting translation table can be thought of as a bilingual dictionary between raw and normalized words and phrases. This statistical dictionary is the one we use to extract rules at the word level for the rule-based step of the proposed approach.
Translation units extracted by the statistical machine translation system are typically of the form "source chunk" $\rightarrow$ "target chunk" (where source and target are raw and normalized Haitian-Créole, respectively), where each chunk can comprise one to several consecutive words. As it can be expected, most of the extracted translation units are indeed identity units, i.e. translation units containing the same source and target chunks. Then, for extracting correction rules, we are only interested in those translation units with different source and target components. However, as the alignment procedure is very noisy, not every translation unit with different source and target components constitutes a good correction rule. According to this, a filtering procedure should be implemented for extracting good correction rules and discarding noisy units.
The implemented procedure for correction rule extraction, which is illustrated in Figure 3, can be summarized in the following four steps:

1. Select, from all translation units contained in the phrase-table, those translation units that contain only one source word.

2. From the set of selected units, discard all identity units (i.e. translation units containing the same source and target chunks).

3. From the remaining set of selected units, discard all those units for which their source words are in the Haitian-Créole dictionary.

4. In the case one source has multiple possible translations, choose the translation with the lowest Levenshtein distance.

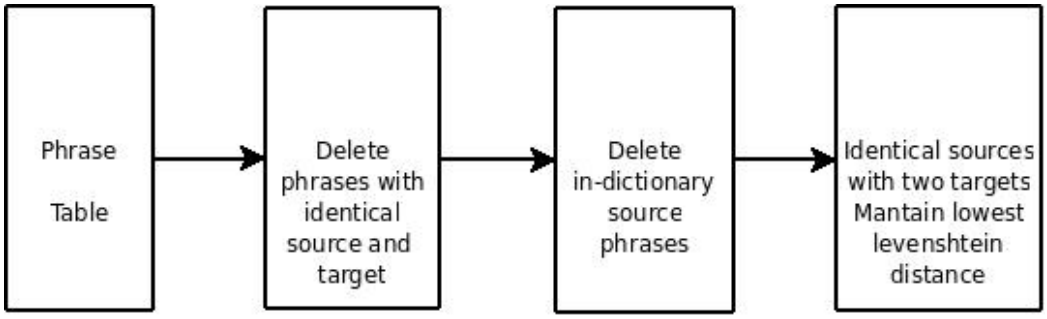

Figure 3: *Generating automatic rules schema.*

The final step for rule extraction described above is very strong and it may produce some errors. For instance, a shorter source word like "w", with possible translations "with" or "who", will be always mapped to "who" because it has the lower Levenshtein distance to "w". However, in practice, we saw that this last filtering step only affects a very small number of rules (around 5%). As further research, we could experiment with some alternatives to this step such as, for example, using a language model to choose the right rule in the right context. We will see along this work that reapplying these rules certainly helps improving the quality of the normalization, even when the statistical translation step has been already applied to the data. This is basically because, in this second step, we are introducing further knowledge such as the one provided by the standard Haitian-Créole dictionary, which allows for distinguishing between real Haitian-Créole words and chat-style speak artefacts.
According to this, and in order to further improve the quality of the normalization procedure, we can concatenate the statistical approach described in the previous section with the rule-based approach described here, as well as iterate over the concatenation of these two methods several times. Figure 4 shows diagram of the iterative process that we can follow.

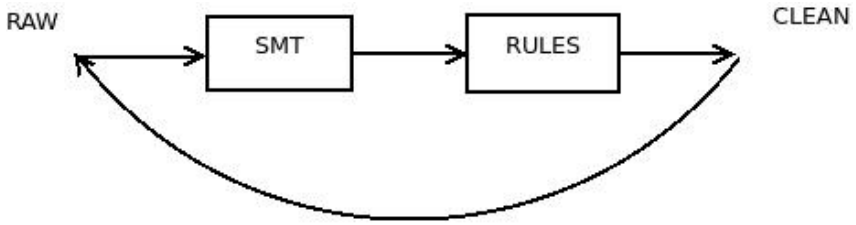

Figure 4: *Iterative process*

5. Experiments

This section describes the experimental work that has been performed to illustrate the proposed short text normalization procedure. First, we describe the specific dataset we have used for the experiments, which is based on the available Haitian-Créole SMS message set. Additionally, we show the experimental details about the proposed statistical approach to short text normalization. Finally, we evaluate the performance of our proposed methodology.

5.1. Data

The data from this work has been obtained from the 6th Workshop on Statistical Machine Translation (WMT 2011). The original task from the workshop was to translate Haitian-Créole SMS messages into English. According to the organizers of WMT 2011, and as mentioned in the workshop's web page [3], these text messages (SMS) were sent by people in Haiti during the January 2010 earthquake. The messages were sent to an emergency response and information service service called "Mission 4636". They were originally written in Haitian-Créole, and were translated into English by some volunteers, so that first responders (many of whom did not speak Haitian-Créole but could speak English) could understand and proceed on them.

Given the specific problem we are addressing in this paper, we will focus our attention only on the raw part of the Haitian-Créole WMT 2011 dataset, which contains the major volume of short text communications. Therefore, we will be working here only with that part of the data mainly containing SMS abbreviations. Table 2 shows the raw corpus that we are using to demonstrate the proposed normalization procedure. The statistics shown in the table where computed after tokenizing the corpus with the standard tokenizer available within the Moses SMT tools. As there is not specific tokenization algorithm for Haitian-Créole, we used the French approximation for taking advantage of the proximity between Haitian-Créole and French.

	Haitian-Créole Raw
Number of Sentences	16,594
Running Words	363.0k
Vocabulary Size	18.9k
Max Sent Length	83
Average Sent Length	21

Table 2: *Tokenized Haitian-Créole raw data.*

Additionally, we also used a subset of data with normalized transcriptions available. This portion of the data provided the necessary raw-clean parallel dataset used for training the statistical machine translation system for the first normalization step. Table 3 shows the raw-clean parallel subset of the Haitian-Créole corpus.

	Haitian-Créole-Raw	*Haitian-Créole-Clean*
Training Sentences	2378	
Words	46192	47924
Vocabulary	5155	4215
Max Sent Length	81	81
Average Sent Length	19.4	20.2
Development Sentences	217	
Words	4196	4366
Vocabulary	1093	971
Max Sent Length	61	68
Average Sent Length	19.3	20.1
Test Sentences	215	
Words	3811	3920
Vocabulary	1060	954
Max Sent Length	66	63
Average Sent Length	17.7	18.2

Table 3: *Haitian-Créole raw-clean parallel dataset.*

Next, we also used a multi-word dictionary of standard Haitian-Créole which statistics are presented in table 4. This resource is a compilation of several dictionaries provided in the WMT 2011 evaluation: a glossary and the haitisurf and krengle dictionaries.

[3] http://www.statmt.org/wmt11/featured-translation-task.html

	Haitian-Créole multi-word dictionary
Instances	49,995
Words	128.6k
Vocabulary	18.2k
Max Sent Length	27
Average Sent Length	2.57

Table 4: *Haitian-Créole multi-word dictionary statistics.*

Finally, table 5 shows some actual examples of manually conducted normalizations extracted from the raw-clean small parallel corpus. Notice that the objective of this task is to normalize terms in order to reduce the vocabulary. The reduction of vocabulary can improve applications such as translation.

	Example
RAW	mwen **beswen imfomasyon tou suit** sou **tan an** .
CLEAN	mwen **bezwen enfomasyon touswit** sou **lameteyo** .
RAW	ki kote **yap** bay manje
CLEAN	ki kote **y ap** bay manje **?**
RAW	nou bezwen **tant** dlo manje nan **kanapeve**
CLEAN	nou bezwen dlo , manje ak tant nan **canape-vert .**
RAW	voye lapolis **jizi** mirak kounye a
CLEAN	voye lapolis **, jezi** mirak kounye a
RAW	yo **e** yon **mesaje** pou di ki kote m 'ap ka ale lopital
CLEAN	yo **se** yon **mesaj** pou di ki kote m 'ap ka ale lopital **?**

Table 5: *Examples of manual normalization of Haitian-Créole*

5.2. Translation system

We used the Moses decoder (Koehn et al., 2007) and all its supporting scripts for implementing and training the translation system. Within this process, the parallel training corpus described in 3 was aligned in both directions by using Giza++ (Och and Ney, 2000), and symmetrizing the alignment by using grow-final-diagonal (Koehn et al., 2005). Phrases were extracted from this symmetric alignment set.

The phrase-table contained all the standard features including the conditional and posterior probabilities, the IBM1 lexical probabilities in both, the source-to-target and target-to-source, directions and the phrase bonus. The target language model was estimated using an n-gram language model of order 3 with Kneser-Ney smoothing. For this, we used the SRILM toolbox (Stolcke, 2002). To compensate the fact that the language model benefits short outputs, we included the word bonus. Reordering was also used, although this task should be monotonic, as we expected the phrase-based system to be able to learn this from the training set.

Finally, we used the Moses implementation of the minimum error training procedure (Och and Ney, 2002) for tuning the system. The minimization of translation BLEU (Papineni et al., 2002) over the development data set was considered for tuning. In this first normalization step we obtained a BLEU of 76.03 over the development set and 72.11 over the test set.

Table 6 shows some examples of the normalization done by the SMT system. Input sentences of the SMT system are named raw and output sentences of the SMT systems are named clean.

5.3. Automatic Rules

The automatic rules are extracted from the translation table derived in the previous subsection by using the procedure described in section 4. We extracted and applied a total of 1130 rules of unit word length, i.e. only rules involving a single source word were considered. Although these rules were contained in the SMT table used in the first step of the normalization procedure, they were not necessarily applied during the raw-to-clean translation procedure as the phrase-based machine translation system makes its decisions during decoding time

	Example
RAW	infomation sou kestion te tranblea ak **lek**
CLEAN	infomation sou kestion te tranblea ak **lekol**
RAW	yo **jwen** yon ti moun 16 ans bo sen gera tou vivan .
CLEAN	yo **jwenn** yon ti moun 16 ans bo sen gera tou vivan .
RAW	ils trvent **par** dieu **ce** qu 'ils ont besoin
CLEAN	ils trvent **ak** dieu **se** qu 'ils ont besoin
RAW	ok dlo kle pa dlo propre
CLEAN	ok dlo kle pa dlo propre **.**
RAW	**seisme** lan pase anko . when did you come . [phonenumber]
CLEAN	**seyism** lan pase anko . when did you come . [phonenumber]
RAW	ske mwen pap fe anyen mwen vle travay mwen pa **jwen** tanpri gade kis
CLEAN	ske mwen pap fe anyen **.** mwen vle travay **.** mwen pa **jwenn .** tanpri gade kis

Table 6: *Examples of automatic normalization in Haitian-Créole using an SMT system*

by taking into account several different models. In this second step of the normalization method, the extracted automatic rules are applied to the texts in a deterministic way.

Source	Target
c	ak
gera	gerard
ske	eske
mka	ka
n.	dwe
ok	oke
kounya	kounye
ecole	lekol
mabite	abite
profesyon	pwofese

Table 7: *Examples of source and target rules.*

Table 7 shows some examples of rules which were extracted from the phrase table and verified for their source parts not to be in the standard Haitian-Créole dictionary while their target parts were. We observe that there are some words in French, as well as some abbreviations or shorter representations, which are quite common in chat-speak style communications.
Table 8 shows some examples of automatic normalization with the SMT system plus rules.
The main difference about Tables 6 and 8 is the introduction of rules. From the examples, we see that when using rules some words are changed by others (see Table 7) but the rules do not generate or eliminate words.

5.4. Results

In order to evaluate the impact of our methodology, we propose to measure the quality of the resulting normalizations in terms of language model perplexity. Therefore, we build a standard language model using the SRILM toolkit with the clean portion of the data (corresponding to the set presented in Table 3). The normalized texts are evaluated and compared among them by using the constructed language model and the standard measure of perplexity. As test dataset we used the same one prepared for the SMT task, which was shown in table 3.
Table 9 shows the quality (in terms of perplexity) of normalized texts obtained by using different variants of the proposed methodology. Additionally, the table also presents the reduction in vocabulary for each case.
The statistical step of the methodology can be concatenated with the rule-based step or the other way around. Additionally, we can apply both techniques several times in an iterative fashion, such as it was described in section 4.. As seen from the results reported in table 9, changing the order of application of the two steps

	Example
RAW	infomation sou kestion te tranblea ak **lek**
CLEAN	infomation sou kestion te tranblea ak **lekol**
RAW	yo jwen yon ti moun 16 ans bo sen gera tou vivan .
CLEAN	yo **jwenn** yon ti moun 16 ans bo sen **gerard** tou vivan .
RAW	ils trvent **par** dieu ce qu 'ils ont **besoin**
CLEAN	ils trvent **ak** dieu se qu 'ils ont **bezwen**
RAW	**ok** dlo kle pa dlo propre
CLEAN	**oke** dlo kle pa dlo propre **.**
RAW	**seisme** lan pase anko . when did you come . [phonenumber]
CLEAN	**seyism** lan pase anko . when did you come . [phonenumber]
RAW	**ske** mwen pap fe anyen mwen vle travay mwen pa **jwen** tanpri gade kis
CLEAN	**eske** mwen pap fe anyen mwen vle travay **.** mwen pa **jwenn** tanpri gade kis

Table 8: *Examples of automatic normalization in Haitian-Créole using a two iteration SMT system plus rules.*

	Perplexity	Vocabulary
Raw	110.7	18.9k
Clean with SMT	107.8	18.4k
Clean with Rules	104.4	18.5k
Clean with SMT + Rules	103.8	18.3k
Clean with Rules + SMT	105.5	18.2k
Two-iterations Clean with SMT + Rules	**102.9**	18.2k

Table 9: *Perplexity and vocabulary measured in the initial raw text and in the processed clean texts. Best perplexity results are in bold.*

(SMT followed by rules, or rules followed by SMT) affects the results. This is mainly because the SMT step is not deterministic and it will behave differently whether the input is the original raw text or it is the cleaned text resulting from applying the rules. Notice that the SMT system was trained with the raw text as source. Therefore, it does not benefit much from the fact of applying rules first. In fact, although the difference is not significance, cleaning with rules has a lower perplexity than cleaning with rules followed by the SMT system.
Additionally, as seen from the table, performing two iterations of the basic normalization procedure does not bring a significant improvement.
All in all, results show that the best configuration is the one that uses both, the statistical and the rule-based, correction methods in that order; in which case only a 6.23% reduction in perplexity is obtained. Slightly, but not significantly, better result is obtained for the configuration that implements a two-pass iteration of the previous system. In this last case a 7.05% reduction in perplexity is achieved. This reduction is the best we could achieve with the scarce resources that we had.
Notice from table 9 that there is a high correlation among reduction of vocabulary and reduction in perplexity. The correlation is 0.82 and $p = 0.047$, which is quite an expected result.

6. Conclusions

This work presented and evaluated a two-steps text normalization method for sms- and chat-like messages. Short texts are typically composed of a small number of words, containing abbreviations, typos and other kind of errors and noise. This kind of textual content must be corrected and normalized in order to make them usable by conventional text analysis and processing techniques.
The proposed text normalization method is based on a combination of statistical and rule-based techniques and is implemented in a two step procedure. In the first step, a statistical machine translation system, which is trained on a small manually corrected dataset, is used for "translating" the raw or noisy text into clean or normalized text. Then, in the second step, a set of deterministic correction rules, which are extracted from the

same manually corrected data set with the help of statistical alignment tools and dictionaries, is used to replace remaining erroneous forms and noisy tokens into normalized dictionary entries.
The methodology has been experimentally tested on a data set of SMS in Haitian-Créole made available through the WMT 2011 challenge task. In order to evaluate the performance of the proposed methodology we measured both language model perplexity and the amount of vocabulary reduction for five different experimental settings. Results show that the best configuration is the one that uses both, the statistical and the rule-based correction methods in that specific order. In this case a 6.23% reduction in perplexity is obtained. Slightly, but not significantly, better result is obtained for the configuration that implements a two-pass iteration of the previous system. In this last case a 7.05% reduction in perplexity is achieved. This reduction is not bad if we take into account the scarce resources that we are working with and the sizes of the involved datasets.
As future work, we intend to test the proposed methodology in some other data sets including different style of communications, such as online chats and twitter messages, as well as different languages, such as English, Spanish and French. Additionally, we plan to improve the performance of the methodology by exploring new alternatives to correction rule extraction and the incorporation of additional on-line resources such as chat-speak dictionaries and some other context-dependent lexicons.

7. Acknowledgments

The authors want to thank the anonymous reviewers for their valuable comments and suggestions which helped improving this paper. The authors also want to thank Barcelona Media Innovation Center and Institute for Infocomm Research for its support and permission to publish this research.
This work has been partially funded by the Spanish Ministry of Economy and Competivity through the *Juan de la Cierva* fellowship program and by the Seventh Framework Programme of the European Comission through the T4ME contract (grant agreement no.:249119).

8. References

A. Aw, M. Zhang, J. Xiao, and J. Su. 2006. A phrase-base statistical model for sms text normalization. In *Proc. of the COLING/ACL on Main conference poster sessions*, pages 33–40, Sydney, Australia.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July.

M. R. Costa-jussà and J. A. R. Fonollosa. 2009. State-of-the-art word reordering approaches in statistical machine translation. *IEICE Transactions on Information and Systems*, 92(11):2179–2185, November.

C. Henriquez and A. Hernández. 2009. A ngram-based statistical machine translation approach for text normalization on chat-speak style communications. In *Proceedings of the CAW2 Workshop*, Madrid, June.

P. Koehn and K. Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

P. Koehn, A. Amittai, A. Birch, C. Callison-Burch, M. Osborne, D. Talbot, and M. White. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proc. of International Workshop on Spoken Languages Translation*, Pittsburgh, October.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

F. J. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proc. of the 18th conference on Computational linguistics*, pages 1086–1090, Morristown, NJ, USA.

F.J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, USA, July.

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, July.

K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, July.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, pages 901–904, Denver, USA, September.

C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proc. of the Human Language Technology Conf., HLT-NAACL'04*, pages 101–104, Boston, May.