Personalized Normalization for a Multilingual Chat System

Ai Ti Aw and Lian Hau Lee
Human Language Technology
Institute for Infocomm Research
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
aaiti@i2r.a-star.edu.sg

Abstract

This paper describes the personalized normalization of a multilingual chat system that supports chatting in user defined short-forms or abbreviations. One of the major challenges for multilingual chat realized through machine translation technology is the normalization of non-standard, self-created short-forms in the chat message to standard words before translation. Due to the lack of training data and the variations of short-forms used among different social communities, it is hard to normalize and translate chat messages if user uses vocabularies outside the training data and create short-forms freely. We develop a personalized chat normalizer for English and integrate it with a multilingual chat system, allowing user to create and use personalized short-forms in multilingual chat.

1 Introduction

Processing user-generated textual content on social media and networking usually encounters challenges due to the language used by the online community. Though some jargons of the online language has made their way into the standard dictionary, a large portion of the abbreviations, slang and context specific terms are still uncommon and only understood within the user community. Consequently, content analysis or translation techniques developed for a more formal genre like news or even conversations cannot apply directly and effectively to the social media content. In recent years, there are many works (Aw et al., 2006; Cook et al., 2009; Han et al., 2011) on text normalization to preprocess user generated content such as tweets and short messages before further processing. The approaches include supervised or unsupervised methods based on morphological and phonetic variations. However, most of the multilingual chat systems on the Internet have not yet integrated this feature into their systems but requesting users to type in proper language so as to have good translation. This is because the current techniques are not robust enough to model the different characteristics featured in the social media content. Most of the techniques are developed based on observations and assumptions made on certain datasets. It is also difficult to unify the language uniqueness among different users into a single model.

We propose a practical and effective method, exploiting a personalized dictionary for each user, to support the use of user-defined short-forms in a multilingual chat system - *AsiaSpik*. The use of this personalized dictionary reduces the reliance on the availability and dependency of training data and empowers the users with the flexibility and interactivity to include and manage their own vocabularies during chat.

2 ASIASPIK System Overview

AsiaSpik is a web-based multilingual instant messaging system that enables online chats written in one language to be readable in other languages by other users. Figure 1 describes the system process. It describes the process flow between *Chat Client*, *Chat Server*, *Translation Bot* and *Normalization Bot* whenever *Chat Client* starts chat module.

When *Chat Client* starts chat module, the *Chat Client* checks if the normalization option for that language used by the user is active and activated. If

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 31–36,
Jeju, Republic of Korea, 8-14 July 2012.

so, any message sent by the user will be routed to the *Normalization Bot* for normalization before reaching the *Chat Server*. The *Chat Server* then directs the message to the designated recipients. *Chat Client* at each recipient invokes a translation request to the *Translation Bot* to translate the message to the language set by the recipient. This allows the same source message to be received by different recipients in different target languages.

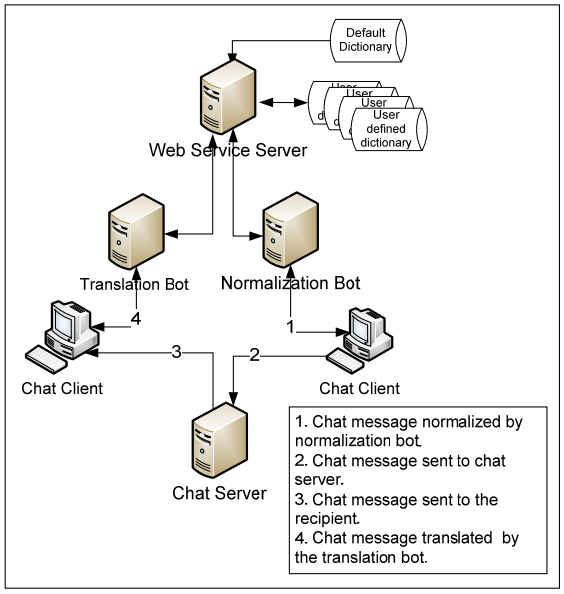

Figure 1. AsiaSpik Chat Process Flow

In this system, we use *Openfire Chat Server* by *Ignite Realtime* as our *Chat Server*. We custom build a web-based *Chat Client* to communicate with the *Chat Server* based on Jabber/XMPP to receive presence and messaging information. We also develop a user management plug-in to synchronize and authenticate user login. The translation and normalization function used by the *Translation Bot* and *Normalization Bot* are provided through *Web Services*.

The *Translation Web Service* uses in-house translation engines and supports the translation from Chinese, Malay and Indonesian to English and vice versa. Multilingual chat among these languages is achieved through pivot translation using English as the pivot language. The *Normalization Web Service* supports only English normalization. Both web services are running on *Apache Tomcat* web server with *Apache Axis2*.

3 Personalized Normalization

Personalized Normalization is the main distinction of *AsiaSpik* among other multilingual chat system. It gives the flexibility for user to personalize his/her short-forms for messages in English.

3.1 Related Work

The traditional text normalization strategy follows the noisy channel model (Shannon, 1948). Suppose the chat message is C and its corresponding standard form is S, the approach aims to find $\arg\max P(S|C)$ by computing $\arg\max P(C|S)$ in which $P(S)$ is usually a language model and $P(C|S)$ is an error model. The objective of using model in the chat message normalization context is to develop an appropriate error model for converting the non-standard and unconventional words found in chat messages into standard words.

$$\hat{S} = \arg\max_{S} P(S|C) = \arg\max_{S} P(C|S)P(S)$$

Recently, Aw et al. (2006) model text message normalization as translation from the texting language into the standard language. Choudhury et al. (2007) model the word-level text generation process for SMS messages, by considering graphemic/phonetic abbreviations and unintentional typos as hidden Markov model (HMM) state transitions and emissions, respectively. Cook and Stevenson (2009) expand the error model by introducing inference from different erroneous formation processes, according to the sample error distribution. Han and Baldwin (2011) use a classifier to detect ill-formed words, and generate correction candidates based on morphophonemic similarity. These models are effective on their experiments conducted, however, much works remain to be done to handle the diversity and dynamic of content and fast evolution of words used in social media and networking.

As we notice that unlike spelling errors which are made mostly unintentionally by the writers, abbreviations or slangs found in chat messages are introduced intentionally by the senders most of the time. This leads us to suggest that if facilities are given to users to define their abbreviations, the dynamic of the social content and the fast

evolution of words could be well captured and managed by the user. In this way, the normalization model could be evolved together with the social media language and chat message could also be personalized for each user dynamically and interactively.

3.2 Personalized Normalization Model

We employ a simple but effective approach for chat normalization. We express normalization using a probabilistic model as below

$$s_{best} = \arg\max_{s} P(s \mid c)$$

and define the probability using a linear combination of features

$$P(s \mid c) \propto \exp \sum_{k=1}^{m} \lambda_k h_k(s,c)$$

where $h_k(s,c)$ are two feature functions namely the log probability $P(s_{i,j} \mid c_i)$ of a short-form, c_i, being normalized to a standard form, $s_{i,j}$; and the language model log probability. λ_k are weights of the feature functions.

We define $P(s_{i,j} \mid c_i)$ as a uniform distribution computed through a set of dictionary collected from corpus, SMS messages and Internet sources. A total of 11,119 entries are collected and each entry is assigned with an initial probability, $P_s(s_{i,j} \mid c_i) = \frac{1}{|c_i|}$, where $|c_i|$ is the number of c_i entries defined in the dictionary. We adjust the probability manually for some entries that are very common and occur more than a certain threshold, t, in the NUS SMS corpus (How and Kan, 2005) with a higher weight-age, w. This model, together with the language model, forms our baseline system for chat normalization.

$$P_s(s_{i,j} \mid c_i) = \begin{cases} \frac{1}{|c_i|} + w & \text{if } |(s_{i,j}, c_i)| \geq t \\ \frac{1}{|c_i|} - \frac{w \times \left\| (s_{i,j}, c_i) \right| \geq t|}{\left\| (s_{i,j}, c_i) \right| < t|} & \text{if } |(s_{i,j}, c_i)| < t \end{cases}$$

To enable personalized real-time management of user-defined abbreviations and short-forms, we define a personalized model $P_{user\_i}(s_{i,j} \mid c_i)$ for each user based on his/her dictionary profile. Each personalized model is loaded into the memory once the user activates the normalization option. Whenever there is a change in the entry, the entry's probability will be re-distributed and updated based on the following model. This characterizes the *AsiaSpik* system which supports personalized and dynamic chat normalization.

$$P_{user\_i}(s_{i,j} \mid c_i) = \begin{cases} P_s(s_{i,j} \mid c_i) \times \frac{N}{N+M} & \text{if } c_i, s_{i,j} \in SD \\ \frac{1}{N+M} & \text{if } c_i \in SD, s_{i,j} \notin SD \\ \frac{1}{M} & \text{if } c_i \notin SD \end{cases}$$

where
SD denotes default dictionary;
N denotes the number of c_i entries in SD
M denotes the number of c_i entries in user dictionary.

The feature weights in the normalization model are optimized by minimum error rate training (Och, 2003), which searches for weights maximizing the normalization accuracy using a small development set. We use standard state-of-the-art open source tools, Moses (Koehn, 2007), to develop the system and the SRI language modeling toolkit (Stolcke,2003) to train a trigram language model on the English portion of the Europarl Corpus (Koehn, 2005).

3.3 Experiments

We conducted a small experiment using 134 chat messages sent by high school students. Out of these messages, 73 short-forms are uncommon and not found in our default dictionary. Most of these

short-forms are very irregular and hard to predict their standard forms using morphological and phonetic similarity. It is also hard to train a statistical model if training data is not available. We asked the students to define their personal abbreviations in the system and run through the system with and without the user dictionary. We asked them to give a score of *1* if the output is acceptable to them as proper English, otherwise a *0* will be given. We compared the results using both the baseline model and the model implemented using the same training data as in Aw et al. (2006).

Table 1 shows the number of accepted output between the two models. Both models show improvement with the use of user dictionary. It also shows that it is very critical to have similar training data for the targeted domain to have good normalization performance. A simple model helps if such training data is unavailable. Nevertheless, the use of a dictionary driven by the user is an alternative to improve the overall performance. One reason for the inability of both models to capture the variations fully is because many messages require some degree of rephrasing in addition to insertion and deletion to make it readable and acceptable. For example, the ideal output for "*haiz, I wanna pontang school*" is "*Sigh, I do not feel like going to school*", which may not be just a normalization problem.

Baseline Model	Baseline + User Dictionary	Aw et al. (2006)	Aw et al. (2006) + user Dictionary
40	72	17	42

Table 1. Number of Correct Normalization Output

In the examples showed in Table 2, '*din*' and '*dnr*' are normalized to '*didn't*' and '*do not reply*' based on the entries captured in the default dictionary. With the extension of normalization hypotheses in the user dictionary, the system produces the correct expansion to 'dinner'.

Chat Message	**Chat Message normalized using the Default dictionary**	**Chat Message normalized with the supplement of user dictionary**
buy *din* 4 urself.	Buy *didn't* for yourself.	Buy *dinner* for yourself.
dun cook *dnr* 4 me 2nite	Don't cook *do not reply* for me tonight	Don't cook *dinner* for me tonight
gtg *bb* ttyl ttfn	Got to go *bb ttyl ttfn*	Got to go *bye talk to you later bye bye*
I dun feel lyk riting	I don't feel *lyk riting*	I don't feel *like writing*
im gng hme 2 mug	I'm going *hme two mug*	I'm going *home to study*
msg me *wh u rch*	Message me *wh you rch*	Message me *when you reach*
so sian I dun wanna do hw now	So *sian* I don't want to do *how* now	So *bored* I don't want to do *homework* now

Table 2. Normalized chat messages
AsiaSpik Multilingual Chat

Figure 2 and Figure 3 show the personal lingo defined by two users. Note that expansions for "gtg" and "tgt" are defined differently and expanded differently for the two users. 'Me' in the message box indicates the message typed by the user while 'Expansion' is the message expanded by the system.

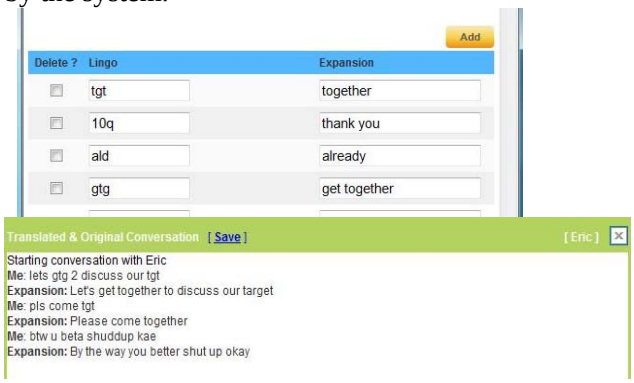

Figure 2. Short-forms defined and messages expanded for user 1

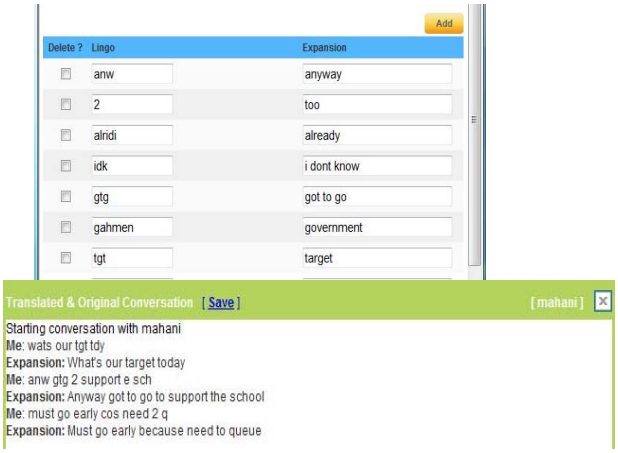

Figure 3. Short-forms defined and messages expanded for user 2

Figure 4 shows the multilingual chat exchange between a Malay language user (Mahani) and an English user (Keith). The figure shows the messages are first expanded to the correct forms before translated to the recipient language.

Figure 4. Conversion between a Malay user & an English user

4 Conclusions

AsiaSpik system provides an architecture for performing chat normalization for each user such that user can chat as usual and does not need to pay special attention to type in proper language when involving translation for multilingual chat. The system aims to overcome the limitations of normalizing social media content universally through a personalized normalization model. The proposed strategy makes user the active contributor in defining the chat language and enables the system to model the user chat language dynamically.

The normalization approach is a simple probabilistic model making use of the normalization probability defined for each short-form and the language model probability. The model can be further improved by fine-tuning the normalization probability and incorporate other feature functions. The baseline model can also be further improved with more sophisticated method without changing the architecture of the full system.

AsiaSpik is a demonstration system. We would like to expand the normalization model to include more features and support other languages such as Malay and Chinese. We would also like to further enhance the system to convert the translated English chat messages back to the social media language as defined by the user.

References

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A Phrase-based statistical model for SMS text normalization. *In Proc. Of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33-40. Sydney.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10:157–174.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *CALC '09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78, Boulder, USA.

Bo Han and Timothy Baldwin. 2011. Leixcal Normalisation of Short Text Messages: Makn Sens a #twitter. *In Proc. Of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 368-378, Portland, Oregon, USA.

Yijue How and Min-Yen Kan. 2005. Optimizing predictive text entry for short message service on mobile phones. *In Proceedings of HCII*.

Philipp Koehn &al. Moses: Open Source Toolkit for Statistical Machine Translation, *ACL 2007*, demonstration session.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X* (pp. 79{86). Phuket, Thailand.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. *In Proceedings of the*

41th Annual Meeting of the Association for Computational Linguistics, Sapporo, July.

C. Shannon. 1948. *A mathematical theory of communication*. Bell System Technical Journal 27(3): 379-423

A. Stolcke. 2003 SRILM – an Extensible Language Modeling Toolkit. *In International Conference on Spoken Language Processing*, Denver, USA.