Abstract

This work reports the benefits of Statistical Machine Translation (SMT) in template messaging domain. SMT has become an actual and practical technology due to significant increment in both the computational power and storage capacity of computers and the availability of large volumes of bilingual data. Through SMT a sentences written with misspelled words, short forms and chatting slang can be corrected. The problem of machine translation is to automatically produce a target-language (e.g., Long form English) sentence from a given source-language (e.g., Short form message) sentence. SMS Lingo is a language used by youngsters for instant messaging or for chatting on social networking websites called chatting slang. Such terms often originate with the purpose of saving keystrokes. This work presents a pioneering step in designing Bi-Gram based back-off decoder for template messages. Among the different machine translation approaches, the Probabilistic N-gram-based system has proved to be comparable with the state-of-the-art phrase-based systems. In N-gram Language Model N words are used to find the context of a word. In this work, Bi-gram LM is used. First LM is trained with bi-lingual parallel word aligned corpus to get Probability Distribution Tables (Bi-gram PDT and Uni-gram PDT). Back-off decoder along with these PDTs is then employed to translate template messages into full form text. Idea behind this work is to deal with text normalization as a translation task with the Bi-gram-based system. The main goal behind this project is to analyze the improvement in efficiency of Language Model as the size of bilingual corpus increases. This work will help researchers as a lead way in the field of N-Gram
Bi-Gram based Probabilistic Language Model for Template Messaging

Probabilistic Machine Translation and Human Computer Interaction. This work will help users to combine multiple languages with larger vocabulary and is a useful tool for small devices like mobile phones. It is also a time saver for those who cannot operate the keys efficiently. Machine learning and translation systems, dictionary and textbook preparations, patent and reference searches and various information retrieval systems are the main applications of the work.

References

- Deana Pennell, Yang Liu, "Toward text message normalization: modeling abbreviation generation", IEEE ICASSP 2011, pp. 5364-5367
- Aarthi Reddy, Richard C. Rose, "Integration of statistical models for dictation of document translations in a machine-aided human translation task", IEEE transactions on audio, speech, and language processing, vol. 18, no. 8, November 2010
- Keisuke Iwami, Yasuhisa Fuji , Kazumasa Yamamoto, Seiichi Nakagawa, "Out-Of-Vocabulary Term Detection By N-Gram Array With Distance From Continuous Syllable Recognition Results", IEEE 2010
- Daniel Jurafsky and James H. Martin, "Speech and Language Processing", 2 / 3
Bi-Gram based Probabilistic Language Model for Template Messaging


Index Terms

Computer Science  
Artificial Intelligence

Keywords

SMS Lingo  Bi-grams  Template messaging