（様式）

論文内容の要旨

氏　名　　OHNMAR HTUN

This dissertation presents a cross-lingual phonetic similarity metrics (CLPSM) that can effectively measure phonetic similarity of words which belongs to different languages. CLPSM is a methodology that integrates the Soundex algorithm, Levenshtein Edit Distance, Stochastic Edit Distance, and Bayesian Alignment Model. The dissertation consists of seven chapters.

Chapter 1 describes the background and motivation of this research. Afterwards, related works are reviewed. In the multilingual webspace, it often happens that more rich information is available in other languages than in the original language. In that case, cross-language information retrieval (CLIR) applications and machine translation (MT) systems are helpful to provide access to those information for users. And CLPSM can support CLIR by providing phonetic borrowing word pairs and can support MT systems by providing transliteration mining tools.

Chapter 2 presents the concepts of linguistic borrowing, the background theories, and the definition of terms employed in this research (e.g., phonetic similarity, linguistic borrowing, phonetic transcription, Soundex phonetic coding system, and edit distance similarity measures).

Chapter 3 describes the development of cross-language sound grouping (CLSG) and CLPSM in two approaches. Firstly, development of CLSG for Asian languages is reported in detail. Secondly, development of CLPSM based on classical edit distance (Levenshtein Edit distance) in a manual designing approach is reported. Thirdly, development of CLPSM based on different stochastic models, namely, stochastic edit

distance (EM), stochastic edit distance with noise (EMn), and Bayesian alignment with noise (BAYESn) in a learning approach are reported.

Chapter 4 presents the experimental results and evaluation in a manual designing approach by using Levenshtein Distance (LD) and Normalization of Levenshtein Distance (NS). The experiment uses the names of 92 chemical elements words in eight Asian language pairs: English-Japanese, English-Korean, English-Malay, English-Myanmar, English-Thai, English-Indonesian, English-Vietnamese, and English-Chinese. The results of two CLSG versions are compared in two metrics (LD and NS).

Chapter 5 presents the experiment results and evaluation in a learning approach by using Stochastic Edit Distance (EM), Stochastic Edit Distance with noise model (EMn) and Bayesian Alignment with noise model (BAYESn). The data for this experiment was prepared by procedures consists of segmentation and romanization of Myanmar language data, and building Myanmar-English bilingual training corpus consisting of 3,100 single word pairs and 14,891 multiple word pairs. The results of the three different models are also compared.

Chapter 6 discusses the analysis of methodologies and experimental results to evaluate the performance of CLSG and compares between manual designing and learning approach. The pros and cons of methodologies in two approaches are presented.

Finally, chapter 7 summarizes the contributions of this work first, then gives brief discussions on limitations of proposed techniques, and future tasks is presented.