Disambiguation Based on Wordnet for Transliteration of Arabic Numerals for Korean TTS

Youngim Jung[1], Aesun Yoon[2], and Hyuk-Chul Kwon[1]

[1]Pusan National University, Department of Computer Science and Engineering,
Jangjeon-dong Geumjeong-gu, 609-735 Busan, S. Korea
{acorn, hckwon}@pusan.ac.kr
[2]Pusan National University, Department of French,
Jangjeon-dong Geumjeong-gu, 609-735 Busan, S. Korea
asyoon@pusan.ac.kr

Abstract Transliteration of Arabic numerals is not easily resolved. Arabic numerals occur frequently in scientific and informative texts and deliver significant meanings. Since readings of Arabic numerals depend largely on their context, generating accurate pronunciation of Arabic numerals is one of the critical criteria in evaluating TTS systems. In this paper, (1) contextual, pattern, and arithmetic features are extracted from a transliterated corpus; (2) ambiguities of homographic classifiers are resolved based on the semantic relations in KorLex1.0 (Korean Lexico-Semantic Network); (3) a classification model for accurate and efficient transliteration of Arabic numerals is proposed in order to improve Korean TTS systems. The proposed model yields 97.3% accuracy, which is 9.5% higher than that of a customized Korean TTS system.

1 Introduction

TTS technologies for naturalness have improved dramatically and have been applied to many unlimited systems in terms of domain. However, improvement on the technique for accurate transliteration of non-alphabetic symbols such as Arabic numerals and various text symbols[1] has been relatively static.

According to the accuracy test results of 19 TTS products by Voice Information Associates, the weakest area of TTS products is in number processing in the following ambiguity-generating areas, as shown in Table 1 [10].

Table 1. TTS Accuracy Test Results Summary

Test area	Accuracy (%)
Number	**55.6**
Word of Foreign Origin	58.8
Acronym	74.1

[1] Since Arabic numerals and text symbols have graphic simplicity and deliver more precise information, the occurrence of Arabic numerals and text symbols is as high as 8.31% in Korean newspaper articles.