pISSN 2005-8063
eISSN 2586-5854
2018. 09. 30.
Vol.10 No.3
pp. 31-39

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2018.10.3.031

코퍼스 기반 프랑스어 텍스트 정규화 평가

Corpus-based evaluation of French text normalization

김 선 희*

Kim, Sunhee

Abstract

This paper aims to present a taxonomy of non-standard words (NSW) for developing a French text normalization system and to propose a method for evaluating this system based on a corpus. The proposed taxonomy of French NSWs consists of 13 categories, including 2 types of letter-based categories and 9 types of number-based categories. In order to evaluate the text normalization system, a representative test set including NSWs from various text domains, such as news, literature, non-fiction, social-networking services (SNSs), and transcriptions, is constructed, and an evaluation equation is proposed reflecting the distribution of the NSW categories of the target domain to which the system is applied. The error rate of the test set is 1.64%, while the error rate of the whole corpus is 2.08%, reflecting the NSW distribution in the corpus. The results show that the literature and SNS domains are assessed as having higher error rates compared to the test set.

Keywords: text normailization, non-standard word (NSW), French, evaluation, corpus, text domain

1. 서론

일반적으로 신문 기사와 같은 텍스트에는 일상생활에서 사용하는 단어 외에도 숫자나 약어와 같이 사전에서 그 의미나 발음을 찾을 수 없는 단어들이 많이 포함되어 있다. 예를 들어, 한국어의 경우, ‘90%’는 /구 십 퍼센트/라고 읽고, ‘10대’의 경우는 앞뒤 문맥에 따라 /십 대 (청소년)/[십때], 혹은 /(자동차) 열 대/[열때]로 읽을 수 있다. 이와 같이, 주어진 단어의 발음을 사전에서 찾을 수 없고 일반적인 발음변환규칙으로 그 발음을 추출해 낼 수 없는 단어를 비표준단어(non-standard word, 이후 NSW)라고 하고, 이러한 NSW를 일반적인 단어 혹은 표준단어(standard word)로 변환하는 과정을 텍스트 정규화(text normalization)라고 한다(Sproat *et al.*, 2001).

텍스트 정규화는 음성합성(speech synthesis, 혹은 text-to-speech) 시스템이나 음성인식(speech recognition, 혹은 speech-to-text) 시스템을 구성하는 주요 언어처리 모듈로서 시스템의 성능에 결정적인 영향을 끼친다. 음성합성 시스템의 경우에 ‘10대 청소년’을 /열 대 청소년/이라고 읽는다면, 비록 합성된 음성의 발음이 정확하고 전반적으로 사람의 음성 같이 자연스럽다고 하더라도, 사용자들에게는 그 시스템의 성능이 매우 좋지 않다고 인식될 수 있다(Ebden & Sproat, 2015; Sproat *et al.*, 2001). 음성인식 시스템의 경우에도 텍스트 정규화의 문제는 인식 성능과 밀접한 관련을 갖는다. 예를 들어, ‘3개’와 ‘세계’는 같은 발음으로 실현되는데, 음성인식 시스템에서 이 두 단어를 분화하기 위해

* 네이버(주), kim.sunhee@navercorp.com, 교신저자

Received 8 August 2018; Revised 9 September 2018; Accepted 27 September 2018

서는 기본적으로 이 두 단어가 모두 음성인식의 후보로 포함되어야 한다(Adda *et al.*, 1997; Adda-Decker, 2001; Adda-Decker *et al.*, 1998).

프랑스어는 복잡한 동사 활용(conjugation), 성(gender), 수(number)의 일치 문제로 인해 어휘의 다양성이 매우 큰 언어로 자소음소 변환(GTP, grapheme-to-phoneme)과 함께 텍스트 정규화의 문제도 매우 복잡한 양상을 보인다(Adda *et al.*, 1997; Yvon *et al.*, 1998; Kim, 2018). 다양한 종류의 규칙성과 불규칙성을 내포하고 있는 프랑스어의 텍스트 정규화 문제는 주로 규칙 기반 방법론이 이용되어 왔는데(Yvon *et al.*, 1998), 최근 크라우드 소싱(crowd sourcing)을 통하여 수집된 데이터를 토대로 통계적 번역 방법(statistical machine translation)을 규칙 기반 방식과 하이브리드로 적용한 경우 가장 좋은 성능을 보인다고 보고되었다(Schlippe *et al.*, 2013). 그러나, 프랑스어의 경우에 있어서 텍스트 정규화를 위하여 기본이 되는 NSW 분류표(taxonomy)는 제안된 적이 없어서 관련 연구 기관의 내부적인 자료에 의존한 것으로 짐작된다.

대부분의 텍스트 정규화 문제는 일반적으로 음성합성을 그 적용 영역으로 하는데, 특히 기본적인 음성합성 시스템의 적용 영역이 뉴스나 책을 읽는 것을 전제로 하고 있어서 테스트셋은 통상 뉴스 문장을 중심으로 구성되거나(Sproat *et al.*, 2001; Yvon *et al.*, 1998), 아니면 일반적인 기계학습의 경우와 같이 일정 분량의 데이터를 학습으로 사용하고 일정 분량의 데이터를 테스트에 사용하는 방식으로 따로 제약을 두지 않았다(Schlippe *et al.*, 2013).

Yvon *et al.*(1998)은 프랑스어의 텍스트 정규화를 포함한 발음열 변환에 대한 평가 방법을 제안하고 이를 토대로 프랑스, 스위스, 캐나다 등 프랑스어 사용국 8개국에서 개발한 시스템들을 평가하였는데, 평가용 테스트셋으로 프랑스 일간지 르몽드지에서 2,000개의 문장을 선정하고, 여기에 200개의 숫자어와 90개의 두문자어(acronyme)와 약어가 포함되도록 테스트셋을 구성하였다. 하지만, 영어나 중국어를 비롯한 대부분의 언어의 텍스트 정규화 연구가 NSW의 체계적인 분류에 기반하고 있는데 반하여 프랑스어의 경우는 NSW에 대한 분류체계가 제안된 적이 없어서 단순히 일정한 수의 두문자어와 약어, 그리고 숫자어를 포함한 테스트셋을 구성하였는데, 이러한 테스트셋이 적절한지에 대한 평가가 어렵다고 할 수 있다. 뿐만 아니라, 비록 테스트셋이 일정한 NSW 범주들을 포함하고 있다고 하더라도 Yvon *et al.*(1998)과 같이 뉴스 영역에서 선정된 테스트셋이 다른 텍스트 영역에 적용되는 경우에도 유효하다고 가정하는 것은 의문의 여지가 있다.

따라서, 본 논문의 목적은 두 가지이다. 먼저, 프랑스어 텍스트 정규화 모듈을 개발하고 평가하기 위하여 필수적이지만 아직까지 제안된 적이 없는 프랑스어의 NSW의 분류표를 제안한다. 그리고, 제안한 NSW 분류체계를 기반으로 하여 테스트셋을 설계하고, 적용 영역이 다른 텍스트의 특성을 반영할 수 있도록 코퍼스에서의 분포를 반영한 새로운 평가 지표를 제안하는 것을 목적으로 한다. 이러한 연구는 특히 프랑스어 음성합성 시스템에 있어서 필수적인 텍스트 정규화 모듈을 개발하고 평가하는 데 기본적인 자료가 되고, 나아가 음성인식을 비롯 한 음성언어처리 분야와 이 분야의 다른 언어들에도 적용될 수 있을 것으로 기대된다.

이후 논문 구성은 다음과 같다. 2장에서는 텍스트 정규화 관련 연구들을 소개하고, 3장에서는 프랑스어 NSW에 대한 분류체계를 제안한다. 4장에서는 3장에서 제안한 분류 체계를 바탕으로 텍스트 정규화 시스템을 평가하기 위한 테스트셋을 설계하고, 이와 함께 코퍼스를 기반으로 한 평가 방법을 제안한 다음, 제안한 방법으로 다양한 영역의 텍스트에 대한 오류율을 추정한다. 마지막으로, 5장의 논의와 결론으로 논문을 마무리한다.

2. 관련 연구

텍스트 정규화에 대한 대표적인 연구로는 Sproat *et al.*(2001)을 들 수 있는데, 영어에서의 뉴스 텍스트를 NSW 관점에서 다른 특성을 보이는 세부 영역으로 분류하고 이를 토대로 NSW 분류표(taxonomy)를 제안하였다. 이 분류표는 이후 텍스트 정규화 영역에 기본 틀로 사용되고 있다(Flint *et al.*, 2017; Zhou *et al.*, 2008). 또한, 그 이전까지는 텍스트 정규화의 문제는 대부분 개별 문제에 상응하는 자의적인 규칙들을 제시하는 방법이 대부분이었는데, n-gram, 결정트리, WFST 등의 한 통계적인 방법론을 제안하였다. 이후 이러한 방법론을 기반으로 한 범용 툴도 제안되었다(Festvox, 2000). 이 연구는 이후 다양한 인터넷 영역인 SMS(Aw *et al.*, 2006; Eisenstein *et al.*, 2013), 이메일(Moore *et al.*, 2010), 트위터(Han & Baldwin, 2011)와 같은 다른 여러 영역에서의 텍스트 정규화 연구에 많은 영향을 끼쳤다.

텍스트 정규화 시스템은 기본적으로 언어나 텍스트에 의존적인 규칙 기반 방법론(Bigi, 2011) 외에도 위에서 언급한 WFST weighted finite-state transducers)나 n-gram과 같은 통계적인 언어 모델이 제안되기도 하였고(Ebden & Sproat, 2015), 기계 번역(statistical machine translation) 방법(Schlippe *et al.*, 2013)이나 Maximum Entropy 등을 이용한 여러 통계적 방법들이 시도되었다(Sproat & Hall, 2014). 최근 DNN 방법론이 음성합성을 비롯한 여러 분야에 적용됨에 따라(Arik *et al.*, 2017; Wang *et al.*, 2017), Sproat & Jaitly(2016)와 같이 텍스트 정규화에도 DNN 방법을 적용한 연구도 보고되었다. Sproat & Jaitly(2016)는 여러 기법의 RNN 기반 텍스트 정규화 모델을 이용하여 정확도 관점에서는 의미 있는 결과를 얻기는 하였으나, 오류가 발생하는 경우 개별 오류를 따로 처리하기 위해 FST 기반의 필터를 이용할 수밖에 없다는 점을 지적하여, 궁극적으로 NSW 문제는 DNN만으로는 해결할 수 없다는 점을 시사하고 있다.

영어의 경우에 최근 연구인 Flint *et al.*(2017)은 Sproat *et al.*(2001)을 토대로 영어의 SNS 텍스트를 포함하기 위한 분류체계를 수정하여 제안하였다. 특히, 영어에 있어서도 영국 영어와 미국 영어를 파라미터를 이용하여 구분하고, 이후 추가되는 지역에 따라 파라미터를 추가할 것을 제안하고 있다. 영어와 같이

알파벳을 기반으로 한 표기 체계(writing system)를 사용하는 언어들과는 달리 독자적인 표기체계를 사용하는 언어들의 경우에는 언어와 문화적인 요인에 의해 텍스트 정규화 문제의 양상이 다양하게 나타난다. 중국어나 한국어, 일본어, 힌디어, 아랍어 등과 같이 각각 다른 독자적인 표기 체계를 갖는 경우, 현대에 와서 많은 영어 단어들이 사용됨에 따라, 특히 영어의 음역(transliteration)을 포함한 텍스트 정규화 문제에서 개별 표기체계를 영어로 변환하는 과정이 주요 부분을 차지하게 되었다. 중국어의 경우 중국어 NSW에 대한 분류 체계를 제안한 Zhou *et al.*(2008)이 있는데, 중국어 NSW의 분류 및 표준단어로의 변환 과정을 규칙 및 통계를 기반으로 제안하였다. Kim(2017)은 Zhou *et al.*(2008)이 제안한 텍스트 정규화 시스템을 본토 중국어와 대만에서 사용하는 대만 중국어에 각각 적용한 연구로, 이 두 언어는 동일한 중국어를 사용하지만 텍스트 정규화 문제와 관련하여 흥미롭게도 NSW의 범주 분포가 다르다고 보고하고 있다.

텍스트 정규화 시스템의 평가 척도로는 일반적으로 단어 오류율(WER, word error rate)이나 토큰 오류율(token error rate)을 사용하거나(Sproat *et al.*, 2001; Sproat & Jaitly, 2016), Precision/Recall을 이용하기도 한다(Zhou *et al.*, 2008). 앞서 언급한 바와 같이 음성합성에 있어 특히, 텍스트 정규화에서 오류가 발생하는 경우 그 영향이 지대하지만, 단순히 단어 오류율이나 토큰 오류율, 혹은 Precision/Recall을 지표로 사용하는 평가가 실질적으로 서비스에 적용되는 경우 그 효과를 보장할 수 없다. 그러므로 좀 더 영역의 특성을 반영한 평가를 위해 코퍼스를 반영하는 평가셋으로 평가할 것을 제안하기도 하였다(Kim, 2017). 그러나, Kim(2017)은 중국어의 코퍼스 분포를 분석하여 이를 반영하는 테스트셋을 만들고 실제 적용하는 영역을 고려하여 평가 방법을 제안하기는 하였으나, 매번 사용할 코퍼스에 대해 분석 결과를 반영한 테스트셋을 제작해야 한다는 제약이 있다. 본 연구는 프랑스어의 텍스트 정규화 시스템을 개발하기 위하여 필요한 언어 지식을 정리하여 필수적인 NSW 분류표를 제안하고, 이를 토대로 텍스트 정규화 평가를 위한 지표로 각 범주별 정확도와 코퍼스에서의 분포를 반영한 새로운 평가 지표를 제안하고자 한다.

3. 프랑스어의 NSW 분류

프랑스어의 텍스트 정규화를 위한 NSW는 Sproat *et al.*(2001)에서 제안한 분류 체계를 바탕으로 문자 범주와 숫자 범주, 그리고 기타 범주의 세 가지 범주로 나누었다. 문자 범주의 경우 다시 하위 범주인 약어와 두문자어(acronyme)로 구분된다. 숫자 범주는 하위 범주로 기수, 소수, 분수, 서수, 로마숫자, 시간 표현, 날짜 표현, 통화/비통화 단위, 기타 숫자로 구분된다. 기타 범주는 하위 범주로 기호와 위에서 언급되지 않은 나머지의 모든 경우를 포함한다.

다음 표 1은 제안한 프랑스어 NSW의 범주 및 각 범주에 대한 예시를 나타낸 표이다. 표 1의 프랑스어 NSW 분류는 Sproat *et al.*(2001)에서 제안한 분류 체계를 기본으로 하고 있으나, 문자 범주에서 하나의 단어로 읽는 경우(ASWD, read as word)를 여기에서는 두문자어(LSEQ)와 통합하였고 오타(MSPL, misspelling)의 경우는 제외하였다. 숫자 범주에 있어서도 Sproat *et al.*(2001)은 좀 더 상세하게 구분하고 있으나 본 연구에서는 Kim(2017)을 근거로 출현 빈도가 낮은 숫자 범주들을 기타 숫자들(NSCAR)로 정의하여 하나로 묶어서 처리하였다.

표 1. 프랑스어 NSW 분류
Table 1. A taxonomy of French NSWs

Tag	Description	Examples
EXPN (약어)	abbreviations	Mme / Mme = madame “Mrs”
LSEQ (두문자어)	letter sequences	ovni = Objet Volant non identifié “ufo (unidentified flying object)”
NUM (기수)	cardinal numbers	321 = trois cent vingt et un
NDEC (소수)	decimal numbers	3,21 = trois virgule vingt et un
NFRAC (분수)	fractional numbers	2/6 = deux sixièmes, deux sur six
NORD (서수)	ordinal numbers	4e= quatrième
NROM (로마숫자)	roman numbers	XXXVII = 37
NTIME (시간)	time	01 h 54 = une heure cinquante-quatre
NDATE (날짜)	date	14 juin 1967 = quatorze juin mille neuf cent soixante-sept
QMON (통화단위)	money	100 USD = cent dollars
QNMON (비통화단위)	non-money	m = mètres, A = ampères
NSCAR (기타숫자)	cardinal umber sequences	587647 = cinquante-huit soixante-seize quarante-sept
SYMB (기호)	symbols	80 % = 80% quatre-vingts pour cent
MISC (기타)	urls/emails	@gmail.com = arobase g mail point com

3.1. 문자 범주 NSW

3.1.1. 약어(EXPN)

문자 범주의 NSW는 약어(abbreviation)와 두문자어(acronym)로 다시 분류할 수 있다. 프랑스어의 약어는 일반적으로 대문자나 소문자, 기호, 혹은 문자와 기호가 혼합된 형태로 다양하게 나타나고, 보통 온점(‘.’)이 동반되기도 하지만 그렇지 않은 경우도 종종 볼 수 있다.

(1) 하나의 문자가 온점과 같이 사용되는 경우
M. = Monsieur ‘Mr.’
(2) 단어의 일부가 생략되는 경우
dict. = dictionnaire ‘사전’
(3) 여러 단어가 결합된 경우
ch. de f. = chemin de fer ‘철도’

(4) 문자와 기호가 혼합된 경우

n° = numéro ‘번호’

자주 사용되는 약어 목록은 다음 두 사이트를 이용하여 정리하였다.

(5) http://www.orthotypographie.fr/volume-I/abreviations-01.html

(6) http://www.les-abreviations.com/civilite.html

3.1.2. 두문자어(LSEQ)

두문자어(acronym)는 일반적으로 주어진 표현을 구성하는 각 단어의 첫 글자로 구성되는데, (7)과 같이 자음 문자와 모음 문자가 혼합된 경우에는 한 단어처럼 읽을 수도 있고, (8)과 같이 자음이나 모음으로만 구성된 경우에는 낱자로 읽게 된다.

(7) sida = Sindrome de l’immunodéficience aquise 'AIDS' [sida]

(8) OUA = Organisation de l’unité africaine

‘아프리카 연합 기구’ [o-y-a]

두문자어의 경우 대문자나 소문자를 구분하여 사용하는 것은 그 의미에 따라 정해진다. 두문자어의 목록은 다음 사이트를 이용하였다.

(9) http://publications.europa.eu/code/fr/fr-5000400.htm

3.2. 숫자 범주 NSW

3.2.1. 기수(NUM)

영어나 한국어의 경우에 4개 이상의 숫자가 연속되는 경우 일반적으로 반점으로 구분하는데 반하여, 프랑스어의 경우 반점을 사용하지 않고 띄어쓰기를 사용한다. 즉, (9)와 같이 한국어에서 ‘1,234,567’ / 백 이십칠만 사천오백 육십칠/은 프랑스어로는 ‘1 234 567’로 나타난다. 따라서, 띄어쓰기 없이 네 자리 이상의 숫자가 연속되는 경우, 예를 들면 ‘(en) 2018’과 같은 연도의 경우나 주소, 혹은 페이지 번호 등은 일반적인 기수가 아닌 경우에 해당하여 기타 숫자(NSCAR)게 되어 따로 처리한다.

(10) 1,234,567 → 1 234 567

= un million deux cent trente quatre mille cinq cent soixante sept

숫자의 경우에 있어 성(gender)에 대한 구분은 없는 것이 일반적이나 ‘하나’를 의미하는 ‘un’이 포함된 숫자의 경우에 후행하는 단어에 따라 /un/ 또는 /une/가 사용된다. 또한, 수에 대한 일치의 문제에 있어서도 일반적으로는 단수나 복수에 대한 구분을 하지 않으나 다수의 예외들이 존재한다. 이런 경우 개별적으로 상세하게 정리하는 것이 필요하기 때문에 본 논문에서는 논외로 한다.

3.2.2. 소수(NDEC)

한국어나 영어에서는 소수점을 온점으로 표시하는 데 반하여, 프랑스어에서는 온점 대신 반점을 사용하고 읽을 때에도 온점(‘point’)이 아닌 반점(‘virgule’)로 읽는다. 또한, 소수점 아래 숫자도 한국어와 달리 숫자열의 수에 따라 다르게 읽는다. 반점 뒤에 3개의 숫자가 오는 경우는 (11)과 같이 기수로 읽고, 4개 이상의 숫자가 오는 경우 두 개씩 기수로 읽게 된다.

(11) 4,258 = Quatre virgule deux cent cinquante-huit

(12) 7, 187369 = Sept virgule dix-huit soixante-treize soixante-neuf

(13) 5,0014 = Cinq virgule zéro zéro quatorze

3.2.3. 분수(NFRAC)

프랑스어에서 분수를 쓰는 방법에는 두 가지가 있다. 첫 번째는 (14)와 같이 ‘기수 + 서수’의 연쇄로 나타내는 방법이 있고, 두 번째는 (15)와 같이 ‘서수 + 전치사 sur + 서수’로 나타내는 방법이 있다. ‘기수 + 서수’로 나타내는 경우 분모가 4 이하일 때는 분모가 2이면 ‘demi’, 3이면 ‘tier’, 4이면 ‘quart’ 같은 독립적인 표현을 사용한다. 이때 분자가 2 이상이면 분모에 복수형 어미인 ‘s’가 붙게 된다.

(14) 7/9 = sept neuvièmes

(15) 7/9 = sept sur neuf

3.2.4. 서수(NORD)

프랑스어의 서수는 ‘기수 + 서수 접미사’로 나타내는데 문법적인 성과 수가 모두 일치되어야 하므로 단수 접미사로는 ‘e/ème/ er/ère,’ 복수 접미사로는 ‘es/èmes/ ers/ ères’를 붙여서 나타낸다. 이때, 이러한 접미사는 일반 글자로 나타내기도 하고 윗첨자로 나타내기도 한다.

(16) 1ère = $1^{ère}$ = 1^{re} = première

3.2.5. 로마 숫자(NROM)

프랑스어에서 로마 숫자는 기수와 서수로 모두 사용된다. 기수인지 서수인지 구분하기 위해서는 아라비아 숫자와 마찬가지로 서수의 경우에는 기수에 서수 접미사를 붙여 나타낸다. 아래 (17)은 기수의 예이고 (18)은 서수의 예이다.

(17) XXV = 25 = vingt-cinq

(18) XVIIe siècle = 17e siècle = dix-septième siècle

3.2.6. 시간 표현(NTIME)

프랑스어 시간 표현은 h (heure) ‘시’, m/min (minute) ‘분’, s/sec (seconde) ‘초’로 나타내거나(19), 콜론을 사용하기도 한다(20). 두 경우 모두 각 단위에서 띄어쓰기를 할 수도 있고, 띄어쓰기 없이 사용할 수도 있다.

(19) 6 h 34 min 12 s = 6h 34min 12s ‘six heures trente-quatre minuites douze secondes ‘6시 34분 12초’

(20) 6 : 34 : 12 = 6:34:12 ‘6시 34분 12초’

오전이나 오후를 표시하기 위해 영어와 같이 am (a.m.) or pm (p.m.)을 사용하는데, 이때 am은 ‘matin’(오전)을 사용하고, pm은 오후 5시나 6시를 경계로 ‘après-midi’(오후)와 ‘soir’(저녁)를 구분하여 사용한다.

3.2.7. 날짜 표현(NDATE)

날짜는 보통 월-일-연도의 순서로 숫자와 단어, 혹은 숫자와 약어로 나타내거나, 온점(‘.’), 슬래시(‘/’), 하이픈(‘-’)을 사용하여 나타낼 수 있다.

(21) 25 décembre 2001 = 25 déc 2001
vignt-cinq décembre deux mille un

(22) 10.01.2014 = 10/01/2014 = 10-01-2014
dix janvier deux mille quartoze

3.2.8. 단위 표현(QMON/QNMON)

단위 표현은 통화 단위와 통화 단위를 제외한 비통화단위로 나눌 수 있다. 통화 단위에는 미국, 프랑스, 영국 등 국가별 통화 단위와 통화 단위와 함께 사용되는 숫자들이 있다. 통화 단위의 경우 숫자의 앞이나 뒤에 나타날 수 있는데, 어떤 경우이건 읽는 방법에 있어 동일하게 ‘숫자(기수) + 단위’로 읽는다(23).

(23) 1,39 $ US = un dollar trente-neuf (centimes)

프랑스의 통화는 유로로 기호 '€'나 약어인 'EUR'가 사용되고, 숫자와 띄어 쓰거나 바로 붙여서 사용된다('3 €' 혹은 '3€'). 영어와 달리 천 단위의 경우 띄어쓰기를 하거나 온점('.')이 사용된다(24). 기본적으로 기수가 사용되고, 반점(',')은 유로로 읽고, 반점 이하의 숫자는 쌍띰 'centime(s)'으로 읽는데, 유로와 쌍띰 사이에 접속사 'et'를 추가한다.

(24) 250 000 € = 250.000€
(25) NN,NN € = NN + euro(s) + et + NN + centime(s)
(26) 2,50€ = deux euros et cinquante (centimes)

유로 외에 프랑스에서 많이 사용되는 통화로 '달러($)'와 파운드가 있는데 각각 'dollars'와 'livres sterling'로 읽는다. '달러'의 경우 사용하는 국가들이 많아 필요한 경우 각 국가를 표시하여 사용한다.

(27) 570 $ = cinq cents soixante-dix dollars
(28) 388 £ = trois cents quatre-vingt-huit livres sterling
(29) US$ / $ = dollar americain, CAN / $CA / $C = dollar canadien

비통화 단위는 통화 단위를 제외한 모든 단위들을 포함하는 것으로 통화 단위와 마찬가지로 ‘숫자(기수) + 단위’로 읽게 된다. (30)은 비통화단위인 도량형의 예이다.

(30) 비통화 단위
a. m = mètres
b. kg = kilogrammes
c. s = secondes
d. A = ampères

3.2.9. 기타 숫자(NSCAR)

위에서 언급한 숫자들을 제외하고도, 전화번호, 주소, 시리얼 번호, 여권 번호, 은행 계좌 번호 등 많은 종류의 연속된 숫자, 혹은 문자와 숫자가 결합된 형태가 있다. 프랑스어에서는 숫자의 연쇄를 읽을 때 기본적으로 한 자리씩이나 두 자리씩 읽고, 최대 세 자리까지 읽는 것이 일반적이다.

전화번호는 보통 두 글자씩 하이픈(‘-’)이나 온점(‘.’) 혹은 띄어쓰기 단위로 나누어 쓰고, 읽을 때에는 기수로 읽는다(31). 주소의 경우 우편번호는 2자리+3자리로 읽고(32), 그 외 길 번호의 경우 두 자리씩 끊어 읽는다(33). 시리얼 번호와 같은 문자와 숫자가 결합된 연쇄의 경우 한 자리씩, 두 자리씩, 혹은 세 자리씩 읽을 수 있다(34).

(31) 40 51 37 65 = quarante cinquant et un trente-sept soixante-cinq

(32) 75015 → 75 015 = soixante-quinze, zéro quinze

(33) 2836, (rue Chambord) → 28 36 = vignt-huit trente-six

(34) W3456B867
→ a. W 34 56 B 8 6 7 Double vé trente-quatre cinquante-six bé huit six sept
→ b. W 3 4 5 6 B 8 6 7 Double vé trois quatre cinq six bé huit cent soixante-sept

3.3. 기타 범주 NSW

3.3.1. 기호(SYMB)

기호들은 텍스트에서 다른 문자들과 나타나거나 혹은 숫자들과 나타나기도 한다. 퍼센트 기호('%')는 선행하는 숫자와 나타나고 이때 숫자는 기수로 읽는다. 수학 기호들과 각 기호의 이름 목록은 다음 사이트 (36)을 참조하였다.

(35) 80 % = 80% = quatre-vingts pour cent

(36) http://fr.wikipedia.org/wiki/Table_des_symboles_math%C3%A9matiques

3.3.2. 기타(MISC)

위에서 언급한 문자나 숫자 이외에도 URL과 이메일 읽기 등이 있다. 이런 경우도 기본적으로 낱자 읽기, 단어 읽기, 기수 숫자 읽기를 기반으로 구성된다.

(37) http://www.google.fr = H T T P deux points double slash W W W point google point F R

(38) Isabelle-1988@gmail.com = Isabelle tiret mille neuf cent quatre-vingt-huit arobase g mail point com

4. 코퍼스 기반 텍스트 정규화 시스템 평가

4.1. 평가 대상 텍스트 정규화 시스템

본 논문에서 평가 대상이 되는 텍스트 정규화 시스템은 Adda *et al.*(1997)과 유사한 전통적인 규칙 기반 시스템으로 NSW 검출, 모호성 처리, 그리고 표준 단어로의 변환의 3개의 하위 모듈로 구성되고, 이 하위 모듈은 순차적으로 적용된다. 각각의 개의 하위 모듈은 언어학적 지식을 기반으로 한 개별 규칙들로 구성되는데, 각각의 규칙은 기본적으로 언어학적 지식을 기반으로 한 정규표현식(regular expressions)과 문맥 의존 다시쓰기 규칙(context-sensitive rewrite rules)을 이용하여 작성된다. 즉, 하나의 규칙은 그 적용되는 조건과 해당 조건에서의 변환으로 표현된다. 예를 들면, 영어에서 'st.'는 경우에 따라 'street' 혹은 'saint'로 변환이 가능한데, 'street'로 변환되는 경우는 앞에 숫자가 나타나는 것을 조건으로 하고, 'saint'로 변환되는 경우는 뒤에 문자가 나타나는 것을 조건으로 하게 되고, 각 규칙은 이러한 조건과 함께 변환 결과를 쌍으로 정의하게 된다.

4.2. 평가 대상 코퍼스

평가를 위해 프랑스어 뉴스 텍스트를 비롯하여 문학작품, 논픽션 텍스트, 소셜 미디어 텍스트(SNS), 그리고 음성 데이터 전사 텍스트를 수집하였다. 코퍼스 전체 크기는 654 MB로 총 5,772,790문장으로 구성되었고, 총 토큰 수는 130,516,738개이며, 여기에는 1,574,545개의 NSW가 포함되어 있다. 다음 표 2는 코퍼스를 구성하고 있는 개별 영역에 대한 통계이다. 그림 1은 3장에서 제안한 NSW의 분류에 따라 전체 코퍼스에 대한 각 NSW 범주의 분포를 나타내었다. 전체 NSW 가운데 문자 범주 NSW인 약어는 13%, 두문문자는 3.3%의 비율로 출현하였고, 숫자 범주 NSW가 그 외 대다수를 차지하는데 기수, 날짜 표현, 단위/도량형 표현 등의 순서의 비율로 출현하는 것을 볼 수 있다.

표 2. 코퍼스 구성
Table 2. Corpus composition

범주	사이즈(MB)	문장 수	토큰 수	NSW 토큰 수
뉴스	325	2,740,370	64,052,321	883,768
문학	162	1,729,895	34,053,220	112,929
논픽션	153	1,138,668	29,141,460	484,758
SNS	15	163,139	3,226,741	91,849
전사	0.2	719	42,996	1,241
전체	654	5,772,791	130,516,738	1,574,545

표 3은 각 영역의 텍스트에서 전체 NSW에 대한 개별 범주 NSW의 분포율을 나타낸 표이다. 각 영역별로 높은 분포를 보이는 상위 세 개의 범주를 굵은 글씨로 표시하였는데, 전반적으로 문자 범주와 기수, 날짜, 그리고 기호 범주가 높은 분포를 보이는 것을 보이고, 테스트셋을 추출한 뉴스 영역이 전체 영역과 유사한 양상을 보이는 것을 볼 수 있다.

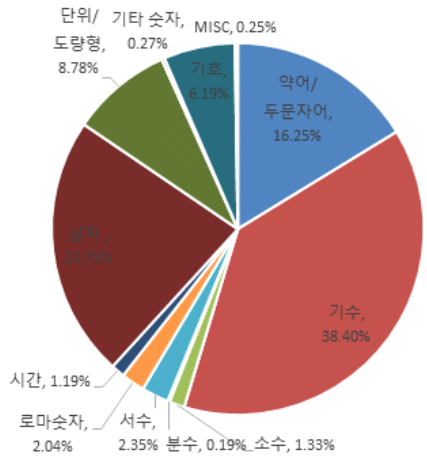

그림 1. NSW 범주별 분포
Figure 1. Distribution of NSW categories

표 3. 코퍼스의 각 영역별 NSW 비율
Table 3. Ratio of each NSW category in the corpus

범주	뉴스 (%)	문학 (%)	논픽션 (%)	SNS (%)	전사 (%)	전체 (%)
약어/ 두문문자	**14.45**	**41.55**	**12.73**	**21.25**	4.83	**16.25**
기수	**43.24**	**12.99**	**37.83**	**25.79**	**63.01**	**38.40**
소수	1.81	0.17	0.91	0.35	0.00	1.33
분수	0.10	0.10	0.19	1.14	0.16	0.19
서수	3.05	2.35	1.09	2.33	0.00	2.35
로마숫자	0.97	4.08	3.82	0.43	0.64	2.04
시간	1.66	0.59	0.38	1.79	0.08	1.19
날짜	**20.81**	6.52	**33.94**	2.50	**6.77**	**22.75**
단위/ 도량형	11.76	1.85	5.37	6.77	2.98	8.78
기타 숫자	0.19	0.14	0.45	0.34	0.00	0.27
기호	1.75	**29.59**	3.03	**36.57**	**21.43**	6.19
MISC	0.21	0.07	0.27	0.75	0.08	0.25

4.3. 테스트셋 구성 및 평가 결과

테스트셋으로는 3장에서 제안한 NSW 분류 체계에 따라 각각의 범주에 해당하는 예가 1개 이상, 70개 전후로 포함된 문장을 뉴스 텍스트를 중심으로 하여 총 1,000문장을 수집하였다. 다음 표 4는 범주별 문장과 토큰 수 및 NSW 토큰 수를 나타낸다. 수집된 1,000문장은 NSW 2,141개를 포함하여 총 18,928개의 토큰으로 구성되었다.

표 4. 테스트셋 구성
Table 4. Statistics of the test set

범주	문장 수	토큰 수	NSW 토큰 수
약어/두문문자	80	1,494	252
기수	80	1,532	326
소수	76	1,578	161
분수	44	901	49
서수	70	1,343	75
로마숫자	70	1,322	100
시간	70	1,248	128
날짜	70	1,437	190
단위/도량형	93	1,648	174
기타 숫자	207	3,887	242
기호	70	1,366	372
기타	70	1,172	72
Total	1,000	18,928	2,141

테스트셋으로 선정한 1,000개의 문장에 텍스트 정규화 시스템을 사용하여 표준단어로 변환한 다음 프랑스어 전문가 검토 후 정답 문장을 만들었다. 표 5는 정답 문장과 시스템 출력 문장을 비교한 결과이다. 비교 결과 1,000문장 가운데 오류를 포함한 NSW는 303개로 전체 NSW 2,141개와 비교했을 때 14.2%의 오류율을 보였다(NSW 오류율). 전체 토큰 18,298개와 비교하면 1.6%의 오류율을 보인다(토큰 오류율). NSW 오류를 포함한 문장은 총 248문장으로 문장 오류율은 24.8%가 된다.

표 5. 테스트셋 평가 결과
Table 5. Evaluation result of the test set

범주	오류 문장수	문장 오류율 (%)	오류 NSW 수	오류 NSW 수 /NSW 수 (%)	오류 NSW 수 /총 토큰 수 (%)
약어/ 두문문자	34	42.5	45	17.9	3.0
기수	20	25.0	25	7.7	1.6
소수	19	25.0	25	15.5	1.6
분수	12	27.3	16	32.7	1.8
서수	16	22.9	19	25.3	1.4
로마숫자	12	17.1	13	13.0	1.0
시간	26	37.1	35	27.3	2.8
날짜	14	20.0	16	8.4	1.1
단위/ 도량형	40	43.0	44	25.3	2.7
기타 숫자	39	18.8	45	18.6	1.2
기호	15	21.4	19	5.1	1.4
기타	1	1.4	1	1.4	0.1
Total/ average	248	24.8	303	14.2	1.6

4.4. 코퍼스 기반 텍스트 정규화 평가 결과

표 5에서 살펴 본 테스트셋 평가 결과는 테스트셋이 실제 시스템이 적용될 영역의 서브 셋(subset)이라는 가정을 바탕으로 한 것으로, 적용될 영역의 각 NSW 범주의 분포를 반영하고 있는 경우에만 본 테스트셋을 이용한 결과가 유효하다고 할 수 있다. 즉, 전체 평가 대상 코퍼스를 테스트하기가 현실적으로 불가능하므로 특정 영역을 기반으로 테스트셋 구성 후 다른 역영의 성능을 코퍼스 내 NSW 분포를 기반으로 유추하는 방법을 제시한다. 이와 같은 테스트셋을 실제 적용 영역에 반영하기 위해서는 대상이 되는 코퍼스에서 각 NSW 범주 분포를 고려한 평가가 필요하다. 따라서, 본 논문에서는 평가 대상이 되는 영역의 NSW 분포를 반영하기 위하여 아래와 같은 수식을 이용하였다.

(39) 타겟 영역에서의 범주별 오류율
= 테스트셋 오류율 / (테스트셋 영역(뉴스)의 각 범주별 NSW 분포율) × (타겟 영역의 각 범주별 NSW 분포율)

표 6은 전체 코퍼스에 표 3에서 제시한 NSW 분포 비율을 반영하여 새로운 추정 오류율을 식 (39)를 적용하여 도출한 결과이다. 뉴스 영역을 기준으로 작성된 테스트셋의 오류율이 1.64%이었으나 전체 코퍼스에 대한 오류율은 이보다 조금 높은 2.08%로 추정되었다. 특히, 분수, 로마숫자, 기호 범주의 경우에 실제 코퍼스에서 NSW의 분포율이 테스트셋보다 높은 경향이 오류율에 반영됨을 볼 수 있다.

표 6. 코퍼스 반영 오류율 추정
Table 6. Corpus-based error estimation

범주	테스트셋 오류율 (%)	테스트셋 NSW 분포율 (%)	전체 도메인 NSW 분포율 (%)	추정 오류율 (%)
약어/ 두문문자	3.00	14.45	16.20	3.36
기수	1.60	43.24	38.40	1.42
소수	1.60	1.81	1.30	1.15
분수	1.80	0.10	0.20	3.60
서수	1.40	3.05	2.40	1.10
로마숫자	1.00	0.97	2.00	2.06
시간	2.80	1.66	1.20	2.02
날짜	1.10	20.81	22.70	1.20
단위/ 도량형	2.70	11.76	8.80	2.02
기타 숫자	1.20	0.19	0.30	1.89
기호	1.40	1.75	6.20	4.96
MISC	0.10	0.21	0.30	0.14
Average/ total	1.64	100.00	100.00	2.08

동일한 방법으로 식 (39)를 적용하여 표 3에서 제시한 각 영역의 NSW 분포 비율을 반영한 각 영역별 오류율을 표 7에 나타내었다. 전체적으로 모든 텍스트에서 오류율이 테스트셋의 오류율보다는 높게 나타났는데, 이 가운데서는 문학 텍스트의 경우 3.56%, SNS의 텍스트의 경우는 5.35%로 두드러진 차이를 보이는 것을 볼 수 있었다. 범주별로 볼 때에는 약어/두문자어, 분수, 기호 등의 범주가 전반적으로 테스트셋보다 높은 비율로 나타났고, 문학 텍스트에서는 약어/두문문자, 로마숫자와 기호가, SNS 텍스트에서는 분수와 기호의 범주가 높은 비율로 나타나고 이에 따른 오류율도 높게 나타나는 것을 볼 수 있다.

표 7. 개별 영역에 대한 코퍼스 반영 오류율
Table 7. Corpus-based error estimation in each doamin

범주	테스트셋 (%)	문학 (%)	논픽션 (%)	SNS (%)	전사 (%)	전체 (%)	Average (%)
약어/두문문자	3.0	8.6	2.64	4.41	1.00	3.36	3.84
기수	1.6	0.5	1.40	0.95	2.33	1.42	1.36
소수	1.6	0.2	0.80	0.31	0.00	1.15	0.67
분수	1.8	1.8	3.42	20.52	2.88	3.60	5.67
서수	1.4	1.1	0.50	1.07	0.00	1.10	0.86
로마숫자	1.0	4.2	3.94	0.44	0.66	2.06	2.05
시간	2.8	1.0	0.64	3.02	0.13	2.02	1.60
날짜	1.1	0.3	1.79	0.13	0.36	1.20	0.82
단위/도량형	2.7	0.4	1.23	1.55	0.68	2.02	1.43
기타 숫자	1.2	0.9	2.84	2.15	0.00	1.89	1.49
기호	1.4	23.7	2.42	29.26	17.14	4.96	13.14
MISC	0.1	0.0	0.13	0.36	0.04	0.14	0.13
Average	1.64	3.56	1.81	5.35	2.10	2.08	2.76

5. 논의 및 결론

본 논문은 프랑스어의 텍스트 정규화 시스템을 개발하기 위하여 필수적인 NSW 분류표를 제안하고, 이를 토대로 텍스트 정규화 평가를 위한 지표로 각 범주별 정확도와 코퍼스에서 NSW의 분포를 반영한 새로운 평가 지표를 제안하였다. 프랑스어의 텍스트 정규화를 위한 NSW의 분류표는 Sproat *et al.*(2001)을 기반으로 제안한 것으로 문자 범주 2가지, 숫자 범주 9가지, 그리고 기타 범주 2가지, 총 13개의 범주로 분류하였다.

프랑스어의 텍스트 정규화 문제를 다룬 Adda *et al.*(1997)은 NSW를 분류하기보다 변환 과정을 기준으로 체계적인 변환(systematic conversion)과 비체계적인 변환(non-systematic conversion)으로 나누어 분류하는데 반해, 본 연구는 기본적으로 Sproat *et al.*(2001)의 영어에 대한 제안을 채용하여 프랑스어에 대하여 기본적인 사항들과 실질적으로 텍스트 정규화 시스템을 개발하는 데 필요한 정보를 포함하여 정의하였다. 이러한 분류표는 프랑스어 합성이나 인식 시스템을 위한 텍스트 정규화 문제와 관련된 언어 정보로서 이를 기반으로 시스템을 개발하는 데 실질적인 도움이 될 것으로 생각한다. 뿐만 아니라, 이러한 분류는 텍스트 정규화에 대한 문제를 접근할 때 다른 언어들과 비교를 좀 더 용이하게 해 줄 수 있을 것으로 보인다.

주어진 텍스트 정규화 시스템의 평가를 위하여 뉴스 텍스트, 문학작품, 논픽션 텍스트, SNS 텍스트, 음성데이터 전사 텍스트로 구성된 코퍼스를 수집하였다. 코퍼스의 전체 크기는 654 MB로 총 5,772,790문장으로 1,574,545개의 NSW가 포함되어 있다. 테스트셋으로는 각 문장이 최소한 한 개 이상의 NSW를 포함하고, 각 범주에 70개 전후의 예들이 포함되도록 총 1,000문장을 선정하였다. 선정된 1,000문장을 텍스트 정규화 시스템을 사용하여 표준단어로 변환한 다음 프랑스어 전문가가 검토하여 정답 문장을 만들어 시스템 결과와 비교하였다. 이와 같이 1,000개의 문장, 혹은 20,000여개의 단어를 이용하여 테스트셋을 구성하는 방식은 이전 연구들과 유사한 방식이나(Adda *et al.*, 1997; Yvon *et al.*, 1998), 테스트셋이 실제로 시스템이 적용될 영역의 서브 셋(subset)이라는 가정을 전제로 하여 실제 적용 영역에 반영될 수 있도록 각 NSW의 비율을 반영하여 오류율을 추정하는 방법을 제안하였다.

제안한 방법에 따라 오류율을 추정한 결과, 전반적으로 모든 텍스트에서 오류율이 테스트셋의 오류율인 1.64%보다는 높게 추정되었고, 전체 코퍼스의 NSW 분포율을 반영한 시스템의 예상 오류율은 테스트셋 결과보다 약간 높은 2.08%의 오류율로 추정되었다. 텍스트 영역 가운데 특히 문학 텍스트와 SNS의 텍스트에서 각각 3.56%와 5.35%로 높은 오류율을 보였는데, 이 두 영역의 텍스트의 경우는 특히 테스트셋을 추출한 뉴스 영역과는 다른 특성으로 인한 것으로 보인다. 또, NSW 범주별로 볼 때에는 약어/두문자어, 분수, 기호 등의 범주에서 상대적으로 오류율이 높게 나타났는데, 이러한 범주의 경우에 좀 더 분류표를 세분화하여 관련 현상을 정리하는 것이 필요할 것 같다. 특히 SNS 텍스트의 경우에 있어서는 신조어와 함께 여러 새로운 기호들에 대한 문제들이 이후에도 증가될 것으로 예상된다.

근래에 음성합성 분야에서 제안된 엔드투엔드 시스템의 경우(Arik *et al.*, 2017; Wang *et al.*, 2017)에 더 이상 이전과 같이 전문적인 지식을 토대로 한 언어모듈이 필요하지 않게 되었다는 주장도 있으나(최연주 외, 2018), 다른 언어처리 모듈들과는 달리 텍스트 정규화 모듈은 여전히 필수적인 모듈로 남게 되어, 그 어느 때보다 그 성능이 전체 시스템에 지대한 영향을 끼치게 되었다고 할 수 있다. 본 논문에서 제안한 코퍼스 기반 텍스트 정규화 평가 방법은 주어진 텍스트 정규화 시스템의 성능을 적절하게 평가하여 시스템의 성능 개선에 기여할 수 있고, 나아가 프랑스어에 국한되지 않은 일반적인 방법론으로 다른 언어들에도 적용될 수 있을 것으로 기대해 본다.

참고문헌

Adda, G., Adda-Decker, M., Gauvain, J. L., & Lamel, L. (1997). Text normalization and speech recognition in French. *Fifth European Conference on Speech Communication and Technology*.

Adda-Decker, M. (2001). Towards multilingual interoperability in automatic speech recognition. *Speech Communication*, 35(1-2), 5-20.

Adda-Decker, M., Adda, G., Gauvain, J. L., & Lamel, L. (1998). On the use of speech and text corpora for speech recognition in French. *First International Conference on Language Resources and Evaluation*. Granada, Spain.

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. Retrieved https://arxiv.org/abs/1702.07825 arXiv preprint arXiv: 1702.07825.on September 27, 2018.

Aw, A., Zhang, M., Xiao, J., & Su, J. (2006). A phrase-based statistical model for SMS text normalization. *Proceedings of the*

COLING/ACL (pp. 33-40). Sydney, Austrailia.

Bigi, B. (2011). A multilingual text normalization approach. *Language and Technology Conference* (pp. 515-526). Cham, Switzerland.

Choi, Y., Jung, Y., Kim, Y., Suh, Y., & Kim, H. (2018). An end-to-end synthesis method for Korean text-to-speech systems. *Phonetics and Speech Sciences,* 10(1), 39-48. (최연주·정영문· 김영관·서영주·김회린 (2018). 한국어 text-to-speech(TTS) 시스템을 위한 엔드투엔드 합성 방식 연구. *말소리와 음성과학,* 10(1), 39-48.)

Ebden, P., & Sproat, R. (2015). The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3), 333-353.

Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language technologies* (pp. 359-369).

Festvox. (2000) Retrieved from http://festvox.org/nsw on September 27, 2018.

Flint, E., Ford, E., Thomas, O., Caines, A., & Buttery, P. (2017). A text normalisation system for non-standard English words. *Proceedings of the 3rd Workshop on Noisy User-Generated Text* (pp. 107-115).

Han, B., & Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 368-378).

Kim, S. (2017). Corpus-based evaluation of Chinese text normalization. *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)* (pp. 1-4). Seoul, Korea.

Kim, S. (2018). A knowledge-based pronunciation generation system for French. *Phonetics and Speech Sciences*, 10(1), 49-55. (김선희 (2018). 지식 기반 프랑스어 발음열 생성 시스템. *말소리와 음성과학,* 10(1), 49-55.)

Moore, S., Buchholz, S., & Korhonen, A. (2010). Annotating the Enron Email Corpus with Number Senses. *Seventh International Conference on Language Resources and Evaluation.*

Schlippe, T., Zhu, C., Lemcke, D., & Schultz, T. (2013). Statistical machine translation based text normalization with crowdsourcing. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8406-8410). Vancouver, Canada.

Sproat, R., & Hall, K. (2014). Applications of maximum entropy rankers to problems in spoken language processing. *Fifteenth Annual Conference of the International Speech Communication Association.*

Sproat, R., & Jaitly, N. (2016). RNN Approaches to text normalization: A challenge. Retrieved https://arxiv.org/abs/1611.00068 on September 27, 2018

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. *Computer Speech & Language*, 15(3), 287-333.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). TACOTRON: TOWARDS END-TO-END SPEECH SYNTacotron: Towards end-to-end speech synthesis. Retrieved https://arxiv.org/abs/1703.10135 on September 27, 2018.

Yvon, F., de Mareüil, P. B., d'Alessandro, C., Aubergé, V., Bagein, M., Bailly, G., Béchet, F., Foukia, S., Goldman, J. F., Keller, E., O'Shaughnessy, D., Pagel, V., Sannier, F., Véronis, J., Zellner, B.. (1998). Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French. *Computer Speech & Language*, 12(4), 393-410.

Zhou, T., Dong, Y., Huang, D., Liu, W., & Wang, H. (2008). A three-stage text normalization strategy for Mandarin text-to-speech systems. *2008 6th International Symposium on Chinese Spoken Language Processing* (pp. 1-4). Kunming, China.

• **김선희 (Kim, Soonee)** 교신저자
네이버(주) 수석연구원
경기도 성남시 불정로 6 그린팩토리
Tel: 031-784-3307 Fax: 02-784-1000
Email: kim.sunhee@navercorp.com
관심분야: 음성학, 음운론, 음성언어처리