단어의 음성학적 특징을 이용한 한국어 기계 번역 데이터 세트 구축 방안

장칭하오○,양홍진,김세린,권혁철
부산대학교 정보컴퓨터공학부
zhangqinghao@pusan.ac.kr, asyhj@pusan.ac.kr, rlagofls12@pusan.ac.kr, hckwon@pusan.ac.kr

Proposed Methodology for Building Korean Machine Translation Data sets Considering Phonetic Features

Zhang Qinghao, Yang Hongjian, Serin Kim, Hyuk-Chul Kwon
Department of Information Convergence Engineering, Pusan National University

요 약

한국어에서 한자어와 외래어가 차지하는 비중은 매우 높다. 일상어의 경우 한자어와 외래어의 비중이 약 53%, 전문어의 경우 약 92%에 달한다. 한자어나 외래어는 중국이나 다른 나라로부터 영향을 받아 한국에서 쓰이는 단어들이다. 한국어에서 사용되는 한자어와 외래어의 한글 표기과 원어 표기를 발음해보면, 발음이 상당히 유사하다는 것을 알 수 있다. 한자어인 도서관(图书馆)을 중국어로 발음해보면 tʰu.ʂu.kwan'로 해당 단어에 대한 한국 사람의 발음과 상당히 유사하다. 본 논문에서는 Source Length, Source IPA Length, Target Length, Target IPA Length, IPA Distance 등 총 5가지의 음성학적 특징을 고려한 한국어-중국어 한국어-영어 단어 기계번역 데이터 세트를 구축하고자 한다.

주제어: 한국어 정보 처리(Korean information processing), 국제음성문자(International Phonetic Alphabet, IPA), 한국어 기계번역 데이터(Korean Machine Translation Data)

1. 서론

국제음성문자(International Phonetic Alphabet, IPA)는 [1]라틴 문자를 기반으로 한 음성 표기법 체계이다. 19세기 말 국제 음성협회가 음성을 문자 형태로 표준적으로 표현하기 위해 고안하였다. 방글라데시(Bangladesh) 언어인 Bangla 기계번역 관련 논문에서는 미등록어(Unknown words)를 번역하기 위해 이 국제음성문자 표기법을 기계번역모델에서 사용하였다. [2] 그리고 인도 남부에 거주하는 드라비다인들이 사용하는 언어인 드라비다어 (Dravidian Language) 기계 번역 관련 논문에서도 이 국제음성문자 표기법을 사용하였다.[3] 이 연구들에 따르면 국제음성문자 표기법을 사용할 경우 기계번역모델을 통해 어휘 및 문장 데이터가 부족한 언어에서 다른 언어로 번역하고자 할 때 이전보다 더 나은 성능을 보이는 것으로 나타났다. 최근에는 레벤슈타인 거리 알고리즘(Levenshtein distance algorithm)을 사용하여 국제음성문자 간의 유사성을 수치적으로 표현할 수 있도록 한 음성 편집 거리(Phonetic Edit Distance, PED) 방법론이 제시되었다. [4]

본 논문은 중국어, 한국어, 영어 이 세 언어 간에 발음의 유사성을 지니는 단어들을 선별하고 각 언어의 국제음성문자 변환도구[5][6][7]를 활용해 한국어-중국어, 한국어-영어의 국제 음성 문자 데이터 세트를 구축하고자 한다.

2. 본론

2.1 한국어 구성 분석

표 1에 제시된 우리말 샘의 한국말 표제어 통계를 보면, 한국말 표제어는 고유어, 한자어, 외래어가 혼합하여 구성된다. 한국의 일상어에서 가장 많이 차지하는 표제어는 고유어로 전체 일상어 중 약 48%에 달한다. 일상어 중 한자어가 차지하는 비중이 약 33%로 두 번째로

높다. 일상어에 비해 전문어에서는 한자어가 차지하는 비중이 더욱 높은데 그 비중이 약 60프로에 달한다. 통계를 보면 알 수 있듯이 한국의 일상어와 전문어에서 한자어가 차지하는 비중이 매우 높으며, 특히 전문어의 경우 고유어보다도 한자어와 외래어의 구성 비율이 약 92%에 달한다.

표 1. 한국 일상어와 전문어의 표제어 통계

원어	일상어		전문어	
	표제어 수	비율(%)	표제어 수	비율(%)
고유어	231,718	48.012	22,907	8.473
한자어	160,660	33.289	160,949	59.535
외래어	5,834	1.209	45,306	16.759
한자어+외래어	2,964	0.614	11,467	4.242
한자어+고유어	79,033	16.376	27,990	10.354
외래어+고유어	1,161	0.241	1,249	0.462
한자어+외래어+고유어	1,252	0.259	475	0.176
합계	**482,622**	**100**	**270,343**	**100**

한국어의 표제어 중 한자어가 차지하는 비중이 높은 만큼, 한국에서 사용하는 한자어의 한국 발음과 해당 단어의 중국 발음은 높은 유사성을 가진다. 표 2에 제시된 예시에서 괄호 안에 나타낸 각 단어의 발음을 보면 두 단어 간의 발음이 매우 유사하다는 것을 확인할 수 있다.

표 2. 한국의 한자어의 한글 및 한자 표기 예시

한글 표기	한자 표기
도서관	图书馆(tʰu.ʂu.kwan)
과학기술	科学技术(kʰɤ.ɕɥœ.tɕi.ʈʂu)
통풍치료제	痛风治疗剂(tʰʊŋ.fɤŋ.ʈʂɚ.ljɑʊ.tɕi)
혈장	血浆(ɕjɛㄱ.tɕjɑŋ)
수혈	输血(ʂu.ɕjɛㄱ)
대사증후군	代谢综合症(taɪ.ɕjɛ.tsʊŋ.xɤ.ʈʂɤŋ)
지방간	脂肪肝(ʈʂɚ.fɑŋ.kan)
중증환자	重症患者(ʈʂʊŋ.ʈʂɤŋ.xwan.ʈʂɤㄱ)

한국어에서 한자어를 제외한 외래어는 대부분 근래에 출현한 것으로 외래어의 약 90%가 영어에서 차용되었다. 일부 서양의 외래어는 임진왜란 때 일본어를 통해 간접적으로 채택되기도 하였다. 부산대 인공지능 연구소의 '한국어 외래어 데이터 세트'에는 이러한 외래어 사례가 4만건 이상 축적되어 있다.[8] 표 3에 8가지의 외래어가 한글과 영어로 표기되어 있다. 각 단어의 괄호 안의 발음을 읽어보면 해당 단어에 대한 두 언어의 발음의 유사성이 매우 높다는 사실을 확인할 수 있다.

표 3. 외래어의 한글 및 영어 표기 예시

한글 표기	영어 표기
프로그램	Program
프로젝트	Project
데이터베이스	Database
컴퓨터	Computer
파이썬	Python
골드	Gold
마스크	Mask
코로나	Corona

2.2 IPA 기능 및 편집 거리

국제음성협회(International Phonetic Association)에 따르면[1] 각 단어로부터 31개의 중요한 발음 특징을 추출할 수 있으며, 각각의 발음 특징은 세 가지 특징값으로 나눌 수 있다. 각 특징에 해당할 경우 '+' 로 표시하고, 해당하지 않으면 '-', 중요하지 않은 특징이면 '0'으로 표시한다. IPA 특징에 관한 리스트는 표 4에 제시되어 있다.

표 4. IPA(International Phonetic Alphabet) 특징

Features list		
Anterior	Approximant	Constricted glottis
Consonantal	Back	Continuant
Coronal	Front	Diphthong
Distributed	Dorsal	Delayed release
Voice	High	Labial
Labiodental	Lateral	Long
Low	Nasal	Round
Segment	Sonorant	Spread glottis
Stress	Strident	Syllabic
Tap	Tense	Front-diphthong
Trill		

표 4에 제시된 리스트에 따라 각 단어의 발음 특징을 계산하여 IPA거리를 도출할 수 있다. IPA거리는 그림 2의 수식에 따라 IPA그룹 간의 IPA 특징의 차이를 총 발음 특징 수(N)로 나누어 계산한다.

$$\text{if } a_i = b_i \text{ , } f(a_i,b_i) = 0, \quad otherwise \ = 1 \qquad distance = \frac{\sum_{i=1}^{N} f(a_i,b_i)}{N}$$

그림 1. IPA 거리 계산 수식

각 단어의 길이가 일치하지 않으면 단어 길이의 차이를 weight에 곱하여 Levenshtein-variant 거리[6]를 계산한다.

$$lev_{a,b}(i,j) = \begin{cases} max\,(i,j) \\ min \begin{cases} lev_{a,b}(i-1,j) + distance \\ lev_{a,b}(i,j-1) + W \\ lev_{a,b}(i-1,j-1) + W_{(a_i \neq b_j)} \end{cases} \end{cases} \quad \begin{matrix} if\ min(i,j) = 0, \\ otherwise \end{matrix}$$

그림 2. Levenshtein-variant 거리 수식

이렇게 Source Length, Source IPA Length, Target Length, Target IPA Length, IPA Distance를 모두 계산한후, 다음과 같은 언어 특징(feature) 5가지를 랜덤 포레스트 분류기(Random Forest Classifier)를 통해 언어 특징 중요도를 추출하고자 한다. 한국어-중국어로 구성된 2000개의 데이터를 랜덤 포레스트 분류기로 분류하는 실험을 하였고, 본 논문에서는 이 랜덤 포레스트 분류기의 언어 특징의 중요도를 시각화하여 그림 3에 제시하고 있다. 그 결과 IPA 거리 특징의 중요도가 다른 특징들에 비해서 높게 나타난다는 것을 알 수 있다.

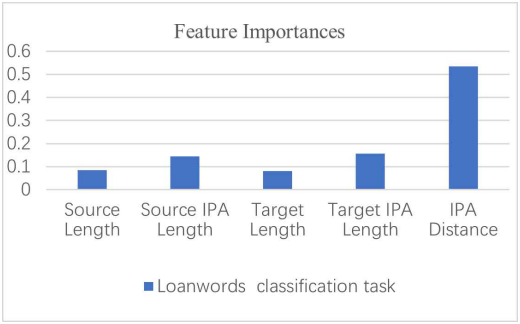

그림 3. Feature Importances

3. 결론

본 논문에서는 한국어 IPA 변환기와 중국어 IPA 변환기 영어 IPA를 통해 한자어 외래어 등에 대한 IPA를 추출하였고, 해당 IPA 정보를 바탕으로 각 단어들 간의 IPA거리 정보를 계산하여 음성학적 특징을 이용한 한국어 기계번역 데이터 세트를 구축하였다. 논문에서 최종적으로 사용된 언어 특징은 Source Length, Source IPA Length, Target Length, Target IPA Length, IPA Distance으로 총 5가지이다. 한국어-중국어로 구성된 2000개의 데이터를 랜덤 포레스트 분류기로 분류하는 실험을 한 결과, IPA 거리 특징의 중요도가 다른 특징들에 비해서 높게 나타난다는 사실을 확인할 수 있었다. 향후 이 데이터 세트을 바탕으로 실제 기계 번역 모델의 성능을 높이는 작업을 진행할 계획이다.

참고문헌

[1] International Phonetic Association, and International Phonetic Association Staff. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press, 1999.

[2] Salam, Khan Md Anwarus, Setsuo Yamada, and Tetsuro Nishino. "How to translate unknown words for English to Bangla Machine Translation using transliteration." Journal of computers 8.5 (2013): 1167-1174, 2013.

[3] Chakravarthi, Bharathi Raja, Mihael Arcan, and John P. McCrae. "Comparison of different orthographies for machine translation of under-resourced dravidian languages." 2nd Conference on Language, Data and Knowledge (LDK 2019). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[4] Ahmed, Tafseer, et al. "Discovering Lexical Similarity Using Articulatory Feature-Based Phonetic Edit Distance." IEEE Access 10 (2021):

1533-1544, 2021.

[5] Lee, Eun-Jung, et al. "IPA Converter of Korean Standard Pronunciation." Proceedings of the Korean Society for Cognitive Science Conference. The Korean Society for Cognitive Science, 2005.

[6] mphilli, et al. Eng-to-Ipa · PyPI. 8 Jan. 2020, https://pypi.org/project/eng-to-ipa/.

[7] Rémi THEVENOUX. “Compute Phonetic Distance between Two IPA String (International Phonetic Alphabet).” GitHub, 3 Aug. 2018, https://github.com/RThevenoux/ipa-distance.

[8] “외래어 - 한글 상호 변환기” 부산대학교 인공지능연구실 2010 http://loanword.cs.pusan.ac.kr/

[9] Sun, Simeng, et al. "Alternative Input Signals Ease Transfer in Multilingual Machine Translation." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

부록. 데이터 샘플

Type	Part	Source	Target	Source IPA	Target IPA	IPA distance	Source length	Target length	Source IPA length	Target IPA length
KRCN		확진	确诊	hoak.tɕin.	tɕʰɥœ.ʈʂənㄱ.	1.87097	2	2	10	12
KRCN	지명	경기	京畿	kjʌŋ.ki.	tɕɪŋ.tɕi.	2.87097	2	2	8	9
KRCN	지명	안양시	安阳市	ŋan.ŋjaŋ.si.	an.jɑŋ.ʂɚ.	4.16129	3	3	12	10
KRCN	전문명사	중심정맥관	中心静脉管	tɕuŋ.sim.tɕʌŋ.mɛk.koan.	ʈʂʊŋ.ɕɪn.tɕɪŋ.maɪ.kwanㄱ.	1.32258	5	5	23	24
KRCN	지명	경북	庆北	kjʌŋ.puk.	tɕʰɪŋ.peɪㄱ.	1.16129	2	2	9	11
KREN		네트워크	network	neˌtʰɯ.ŋuʌ.kʰɯ.	nɛtˌwərk.	4.77419	4	7	4	9
KREN	지명	예멘	yemen	ŋje.men.	jɛmən.	2.09677	2	5	8	6
KREN	지명	티베트	tibet	tʰi.pe.tʰɯ.	təbɛt.	2.16129	3	5	11	6
KREN	전문명사	루트	root	lu.tʰɯ.	rut.	2.06452	2	4	7	4
KREN		샘플	sample	sɛm.pʰɯl.	sæmpəl.	0.16129	2	6	9	7