다국어 공통 음소 체계를 이용한 금기어 매칭 시스템

*김다희 신사임 장달원 이종설 장세진
전자부품연구원
*daheiy@naver.com

Taboo Word Matching System Using a Common Multilingual Phoneme System

*Kim, Da-Hee Shin, Sa-Im Jang, Dal-Won Lee, Jong-Seol Jang, Sei-Jin
Korea Electronics Technology Institute

요약

단어의 유사도 측정 알고리즘은 DB 인덱싱, 필터링, 소스코드 분석 소프트웨어, 음성 인식 등 다양한 분야에서 활용되고 있다. 하지만 기존의 단어의 유사도만 비교하는 시스템에는 발음이 비슷한 유사단어나 오타가 있는 유사단어들은 측정을 못하는 단점이 있다. 언어의 유사도 측정에서는 알파벳만으로 볼게 아니라 언어 발음의 발화적 특성 또한 고려되어야 한다. 본 논문에서는 글로벌 시장에서의 다국적 기업들의 제품이나 문화 수출 등의 도움이 되는 각 나라의 금기어와의 발화적 특성까지 고려한 단어 유사도를 측정 할 수 있는 시스템을 제안한다. 11개국의 4개 언어 총 21487개의 금기어 단어를 금기어 데이터로 사용하였다. 제안하는 방법의 성능을 평가하기 위하여 타 알고리즘과의 성능비교와 여러 나라의 다양한 언어의 사용자들로부터 사용자 평가를 수행하였고 제안하는 방법이 발음 유사도를 측정하지 않는 알고리즘보다 우수한 성능을 보임을 확인하였다.

1. 서론

요즘 들어 신 한류라는 용어가 생겨날 만큼 한국 기업의 물품과 한국 문화가 세계적으로 관심을 받고 있다[1]. 이러한 현상으로 우리나라의 많은 기업들이 해외시장을 타겟으로 물건과 문화를 수출할 수 있는 기회가 많아졌다. 이렇게 진출한 글로벌 시장에서 문제가 되는 부분은 다문화간의 단어의 발음이 같으나 의미가 달라서 생기는 언어의 문제이다. 대표적으로 싸이가 2012년 강남스타일로 전 세계적으로 유명해질 때 싸이의 이전 노래인 '챔피언'에 가사에 '니가'(네가)라는 단어 때문에 흑인 인종차별 논란이 생기도 했었다. 이렇게 다른 언어이지만 우리나라의 문화가 세계시장에 진출했을 때 발음이 같아서 인종차별 논란까지 생기는 일이 생기고 있다. 이렇게 발음상의 문제가 되는 '니가'라는 단어를 본 논문에서는 발음상 유사어라고 정의한다.

발음상 유사어는 다른 언어권의 발음은 유사하나 단어의 뜻은 다른 단어들을 말한다. 발음은 유사하나 뜻은 전혀 다른 뜻을 갖고 있기 때문에 타국에서의 발음상 유사어로 인해 해프닝이 생길 수 있다. 발음상 유사어를 예를 들면 우리나라의 "뽀로로"라는 애니메이션이 전 세계 아이들에게 인기가 좋은데 "뽀로로"라는 단어의 발음이 "포르노"와 비슷하기 때문에 발음상 주는 어감이 영어문화권 사용자들에게는 부정적으로 들릴 수 있다. 또한 우리나라에 들어온 "양키 캔들"은 우리나라에서 "양키"라는 단어가 외국인을 비하하는 목적으로 쓰이기 때문에 이 상호를 처음 듣는 사람들에게는 부정적 인식을 심어줄 수 있다. 이렇게 브랜드의 명칭이 다른 문화권의 나라에 진출할 때 그 나라의 단어의 발음과의 문제를 다국어 공통 음소 체계를 이용한 금기어 매칭 시스템으로 해결하고자 한다.

2. 관련 연구

단어 유사도 비교 연구는 여러 국가에서 다양한 언어로 연구되고 있다. 맞춤법 보정, 어원 식별 등 응용 시스템 또한 다양한 방면으로 연구되고 있다. 그러나 복수의 언어에 대해서 단어 유사도를 분석하는 논문은 부족하며 복수의 언어를 음소로 변환하여 유사도 비교하는 연구는 더욱 부족한 실정이다. 관련 연구들은 표 1에서 확인 할 수 있다.

표 1. 단어 유사도 관련 연구

저자	연도	도메인	특징
Zhang [2]	2008	영어	맞춤법 오류 검사 및 보정 시스템
Kim [3]	2014	한국어	특수문자나 틀린 단어도 음소 간 유사도 자동 계산
Songyot [4]	2014	중국어, 영어, 아랍어	유사도 계산을 이용한 단어 정렬
Cho [5]	2014	프로그래밍언어	서열정렬 알고리즘을 이용하여 악성코드 분석 및 분류

3. 제안하는 시스템

본 논문에서 제안하는 시스템은 그림 1과 같다. 사용자가 금기어와 유사도 비교를 원하는 단어를 입력하면 입력된 단어는 음소 변환 모듈를 통해 후보 음소들로 변환된다. 음소 변환 모듈을 통해 변환된 후보들 중 사용자가 원하는 음소를 선택하면 서열정렬 알고리즘과 음소 발음의 다차원 수치화 알고리즘을 통해 11개국 4개 언어의 DB와 금기어 유사도 비교를 한 결과가 도출된다.

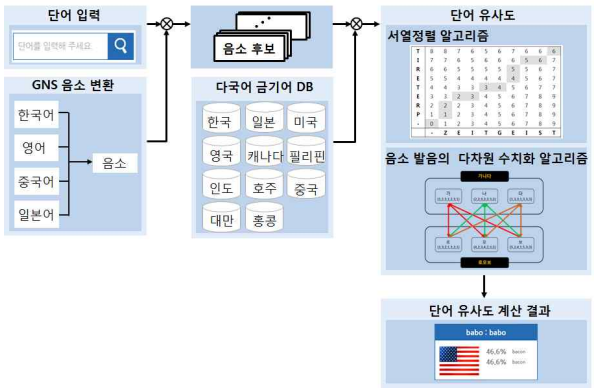

그림 1. 금기어 매칭 시스템

3.1 음소 변환 모듈

사용자가 입력한 단어를 다국어의 음소기반 유사도 알고리즘으로 계산하기위해서는 단어를 음소로 변환이 필요하다. 본 논문에서는 하나의 언어가 아니라 4개의 언어를 모두 같은 형태의 음소로 변환한다. 따라서 다국어 기반의 통합된 음소 변환 모듈이 필요하다.

본 논문에서의 음소 변환 모듈은 다국어의 발음 유사도를 비교할 수 있는 통합된 음소 변환 규칙이다. 여러 국가의 언어들 모두 다른 언어 체계를 갖고 있다. 따라서 서로 다른 발음 체계와 음소 변환 규칙을 통일하여 다국어간의 통합된 규칙이 있어야한다. 본 논문에서의 사용된 언어는 총 4개 언어권의 나라 11개국의 언어이다. 4개 언어는 한국어, 일본어, 중국어, 영어이다. 한국어와 일본어를 제외한 중국어는 중국, 대만, 홍콩 3개국이며 영어는 미국, 영국, 캐나다, 호주, 필리핀, 인도 6개국이다. 이렇게 총 11개국의 언어를 하나의 음소 체계로 통합하여 변환한다. 음소 변환 규칙은 한글의 자음과 모음의 변환 규칙과 알파벳의 변환 규칙이 있다. 총 91개의 규칙을 이용하여 총 24개의 음소로 변환된다. 24개의 음소는 표 2와 같다.

표 2. 24개의 음소별 발음기호 및 한글 자음/모음

음소	발음 기호	자음	음소	발음 기호	모음
b	b	ㅂ	a	a	ㅏ
C	tʃ	ㅊ	e	e	ㅐ
d	d	ㄷ	i	i	ㅣ
f	f	ㅍ	o	o	ㅗ
g	g	ㄱ	u	u	ㅜ
h	h	ㅎ	E	ə	ㅓ
j	j	ㅈ	U	-	ㅡ
k	k	ㅋ			
l	l	ㄹ			
m	m	ㅁ			
n	n	ㄴ			
p	p	ㅃ			
r	r	ㄹ			
s	s	ㅅ			
t	t	ㅌ			
T	ɵ	ㄸ			
N	ŋ	ㅇ			

3.2 유사도 비교 모듈

유사도 비교 모듈에는 크게 2가지 알고리즘을 사용한다. 먼저 DB의 단어들을 첫 번째로 유사도 비교하는 서열정렬 알고리즘을 사용한다. 서열정렬 알고리즘은 DB 내에서 유사도 비교한 결과를 정렬한 뒤 0이하의 정확도는 필터링하는 기능을 갖는다. 또한 서열정렬 알고리즘을 이용하여 사용자의 입력 단어와 금기어 DB 간의 최적의 매칭 패쓰(path)를 구할 수 있다.

서열정렬 알고리즘으로 필터링 된 금기어 DB는 음소 발음의 다차원 수치화 알고리즘을 통해서 서로 상이한 음소의 발음차이를 반영하는 결과를 도출할 수 있다. 각 음소별 수치화 값은 표 3과 4를 통해 확인할 수 있다. 표 3과 4로 수치화된 음소를 사용자 입력 단어의 음소들 간의 거리 계산을 통해 유사도 결과를 얻는다. 유사도 결과는 0에서 100사이의 값을 갖는다.

표 3. 모음의 조음 위치, 자음서 정도, 무성/유성 수치

음소	한글 자음	조음위치	자음성정도	무성/유성
b	ㅂ	1.5	1.1	1.0
C	ㅊ	-0.7	0.7	-1.0
d	ㄷ	-0.1	1.1	1.0
f	ㅍ	1.0	-0.3	-1.0
g	ㄱ	-1.2	1.1	1.0
h	ㅎ	-1.2	-0.3	-1.0
j	ㅈ	-0.7	-1.7	-1.0
k	ㅋ	-1.2	1.1	-1.0
l	ㄹ	-0.1	-0.8	1.0
m	ㅁ	1.5	-1.2	1.0
n	ㄴ	-0.1	-1.2	1.0
p	ㅃ	1.5	1.1	-1.0
r	ㄹ	-0.1	-0.8	1.0
s	ㅅ	-0.1	0.2	-1.0
t	ㅌ	-0.1	1.1	-1.0
T	ㄸ	-0.1	1.1	-1.0
N	ㅇ	0.0	0.0	0.0

표 4. 자음의 조음 위치, 자음성 정도, 무성/유성 수치

음소	한글 모음	조음 위치	자음성 정도	무성/유성
a	ㅏ	-1.4	-0.3	-0.6
e	ㅐ	-0.5	0.3	-0.6
i	ㅣ	0.5	0.8	-0.6
o	ㅗ	-0.5	-0.8	1.3
u	ㅜ	0.5	-1.3	1.3
E	ㅓ	-0.5	0.3	-0.6
U	ㅡ	0.0	0.0	0.0

4. 구현 및 실험

4.1 시스템 구현

3절에서 제안하는 시스템을 본 논문에서는 홈페이지 형식으로 구현하였다. 그림 2는 제안하는 시스템의 첫 홈페이지의 페이지를 보여준다. 그림 3은 사용자가 '바보'라는 단어를 입력 했을 때의 음소 변환 모듈의 결과인 음소 후보들을 보여주는 페이지이다. 그림 4는 사용자가 선택한 음소와 금기어 DB간의 유사도 비교 결과를 보여주는 페이지이다. 이때 사용자가 원하는 다른 나라의 국기를 선택하면 원하는 나라의 언어로 유사도 비교가 가능하다. 그림 5는 단어 금기어 매칭뿐만 아니라 가사와 같은 긴 글을 위한 텍스트 분석의 결과를 보여주는 페이지

이다.

그림 2. 금기어 매칭 시스템 구현 페이지

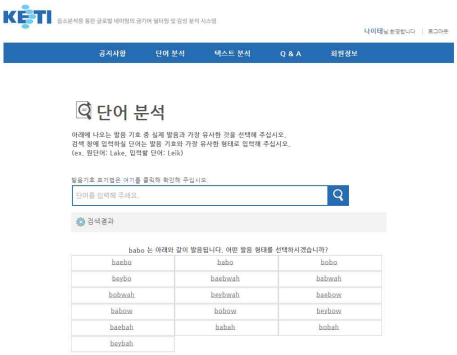

그림 3. 음소 변환 모듈 결과 페이지

그림 4. 단어 금기어 매칭 시스템 구현 페이지

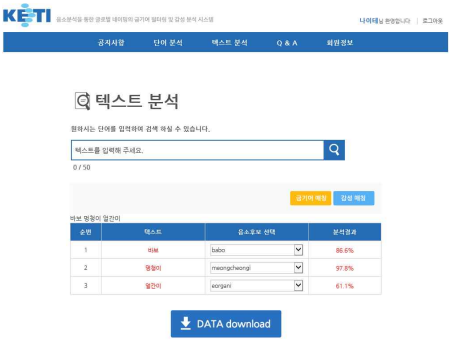

그림 5. 텍스트 금기어 매칭 시스템 구현 페이지

4.2 시스템 실험 결과

4.1 절에서 구현된 시스템을 통해 성능 평가를 실험을 하였다. 기존에 단어 매칭 시스템으로 많이 사용되지만 본 논문에서 제안하는 방법처럼 음소간의 발화적 특징까지는 구분하지 못했던 서열정렬 알고리즘과 본 논문에서 제안하는 시스템의 성능을 비교 실험하였다.

실험 데이터는 4개국 11개 언어 금기어 데이터를 총 21487개 사용했다. 총 9명의 사람이 금기어 데이터를 보고 평가 데이터를 50개 만들었다. 이 평가 데이터는 입력 단어와 금기어 매칭 순위 결과를 포함한다. 정답 데이터는 금기어 데이터 21487개를 사용하였다. 평가 데이터를 입력하였을 때 서열정렬 알고리즘의 금기어 매칭 결과와 본 논문에서 제안하는 시스템의 매칭결과를 비교하여 정답이라고 생각하는 등수가 올라가는 경우와 오답의 등수가 내려가는 경우 2가지를 계산하였다. 실험의 결과는 아래 그림 6과 같다.

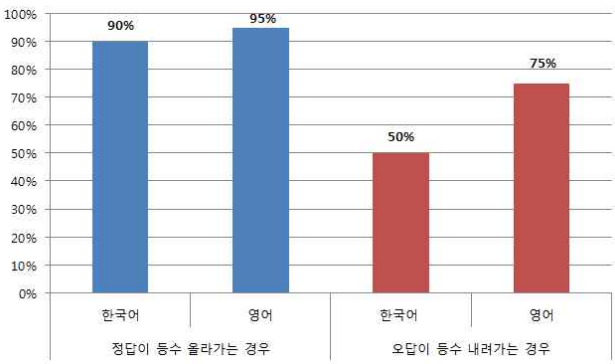

그림 6. 실험 결과 그래프

5. 결론

본 논문은 글로벌 시장에 진출하는 우리나라의 기업이나 문화에 도움이 되고자 다국어 공통 음소 체계를 이용한 금기어 매칭 시스템을 구현하였다. 다국어 공통 음소 체계를 이용한 금기어 매칭 시스템은 11개국의 4개 언어를 사용하는 금기어 매칭 시스템으로 홈페이지 형태로 구현되었다. 총 21487개의 금기어 데이터를 기반으로 기존의 서열정렬 알고리즘과 본 논문이 제안하는 시스템의 비교 실험을 하였다. 실험을 통해 제안하는 시스템이 서열정렬 알고리즘에 비해 발화적 특성까지 고려하여 유사도 측정에 우수하다는 것을 입증하였다.

감사의 글

이 논문은 2014년도 정부(산업통상자원부)의 재원으로 산업융합기반 구축개발사업의 지원을 받아 수행된 연구임(No. 10037244).

6. 참고문헌

[1] Kuwahara, Yasue, ed. “The Korean Wave: Korean Popular Culture in Global Context,” *Palgrave Macmillan*, 2014.

[2] Y. Zhang, P. He, W. Xiang, and M. Li “Discriminative reranking approach to spelling correction,” *Ruan Jian Xue Bao/Journal of Software*, vol. 19, no. 3, pp. 557-564, 2008.

[3] S.-H. Kim, and H.-G. Cho, “A similarity calculation method of deformed string sets with multiple sequence alignment,”*KCC Conference*, vol.40, no. 1, pp.53-60, 2014.

[4] T. Songyot and D. Chiang, "Improving Word Alignment using Word Similarity," *Proc of Conf. on Empirical Methods in Natural Language Processing*, pp. 1840-1845, 2014.

[5] I.-G. Cho, and E.-G. Im, "Malware similarity analysis and classification by using sequence alignment algorithm," *KCC Conference*, pp. 940-942, 2014.