한류 콘텐츠를 위한 발음 기반 금기어 검색 시스템 개발

이종설 신사임 김다희 장달원 장세진 임태범

전자부품연구원

leejs@keti.re.kr

Development of tabooed words search system for Korea wave contents based on pronunciation

Lee, JongSeol Shin, Saim Kim, Dahee Jang, Dalwon Jang, Sei-jin LIm, Tae-Beom

KETI

요약

최근 한류(韓流) 콘텐츠의 글로벌화로 인해 콘텐츠가 전 세계로 수출됨에 따라 글로벌 환경에 적합한 콘텐츠에서의 단어 선택이 매우 중요하게 되었다. 멀티미디어 콘텐츠에서의 글로벌 단어 선택을 위해서는 각 나라의 비속어나 욕설 단어를 회피하고 오해하지 않을 말과 단어를 선택하는 것이 매우 중요하다.

이에 본 논문에서는 글로벌 콘텐츠를 위한 금기 언어 검색 시스템을 개발한다. 이를 위하여 한국어를 영어로 변환하기 위한 음소 변환 모델을 개발하고, 변환된 음소와 금기 언어 검색 데이터베이스를 개발하였다.

1. 서론

이십여년 전 미국, 일본, 홍콩 등에서 방송 콘텐츠를 수입하던 우리나라 방송 콘텐츠는 불과 몇 년 사이에 게임, 드라마, 음악 등 다양한 분야에서 글로벌화 되어 가고 있다. 특히 음악과 드라마 등에서는 K-POP, K-DRAMA란 명칭으로 한류(韓流) 콘텐츠의 핵심 아이템으로 그 중요성이 커져가고 있다. 이와 동반하여 한류 콘텐츠에 사용되는 한국어의 발음이 오해되거나 각국의 문화, 습관, 언어등과 비교하였을 때 문제가 발생한다면 이를 통한 유 무형 손실은 이루어 말할 수 없다. 얼마 전 가수 싸이의 경우 대표곡 '챔피언'에서 '니가 챔피언'이라는 가사를 'Nigga'로 표기하여 일부 외국 네티즌은 흑인을 비하하는 'Nigger'로 오해하는 해프닝이 발생하였다. 이와 같이 한국어로는 문제없는 콘텐츠의 가사/대사가 외국에서는 문제가 발생할 수 있는 가능성이 존재한다. 이에 본 논문에서는 한국어로 작성된 멀티미디어 콘텐츠의 가사/대사에 대한 글로벌 금기어를 사전에 검토하기 위한 시스템을 개발한다. 본 연구는 언어학적 분석을 통하여 구축된 다국어의 금기어 검색에 필요한 유용한 정보들을 통합 제공하는 시스템 개발을 목적으로 한다.

2. 시스템 구성

글로벌 금기 언어 검색 시스템은 우선적으로 영어와 한국어를 지원한다. 이를 위하여 한국어 및 영어 단어를 입력 받아서 이를 음소로 변환하고 추출된 음소를 통해 금기어 데이터베이스의 금기어와 유사도 매칭을 수행한다. 매칭된 결과는 순위화 하여 사용자에게 제공한다.

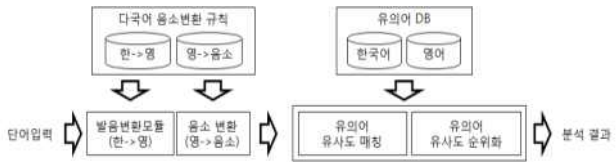

그림 1. 금기 언어 검색 시스템

가.음소 변환 모듈 구성

사용자가 입력한 단어를 음소 단위의 정보 분석을 수행하기 위한 전처리 단계로 본 연구는 금기어 및 감성 매칭 알고리즘 적용을 위한 언어 별 음소 변환을 수행한다. 그림 1에서는 본 연구에서 구축한 시스템의 다국어 음소 변환 과정을 보여준다. 한국어 입력 단어의 경우 영어로 변환하는 발음 변환 과정을 거치게 되는데, 본 연구에서는 국립국어원에서 공표한 '로마자 표기법'에 의한 한국어에서 영어로의 발음 변환을 구현하여 적용하고 있다. 음소 변환 과정의 각 단계 - 발음 변환 및 음소 변환 -에서 변환 규칙의 복수 적용으로 인하여 입력된 단어는 다수 개의 복수 변환 결과의 도출이 가능하다. 따라서 음소 변환 과정에서 도출된 복수 개의 음소 변환 결과들 중 사용자들은 분석을 원하는 정확한 음소 변환 결과를 선택하여 분석 알고리즘에 분석을 요청하게 된다.

영어 단어에서 발음되는 음소들을 추출하여 나열하기 위한 영어에서 음소로의 변환 규칙은 기 구축된 금기어 데이터의 변환 패턴 유추를 토대로 구축되었다. 본 시스템에 적용된 음소 변환 규칙의 내용은 다음과 같다.

영어표기	발음음소후보	영어표기	발음음소후보
a	oe, a, o, ey, ae	n	n
b	b	o	o, u, wah, ow, ah
c	s, k	p	p
d	d	q	k
e	iy, e, eh, x, i	r	r
f	f	s	s, z
g	g, jh, j	t	t
h	h	u	uh, ui, yu, ah
i	ai, iy, i	v	v
j	jh, j	w	wah, er, u, w
k	k	x	iks, eks
l	l	y	iy, i
m	m	z	z

나. 음소별 유사도 점수화

음소 기반 금기어와의 유사도 비교를 위해서는 음소 간의 유사도 값의 수치화가 필요하다. 본 연구에서 적용한 음소별 유사도값은 다음과 같은 특징을 보인다.

- 거리의 1차원화: 음소 간 발음 유사도 거리값은 1차원 정수 값으로 표현한다.
- 거리값의 상대화: 음소와 음소 간의 발음 유사도 거리값은 결정된 절대값이 아닌 두 음소 사이의 상대적인 거리값으로 환산되어 측정된다. 측정된 거리의 값의 절대값의 크기 정도보다 두 음소 간의 거리의 크기 정도를 가늠하는 것으로도 음소 사이의 거리 정도를 분석 가능하다.

위와 같은 특성을 보이는 음소간의 거리 유사도값 측정을 위한 영어와 한국어의 음소 별 / 자음 및 모음 별 1차원 거리좌표는 다음과 같이 보여질 수 있다.

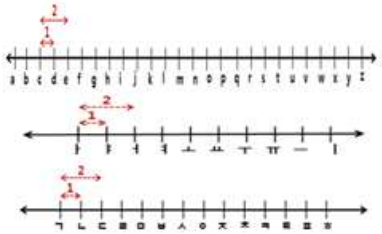

그림 2 영어/한국어모음/한국어자음의 유사도값 도출을 위한 거리좌표

각 언어 별 좌표값에 의한 음소 간 유사도값 도출을 통하여, 단어를 구성하는 이들 음소와 각 금기어 사이의 유사도값을 통한 유사도 계산을 수행한 후, 금기어 매칭 결과를 분석하게 된다.

다. 금기어 매칭 알고리즘

본 논문에서는 질의어와 금기어와의 거리 측정을 위해 서열 정렬 알고리즘을 이용한다[2]. 서열 정렬 (Sequence alignment)은 본래 생물 정보학 분야에서 두 유전체 서열의 유사도를 측정하거나 유사 구간을 찾기 위해 사용되는 기법이다. 유전체 서열의 특성 상 이는 문자열로 표현되며, 따라서 임의의 문자열 간의 유사도 및 유사 구간을 판별하는데 곧바로 활용이 가능하다. 두 문자열에 대한 정렬은 각각의 문자열에 적당한 갭 (gap)을 삽입하여 서로 대응되는 문자간의 유사도 합이 최대가 되는 조합을 찾는 문제이다. 두 문자열 $x = x_0 x_1 \ldots x_{m-1}$, $y = y_0 y_1 \ldots y_{n-1}$에 대한 서열 정렬을 통한 유사도 점수 GA (x, y)는 다음의 동적 계획법식을 풀어 구할 수 있다.[1]

$$GA(x, y) = M(m, n; x, y)$$
$$M(0,0; x, y) = 0$$
$$M(i, 0; x, y) = M(i-1, 0; x, y) + \sigma(x_{i-1}, -)$$
$$M(0, j; x, y) = M(0, j-1; x, y) + \sigma(-, y_{j-1})$$
$$M(i, j; x, y) = \max\begin{cases} M(i-1, j; x, y) + \sigma(x_{i-1}, -), \\ M(i, j-1; x, y) + \sigma(-, y_{j-1}), \\ M(i-1, j-1; x, y) + \sigma(x_{i-1}, y_{j-1}) \end{cases}$$

4. 결론

본 논문에서는 한류 콘텐츠에 대한 금기어 검색 시스템을 개발하였다. 개발한 시스템은 단순한 텍스트 매칭이 아닌, 발음 환경에 기반한 검색 환경을 제공한다. 이를 통하여 콘텐츠 생산자는 글로벌 환경에서 발생할 수 있는 발음으로 인한 문제를 사전에 예방이 가능하다. 앞으로 금기어 데이터베이스의 확충 및 다양한 언어로 확장이 필요하다.

그림 3. 단어 기반 검색 결과

[1]윤태진, 조환규, “버로우즈-휠러 변환과 다단계 정렬을 이용한 고속 한글 문서 탐색 기법”,정보과학회논문지, 2012

[2] J. L. Donaldson, A.-M. Lancaster and P. H. Sposato, "A Plagiarism Detection System," In Proc. of ACM SIGSCE, pp.21-25, 1981

[3] T. Yoon, H. G. Cho, W. Chung, "A Phonemebased Approximate String Searching System for Restricted Korean Character Input Environments," Journal of KIISE: Software and Applications, vol.37, no.10, pp.788-801, Oct. 2010. (in Korean)

[4] E. Chavez, G. Navarro, R. Baeza-Yates, J. L. Marroquin, "Searching in Metric Spaces," ACM Computing Surveys, vol.33, no.3, Sept. 2001.

[5] S. H. Kim, H. G. Cho, "Proximity Word Filtering by Hierarchical Clustering," in Proc. of 37th KIPS Spring Conference, vol.19, no.1, pp.1101-1104, 2012. (in Korean).