트위터에서 이슈가 되고 있는 중국어-한국어 교차언어 뉴스 탐지

조승남, 촐몽 바야르, 이경순, 이용석
전북대학교 컴퓨터공학부/영상정보신기술연구센터
e-mail : shengnan21@hotmail.com, yslee@chonbuk.ac.kr

Chinese and Korean Cross Lingual News Detection in Twitter

Shengnan Zhao, Bayar Tsolmon, Kyung-Soon Lee, Yong-Seok Lee
Dept. of Computer Engineering, Chonbuk National University

요 약

국제적으로 이슈가 되고있는 사건들의 뉴스는 보도당국의 입장차이에 따라 동일 이슈에 대한 관점의 차이를 나타낸다. 교차언어 연구에서는 번역하는 과정이 중요하다. 본 논문에서는 중-한 어휘번역에서 발생하는 오류 및 모호성을 해결하기 위해 키워드를 중심으로 문맥 어휘를 이용해서 번역한 후 번역결과에서 빈도가 높은 한국어 어휘를 선택하는 방법을 제안한다. 제안 방법의 유효성을 검증하기 위해 소셜 이슈 3 개에 대한 트윗 데이터에서 실험하여 추출된 중-한 이슈 뉴스 결과에서의 정확도 85.8%의 성능을 보였다. 실험을 통해 제안 방법이 중-한 교차언어 트위터 데이터에서 동일한 이슈와 관련된 뉴스를 찾는데 효과적인 방법임을 알 수 있다.

1. 서론

트위터는 다국어 지원과 지역에 대한 제한이 없이 전 세계 이용자와 실시간으로 대화와 소통이 가능하여 트윗(tweet)으로 자신의 의견을 표현하고, 리트윗(retweet)을 통해 의견을 공유하기 때문에 정보가 빠르게 전파되는 특징이 있다. 트위터는 정보 정달하는데 사용되어 이슈자질을 포함한 트윗의 내용에 85%가 해당하는 뉴스 기사에 관한 트윗이였다고 하였다[5]

국제적으로 이슈가 되는 뉴스는 트위터를 통해 각 국가의 사람들에게 전파가 되면서 입장차이에 따라 동일이슈에 대한 관점의 차이가 있다. 국제적으로 이슈가 되고 있는 사건들의 뉴스 내용들은 보도당국의 입장차이에 따라 동일 이슈에 대한 관점의 차이를 있다.

교차언어에 대한 연구로는 NTCIR[8]에서 교차 언어를 구사하는 링크를 발견 (CLLD: Cross Language Link Discovery) 하여 자동으로 다른 언어의 문서 사이의 잠재적인 링크를 찾는 방법에 대한 연구가 진행되고 있다.

영어 원본 문서와 한국어, 중국어, 일본어의 대상 문서간의 연결에 초점을 맞추어 자동 링크 검색에 대한 시스템을 구축하고 재사용할 수 있는 방법을 제안하였다[1]. 핵심어를 추출하는 연구로는 소스 언어에서 각 링크에 해당하는 앵커(anchor) 텍스트의 빈도수를 이용하여 영어, 중국어, 한국어, 일본어등 다국어 핵심어를 추출 및 링크 추정 시스템을 제안하였다[2]. 교차 언어 번역 과정에 관한 연구로 위키 백과, 온라인 백과 사전을 이용하여 중국어로 영문이름의 번역 문제를 해결하기 위해 음역 매핑 단어를 이용하였다[3]. 또한 공동 발생 방법을 사용하여 한국어 질의에서 중국어 질의를 번역하였다. 위키백과는 다국어 온라인 백과 사전이며 다국어에 대한 다양한 정보를 포함하고 있다. 언어별로 구성되어 있는 페이지들은 서로 관련있는 데이터의 링크정보 가지고 있지만 교차언어 링크 정보가 없는 문제를 해결하는 방법을 소개하였다[4].

트위터에 관한 연구로는 사용자가 신뢰할 수 있는 정보를 게시하고있는 트윗을 통하여 사용자의 흥미를 확인할 수 있는 방법을 제안하였으며 트위터에서 사건지역을 통해 뉴스를 추출하는 시스템에 대한 연구도 있다[6]. 트위터는 정보를 전달하고 공유하는 특징을 통해 이중 언어 및 다국어 사용자의 언어 선택 기반 뉴스 추출 방법을 소개하였다. 동시에, 그것은 교차 언어 정보의 흐름을 활성화할 수 있는 특성을 식별하는 목적이었다[7].

본 논문에서는 중-한 어휘번역에서 발생하는 오류 및 모호성을 해결하기 위해 키워드를 중심으로 문맥 어휘를 이용하여 번역한 후 번역결과에서 빈도가 높은 한국어 어휘를 선택하는 방법을 제안한다. 중국어 트위터에서 많이 인용된 뉴스 URL 을 선택하여 이슈를 추출하여 관련된 중국어, 한국어 뉴스 기사를 추출한 다음 서로간의 유사도를 측정하는 방법을 제안한다. 소셜 이슈 3 개에 대한 트윗 데이터에서 실험하여 이슈와 관련된 중-한 뉴스 기사를 탐지하였다.

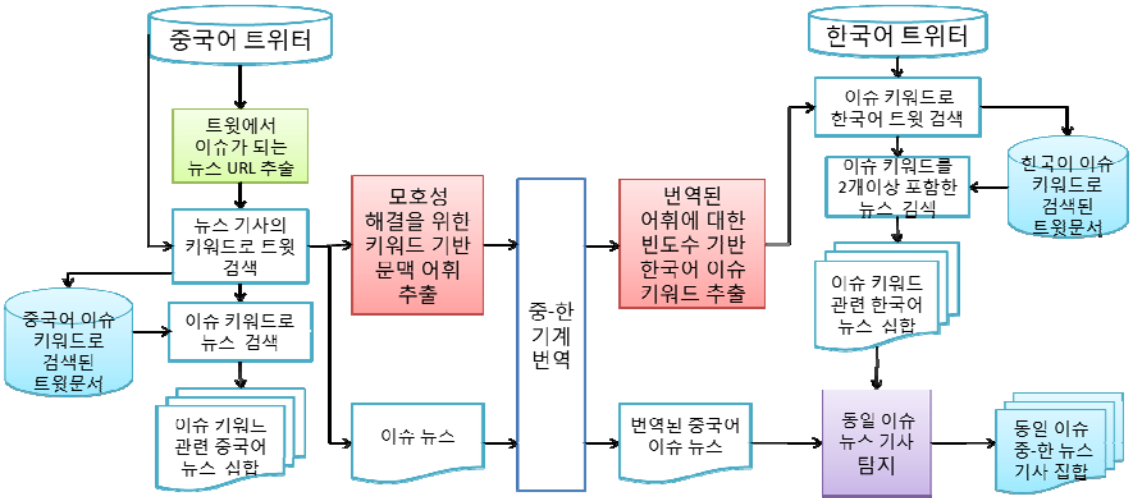

그림 (1) 트위터에서 이슈가 되고 있는 중-한 뉴스 탐지 시스템 구조

2. 중-한 교차언어 트위터 데이터에서 이슈가 되고 있는 뉴스 추출

중-한 교차언어 트위터 데이터에서 동일 이슈가 되고 있는 뉴스 및 트윗을 추출하는 시스템 구조는 그림 (1)과 같다.

중국어 트위터에서는 이슈가 되는 뉴스 기사의 URL 을 추출하고 뉴스 키워드 관련 트윗 추출과 관련 뉴스를 수집한다.

번역과정에서 발생하는 모호성과 오류를 줄이기 위하여 키워드를 중심으로 문맥 어휘를 이용한다.

번역된 이슈 키워드로 한국어 트위터에서 관련 트윗 및 이슈 키워드 두개 이상 포함하는 뉴스 기사를 추출한다.

추출된 중-한 동일한 이슈 뉴스 기사 간의 유사도를 측정한다

2.1 중국어 트위터에서 이슈가 되는 뉴스 추출

현대사회에서 일어나는 특정한 사건이나 사회 현상은 뉴스 기사의 형태로 우리에게 전달된다. 이슈가 되고 있는 사건이 일어났을 때 그 사건과 관련된 트윗 수는 폭발적으로 증가한다.

트위터는 140 자 이내의 문자로 한정되어있어 뉴스 기사의 내용을 전반적으로 표현하기 어렵기때문에 사용자들은 뉴스기사의 제목이나 내용을 요약한 후 URL 을 부가정보로 추가하는 경향이 있다. 이 과정에서 문자가 제한된 트윗에서 URL 이 길면 트윗내용을 작성하는데 제약이 되므로 트위터에서 사용자들에게 긴 URL 정보를 다음과 같이 짧은 URL 로 바꾸어 사용할 수 있도록 제공해준다.

- 원본 URL 주소:
  http://news.chosun.com/site/data/html_dir/2012/09/21/20120921 01123.html
- 축소 된 URL 주소:
  http://t.co/j606UiL7

이러한 관찰을 통해 중국어 트윗에서 이슈가 된 사건들을 추출하기 위해 먼저 URL 이 달린 트윗을 추출한다.

축소 된 트위터 URL 의 주소가 달라도 원본 URL 주소는 같은 경우가 있기때문에 축소 된 URL 을 추출하여 원본 URL 주소로 변경 시켜 빈도수가 높은 순서로 순위화한다. 대부분의 상위 URL 은 동영상 또는 사진, 광고 등 뉴스기사가 아닌 URL 제거한 후, 재순위화하여 중국어 트위터에서 해당 날짜에 올라온 뉴스 URL 빈도수가 높은 뉴스 기사 1 개를 이슈 뉴스 기사로 추출한다.

2.2 중국어 이슈 관련 트윗 및 뉴스 검색

추출한 뉴스 URL 빈도수가 높은 중국어 이슈 뉴스의 소스 파일에 들어가면 그 뉴스에 해당되는 키워드들을 볼 수 있다. 하지만 같은 내용의 뉴스 기사라도 신문사에 따라 제공하는 키워드들이 약간의 차이를 나타낸다.

이슈 키워드를 이용하여 중국어 트위터에서 이슈 키워드를 2 개 이상을 포함하고 있는 트윗에서 빈도수가 높은 URL 순서로 해당하는 뉴스를 이슈와 관련된 중국어 뉴스로 추출한다.

2.3 중국어 키워드 기반 문맥어휘를 이용한 중-한 기계번역

번역 과정에서 발생하는 오류 및 애매모호성을 해결하기 위해 키워드 중심 문맥어휘를 사용한다.중국어의 특성은 띄어쓰기가 없기 때문에 추출한 이슈 뉴스의 제목과 내용을 중국어 단어 분리 오픈 소스 프로그램[9]을 이용하여 분리하였다.

기계 번역 과정에서 단어가 하나일 경우는 의미적으로 적합하게 번역이 되지 않는 문제점을 보였다.

예를 들면, 중국어 '右翼 (우익)' 이라는 단어 하나를 기계 번역하였을 경우 '오른쪽 날개' 로 번역되였다. 사실 '우익' 이라는 단어는 '오른쪽 날개' 의 의미도 갖고 있지만 정치사상의 경향을 나타내는 뜻으로 많이 쓰이기에 적합하게 번역되지 않았

다고 볼 수 있다.

번역 과정에서 나타나는 오류 및 애매모호성을 해결하기 위해서 사용한 문맥정보는 다음같이 중-한 기계번역기에 입력하였다.(키워드 단어를 w_i로 표시함)

- w_i
- w_{i-1} w_i
- w_i w_{i+1}
- w_{i-1} w_i w_{i+1}

문맥 어휘정보를 이용한 예를 들면, '右翼 (우익)'으로 기계 번역하기 보다는 '日本右翼分子(일본우익분자)'로 번역하면 '일본 우익'이 출력된다. 이렇게 번역된 결과는 단일 키워드 번역보다는 효과적이다.

번역된 결과에 대해서는 각 키워당 각 문맥 어휘에서 빈도수가 높은 단어를 추출하여 한국어 키워드로 설정하였다. 중국어에서 한국어로 키워드를 번역하는 과정에서는 Microsoft 에서 제공하는 번역 API 를 사용하였다.

추출된 중국어 이슈 키워드를 기계번역기로 돌린 다음에 한국어 이슈 키워드를 추출한 결과는 <표 1>과 같다.

<표 1> 추출한 이슈 및 키워드

이슈	일본 중국 영토 분쟁	일본 민주당 선거	필리핀 대통령
중국어 뉴스에서 추출한 중국어 이슈 키워드	习近平 领土争端 日本	日本总理 民主党 右翼	南海 菲律宾 总统特使
기계 번역한 한국어 키워드	XI Jinping 시진핑 영토 분쟁 일본	일본 총리 민주당 오른쪽 날개	남쪽 바다 필피핀 대통령 특사
추출된 한국어 이슈 키워드	시진핑 영토 분쟁 일본	일본 총리 민주당 일본 우익	남중국해 필리핀 대통령 특사

2.4 번역된 한국어 키워드를 이용한 한국어 트윗 및 뉴스 검색

추출한 한국어 이슈 키워드를 이용하여 한국어 트위터에서 키워드와 관련된 트윗을 검색하여 저장한다.

검색된 결과내에서 빈도수가 높은 뉴스 URL 을 찾아 이슈와 관련된 뉴스로 순위화한다.

2.5 동일 이슈 뉴스의 중-한 뉴스 기사 탐지

이슈 키워드와 관련된 중국어 뉴스 기사와 한국어 뉴스 기사간의 유사도를 측정하기 위해 코사인 유사도를 사용한다.

중국어 이슈 뉴스를 번역한 다음 결과에서 단어빈도를 측정한다.

한국 이슈 키워드와 관련된 뉴스 기사 집합을 한국어 형태소 분석기를 통하여 단어 빈도를 측정한다.

번역된 중국어 이슈 뉴스에서 키워드를 포함한 빈도수가 높은 상위 단어 10 개를 선택하고 추출된 한국어 뉴스에서 나타나는 같은 단어들의 빈도의 코사인 유사도값이 0.4 이상인 중-한 동일뉴스와 관련된 중한 기사를 추출한다.

3. 실험 및 결과

3.1 실험 집합

본 논문에서 제안한 방법을 실험하기 위해 2012 년 9 월 21 일 중-한 트위터에 올라온 모든 트윗을 Twitter Streaming API[10] 를 이용하여 수집한 결과는<표 2>와 같다.

<표 2> 실험 데이터 집합

	중국어 트윗	한국어 트윗
트윗 개수	244,361	1,543,291
URL 를 포함한 트윗 수	73,341	296,771

3.2 실험 결과

실험하기 위해 추출된 뉴스 URL 빈도수가 높은 상위 세개의 이슈<표 3>와 같다.

<표 3> 추출한 이슈

이슈	이슈 뉴스
1	일본 중국 영토분쟁 문제
2	일본 민주당 총리 선거
3	필리핀 대통령 남중국해 분쟁

중국어 트위터에서 해당 날짜에 이슈가 되는 사건에서 중국어 뉴스에서 제공하는 키워드를 단일 키워드로 번역기에 넣어 번역한 결과 와 단일 키워드 문맥 어휘를 이용해 추출한 단어를 번역하여 단어 빈도수가 제일 높은 중국어 키워드 가 번역된 결과를 비교하였을때 단일 단어가 단어 조합으로 키워드를 번역하는것이 효과적이였다.

<표 4> "일본 중국 영토" 이슈 키워드 개수에 따른 중한뉴스 검색 결과

	키워드		
	1 개	2 개	3 개
	영토	일본 영토	시진핑 일본 영토
중국어 URL 을 포함하는 트윗 수	2506	1482	306
중국어 뉴스 수	54	28	18
한국어 URL 을 포함하는 트윗 수	3824	2104	24
한국어 뉴스 수	34	24	4

<표 6> 중-한 이슈 키워드를 포함한 트윗 및 관련 뉴스 예

	중국어 뉴스 키워드	문맥을 반영한 중국어 키워드	추출한 한국어 이슈 키워드	이슈가 된 뉴스를 포함하는 트윗 예	뉴스의 제목과 원본 URL 주소
중국어 뉴스	习近平 领土争端 谈判	副主席习近平 习近平出席 副主席习近平出席 解决领土争端 通过谈判	-	*习近平：中国致力于通过谈判和平解决领土争端: 9 月 21 日，第九届中国-东盟博览会在广西南宁开幕，中国国家副主席习近平出席开幕式。中新社发盛佳鹏摄中新网 9 月 21 日电"我们永远不争霸，永远不称霸。"中国国家副主席习近平今日强调 ， 中国始 http://t.co/nMarEoGp*	제목 : **习近平**：中国致力于通过**谈判和平**解决**领土争端** 번역된 제목: "시진핑: 중국은 평화 담판은 통하여 영토분쟁을 해결하려한다" http://news.qq.com/a/20120921/001093.htm?utm_source=feedburner&utm_medium=twitter&utm_campaign=Feed%3A+baidu%2FCGwj+%28%E4%B8%AD%E5%9B%BD%E5%9B%BD%E5%86%85%E7%84%A6%E7%82%B9%E6%96%B0%E9%97%BB%29
한국어 뉴스	-	-	시진핑 영토 분쟁 해결 협상	*中시진핑 "우호 담판 통해 영토분쟁 해결":중국의 차기 최고지도자로 유력한 시진핑(習近平) 중국 국가주석이 21 일 일본과의 센카쿠(중구명 댜오위다오) 영토 분쟁과 관련해 "우호 담판을 통해 이웃 국... http://t.co/j606UiL7*	제목 : 中**시진핑** "**우호 담판** 통해 **영토분쟁** 해결 " http://news.chosun.com/site/data/html_dir/2012/09/21/2012092101123.html

<표 5> 중-한 동일 이슈 뉴스 탐지 결과(정확도)

이슈	이슈 1	이슈 2	이슈 3	평균
중국어 뉴스수	15/18	9/13	5/7	77.2%
한국어 뉴스수	4/4	5/6	3/3	94.4%
정확도	91.6%	76.2%	85.7%	85.8%

추출한 이슈 키워드를 이용한 중-한 이슈 뉴스 추출 결과는<표 4>과 같다. 검색어로 이슈 키워드 한개를 넣었을 때의 문제점은 이슈 키워드 '영토'와 같이 '시진핑' 단어가 나올 수 있고 또는 '독도'라는 단어가 나올 수 있다. 이러한 문제점을 해결하기 위해 2 개 이상의 이슈 키워드 단어로 검색하였으며 추출된 관련 뉴스 기사의 정확도가 성능이 좋았다.

중-한 트위터에서 추출한 뉴스의 개수가 각각 달라 실험 평가를 하기 위해 추출한 모든 뉴스에서 정확도를 계산한 결과는 <표 5>과 같다. 이슈 1 에 대한 정확도가 91.6%였으며 평균적으로 3 개의 이슈에서 85.5%의 성능을 보았다.

중-한 이슈 키워드를 포함한 트윗 및 관련 뉴스에 대한 결과 예는 <표 6>과 같다.

4. 결론

본 논문에서는 중-한 어휘번역에서 발생하는 오류 및 모호성을 해결하기 위해 키워드를 중심으로 바이그램 어휘를 이용해서 번역한 후 번역결과에서 빈도가 높은 한국어 어휘를 선택하는 방법을 제안하였다. 실험 결과, 소셜 이슈 3 개에 대한 트윗 데이터에서 정확도 85.8%의 성능을 보였다. 실험을 통해 제안 방법이 중-한 교차언어 트위터 데이터에서 동일한 이슈와 관된 뉴스를 찾는데 효과적인 방법임을 알 수 있었다.

향후 연구로는 한국어-중국어-영어 트위터에서 이슈가 되고 있는 사건을 추출하기 위해 위키피디아 정보를 이용한 방법을 제안할 계획이다.

참고문헌

[1] L.X. Tang, S. Geva, A. Trotman, Y. Xu and K.Y. Itakura, "Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery", Proceedings of NTCIR-9, 2011

[2] Y.H. Lee, C.Y. Chuang, C.C. Chen and W.L. Hsu "Discovering Links by Context Similarity and Translated Key Phrases for NTCIR9 CrossLink", Proceedings of NTCIR-9, 2011

[3] Y.C. Wang, R. T. H. Tsai, Hsu-Chun Yen, W.L. Hsu "Korean-Chinese Cross-Language Information Retrieval Based on Extension of Dictionaries and Transliteration", Proceedings of the 19th Conference on Computational Linguistics and Speech Processing, 2007

[4] L.X. Tang, D. Cavanagh, A. Trotman, S. Geva, Y. Xu and L. Sitbon "Automated Cross-lingual Link Discovery in Wikipedia", Proceedings of NTCIR-9, 2011

[5] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a social network or a news media?", Proceedings of the 19th international conference on World wide web, 2010

[6] A. Jackoway, H. Samet and J. Sankaranarayanan, "Identification of live news events using Twitter", Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, 2011

[7] I. Eleta, "Multilingual use of twitter: social networks and language choice", Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion, 2012

[8] NTCIR: NII Test Collection for IR Systems http://ntcir.nii.ac.jp/

[9] ICTCLAS(Chinese Academy of Sciences, CAS) 중국어 단어 분리 오픈소스 http://ictclas.org/ictclas_download.aspx

[10] 트위터 개발자 사이트 https://dev.twitter.com/docs/streaming-apis