동의어와 용어에 대한 번역 신뢰도 개선 방법

임지연*, 윤재열*, 김이준*, 김응모*
*성균관대학교 컴퓨터공학과
e-mail:limjy@skku.edu

The Way to Improve Credibility of Translation for Synonyms and Terms

Ji-Yeon Lim*, Jae-Yeol Yoon*, Iee Joon Kim*, Ung-Mo Kim*
*Dept of Computer Engineering, SungKyunKwan University

요 약

인터넷의 비약적인 발전으로 우리는 생활에 필요한 많은 것들을 인터넷을 통해 얻는다. 날씨, 뉴스, 만화, 나아가서는 전공 공부까지 많은 정보를 인터넷에서 얻을 수 있다. 또한 이러한 여러 서비스를 제공하고 있는 포털사이트에서는 무료 번역기 또한 무료로 제공하고 있다. 하지만 무료로 제공하는 만큼 번역기의 신뢰도가 높지 않아, 실제 사용함에 있어 실제 번역에 제약이 있기 마련이다. 이러한 점에서 본 논문에서는 수많은 사람들이 작성한 정보를 통해 문맥 속에서 쉽게 틀릴 수 있는 전공 용어의 올바른 해석을 위해 오피니언 마이닝 기법 중 PMI-IR 수식을 이용하여 번역기의 신뢰도를 개선하는 방법을 제안한다.

1. 서론

인터넷의 생활화에 따라서 우리는 엄청난 자료를 습관적으로 접할 수 있고, 접하고 있다. 이미 많은 기존 논문에서 거론되었지만, 이렇게 많은 자료 중에서 본인에게 알맞고 가치 있는 정보를 찾아내는 작업이 필요하다. 자료를 정보로 만드는 작업은 오늘날까지 많은 연구를 통해 방법이 소개되고 있으며, 본 논문에서는 PMI-IR을 통하여 번역하려는 문장 내에서 전문 용어를 찾아내어, 번역기의 신뢰도를 높이고자 한다.

PMI-IR[1] 수식은 TOEFL(Test Of English as a Foreign Language)의 문제에 중 특정 단어와 보기 단어의 연관성을 계산한 값을 비교하여 연관성이 큰 단어를 정답으로 추천하고 있다. 본 논문에서는 번역기에 입력된 문장에서 단어 단위로 추출된 두 단어의 연관성을 계산한다.

본 논문의 2절에서는 관련 연구를 조사하고, 3절에서 데이터 마이닝을 통한 예측에 대해 설명하고, 실험 결과를 도출한다. 마지막으로 4절에서 결론을 맺는다.

2. 관련연구

2.1 데이터 마이닝(Data Mining)

하루가 다르게 늘어가는 방대한 자료와 다르게 인간이 읽고, 이해할 수 있는 정보에는 한계가 있다. 이를 위해서 방대한 구조화 되어있지 않은 자료(Unstructured Text)에서 가치 있는 정보를 생산하는 작업이 필요하다. 이를 위해 사용되는 것이 데이터 마이닝이다. 데이터 마이닝은 정보 추출(Information Extraction), 문서 분류(Text Classification)/ 문서 클러스터링(Text Clustering), 토픽 트래킹(Topic Tracking), 웹 마이닝(Web Mining) 등 여러 가지 기법[2]이 있고, 우리는 그 중 동적인 변화가 크게 영향을 미치는 웹 마이닝 기법을 이용하기로 한다. 웹 마이닝을 크게 웹 서버에서 보관되어 있는 로그를 분석하는 기법, 하이퍼링크로 연결된 웹 페이지들 간의 관계 분석 기법, 웹 페이지의 문서 데이터에 대해 다양한 데이터 마이닝 알고리즘을 적용하는 기법이 있다.

우리는 웹 마이닝 기법 중 아래 수식 1과 같은 PMI-IR 수식[1][3][4]을 이용하여 두 단어가 같은 페이지 내에서 동시에 출현한 빈도수를 측정하여 두 단어 사이의 연관성을 계산한다.

$$PMI-IR(word_1, word_2) = \log 2\left(\frac{p(word1 word2)}{p(word_1)p(word_2)}\right)$$

<수식 1> PMI-IR 감정어 연관도 추출 수식

2.3 번역기

현재 대부분의 포털사이트에서 번역기를 제공하고 있다. 앞서 언급했듯이, 이러한 무료 번역기의 경우 비교적 높지 않은 신뢰도에 따라 공식적인 문서의 번역을 하는데 있어 한계가 있다.

이러한 번역기는 동의어(Synonyms), 품사(Parts of speech), 용어(Terms) 등을 구분하지 못하여, 번역기의 신뢰도를 떨어뜨리는 것으로 나타났다.

3. 제안내용 및 실험

본 논문에서는 동의어(Synonyms), 용어(Terms) 등을 구분하지 못하는 번역기의 한계를 극복하고, 번역된 결과의 신뢰성을 높이기 위하여 문장을 단어 단위로 구분하여 추출한 뒤, 두 단어 사이의 계산된 연관성을 통해 전문 용어인지 구분한다.

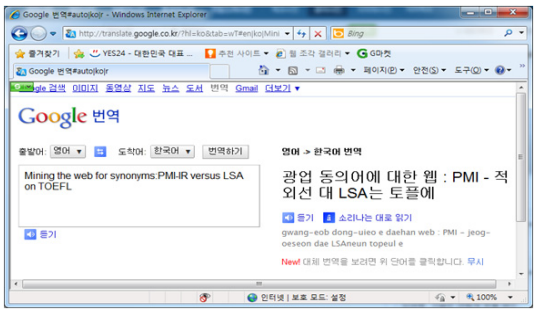

그림 1 Google 번역기에서의 틀린 번역 예

그림 2에서 볼 수 있듯이, 구글 번역기는 영한 번역시 마이닝(Mining)이라는 단어를 광업(Mining)이라고 해석했다. 마이닝(Mining)은 '채굴, 채광, 광(산)업' 이라는 뜻이므로 직역을 하자면 틀린 번역은 아니지만, 글의 문맥상 틀린 번역이 된다. 또, PMI-IR의 경우 마이닝 기법 중 연관도를 계산하는 수식이므로 번역하지 않는 것이 매끄럽다.

#	단어 1	단어 2
1	PMI	IR
2	PMI	적외선
3	Mining The Web	마이닝
4	Mining The web	광업

표 2 문장 내 단어 추출

우리는 표 1과 같이 문장 내에서 번역에 매끄럽지 못한 단어와 그 단어에 대한 번역 결과, 그리고 또 다른 단어에 대한 각각의 검색 결과를 추출했다.

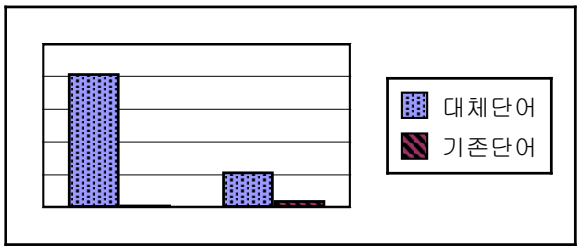

그림 2 문장 내 단어 연관도 검색결과

추출된 두 단어의 같은 페이지 내 발생 빈도에 PMI-IR 수식을 적용하여 연관도를 측정하였다. 측정된 결과는 그림 3과 같다.

기존 문장 번역기 결과	광업 동의어에 대한 웹 : PMI - 적외선 대 LSA는 토플에
PMI - IR을 이용한 개선 결과	동의어에 대한 웹 마이닝 : PMI - IR 대 LSA는 토플에

표 3 실험 결과

그림 3에서 볼 수 있듯이, 번역된 단어와 기존 단어의 연관도보다 대체 단어의 연관도가 현저하게 높다는 것을 알 수 있다. 용어 그대로 썼을 때를 나타내는 'PMI'와 'IR'의 검색 결과는 8189건이지만, 'PMI'와 'IR'의 번역 결과인 '적외선'의 검색 결과는 184건으로 굉장한 차이를 보였고, 'Mining The Web'에 대한 검색 결과 또한 '마이닝' 2204건, '광업' 454건으로 그 차이가 컸다. 이와 같은 검색 결과를 이용하여 PMI-IR 수식으로 단어 간의 연관도를 추출해 봤을 때, 표 3과 같은 결과를 보여준다.

4. 결론 및 향후 연구 방향

본 논문을 통해 PMI-IR 수식을 이용하여 정보를 추출해봄으로써 방대한 자료에서 가치 있는 정보를 재생산할 수 있다는 결론을 얻었다. 그리고 이렇게 재생산된 정보를 실생활에 접목해 봄으로써 더욱 유용하게 사용할 수 있다. 우리는 실험을 통해 완벽한 결과를 낼 수 없었으나, 번역기의 높은 신뢰성을 위해서 더욱 다양한 데이터 셋(Data set)을 확보하거나, 자연어 처리에 대한 다양한 연구가 필요하다.

참고문헌

[1] P turney, Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, Proceedings of the Twelfth European Conference on Machine Learning, pp. 491-502, 2001

[2] A Brief Survey of Text Mining, Andreas Hotho and Andreas Nu″rnberger and Gerhard Paaß, Journal for Computational Linguistics and Language Technologie 2005

[3] 강한훈, 유성준, 한동일, PMI-IR 기법을 이용한 한국어 감정어 자동 추출 및 성능 개선 방법, Proceedings of KIIS Spring Conference, 2010

[4] 임지연, 김이준, PMI-IR을 이용한 국내 소셜커머스 상품 평가, 한국정보과학회, 2011