제25회 한국정보처리학회 춘계학술발표대회 논문집 제13권 제1호 (2006. 5)

웹 검색을 이용한 한글대역어에 대한 영어약어의 중의성 해소

구희관*, 정한민**, 강인수**, 성원경**
*과학기술연합대학원대학교 응용정보과학전공
**한국과학기술정보연구원 차세대유통개발팀
e-mail : {hkkoo, jhm, dbaisk, wksung}@kisti.re.kr

Web-based disambiguation of English Abbreviation for Korean Term

Hee-Kwan Koo*, Hanmin Jung**, In-Su Kang**, Won-Kyung Sung**
*Practical Information Science, UST
**Information System Research Lab., KISTI

요 약

특정 신문은 해당 도메인의 언어자원을 구축하는데 필요한 자원이며, 한글과 영어의 괄호를 통해 표현되는 대역어구는 다국어 정보로 언어자원 구축에 이용된다. 그러나, 실제로 신문에서 사용되는 한영대역어의 구성은 한글대역어와 영어약어로 구성된 비율이 80%이상을 보인다. 신문을 대상으로 대역어사전 등을 구축하기 위해서는, 영어약어의 완전한 형태인 영어비약어 정보가 필요하다. 본 논문은 영어비약어 정보를 획득하기 웹검색을 통해 영어비약어를 획득하고, 영어약어를 이용해 영어약어와 영어비약어의 관계를 이용하는 방법을 제안한다.

1. 서론

기존 언어자원의 구축 및 유지의 가장 큰 이슈 중에 하나는 최신성의 반영이다. 과학 기술이 빠르게 변화, 발전하면서 언어자원의 최신성을 잘 반영할 수 있는 방법으로 특정 도메인의 신문 말뭉치로부터 정보를 추출하는 것이 필요하다. 또한 말뭉치에서 대역 패턴을 이용하여 추출한 한영대역쌍들은 대역어사전으로 구축하여 활용성을 증가시킬 수 있다.

신문에서 추출한 대역어쌍들은 영어약어의 사용비율이 크고, 중의성이 높기 때문에 언어자원구축에 이용하기에는 어려움이 있다. 이렇게 영어약어가 빈번하게 신문에서 출현하는 이유는 크게 신문 글자수 제한과 의미 전달의 집약된 형태 유지를 하기 위해 쓰는 것으로 보인다. 예를 들어, “세계무역기구(WTO), 세계관광기구(WTO)”와 같은 형태로 신문에서 출현하는 것을 자주 볼 수 있다. “세계무역기구”와 “세계관광기구”는 다른 의미의 대역어이지만 영어약어가 같기 때문에 영어약어를 기준으로 군집화를 하면 같은 군집에 포함된다. 이들의 의미를 구별하기 위해서 대역어에 대응하는 영어비약어가 필요하다.

영어약어를 기준으로 대역어들을 정렬했을 때 보여주는 또 다른 특징은, 하나의 영어약어에 여러 개의 비슷한 의미의 다른 한글대역어가 생성이 된다는 것이다. 그 이유는 과학기술과 관련해서 기사를 작성한 사람이 한 사람이 아니라 여러 전문가들이 초고를 작성하기 때문으로 보인다. 예를 들어, “NII”를 기준으로 정렬을 해보면 “국가정보기반구조”, “국가정보기간망”, “국가정보인프라”, “국가정보통신기반구조”, “정보통신기반구조”와 같은 대역어들이 모두 같은 영어비약어에 대해 대역어로 사용되고 있는 것을 알 수 있다. 또한 영어비약어를 획득여부를 기준으로 신문에서 자동 추출된 한글대역어들의 오류들도 줄이는 역할을 한다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련연구, 3 장에서는 웹검색을 이용한 대역어 원어비약어 획득, 4 장에서는 영어약어와 영어비약어의 관계 확인, 5 장에서는 실험 및 결과에 대해 기술한다. 6 장에서는 결론 및 향후 연구과제에 대해 논의한다.

2. 관련연구

이재성[1]에서 제안한 대역어구 인식방법은 기본적으로 사전을 이용하는 방법을 바탕으로 사전에 존재하지 않는 단어들을 처리하기 위해 음운유사도 일치, 대역어 부분 일치 등의 방법을 제안하였다. 이 대역어구 인식방법에는 기본적으로 사전이 필요하며, 대응되는 영어의 형태가 완전해야만 한다. 그러나 신문에서 추출된 대역 관계에서는 영어약어로 구성되는 비율이 높기 때문에 신문말뭉치에는 바로 적용하기 힘들다.

강재호[2]은 한글과 영어대역어의 웹검색결과 내에 대역관계가 표현된 형태를 이용하여 검색결과 페이지에서 대역어구를 추출하는 방법을 제안하였다. 이 방법은 무작위적인 대역관계를 추출하기 때문에 특정 영어비약어를 추출 할 때는 적합하지 않다.

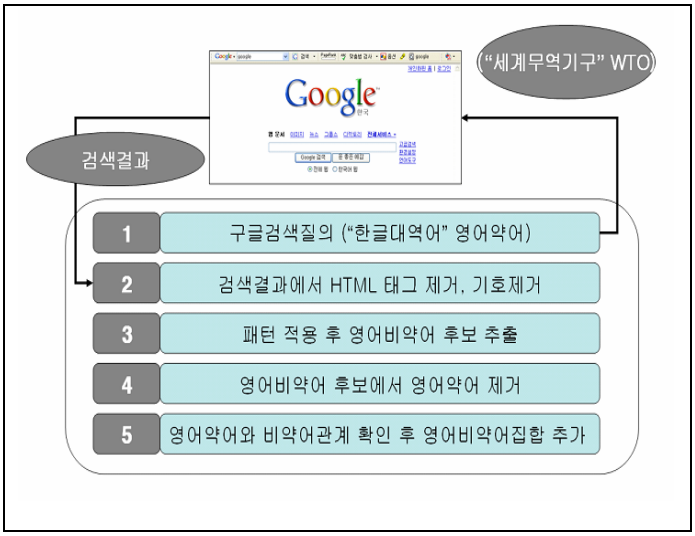
Manuel Zahariev[4]은 언어적인 정보를 이용한 영어약어와 영어비약어의 관계에 대한 확인 방법을 제안하였다. 두 개의 문자열 간에 최대의 문자열을 측정하는 방법인 최대공통문자열(Longest Common Subsequence)을 기반으로 언어적인 음절 별 가중치를 계산하여, 이를 순위화하는 방법을 제안하였다. 이 방법은 정확하게 약어의 철자가 영어비약어의 어느 철자와 대응되는지를 판별하는 방법으로 영어약어와 영어비약어의 판별에는 적합하지 않다.

3. 웹검색을 이용한 대역어 원어비약어 획득

대역어의 원어비약어 획득에는 구글[6][1]을 이용하여 검색결과에서 패턴을 적용하고 영어비약어를 수집한다. 구글의 검색어는 한글대역어와 영어약어를 함께 질의하며, 한글 대역어는 검색어를 따옴표로 묶어 사용하는 문장검색을 이용하여 질의한다. 예를 들어, WTO군집의 세계무역기구 질의어는 ""세계무역기구" WTO"이다.

그림 1 은 한글대역어, 영어약어, 검색페이지의 수를 인자로 받아와 영어비약어를 획득하는 과정을 단계별로 보여준다. 한글대역어로 검색을 하고, 검색결과 페이지(result)를 가져와 HTML 태그를 제거하고 기호(> < " · &)를 변환한다. 검색결과만을 이용한 이유는 검색어를 중심으로 비교적 가까운 거리에 영어비약어가 존재한다고 가정하였기 때문이다[5]. 또한, 구글 검색결과는 구글이 미리 검색어에 대해 전처리를 하고 검색결과에서도 기호와 공백은 무시하기 때문에 한글결과에 대한 공백에 대해 고려가 필요했고, 이후 검색결과에 적용할 패턴에서 공백을 고려한 정규식을 사용한다.

그림 1 에 3 번째 단계는 대역패턴을 적용해서 영어비약어를 추출하고 영어비약어 후보집합에 추가한다. 그림 1 의 4 번째 단계는 영어비약어 후보집합 내 요소들에 존재하는 영어약어를 확인하여 제거한다. 이렇게 따로 영어약어를 제거해야 하는 이유로는 대역패턴으로 검출된 영어비약어 후보에 영어약어와 영어비약어가 함께 기술되어 있는 경우가 많았기 때문이다. 예를 들어, 구글의 검색을 "세계무역기구"와 "WTO"를 질의했을 경우 패턴에 의해 검색된 결과를 보면, 다음과 같다. "세계무역기구(World Trade Organization, WTO), 세계무역기구(World Trade Organization : WTO), 세계 무역기구(WTO ; World Trade Organization), 세계무역기구(WTO : World Trade Organization), 세계무역기구(WTO, World Trade Organization)" 등의 결과를 보면 영어약어가 영어비약어 후보에 포함 되는 경우에 대해 약어정보를 이용해 제거한다. 또한, 위의 예에서 보듯, 영어비약어후보 내에 기호들(":", ";", ",", "-")은 구분자로만 가정하여 사용한다. 예를 들어 "아시아 태평양 경제협력체"의 영어비약어후보가 "Asia-Pacific Economic Cooperation" 와 같이 추출이 되었다면 "-"은 전처리되어 "Asia Pacific Economic Cooperation"과 같이 변환되어 처리된다.



(그림 1) 영어비약어 획득 과정

```
# 패턴1 = 대역어(영어비약어)
# 예제) 세계무역기구(World Trade Organization)
KOREAN(UNABBREVIATED_FORM_WITHOUT_SYMBOL)
# 패턴2 = 대역어(영어약어 심볼1 영어비약어)
# 예제) 세계무역기구(WTO : World Trade Organization)
KOREAN(ABBREVIATED_FORM SYMBOL1 UNABBREVIATED_FORM_WITHOUT_SYMBOL)
# 패턴3 = 대역어(영어비약어 심볼1 영어약어)
# 예제) 세계무역기구(World Trade Organization : WTO)
KOREAN(UNABBREVIATED_FORM_WITHOUT_SYMBOL SYMBOL1 ABBREVIATED_FORM)
# 패턴4 = 영어비약어(대역어)
# 예제) World Trade Organization(세계무역기구)
UNABBREVIATED_FORM_WITH_CONSTRAINTS(KOREAN)
# 패턴5 = 영어약어 (영어비약어) 심볼1 대역어
# 예제) WTO(World Trade Organization) : 세계무역기구
ABBREVIATED_FORM(UNABBREVIATED_FORM_WI
```

[1] 구글 웹검색을 사용할 수 있도록 구글이 API를 제공하고 있다. 한 아이디당 하루에 1000 건의 검색이 가능하다.

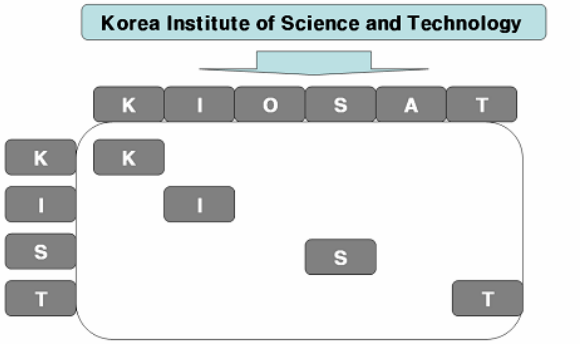
```
TH_CONSTRAINTS) SYMBOL1 KOREAN
# 패턴6 = 영어약어 (대역어) 심볼1 영어비약어
# 예제) WTO(세계무역기구) : World Trade Organization
ABBREVIATED_FORM(KOREAN) SYMBOL1
UNABBREVIATED_FORM_WITH_CONSTRAINTS
# 패턴7 = 영어비약어. 대역어
# 예제) World Trade Organization. 세계무역기구
UNABBREVIATED_FORM_WITH_CONSTRAINTS.
KOREAN
# 패턴8 = 영어약어 (대역어) 영어비약어
# 예제) WTO(세계무역기구) World Trade Organization
PATTERN8 = ABBREVIATED_FORM(KOREAN)
UNABBREVIATED_FORM_WITH_CONSTRAINTS
```

(그림 2) 영어비약어 추출 대역패턴

그림 2 는 그림 1 의 3 번째 단계에서 구글검색결과에 패턴을 적용할 때 사용하는 패턴을 보여준다. 이 패턴들은 패턴 1 을 기준으로 검색결과에서 보이는 패턴들을 수작업으로 추가한 것이다. 패턴들은 로딩 이후, 정규식으로 변환하여 검색결과에 적용된다. 정규식으로 변환되는 예를 살펴보면, WTO 가 영어약어인 "세계무역기구"를 구글에 질의한 결과에 패턴 1 을 적용하여 영어비약어를 추출할 때, KOREAN 은 "세[ ]?계[ ]?무[ ]?역[ ]?기[ ]?구"로 치환된다. 그리고 패턴에 의해 UNABBREVIATED_FORM_WITHOUT_SYMBOL 은 "[a-zA-Z -]*" 으로, UNABBREVIATED_FORM_WITH_CONSTRAINT 는 "[Ww][a-zA-Z]* [Tt][a-zA-Z]* [Oo][a-zA-Z]*"으로, SYMBOL1 은 "';', ',', ':', '-'"으로, ABBREVIATED_FORM 은 "WTO" 등으로 치환이 된다.

4. 영어약어와 영어비약어의 관계확인

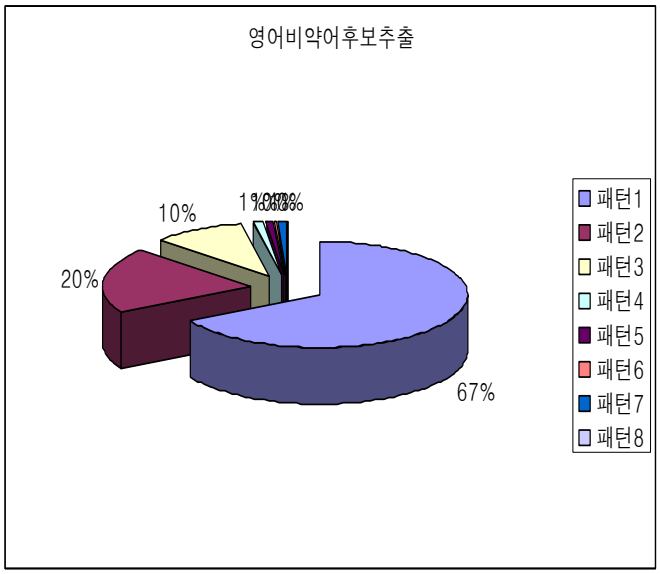
영어비약어의 머리글자와 영어약어가 최대공통문자열을 만족하는 경우를 대역어라 인정한다. 예를 들어, 영어약어가 WTO 이고, 영어비약어 후보가 "World Trade Organization"일 때, 영어약어인 WTO 는 영어비약어의 첫머리 글자와 최대공통문자열이므로 "World Trade Organization"는 WTO 의 영어비약어가 된다. 또한, 영어비약어에 포함되는 전치사와 접속사는 영어약어에 포함되지 않는 경우가 있으므로 영어비약어의 머리글자의 구성에 영어비약어를 구성하는 모든 머리글자를 포함하도록 한다. 그림 3 에서 예로 보이듯이, "Korea Institute of Science and Technology"의 머리글자는 "KIOSAT"이며, "KIOSAT"와 "KIST"의 최대공통문자열이 영어약어를 만족하므로 영어비약어로 인정한다.



(그림 3) 영어약어를 이용한 영어비약어 확인

5. 실험 및 결과

전자신문 말뭉치로부터 괄호를 이용해 표기된 한글과 영어의 대역어쌍을 자동 추출하여 영어를 기준으로 군집된 대역어집합을 구했다. 전체 1,806 개의 대역어 집합 중 한글대역어의 개수가 5 개 이상인 영어약어기준 대역어 집합 100 개를 선택하고, 이 대역어 집합에 포함되어 있는 대역어 1,211 개를 선택해 질의를 하였다. 이 중 영어비약어를 검색이 된 466 개의 한글대역어의 검색결과에 패턴을 적용하여 3701 건의 영어비약어 후보를 수집했다. 그림 3 은 각 패턴별 검색결과내에 영어비약어후보 추출의 발생빈도를 보여주고 있다. 전체 영어비약어후보 추출비율 중 가장 높은 비율을 차지하는 것이 패턴 1 이다.



(그림 3) 영어비약어후보 추출 별
대역패턴 별 발생빈도 (총 3,701 건)

(표 1) 패턴 별 추출건수 및 비율

	영어비약어후보추출 (3단계)	영어비약어추출 (5단계)	정답 개수	정답 비율
패턴1	2475 (66.874%)	324 (22.978%)	318 (23.043%)	98.148%
패턴2	746 (20.157%)	709 (50.283%)	695 (50.362%)	98.025%
패턴3	366 (9.889%)	263 (18.652%)	257 (18.623%)	97.718%
패턴4	42 (1.135%)	42 (2.979%)	41 (2.971%)	97.619%
패턴5	30 (0.810%)	30 (2.128%)	28 (2.029%)	93.333%
패턴6	11 (0.297%)	11 (0.780%)	11 (0.797%)	100%
패턴7	30 (1.135%)	30 (2.128%)	29 (2.101%)	96.667%
패턴8	1 (0.02%)	1 (0.071%)	1 (0.072%)	100%
계	3701	1410	1380	97.689%

표 1 은 패턴별 추출 건수 및 비율을 보여준다. 기존의 주관심 패턴이었던 패턴 1 이 전체 영어비약어후보 추출 중에 높은 비율을 차지하지만 영어비약어를 추출하는 데 효율이 좋지 못한 것을 보여준다.

정답비율은 영어비약어 추출 분에 정답개수를 나타낸 것이다. 97.68%로 나온 이유 중에 하나는 영어약어를 이용하여 영어비약어 추출과정 중에 제약을 걸었기 때문이라고 파악이 된다. 패턴 1 에서 영어비약어후보에서 영어비약어로 추출되지 않은 대부분을 차지하는 것은 영어약어로만 구성되어 있는 것들이며, 그 외에 상당부분 괄호를 이용한 의미적인 연결이 대등하지 않은 경우였다.

전체 대역어 1,211 개 중에 466 개의 영어비약어를 추출하는 데에 그쳤는데, 그 이유를 보면 신문에서 자동추출된 대역어와 영어약어의 추출 당시의 규칙을 괄호 앞의 명사들의 결합을 이용한 추출오류가 상당부분 작용을 하여 검색결과에서 발견할 수 없는 경우가 대부분이 이었다. 예를 들어 WTO 군집에서는 “반도체시장개방세계무역기구, 세계경제가세계무역기구” 등은 영어비약어가 발견되지 않아 자동추출의 오류를 줄일 수 있는 영어비약어의 발견부분이 상당부분 순기능적인 역할을 하는 것으로 보인다.

영어비약어의 획득으로 인해 100 개의 군집 중 22 개의 군집에 한글대역어들을 분리해 낼 수 있었다. 예를 들면 BIS 군집의 “국제결제은행(Bank for International Settlements) “과 “버스 정보 시스템 (Bus Information System)” 등을 들 수 있다. 다만 DVD 군의 “디지털 다기능 디스크(Digital Versatile Disc)”과 “디지털 비디오 디스크(Digital Video Disc)”는 같은 의미로 사용되나 영어비약어 추출로 인해 분리되는 결과가 나왔다.

100 개 군집의 한글대역어를 영어비약어 추정을 통해 183 개의 원어를 파악해 보면, 영어비약어를 처리할 때, 전처리의 필요성이 있는 기호가 ‘-‘ 이었다. 예를 들어, ASIC 이나 EMC 등은 “Application-Specific Integrated circuit”이나 “Electro-Magnetic Compatibility” 와 같은 구분자의 역할을 하는 경우가 7 건이었고, BPR 과 “Business Process Re-engineering” 같이 구분자로 처리하지 않아야 하는 경우도 2 건 있었다. 전체 비중이 제일 많은 형태는 두문자어 형태로 총 156 건이었으며, 형태는 “WTO”와 “World Trade Organization” 가 그 예에 든다. 이것과 유사한 생략된 두문자어형태가 있는데 그 예로는 “PACS”와 “Picture Archiving and Communication System”과 같은 전치사, 접속사 등이 포함된 원어에서 생략된 경우가 9 건이었다. 그리고 생략된 두문자어 형태 중 명사가 생략된 형태가 2 건 있었다. 위의 모든 경우는 원어비약어의 머리글자와 영어약어의 최대공통문자열로 파악이 가능하다. 그러나 영어비약어의 머릿글자보다 영어약어의 글자수가 적을 때를 검출하지 못했는데 총 9 건이 이에 해당 되었다. 예를 들어 “DBMS”는 “Database Management System”의 약어인데, 현재의 머리글자 비교방법을 보면 “DBMS”를 “DMS”와 비교하게 되어 지금의 알고리즘으로는 판별할 수 없었다. 향후 이 문제점은 더 세분화 하여 처리해야 할 필요가 있다.

6. 결론 및 향후 연구

본 논문은 신문 말뭉치로부터 대역 패턴을 이용하여 추출한 영한 대역쌍들에 대해 영어약어의 완전한 형태인 영어비약어 정보를 획득하기 위해 웹검색을 통해 영어비약어를 획득하고, 영어약어를 이용해 영어약어와 영어비약어의 관계를 이용해 검증하는 방법을 제안하였다. 연구 결과를 대역 사전과 다국어 시소러스를 포함하는 다양한 지식 기반에 적용할 예정이다.

참고문헌

[1] 이재성, 서영훈 “한영 혼용문에서 괄호 안 대역어 구의 자동 인식”, 한국정보처리학회 논문지 B, 9 권, 4 호, pp.445-452, 2002

[2] 강재호, 김종성, 류광렬 “패턴생성을 통한 인터넷 문서의 한글-영문용어 추출”, 30 권, 2-1 호, pp. 148-150, 한국정보과학회 2003 년 추계학술대회, 2003

[3] 구희관, 정한민, 이미경, 성원경 “형태정보를 이용한 대역어 군집화 및 적합대역어 선정”, 32 권, 2 호, pp.532-534, 32 회 한국정보과학회 2005 년 추계학술대회, 2005

[4] Manuel Zahariev “A Linguistic Approach to Extracting Acronym Expansions from Text”, Knowledge and Information Systems, V.6,pp.366-373, 2004.

[5] Masaaki Nakata “Using the Web as a Bilingual Dictionary”, ACL 2001 Workshop Data-Driven Methods in Machine Translation, pp.95-102, 2001

[6] Google Web API, http://www.google.com/apis/