LCS알고리즘을 이용한 한-영 대역어 추출 연구

박은진*, 양성일*, 김영길*
*한국전자통신연구원 음성/언어연구센터 언어처리연구팀
e-mail : {ejpark, siyang, kimyk}@etri.re.kr

A Study on extraction for Korean-English word pair by using LCS algorithm

Eun-Jin Park*, Seong-Il Yang*, Young-Kil Kim*
* NLP Team, Speech/Language Technology Research Center, ETRI

요 약

매일 생성되는 웹 신문에서 독자가 접해보지 못한 단어는 독자의 이해를 돕기 위하여 괄호를 사용한다. 괄호를 사용하여 표기된 웹 신문의 한국어-영어 대역쌍은 특정 기사에는 출현빈도가 낮지만 전체적으로 여러 신문의 기사를 봤을 때, 최소한 한번 이상 출현하게 된다. 즉, 괄호 안의 동일한 영어 용어 두 개 이상의 문장을 최장일치법 알고리즘에 적용하면 한국어 단어 경계를 자동으로 인식할 수 있다. 본 논문에서는 이런 웹 신문의 괄호 표기 특성을 이용하여 한-영 대역어쌍을 추출하는 방법을 제안한다. 웹 신문 기사 43,648 건에서 최대 2,087 개의 한-영 대역어를 추출하였다. 3 개의 서로 다른 테스트 그룹으로 실험한 결과 최대 84.2%의 정확도를 보였다.

1. 서론

일반적으로 문서 내에서 괄호는 앞의 내용을 좀더 서술하거나, 약어를 풀어 쓴 경우, 동의어를 나타낸 경우, 하나의 분리 단위로 나타내기 위한 경우, 수식의 일부로 사용하는 경우, 한글 단어의 영어 대역어를 사용한 경우 등이 있다[1]. 또한 웹 신문에서는 통계 수치를 나타낸 경우, 한자, 새로운 용어를 설명한 경우, 회사의 대표를 표시한 경우, 회사 홈페이지 주소를 표시한 경우 등이 더 있다.

일반적으로 웹 신문에서는 독자가 접해보지 못한 단어를 표기할 때, 독자의 이해를 도울 목적으로 괄호를 사용하여 새로운 용어를 표기한다. 예를 들어 <표 1>과 같이 신문 기사에서는 'HSDPA'나 '첩 스펙트럼 확산'이라는 새로운 용어를 설명하기 위하여 부가적으로 괄호 안에 한글 혹은 영어를 표기한다.

<표 1> 웹 신문 기사[1]

SK 텔레콤(대표 김신배)은 KT 에 이어 와이브로와 HSDPA(고속하향패킷접속) 서비스를 동시에 이용할 수 있는 'T 로그인(와이브로 · HSDPA) 통합단말기' 를 출시한다고 5 일 밝혔다. 통신장비 전문업체 오소트론(대표 이경국)은 자체 개발한 2.4GHz 첩 스펙트럼 확산(Chirp-Spread-Spectrum, CSS) 기술이 지난달 말 IEEE 표준화위원회 심의에서 물리계층의 새 표준으로 최종 승인됐다고 발표했다.

이러한 신조어는 하나의 기사에는 한번 이상 나타나기 어렵지만 전체적으로 여러 신문에서는 최소한 한번 이상 나타난다.

여기서, 최소 한번 이상 동일한 괄호 안 용어(영어)를 설명하는 문장에서 두 개의 문장의 겹치는 정도를 측정하면 한국어 단어 경계를 인식할 수 있다. 예를 들어 <표 2>에서와 같이 괄호 안 단어 'ERP'에 대하여 (1-4)번 문장의 겹치는 정도를 측정하면 한국어 대역어 '전사자원관리'를 인식할 수 있다.

<표 2> 용어 '전사자원관리'가 나타난 기사

(1) … KRG 는_국내기업들의_**전사자원관리**(ERP) …
(2) …중견기업을_대상으로_한_**전사자원관리**(ERP)…
(3) …중국에서_**전사자원관리**(ERP)…
(4) …성장을_거듭해왔던_**전사자원관리**(ERP)…

이렇게 (1-4)번 문장을 역으로 두 문장씩 비교하며 가장 겹치는 부분이 많은 문자열을 추출하는 방법을 최장일치법(LCS) 알고리즘[2]이라고 한다.

본 논문에서는 이러한 웹 신문의 신조어 표기 형식 특성과 최장일치법 알고리즘을 이용하여 한-영 대역어쌍을 추출하는 방법을 제안한다.

본 연구와 관련된 연구로, 괄호 안 영문에 대응하는 한국어의 범위를 인식하는 문제에 관한 연구가 있었다[1]. 이 연구에서는 한국어 범위 인식률을 높이기 위하여 음운유사도와 복합어를 포함한 대역어 부분일치법을 혼용하여 사용하였다. 다른 연구로는 한글-영문용어를 추출하면서 이를 패턴화하고, 생성된 패턴을 이용하여 인터넷 문서에서 번역용어사전을 자동으로 추출하는 연구가 있었다[2].

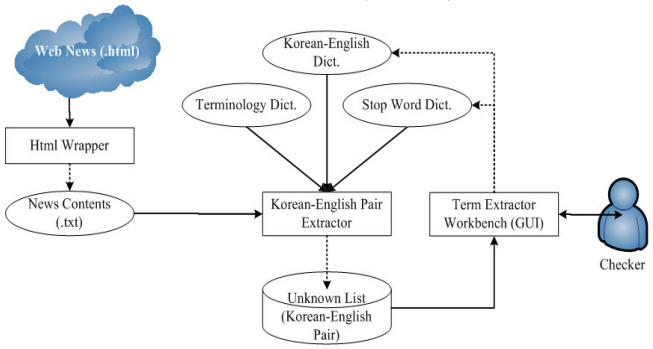
본 논문의 구성은 우선 2 장에서 한-영 대영어 추출 과정에 대해 설명하고 3 장에서는 실험 및 향후 연구에 대하여 언급한다.

[1] http://www.etnews.co.kr/

[2] LCS(Longest common subsequence) 알고리즘은 두 개의 임의의 문자열 사이의 최대 공통 분모를 찾는 알고리즘이다

2. 한-영 대역어 추출 과정

한-영 대역어 추출 과정은 (그림 1)과 같다.



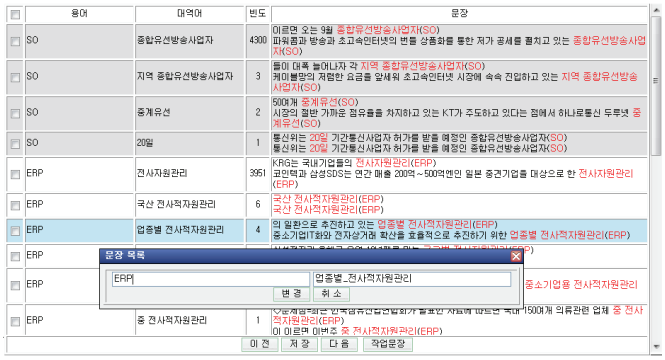
(그림 1) 한영 대역어 추출 과정

(그림 1)에서 Html Wrapper 는 웹 텍스트 형식의 문서에서 우리가 필요로 하는 기사 원문만을 추출하여 이를 텍스트 형식으로 저장한다. Korean-English pair Extractor 는 앞에서 추출한 텍스트 형식의 기사에서 <표 3>의 과정으로 괄호에 표기된 새로운 용어를 추출한다.

<표 3> 한영 대역어 추출 과정

1. 문장에서 닫힘 괄호())와 인용문 (‘’, “”) 기준으로 문장을 분리
2. 괄호 안의 불용어[3] 제거
3. 최장일치법 알고리즘 적용
4. 전문용어 및 한-영 대역어 사전의 표제어 제거

최장일치법 알고리즘을 적용할 때, 괄호 안의 영어 단어를 기준으로 가능한 모든 조합을 이용한다. 즉, <표 2>에서 (1-2),(1-3),(1-4),(2-3),(2-4),(3-4)와 같은 순으로 비교하여 한글 대역어를 추출한다. 이렇게 추출한 한-영 대역어 쌍을 (그림 2)와 같은 UI 를 통해서 한-영 대역 사전에 추가한다.



(그림 2) 검수자 평가 GUI

(그림 2)에서 선택 후 저장한 단어는 Korean-English Dict.(한-영 대역어 사전)에 저장되고 나머지는 Stop Word Dict.(불용어 사전)에 저장된다. 나중에 Stop Word Dict.은 추가 추출 데이터에서 정제되어 대역어 추출 정확도를 높이는데 사용한다. 평가를 목적으로 검수자가 검토할 때, 적합한 대역어가 아닐 경우, 용어나 대역어를 선택하고 적합한 대역어를 입력할 수 있도록 하였다.

3. 실험 및 향후 연구

본 논문에서는 동아일보(21,569 건), 조선일보(2,189 건), 전자신문(19,890 건)의 웹 신문 기사를 실험에 사용하였다. 이 실험 데이터에서 출현한 단어쌍은 모두 133,715 개였고 <표 3>의 과정으로 추출한 한-영 대역어 쌍은 43,564(32.5%)개였다. 괄호 안 영어가 한번 출현한 단어는 14,006(32.1%)개였다. 대역어가 한번 출현한 단어는 최장일치법 알고리즘을 적용하지 못하므로 나머지 29,558(67.9%)개에 대해서 최장일치법 알고리즘을 적용하였다. 적용 결과 최대 2,087 개의 대역어가 추출되었고 이를 (그림 2)와 같은 사용자 GUI 로 사람이 평가하였다.

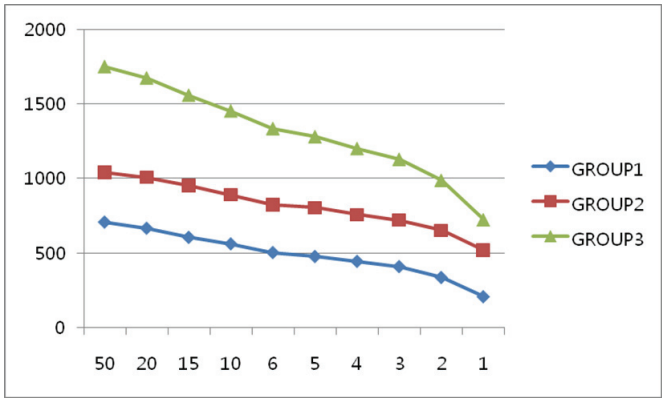
사람이 평가한 결과 중 정확히 자동으로 추출된 746 개의 용어의 특성을 분석해 본 결과, 고유명사 축약 형태가 가장 많은 504(64.3%)개, 순수 한-영 대역어가 203(27.2%)개, 인명이 39(5.2%)개 순으로 나타났다.

자동으로 추출된 한-영 대역어를 분석한 결과, 출현빈도가 낮거나, 하나의 영어 용어에 대해 여러 개의 한글 대역어가 나타났다. 이러한 특성을 고려하여 <표 4>와 같이 3 개의 테스트 집합으로 나누어 실험하였다.

<표 4> 테스트 집합

- GROUP1: 최소 3 개 이상의 문장에서 추출
- GROUP2: 2 개 문장에서만 추출
- GROUP3: 2 개 이상의 문장에서 추출

GROUP1 은 괄호 안 영어 단어가 최소한 3 번 이상 신문 기사에 나타난 집합이고, GROUP 2 은 최장일치법 알고리즘을 적용할 수 있는 최소한의 수준인 2 번만 나타난 용어만을 모은 집합이다. GROUP3 는 GROUP1 과 GROUP2 를 모두 합한 집합이다. (그림 3)은 각 실험 집합 별 추출된 한-영 대역어의 수를 나타낸다.



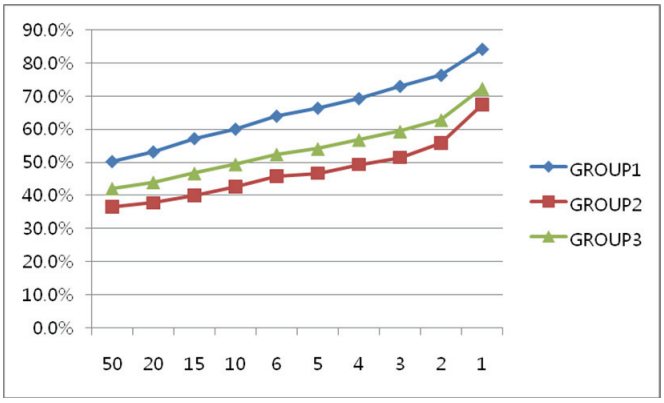
(그림 3) 테스트 집합 별 추출 용어 개수[4]

(그림 4)는 식 (1)을 사용하여 각 실험 집합 별 알고리즘의 성능을 측정한 결과이다.

[3] 웹 주소, 도량형 단위(㎛,㎚, μ ,㎜,㎠,㎱,㎂,㎓,㏀,㎌,㏅), 통계 비율(#,%), 한글-한글 쌍 등

[4] 가로축은 하나의 영어에 대하여 대응되는 한글 대역어의 개수를 나타낸다.

$$\text{Accuracy} = \frac{\text{정확한 대역쌍}}{\text{기계가 추출한 대역쌍}} \quad (1)$$



(그림 4) 테스트 집합 별 성능

(그림 4)의 GROUP1 을 보면, 최소한 3 번 이상 나타난 문장에서 추출하고, 하나의 영어 대역어에 하나의 한글만이 추출된 한-영 대역어의 정확도가 84.2%로 가장 높게 나타났다.

본 논문에서는 매일 생성되는 웹 신문에서 한번 이상 출현한 새로운 용어를 최장일치법 알고리즘을 이용하여, 한국어 대역어 부분을 추출하는 방법을 제안하였다. 웹 신문 기사 43,648 건에 대하여 본 논문에서 제안한 방법으로 추출한 결과, 최대 84.2%의 정확도를 보였다. 앞으로, 이 연구를 바탕으로 RSS 를 이용한 실시간 추출 시스템 연구 등의 연구가 필요할 것이다.

참고문헌

[1] 이재성, 서영훈, “한영 혼용문에서 괄호 안 대역어구의 자동 인식”, 한국정보처리학회논문지 제 9-B 권 제 4 호, 2002. 8, pp. 445-452.

[2] 강재호, 김종성, 류광렬, “패턴생성을 통한 인터넷 문서의 한글-영문용어 추출”, 한국정보과학회 제 30 권 제 2-1 호, 2003.20, pp.148-150.