조어법 정보를 이용한 전문용어의 영/한 번역 시스템 개발

서충원[O] 배선미 최기선
한국과학기술원 전산학과/전문용어언어공학센터/언어자원은행
{cwseo[O], sBae, kschoi}@world.kaist.ac.kr

English/Korean Terminology Translation System Using Word Formation

Chung-Won Seo[O] Sun-Mee Bae, Key-Sun Choi
Dept. of EECS, KAIST/KORTERM/Bank of Language Resource

요 약

전문용어 조어법 분석은 기존의 전문용어들의 어휘의 구성과 구조를 파악하여 전문용어 생성의 원리를 밝혀 여러 응용시스템에 이용하기 위한 기초 작업이다. 조어법 정보를 이용한 전문용어 번역 시스템은 조어법 분석 결과의 조어단위 정렬과 색인을 통하여, 새로운 영어 용어에 대한 한국어 대역어 후보 집합을 생성한다. 생성된 후보들은 언어 모델의 정보량의 차이를 이용한 가중치에 의하여 순서화된다. 본 논문에서 제안하는 가중치 방법을 이용하여 조어법 분석 결과에 포함되지 않은 용어들을 대상으로 성능을 평가했을 때, 영-한 조어단위 번역의 n-best 정확률에서 1순위 정확률은 약 61%, 10순위 정확률은 97%의 성능을 보였다.

1. 서 론[1]

매일 새로운 기술의 개발과 함께 수많은 분야에서 전문용어들이 생성되고 있다. 전문분야 문서의 번역이나 정보 검색, 정보 추출을 위하여서는 새로 생성되는 전문용어의 효과적인 처리가 필요하다. 새로 생성되는 전문용어의 경우 새로운 형태의 단어보다는 기존의 용어가 쓰임을 달리하거나 조합되어 생성되는 경우가 많다.

전문용어 조어법 분석은 기존의 전문용어들의 어휘의 구성과 구조를 파악하여 전문용어 생성의 원리를 밝혀 여러 응용 시스템에 사용하기 위한 기초작업이다. KORTERM[2]의 전문용어 정비 사업에서는 물리, 화학, 생물 분야의 용어들을 선정하여 조어법 분석을 수행하였고, 다른 분야로 확대하여가고 있다 [1].

전문용어 조어법 검색 시스템[4]은 새로운 용어가 들어 왔을 때, 기존의 조어법 정보를 활용하여 적절한 한국어 대역어 후보를 생성하여 준다. 이는 작은 범위의 명사구의 번역이라고 할 수 있다. 예를 들어 "air column resonance apparatus"와 같은 용어에서, "column"은 기존의 용어에서, "관", "기둥", "탑" 과 같은 한국어로 번역이 되고 있으며, "공기 기둥 공명장치" 나 "공기 관 공명 장치" 가 될 수 있다. 이 때, 둘 중 어떤 용어가 적합한지에 대한 문제는 의미 애매성 해소의 문제라기 보다는 자연스러운 한국어 표현을 찾는 문제가 된다.

본 논문에서는 전문용어의 번역을 위하여 조어법 정보로부터 수집된 영/한 조어 단위 대역 정보를 이용하여 영어 용어에 대한 한국어 대역어 후보를 생성하고, 언어 모델을 사용하여 순서화하였다. 대역어 후보들에 대하여 n-best 정확률을 평가 했을 때, 1순위 용어들의 경우 약 62%의 정확도를 보였으며, 10 순위 이내의 용어에 대하여 97.7%의 정확도를 보였다.

2. 관련연구

2.1 조어후보 가중치 부여

복합 명사의 번역을 위한 방법은 '의미 해석과 구조 변환 과정을 통하여 번역을 하는 깊은 처리 (deep-processing)'와 '원어의 어휘 구성 정보만을 사용하여 번역을 하는 얕은 처리 (shallow processing)'[2]로 나눌 수 있다.

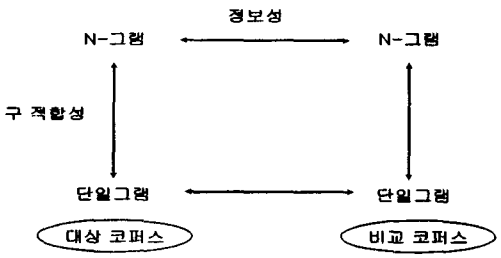
Tanaka 등[2]은 깊은 처리를 통하여 복합 명사의 번역을 수행하였다. '메모리기반 기계번역 (Memory Based Machine Translation) 방법'과 '단어 대 단어 기계번역 방법'으로 일본어와 영어의 복합 명사 번역에서 얕은 처리의 안정성을 검증하였다.

Tomokiyo 등[3]은 단일 언어문서에서 중심구 (key-phrase)를 추출하기 위하여 언어 모델을 통한 가중치 값을 사용하였다. 대

[1] 이 논문은 과학기술부, 과학재단의 지원을 받았음.
[2] KORTERM(Korea Terminology Research Center for Language and Knowledge Engineering)

상 코퍼스 (foreground corpus)와 비교 코퍼스 (background corpus)로부터 얻어진 각각의 언어 모델에서 구한 정보량의 차이로 중심 구의 정보성 (Informativeness)과 구 적합성 (Phraseness)을 구하고 있다. 정보성은 해당 구가 얼마나 분야 특성적인 정보를 포함하고 있는지를 나타내며, 구 적합성은 추출된 구가 실제 쓰임에서 구로 묶여서 사용되는 정도를 나타낸다.

그림 1 정보성과 구 적합성



전문용어의 번역에서 좋은 대역어 후보를 찾는 문제는, 정보의 전달이 명확하고 실제 코퍼스에서 자주 나타나는 표현을 찾는 문제다. 대상 코퍼스를 해당 전문 분야로 한정했을 때, 정보성과 구 적합성은 이 두 가지를 반영한다고 할 수 있다.

3. 전문용어 조어법

3.1 조어법 데이터 정렬

조어법 데이터는 표1과 같은 형식으로 구축되어 있다.

표 1 조어법 데이터[3]

영어 용어	한글 용어	분석	형태	원어		분석	형태	원어
tunable dye laser	가변 영료 레이저	가변	nn	ch	...	레이저	nn	ie

다음과 같은 가정을 사용하여 영어와 한국어의 조어 단위를 정렬하였다.

가설 1. 교차 정렬은 존재하지 않는다.

가설 2. 조어 단위는 영어 단어가 중심이 된다.

이를 용어 어휘 구설 정렬의 가정이라고 한다.

전체 용어의 정렬을 위하여서는 영어와 한국어의 어순의 차이에 의한 교차 정렬도 생각하여야 한다. 하지만, 명사구만을 대상으로 했을 경우 구조가 거의 유사하며, 실제 데이터에서 영어와 한국어의 구조가 다르게 나타나는 용어는 5% 이하였다.

3.2 조어 후보의 순서화

조어법 검색 결과의 대역어 후보는 순위 없이 나열된다. "ideal gas equation (이상 기체 방정식)" 의 출력 결과는 다음과 같다.

이상 기체 방정식

이상 기체 식

이상 기체 상태

이상 가스 방정식

이상 가스 식

이상 가스 상태

...

"*ideal gas equation*"는 한국어 용어로 "이상 기체 방정식"이며, 첫 번째 후보로 나타나고 있다. 하지만, "air cycle refrigeration (공기 순환 냉동)"의 결과는 다음과 같다.

공기 고리 냉동

공기 고리 냉각

공기 고리 냉장

공기 순환 냉동

...

이 경우는 4번째에 정답이 나타나고 있다. 여기서, "공기 고리 냉동"이나 "공기 고리 냉각" 같은 경우는 의미적으로는 유추가 가능하지만, 어색한 표현이다. 따라서 "공기 순환 냉동"과 같은 용어를 찾아내는 일이 필요하다.

본 논문에서는 언어 모델 사이의 정보량의 차이를 사용하여 정량화한 정보성과 구 적합성을 사용하여 용어를 순서화 하였다. 정보량의 차이는 점마다 KL-발산 (pointwise KL-divergence) (식 (1))[3]을 사용하여 계산 하였다.

단어 열 w에 대한 서로 다른 확률 모델 p, q 의 점마다 KL-발산 δ는 다음과 같은 식에 의하여 정의 된다. 이는 w가 q의 분포를 갖는다고 가정했을 때, 실제로는 p의 분포로 나타날 경우 생기는 확률 모델 사이의 정보량의 차이를 나타낸다.

$$\delta_w(p \| q) \stackrel{def}{=} p(w)\log\frac{p(w)}{q(w)} \quad (1)$$

이때, 구 적합성과 정보성은 점마다 KL-발산 δ에 의하여 다음과 같이 나타낼 수 있다.

구 적합성:

$$\delta_w(LM_{fg}^{N} \| LM_{fg}^{1}) \quad (2)$$

[3] nn: 명사(품사), ch: 한자어(원어), ie: 외래어(원어)

정보성:

$$\delta_w(LM_{fg}^N \| LM_{bg}^N) \quad (3)$$

여기서 LM_{fg}는 대상 분야에서 수집한 언어 모델이고, LM_{bg}는 비교 분야에서 수집한 언어 모델이다. 구 적합성은 대상 코퍼스의 N-그램과 단일그램의 정보량의 차이로 나타낸다. 정보성은 대상 코퍼스와 비교 코퍼스의 N-그램의 정보량의 차이로 나타낸다.

4. 실험 및 결과

4.1 실험 환경

평가를 위하여서 물리학 용어 사전의 3단어 이상으로 구성된 영어 표제어 중에서 조어법 분석 결과에 포함되어 있지 않은 용어 50개를 무작위로 추출하였다.

대상 코퍼스는 물리분야 전문 분야 문서를 형태소 분석기를 이용하여 태깅한 약 94만 어절의 문서를 사용하였다. 비교 코퍼스는 수동으로 품사 태깅된 100만 어절 일반 분야 코퍼스와 형태소 분석기를 이용하여 품사 태깅한 19만 어절 분량의 화학, 생물 분야 전문분야 문서를 사용하였다.

4.2 평가

정보성과 구 적합성의 계산을 위하여 단일그램, 이진그램, 삼진그램을 수집하여 조합을 하여 사용하였다. (표2) 기준치(baseline) 은 대상 분야의 이진그램만으로 가중치를 부여한 모델을 사용했다. 대역어 후보 중에서 한국어 용어와 일치하는 것이 없는 것은 8개로, 적용률은 84%이다. 평가는 후보 집합에 정답이 포함된 42개의 용어를 대상으로 n-best 정확률을 계산하였다.

표 2 n-best 정확률 (In: 정보성 Ph:구 적합성)

	1 순위	3 순위	5 순위	10순위
Baseline	59.52 (%)	71.43	80.95	85.7
Ph (bi)	58.14	78.57	85.71	97.62
Ph (tri)	59.52	76.19	80.95	92.86
In (uni)	19.04	50.00	69.05	78.57
In (bi)	45.24	69.05	71.43	88.10
In (tri)	61.90	73.81	83.33	90.48
Ph (bi)+In (tri)	42.86	73.81	82.93	88.10

1순위 정확도만 본다면 삼진그램을 사용한 정보성이 가장 높은 성능을 보인다. 3순위 이상의 정확도에서는 이진그램을 사용한 구 적합성이 가장 높은 성능을 보였다.

전체적으로 보았을 때, 정보성이 구 적합성에 비하여서 정확도가 낮다. 이는 대상으로 하는 대역어 후보들이 물리용어에서 주로 사용되는 형태소로 구성이 되어 있어서 정보성의 차이가 크지 않기 때문이다.

5. 결론 및 토의

KORTERM 전문용어 정비사업의 일환으로 구축된 전문용어의 조어법 정보[1]를 이용하여 전문용어의 한국어 대역어를 제시하는 시스템을 구축하였다. 기존의 조어법 정보에서 수집된 영/한 대역 정보를 이용하여 한국어 대역어 후보를 제시하였고, 이를 가중치 부여를 통하여 순서화하였다. 한국어 대역어 집합에 대한 n-best 정확률을 계산했을 때, 1순위 정확률에서 약 62%의 정확률을, 10순위 정확률에서는 97.6%의 정확률을 보였다.

본 논문에서는 대상 코퍼스와 비교 코퍼스의 언어모델 사이의 정보량의 차이를 이용한 정보성과 구 적합성을 통하여 전문용어의 대역어 순위화에 대한 정량적인 기준을 제시하였다. 단순히 언어모델만을 사용하여 가중치를 부여한 결과보다 1-best 정확도에서는 2.4%, 10-best정확도의 경우 11.9%의 정확도를 향상시켰다.

오류 중에서 조어정렬의 실패로 대역어를 생성하지 못하는 경우와 정렬 오류로 인한 잘못된 대역어 선택은 적용률을 낮춘다. 적용률을 높이기 위하여 일반 분야의 영-한 대역 사전 등을 이용하는 방법의 적용이 필요하다.

향후 연구로, 조어법 정렬의 확장과 성능을 개선하기 위한 연구가 필요하다. 본 논문에서 배재하고 있는 교차 정렬의 문제는 후치 수식어구에 의한 차이와 어휘 특성에 따른 결합의 선호에 의한 어순 변화가 있다. 이는 전문용어의 구조에 대한 연구를 통하여 교차 정렬의 문제를 해결하여야 할 것이다.

참고 문헌

[1] 최기선, 전문용어의 정비, 21세기 세종계획 과제보고서, 문화관광부 국립국어연구원, 2002.

[2] Tanaka, Takaaki and Timothy Baldwin, "Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing", ACL-2003 workshop on Multiword Expression. 2003

[3] Tomokiyo, Takashi and Matthew Hurst, "A Language Model Approach to Keyphrase Extraction", ACL-2003 workshop on Multiword Expression. 2003

[4] http://gensum.kaist.ac.kr/~cwseo/MunHwa, 조어법 검색 시스템