정렬기법을 이용한 전문분야 조어단위 대역쌍 추출

오종훈°, 황금하, 최기선
한국과학기술원 전자전산학과/전문용어언어공학연구센터
{rovellia°, hgh, kschoi}@world.kaist.ac.kr

An Alignment method for Extracting English-Korean translations of term constituents

Jong-Hoon Oh°, Jin-Xia Huang, Key-Sun Choi
Department of EECS
Korea Advanced Institute of Science and Technology/KORTERM

요 약

전문용어는 전문분야의 개념을 표현하는 언어적 표현이다. 전문용어의 조어단위는 전문용어를 구성하는 최소의 형태적 단위이다. 이러한 조어단위는 전문용어의 의미를 파악하는데 중요할 뿐만 아니라 전문분야 문서에 대한 기계번역과 같은 작업에 중요한 언어자원으로 사용될 수 있다. 하지만 '조어단위와 개념단위의 불일치 문제', 조어 단위의 '동형이의어', '동의이형어' 문제 등으로 인하여, 하나의 전문분야 개념을 나타내는 조어단위들의 덩어리를 파악할 필요가 있다. 본 논문에서는 이러한 문제점을 조어분석된 한영 대역 전문용어사전에 대한 한국어-영어 조어단위 정렬문제로 해결하고자 한다. 본 논문의 기법은 97%의 정확률로 조어단위 간의 정렬을 수행하였다.

1. 서론

전문용어는 전문분야의 개념을 표현하는 언어적 표현이다. 전문용어의 조어단위는 전문용어를 구성하는 최소의 형태적 단위이다[1]. 전문용어는 전문용어를 구성하는 조어단위의 개수에 따라 '단일 용어'와 '복합용어'로 나누어진다[2]. '단일용어'는 하나의 조어단위로 구성된 용어인 반면, '복합용어'는 두 개 이상의 조어단위로 구성되는 용어이다. 한국어 전문용어의 대부분은 복합용어의 형태를 가진다[3]. 대부분의 복합용어는 조어단위들의 결합에서 합성성을 따르는 투명한 용어이다. 즉, 조어단위의 개념들의 결합으로 해당 전문용어의 의미를 유추할 수 있다. 복합용어의 조어양상을 파악한다는 것은 전문용어를 구성하는 조어단위의 개념간의 연결관계를 파악하는 것을 의미한다. 따라서 전문용어의 의미를 파악하기 위해서는 이러한 조어단위에 대한 개념을 올바르게 파악하는 것이 중요하다.

하지만 한국어 전문용어 조어단위의 특성으로 인하여 나타나는 몇 가지 문제점으로 인하여, 한국어 조어단위만을 이용한 전문용어의 의미파악은 쉽지 않다.

첫 번째 문제점은 '조어단위와 개념단위의 불일치 문제'이다. 개념단위란 하나의 전문분야 개념을 표현하는 언어적 단위로 정의된다. 하나의 조어단위는 그 자체로 개념단위가 될 수 있지만, 여러 개의 조어단위가 개념단위로 사용되는 경우가 많다. 또한, 조어단위의 결합 양상에 따라 같은 조어단위라도 그 자체가 개념단위로 사용되는 경우와 다른 조어단위와 결합하여 개념단위로 사용되는 경우가 있다. 예를 들어 '효소'는 그 자체로 'enzyme'을 나타내는 개념단위가 되지만, '가수분해효소'에서는 '가수', '분해', '효소'가 결합하여 'hydrolase'를 표현하는 개념단위를 형성한다. 따라서 복합용어에서 전문용어의 의미를 효과적으로 파악하기 위해서는 개념단위를 파악하는 작업이 필요하다.

두 번째 문제점으로 조어단위의 '동형이의어 문제'이다. 한국어 전문용어의 많은 부분은 한자어나 외래어로 구성되어 있다. 특히 한국어 전문용어는 영어 용어와 비교할 때 한자어 조어력으로 인해 구 용어보다 단어 형태가 더 선호된다. 이는 한자어에 있어서 단일 명사뿐만 아니라 한자어 접사가 전문용어의 개념요소로서 사용됨을 의미한다. 예를 들어, '기'와 같은 접사는, 생물학 분야에서 표 1과 같은 다섯 가지 의미로 사용된다. 표 1에서와 같이 접사 '기'의 올바른 의미를 파악하지 않으면, 전문용어의 의미를 올바르게 해석할 수 없다. 따라서 조어단위의 의미모호성 해결은 전문용어의 의미파악에 매우 중요하다.

표 1 생물학 분야에서 접사 '기'의 의미

의미	한국어	영어
Group	아미노기	Amino *group*
Period	수축기	Contraction *period*
Stage, phase	생장기	Growth *phase*
Organ	후각기	Olfactory *organ*
器 (도구)	반응기	Effect*or*

세 번째 문제점은 조어단위의 '동의이형어 문제'이다. 동의이형어는 전문용어의 많은 부분이 영어 등과 같은 외국어에 기원을 두고 있기 때문에 발생한다. 외국어에 기원을 둔 전문용어는 다양한 방법으로 한국어로 번역되거나 표기된다 - 1) 고유어로 번역 2) 한자어로 번역 3) 음차표기 4) 원어. 이러한 다양성으로 인해 같은 의미를 가진 조어단위가 여러 가지 형태로 사용되는 경우가 있다. 예를 들어, 'abdominal'은 표 2와 같이 '복부', '복', '배' 등으로 다양하게 번역된다.

표 2 생물학 전문용어사전에서 'abdominal'의 번역형태

'abdominal'의 번역어	한국어	영어
복부	복부부속지	abdominal appendage
복	복강	abdominal cavity
배	배지느러미	abdominal fin

본 논문에서는 이러한 문제점을 조어분석된 한영 대역 전문용어사전에 대한 한국어-영어 조어단위 정렬문제로 해결하고자 한다. 첫째, 개념단위 인식 문제는 한국어-영어 용어의 조어단위 간의 대응관계를 파악하는 문제로 정의될 수 있다. 즉, 영어 조어단위를 개념단위로 정의하고, 영어 조어단위에 대응되는 한국어 조어단위의 집합을 파악하는 문제로 정의된다. 이는 영한 전문용어 사전 표제어에 대한 영어-한국어 조어단위 정렬문제로 변환할 수 있다. 본 논문에서는 정렬 모델을 이용하여 개념단위를 인식하였다. 둘째, '동형이의어' , '동의이형어' 문제는 대응된 개념단위와 조어단위 간의 관계를 통하여 같은 개념단위에 나타난 조어단위의 집합을 파악함으로써 해결할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 기술하고, 3장에서는 본 논문의 기법에 대하여 설명한다. 4장에서는 실험과 실험결과에 대하여 기술하고 5장에서는 결론을 맺는다.

2. 관련 연구

단어 정렬은 통계적 기계번역의 번역확률을 계산하는 모델로서 처음 소개되었다[4]. 이후 단어정렬은 이중언어에 대한 지식을 획득하기 위한 방법으로 많은 연구가 있어왔다. 예를 들어 정렬은 대역어 추출, 대역규칙, 문장단위 파악을 위한 분류정보 추출 등을 위한 연구의 방법으로 사용되어 왔다.

단어정렬의 연구는 크게 확률적 방법과 통계기반 방법으로 나누어진다. 확률적 방법에서는 주어진 원문 S에 대하여 대역문 T로 번역될 확률 P(T|S)에 정렬의 개념을 도입하여 식 (1)과 같이 번역확률을 정의하였다[4]. 식 (1)에서 A는 S와 T에 대하여 가능한 모든 정렬의 집합을 나타낸다.

$$p(T|S)=\sum_{a\in A} p(T,a|S) \qquad (1)$$

Brown 등[4]은 식 (1)을 기반으로 다섯 가지 영-불 정렬모델을 제안하였다. 모델 1은 P(F|E)가 오직 단어간 대역확률 $t(f_j|e_i)$에만 의존한다는 가정과 1 :1 정렬만을 가정하고 식 (2)에 의해 정렬을 수행하였다. [4]에서는 대역확률 $t(f_j|e_i)$을 EM 알고리즘에 의해 계산했다. 식 (2)에서 m은 F의 길이, l은 E의 길이 $C_{l,m}$은 l,m에 의해 결정되는 상수를 각각 나타낸다.

$$p(F|E)=C_{l,m}\prod_{j=1}^{m}\sum_{i=1}^{l} t(f_j|e_i) \qquad (2)$$

모델 2에서는 식 (2)를 기반으로 문장내의 위치정보를 추가하였으며, 모델 3에서는 1 :n 정렬까지 고려하여 정렬된 단어사이의 거리정보를 확률식에 포함하여 사용하였다. 모델 4, 5는 단어단위 정렬을 구 단위 정렬로 확장한 모델이다.

Dagan 등[5]은 [4]의 모델 2를 변형하여 문장단위로 정렬되지 않은 코퍼스에서도 단어단위 정렬이 가능하도록 하면서 파라메터 수를 줄인 모델을 제안하였다.

통계기반 모델에서는 카이 제곱법이나 로그우도와 같은 통계기법을 이용하여 대역어 간의 연관도를 측정하여 정렬을 수행하였다. 이러한 연관도는 단어 정렬에서 제약조건으로 사용되었다[6, 7].

본 논문에서는 [4]의 모델을 기반으로 조어단위 정렬을 수행한다. 또한, 전문용어의 대역 특성을 정렬모델에 반영하여 정렬을 수행한다. 기존의 단어 정렬이 양국어 대역 문장에 나타난 단어간의 정렬을 수행한 것에 반해, 본 논문에서는 주어진 대역 전문용어에 대하여 전문용어를 구성하는 한-영 조어 단위간의 정렬을 수행한다.

3. 한-영 조어단위 정렬

3.1 문제 정의

한국어-영어 조어단위 정렬 문제는 영어 용어를 구성하는 조어단위와 대역되는 한국어 용어를 구성하는 조어단위 간의 대응관계를 파악하는 작업으로 정의된다. 즉, 주어진 영어 전문용어 $E=e_1,...,e_n$와 한국어 전문용어 $K=k_1,...,k_m$에 대하여, 식 P(A|K,E)를 최대화하는 정렬집합 A을 찾는 문제로 정의된다. 이는 식 (2)와 같이 표현된다.

$$A^* = \arg\max_{A} P(A|K,E) \qquad (2)$$

3.2. 확률적 모델링

본 논문에서는 문제를 간단화하기 위하여 영어-한국어 전문용어 대역쌍에 나타나는 세 가지 특성을 이용하여 조어단위 정렬의 제약조건을 설정하고 이를 이용하여 조어단위 정렬을 수행한다. 사용되는 제약조건은 다음과 같다.

제약조건 1) *하나의 영어 조어단위(e_i)는 하나 이상의 한국어 조어단위(k_j)와 정렬된다.*

전문분야 사전에서 영어조어단위가 단어 수준임에 반해, 한국어 조어단위는 단어 수준이 아니라 형태소 수준이다. 한국어 형태소는 영어의 단어보다 더 작은 단위이기 때문에 영어 한 단어가 여러 한국어 형태소로 정렬되는 경우가 많다[8]. 이로 인해 하나의 영어조어단위에 대하여 여러 개의 한국어 조어단위가 대응되는 양상을 나타낸다. 본 논문에서 사용한 생물학 분야 사전 약 14,000여 개의 영어 용어에 대하여, 약 95% 정도가 이러한 특성을 나타내었다.

제약조건 2) *교차정렬을 허용하지 않는다.*

원어의 정렬단위 s_i, $s_j(i<j)$와 번역어 정렬단위 t_l, t_m $(l<m)$에 대하여, 정렬 $a_i=\{s_i,t_m\}$, $a_j=\{s_j,t_l\}$와 같이 나타날 때 이를 교차정렬이라 한다. 여기에서, i, j는 원어에서의 위치정보 l, m은 목적어에서의 위치정보를 나타낸다. 한국어와 영어는 문장의 구조가 다르기 때문에 단어들의 정렬에 있어 교차정렬이 빈번하게 나타난다. 하지만 전문용어의 대부분을 차지하는 명사구의 경우, 영어와 한국어 모두 수식어-피수식어 구조의 형태를 가지기 때문에 구조적 유사성을 가진다. 이로 인해 위치정보가 유용한 정렬정보로 사용될 수 있다. 예외적인 경우로 영어 용어 'of'에 인해 교차정렬이 발생하는 경우가 있는데, 본 논문에서는 이러한 경우 'of'를 중심으로 치환을 수행한 후 정렬을 수행한다. 예를 들어 영어용어 'clotting of blood'와 대역어 '혈액 응고'에 대하여 'of'를 중심으로 치환한 결과인 'blood clotting'을 정렬대상으로 한다. 교차정렬을 허용하지 않은 제약으로 인하여 정렬은 순차정렬에 대한 문제로 간단화 된다.

제약조건 3) *모든 영어 단어와 모든 한국어 조어단위가 정렬된다.*

정렬의 대상이 영어 전문용어와 대역어인 한국어 전문용어이기 때문에 정렬결과에서 영어의 모든 조어단위가 한국어의 조어 단위로 대응되는 것을 가정한다. 즉 정렬에서 널 정렬이 없음을 나타낸다. 본 논문에서는 널 정렬이 나타나는 대역쌍에 대해서는 정렬을 수행하지 않는다. 이는 전체 대역쌍 중 약 50개 정도가 해당한다. 예를 들어, 전문용어 대역쌍 'Dutch elm disease: 네덜란드 느릅나무 체관 병'에서는 한국어 조어단위 '체관'에 해당하는 영어 조어단위가 없기 때문에 널 정렬이 존재한다고 판별되며, 정렬대상에서 제외된다.

제약조건에 의해, 식 (2)는 식(3)과 같이 표현할 수 있다. 여기에서, e_i와 k_j가 번역쌍이거나 번역쌍의 부분[1]일 때, $a(e_i,k_j)$와 같이 표현한다. 이를 이용하여 정렬집합 A는 $A=\{a_1,..,a_i,..a_m=a(e_{im},k_{jm})\}$와 같이 표현된다. 식 (3)에서 제약 조건 1), 3)에 의해 $A=\{a_1,..,a_t\}$, $K=\{k_1,...,k_t\}$와 같이 대역쌍에서 가능한 조어단위 정렬의 개수는 한국어 조어단위의 개수와 같게 된다. 식 (3)에서 $a(i|j,n,t)$는 위치정보를 나타내며, 제약조건 2)에 의해 교차정렬로 나타나는 정렬 a_m에 대해서는 $a(i|j,n,t)=0$의 값을 가지며, 나머지 경우에는 1의 값을 가진다. 또한, 제약조건 2), 3)에 의하여 $p(a_1|k_{11}, e_{11})=p(a_t|k_{tt},e_{nt})=1$의 값을 가진다.

$$P(A|K,E)=\prod_{m=1}^{t} p(a_m|k_{jm},e_{im})\times a(i|j,n,t) \qquad (3)$$

$p(a_m|k_{jm},e_{im})$은 한국어조어단위의 어휘정보만을 사용한 경우와 어휘정보와 품사정보를 모두 이용한 경우에 의해 식 (4)와 식 (5)로 나타내어진다. 본 논문에서는 식 (4)와 식 (5)를 이용한 조어단위 정렬모델을 각각 모델 1, 모델 2라고 정의하고, 한-영 조어단위 정렬의 성능을 비교 실험한다. kw_j kt_j를 각각 j번째 한국어 조어단위의 어휘정보와 품사정보로 정의하면, 식 (4)에서

---

[1] 본 논문에서는 번역쌍의 부분을 '원용어의 여러 조어 단위가 대상용어의 하나의 조어단위로 정렬될 때, 대상용어의 하나의 조어단위에 대응되는 원용어의 각 조어단위 간의 대응관계를 대역쌍의 부분이라 정의한다. 예를 들어, 전문용어 대역쌍 'deamilase 탈+아미노+화+효소'에서 'deamilase/탈', 'deamilase/아미노', 'deamilase/화', 'deamilase/효소'를 번역쌍의 부분이라 정의한다.

k_{jm}={kw_{jm}}으로 나타내어지며, 식 (5)에서는 k_{jm}={kw_{jm} kt_{jm}}로 나타내어 진다.

$$p(a_m \mid k_{jm}, e_{im}) = p(a_m \mid kw_{jm}, e_{im}) \quad (4)$$

$$p(a_m \mid k_{jm}, e_{im}) = p(a_m \mid kt_{jm}, kw_{jm}, e_{im}) \quad (5)$$

그림 1은 제안하는 정렬모델에 의한 대역쌍 'adenine deamilase : 아데닌+탈+아미노+화+효소'에 대한 조어단위 정렬예를 나타낸다. 제약조건에 의해 한국어의 첫번째와 마지막 조어단위는 영어 조어단위 'adenine'과 'deamilase'와 정렬이 되며, 교차 정렬이 허용되지 않아 영어 조어단위의 역순으로 정렬되는 연결관계는 존재하지 않는다.

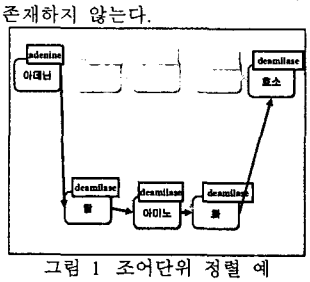

그림 1 조어단위 정렬 예

3.3 EM 알고리즘을 이용한 파라메터 학습

식 (4), (5)에서의 각 파라메터는 EM 알고리즘에 의해 학습된다. EM 알고리즘의 초기 파라메터는 하나의 조어단위로 구성된 영어 전문용어에 대역되는 한국어 조어단위들을 이용하였다. 이는 그 자체로 정렬된 결과를 나타내므로 정렬의 학습데이터로 사용할 수 있다. 초기 학습된 파라메터를 이용하여 두 개 이상의 조어단위로 구성된 영어단어와 그에 대역되는 한국어 조어단위에 대하여 EM 알고리즘을 통하여 순환적으로 파라메터를 학습하고 정렬을 수행한다.

4. 실험

4.1 실험데이터 및 평가 방법

실험을 위하여 14,000여 표제어 수준의 생물학분야 전문용어 사전을 사용하였다. 사전은 영어와 한국어에 대하여 조어 분석된 결과를 포함하고 있다. 이 중 하나의 조어단위로 구성된 영어용어는 약 8,800개이며 두 개 이상의 조어단위로 구성된 영어용어는 약 5,500개이다. 정렬결과에서 두 개 이상의 조어단위로 구성된 영어용어에 대하여 정렬결과를 평가한다. 평가를 위하여 Baseline, 모델1, 모델2, Upper Bound의 네 가지 시스템에 대한 실험을 수행한다. Baseline은 한국어 조어단위에 대한 기저명사구파악 후 영어 조어단위와 한국어 명사구의 개수가 같을 때, 순서대로 정렬하는 시스템이다. 모델 1과 모델 2는 식 (2),(4),(5)에 의해 본 논문에서 제안하는 조어 단위 정렬 모델으로 정렬하는 시스템이다. Upper Bound는 실험집합에서 제약조건을 만족하는 전문용어 대역쌍에 대하여 완벽하게 정렬하는 시스템을 말한다. 따라서 Upper Bound는 본 논문에서 제시하는 조어단위 정렬모델 성능의 상한선을 나타내는 시스템이라 할 수 있다. 실험결과에 대한 평가는 식 (6)의 용어단위 정확률과 조어단위 정확률로 평가한다.

$$\text{조어단위정확률} = \frac{\text{올바르게 정렬된 조어단위의수}}{\text{조어단위의수}}$$

$$\text{용어단위정확률} = \frac{\text{모든 조어단위가 올바르게 정렬된 용어의수}}{\text{전체용어의수}} \quad (6)$$

4.2. 실험결과

표 3은 조어단위 정렬의 실험결과를 나타낸다. 실험결과에서 모델 1은 Baseline보다 약 26%의 용어단위 정확률 향상과 약 38.5%의 조어단위 정확률 향상을 나타내며, 모델 2는 Baseline보다 약 29%의 용어단위 정확률 향상과 약 40%의 조어단위 정확률 향상을 나타냄을 알 수 있다. 또한, 모델 2는 모델 1보다 용어단위 정확률과 조어단위 정확률 모두 높게 나타남을 알 수 있다. 이는 품사정보가 조어단위의 경계를 파악하는데 유용하게 사용되기 때문으로 분석된다. 예를 들어, 접미사는 영어 조어단위의 시작부분에 정렬되지 못하며, 접두사는 영어 조어단위의 끝부분에 정렬되지 못한다는 정보에 의해 모델 2의 성능이 모델 1의 성능보다 높게 나타나는 것으로 분석된다. 실험결과에서 모델 2는 Upper Bound와 비슷한 결과를 나타냄을 알 수 있다. 이를 통하여 본 논문에서 제시하는 정렬 모델은 조어단위정렬에 효과적으로 사용됨을 알 수 있다. 실험결과에서 N:1 정렬 (제약조건 1의 위배)로 인한 오류는 전체 오류의 7% (19개 용어)였으며, 교차정렬 (제약조건 2의 위배)로 인한 오류는 전체오류의 16% (42개 용어)였다.

표 3. 한-영 조어단위 정렬의 실험 결과

실험방법	용어단위 정확률	조어단위 정확률
Baseline	73.92%	69.41%
모델 1	93.59% (+26.60%)	96.16% (+38.54%)
모델 2	95.47% (+ 29.15%)	97.16% (+39.98%)
Upper bound	97.02% (+ 31.25%)	97.76% (+40.84%)

5.결론

본 논문에서는 조어 분석된 한-영 전문용어 표제어에 대한 조어단위 정렬방법을 제안하였다. 본 논문의 기법은 약 95% 용어 정확률과 약 97% 조어정확률로 효과적으로 조어단위를 정렬하였다. 하지만 제약조건으로 인하여 N:1 정렬, 교차정렬, 널 정렬을 포함하는 전문용어 대역쌍에 대한 효과적인 처리가 어려웠다. 따라서 향후 이에 대한 보완이 필요할 것으로 생각된다.

본 논문의 기법은 새로운 영어 전문용어에 대한 한국어 조어의 양상을 파악할 수 있는 기반 기술로 사용될 것으로 기대된다. 또한 정렬결과는 개념단위에 기반한 전문용어의 조어 패턴과 전문용어의 변이파악을 위한 자원으로 사용될 수 있을 것으로 기대된다.

감사의 글

본 연구는 한국과학기술기획 평가원 국책연구개발사업 (M1-0107-00-0018)과 한국과학재단 특성장려연구사업 (R21-2003-000-10042-0)의 지원으로 수행되었습니다.

참고문헌

[1]조은경, 서상규 (2000), "전문용어연구를 위한 복합용어 분석의 단위" 제 3회 전문용어언어공학심포지움

[2] Sager, J.C. (1997), "Section 1.2.1 Term formation", in Handbook of terminology management Vol.1, John Benjamins publishing company

[3]조은경, 서상규 (2001), "전문용어의 조어 분석을 통한 개념 분석" 제 4회 전문용어언어공학심포지움

[4] Brown P.F., V.S.A. Della Petra, V.J. Della Pietra and R.L. Mercer, (1993), "The mathematics of statistical machine translation: parameter estimation", Computational Linguistics, Vol. 19 No 2, pp 263-311

[5] Dagan, I., K. Church and W. Gale, (1993), "Robust bilingual word alignment for machine aided translation", In Proceedings of the workshop on Very Large Corpora. pp. 1-8

[6] Melamed I. Dan, (2000), Models of translational equivalence among words, Computational Linguistics, 26(2): 221-249

[7] Cherry Colin and Dekang Lin, (2003), "A Probability Model to Improve Word Alignment", In Proceedings of 41st Annual Meeting of the Association for Computational Linguistics

[8] 신중호 (1996), "한국어/영어 병렬 코퍼스에 대한 단어단위 및 구단위 정렬 모델", 한국과학기술원 전산학과 석사학위 논문