영-한 조어단위 대역쌍 추출을 위한 조어단위 정렬 모델

(An Alignment Model for Extracting English-Korean Translations of Term Constituents)

오 종 훈 † 황 금 하 †† 최 기 선 †††
(Jong-Hoon Oh) (Jin-Xia Huang) (Key-Sun Choi)

요 약 전문용어는 전문분야의 개념을 표현하는 언어적 표현이다. 전문용어의 조어단위는 전문용어를 구성하는 최소의 형태적 단위이다. 따라서 조어단위는 전문용어의 의미를 파악하는데 중요한 요소이다. 하지만 조어단위를 이용한 전문용어의 의미파악은 '조어단위와 개념단위의 불일치 문제', 조어 단위의 '동형이의어', '동의어' 문제 등으로 인한 어려움이 있다. 이러한 문제를 해결하기 위해서는 하나의 개념을 나타내는 조어단위의 덩어리인 개념단위를 파악하는 작업이 선행되어야 한다. 본 논문에서는 영어의 조어단위를 하나의 개념단위로 정의하고 개념단위에 대응되는 한국어 조어단위의 집합을 개념단위로 인식한다. 개념단위의 파악과정은 영한 대역 전문용어사전에 대한 영어-한국어 조어단위 정렬문제로 해결하고자 한다. 본 논문의 기법은 물리, 화학, 생물 분야에 대한 조어정렬 실험을 수행하였으며, 평균 약 93%의 정확률로 조어단위 간의 정렬을 수행하였다.

키워드 : 조어단위, 전문용어, 조어, 정렬, 개념

Abstract *Terms are linguistic realization of technical concepts. Term constituents are important elements used for representing the concept. Since many new terms are created from the modification or combination of existing constituents, it is important to analyze term constituents for understanding the concept of the term. It means that term constituents offer clues for understanding the concept of terms. However, there are a couple of difficulties in matching concept unit and term constituents such as mismatching between a term constituent and a concept unit, homonym of term constituents and synonym of term constituents. To solve them, it is necessary to recognize concept units of term constituents. In this paper, we define an English term constituent as the concept unit and use an alignment algorithm between English-Korean term constituents in order to recognize concept units of term constituents. By our alignment algorithm we recognize Korean term constituents corresponding to an English term constituent with about 93% precision.*

Key words : term constituent, terminology, term formation, alignment, concept

1. 서 론

전문용어는 전문분야의 개념을 표현하는 언어적 표현이다[1]. 전문용어의 조어단위는 전문용어를 구성하는 최소의 형태적 단위이다[2]. 전문용어는 전문용어를 구성하는 조어단위의 개수에 따라 '단일 용어'와 '복합용어'로 나누어진다[1,2]. 단일용어는 하나의 조어단위로 구성된 용어인 반면, 복합용어는 두 개 이상의 조어단위로 구성되는 용어이다.

한국어 전문용어의 대부분은 복합용어의 형태를 가진다[3]. 대부분의 복합용어는 조어 단위들의 결합에서 합성성을 따르는 투명한 용어이다[1]. 즉, 조어단위가 나타내는 개념들의 결합으로 해당 전문용어의 의미를 유추할 수 있다. 복합용어의 조어양상을 파악한다는 것은 전문용어를 구성하는 조어단위의 개념간 연관관계를 파악하는 것을 의미한다. 예를 들어, "향축+성 분열 조직"은 "향축+성(adaxial)"과 "분열 조직(meristem)"의 두

· 본 연구는 과학기술부와 한국과학재단 특성장려연구사업(R21-2003-000-10042-0)의 지원으로 수행되었음
† 학생회원 : 한국과학기술원 전산학과
rovellia@world.kaist.ac.kr
†† 비 회 원 : 한국과학기술원 전산학과
hgh@world.kaist.ac.kr
††† 종신회원 : 한국과학기술원 전산학과 교수
kschoi@world.kaist.ac.kr
논문접수 : 2004년 10월 6일
심사완료 : 2005년 2월 11일

개의 개념단위로 구성된다. 또한 두 개념단위와 이들간의 연관관계로부터 "향축지역에 존재하는 잎세포의 분열조직"으로 해석된다. 하지만 "향축(adaxial)", "성(sexual)", "분열(division)", "조직(tissue)"과 같이 개념단위를 잘못 파악하게 되면 전문용어의 의미를 올바르게 유추할 수 없다. 즉, 용어의 조어양상을 파악한다는 것은 해당 용어의 개념을 나타내는 "향축+성"과 "분열조직"과 같은 개념단위를 올바르게 파악하고 파악된 개념단위들간의 연관관계("향축+성" + "분열 조직")를 파악하는 것으로 해석된다. 따라서 전문용어의 의미를 파악하기 위해서는 이러한 조어단위에 대한 개념을 올바르게 파악하는 것이 필요하다.

또한 많은 한국어 전문용어가 외국어에 기반하여 생성되기 때문에 일관성 있는 한국어 번역용어의 파악과 기계번역 사전의 커버리지를 높이기 위해서도 조어단위의 개념파악이 필요하다. 전문분야마다 용어를 이루는 조어단위가 다른 의미나 다른 형태로 번역되는 경향이 있다. 예를 들어, *cell*은 물리학이나 화학 분야에서는 주로 "전지"로 생물학 분야에서는 "세포"로 번역되는 경향이 있다. 이처럼 전문분야마다 조어단위별 번역양상이 다르게 나타나기 때문에 전문용어에 대한 일관적이면서 효과적인 기계번역을 위해서는 분야별 조어단위 번역정보가 필요하다. 또한 전문용어는 전문용어 조어단위간의 결합에서 합성성을 나타내기 때문에 조어단위의 번역정보를 이용하여 기계번역사전(전문용어의 영-한 대역쌍)에 나타나지 않은 전문용어를 번역할 수 있다. 조어단위 정보를 이용하여 기계번역사전의 커버리지를 향상시킬 수 있다. 예를 들어, 생물학 분야 기계번역사전에 <embryo sac mother cell, 배낭(胚囊) 모세포(母細胞)>라는 번역정보만이 존재하고 각 조어단위에 대한 번역쌍이 존재하지 않을 경우, 해당 조어단위로 구성된 전문용어인 *embryo sac cell, mother cell, embryo sac* 등에 대한 효과적인 번역이 어렵다. 이 때, <embryo, 배(胚)>, <sac, 낭(囊)>, <embryo sac, 배낭(胚囊)>, <mother, 모(母)>, <cell, 세포(細胞)>, <mother cell, 모세포(母細胞)> 등과 같은 조어단위간의 번역정보를 이용하면 *embryo sac cell, mother cell, embryo sac*의 번역인 "배낭 세포", "모세포", "배낭"을 효과적으로 파악할 수 있다.

본 논문에서는 전문용어의 의미파악과 일관성 있는 한국어 번역용어 생성에서 나타나는 문제점을 기술하고 이러한 문제점을 해결하기 위해 필요한 영-한 조어단위 대역쌍을 추출하는 알고리즘을 제안한다.

한국어 전문용어 의미파악에서 **'조어단위와 개념단위의 불일치 문제'**, 조어 단위의 **'동형이의어(homonym)'**, **'동의어(synonym)'** 문제 등으로 인하여, 한국어 조어단위만을 이용한 전문용어의 의미파악은 쉽지 않다. 본 논문에서는 이러한 문제들은 영-한 조어단위 정렬을 통한 조어단위 대역쌍을 추출하여 해결한다. 즉, 영어용어를 구성하는 단어를 하나의 개념단위와 하나의 조어단위로 정의하고, 영어 단어에 대응되는 한국어 조어단위의 집합을 파악하여 영어조어단위와 한국어 조어단위의 개념단위를 파악한다.

첫번째 문제점인 **'조어단위와 개념단위의 불일치 문제'**는 하나의 조어단위가 전문용어의 구성형태에 따라 개념단위로 사용되기도 하지만 그렇지 않을 경우도 발생하기 때문에 나타난다. **'개념단위'**란 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 <정단 세포, *apical cell*>에서 '세포'는 그 자체로 *cell*을 나타내는 개념단위가 되지만, <난(卵)+세포, *oocyte*>에서는 '세포'만으로 개념단위가 되지 않고, 조어단위 '난(卵)'과 결합하여 *oocyte*[1]을 표현하는 개념단위를 형성한다. 따라서 복합용어에서 전문용어의 의미를 효과적으로 파악하기 위해서는 개념단위를 파악하는 작업이 필요하다.

두 번째 문제점은 조어단위의 **'동형이의어(homonym) 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 네 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 (基)	아미노기	amino group
period (紀)	수축기	contraction period
stage, phase(期)	생장기	growth phase
organ (器)	후각기	olfactory organ

1) "oocyte"는 "oo(egg)+cyte(cell)"로 분석되기 때문에 두 개의 개념단위로 구성된 용어이다. 하지만 본 논문에서는 영어 조어단위는 하나의 개념단위를 나타내는 단일용어라고 가정한다.

세 번째 문제점은 조어단위의 **'동의어(synonym)'** 문제이다. 외국어에 기원을 둔 전문용어는 고유어나 한자어로 번역되거나, 음차표기 등의 방법으로 한국어로 표기된다. 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

abdominal의 번역어	한국어 용어	영어 용어
복부	복부부속지	abdominal appendage
복	복강	abdominal cavity
배	배지느러미	abdominal fin
복, 복부	복공, 복부공	abdominal pore

본 논문에서는 상기에 언급한 문제점을 조어분석된 영한 대역 전문용어사전에 대한 영어-한국어 조어단위 정렬문제로 해결하고자 한다. 첫째, 개념단위 인식 문제는 영어-한국어 용어의 조어단위 간의 대응관계를 파악하는 문제로 정의될 수 있다. 이는 영한 전문용어 사전 표제어에 대한 영-한 조어단위 정렬 문제로 변환할 수 있다. 예를 들어, <정단 세포, *apical cell*>에서는 '세포'가 *cell*과 대응되므로 '세포'를 개념단위로 인식할 수 있으며, <난세포, *cocyte*>에서는 '난세포'가 *oocyte*와 대응되기 때문에 '난세포'를 하나의 개념단위로 인식할 수 있다. 둘째, '동형이의어', '동의어' 문제는 대응된 개념단위와 조어단위 간의 관계를 통하여 같은 개념단위에 나타난 조어단위의 집합을 파악함으로써 해결할 수 있다. 예를 들어, 동형이의어 문제를 야기하는 접사 '-기'는 '-기'와 대응되는 영어조어단위에 의해 <-기(基), *group*>, <-기(紀), *period*>, < -기(期), *stage or phase*>, < -기(器), *organ*>으로 그 의미를 구분할 수 있다. 또한 '동의어' 문제를 야기하는 *abdominal*은 *abdominal*에 대응되는 한국어 조어단위의 집합인 *abdominal* = {복, 복부, 배}를 통하여 여러 다른 형태로 나타나는 같은 의미의 한국어 조어단위를 파악할 수 있다.

복합용어 형태의 외국어 전문용어를 한국어로 번역할 때 복합용어의 합성성으로 인하여 영어 조어단위의 한국어 번역을 사용하는 경향이 있다. 예를 들어, *suppressor gene*은 표 3과 같은 기존의 전문용어의 조어단위 번역정보 <*suppressor*, '억제'>, <*gene*, '유전자'>로부터 한국어 번역 '억제 유전자'를 생성할 수 있다.

영어 전문용어를 일관성 있게 한국어로 자동 번역하기 위해서는 각 조어단위의 번역정보로부터 생성 가능한 한국어 번역후보를 생성하는 **"번역후보 생성 과정"**과 생성된 번역후보 중 적절한 **"번역후보의 선택과정"**이 필요하다[4]. 번역후보 생성과정의 목적은 적합한 번역후보를 포함하는 번역후보를 생성함과 동시에 번역후보의 개수를 최소화하는 것이다. 이를 위해서는 분야 특성적인 조어단위 번역정보가 필요하다. 예를 들어, 물리분야 전문용어 *absolute refractive index*는 한국어 전문용어 '절대 굴절률', '절대 꺾임율'로 번역된다. 하지만 일반분야 사전[5]으로는 해당 전문용어에 대한 적합한 후보를 생성할 수 없다. 표 4는 일반분야사전에 의한 조어단위 번역정보와 전문분야 조어단위 번역정보를 나타낸다. 표 4에서 일반분야사전 기반 번역후보의 총 수는 100개가 생성되며, 이 중 올바른 번역어는 존재하지 않는다. 이는 조어단위 *index*의 번역어인 '율' 또는 '률'에 대한 번역 정보가 일반분야 사전에 존재하지 않기 때문이다. 이와 반대로 전문분야 조어단위 번역정보에 의해서는 총 8개의 후보가 생성되며, 이 중 2개가 올바른 번역어이다. 전문분야 조어단위 번역정보를 사용한 경우는 분야 특성적인 번역정보만을 사용하기 때문에 조어단위의 의미모호성이 줄어들 뿐만 아니라 일반분야에서 사용되지 않는 번역정보를 획득할 수 있기 때문에 보다 적합한 번역 후보를 생성할 수 있다. 물리 분야 전문용어 50개에 대하여 전문분야 조어단위 번역정보를 이용하여 대역후보를 생성한 결과 총 603개의 번역후보를 생성하였으며, 이 중 42개 용어에 대해서 적합한 번역어를 포함하였다 - 84%(42/50)의 적용률과 6.97%(42/603)의 정확도를 나타내었다. 일반분야 사전을 이용한 경우는 총 19,437개의 번역후보를 생성하였으며, 이 중 10개만이 올바른 번역 후보를 생성하여 20%(10/50)의 적용률과 0.05%(10/19,437)의 정확도를 나타내었다. 따라서 효과적인 번역후보를 생성하기 위해서는 전문분야 조어단위 번역정보가 필요하다.

"번역후보 선택 과정"을 위해서는 번역어간의 연관관계를 파악해야 한다. [4]에서는 이러한 연관관계를 조

표 3 영어 전문용어에 대한 한국어 전문용어의 조어단위 번역 예

영어 용어	한국어 용어	조어단위 번역
Suppressor mutation	억제 돌연변이	<suppressor, '억제'>, <mutation, '돌연변이'>
Integrative suppression	통합 억제	<integrative, '통합'>, <suppression, '억제'>
Complementary gene	보족 유전자	<complementary, '보족'>, <gene, '유전자'>
Cumulative gene	누적 유전자	<cumulative, '누적'>, <gene, '유전자'>

표 4 absolute refractive index에 대한 조어단위 번역정보

	일반분야 사전기반 번역정보	전문분야 조어단위 번역정보
Absolute	**절대(의)**, 절대적인, 완전무결한, 완전한, 철저한, 순수한, 무제한(의), 무조건(의), 명백한, 본질적인	**절대**
Refractive	**굴절(의)**	**굴절, 꺾임**
Index	색인, 목록, 표시, 징후, 조짐,지침, 지표, 바늘, 눈금, 지수	지표, 지수, **율, 률**
총 대역후보 수	100	8

어단위의 바이그램, 트라이그램에 기반하여 파악하였다. 조어단위의 *n*-그램을 파악하기 위해서는 영-한 조어단위간의 대응관계를 파악하는 것이 선행되어야 한다. 예를 들어 *absolute refractive index*의 번역용어 '절대굴절률'을 올바른 결과로 선택하기 위해서는 '절대+굴절', '굴절+률', '절대+굴절+률'과 같은 한국어 조어단위 *n*-그램 정보와 'absolute/절대+refractive/굴절', 'refractive/굴절+index/률'과 같은 영어 및 한국어 조어단위의 *n*-그램 정보가 필요한다. 이를 위해서는 '절대', '굴절', '률'이 각각 *absolute, refractive, index*와 대응되는 조어단위라는 것이 파악되어야 한다. 본 논문에서는 **"번역후보 생성과정"**과 **"번역후보 선택 과정"**에 필요한 전문분야 조어단위 번역 정보를 영-한 조어단위 정렬 기법을 이용하여 획득한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 기술하고, 3장에서는 본 논문에서 제안하는 조어단위 정렬기법에 대하여 설명한다. 4장에서는 실험결과에 대하여 기술한다. 그리고 5장에서 결론을 맺는다.

2. 관련연구

2.1 단어기반 통계적 기계번역(단어정렬)

단어 정렬은 통계적 기계번역의 번역확률을 계산하는 모델로서 처음 소개되었다[6]. 이후 단어정렬은 이중언어에 대한 지식을 획득하기 위한 방법으로 많은 연구가 있었다. 예를 들어 정렬은 대역어 추출, 대역규칙, 문장단위 파악을 위한 분류정보추출 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[6]. 식 (1)에서 A는 S와 T에 대하여 가능한 모든 정렬의 집합을 나타낸다.

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

Brown 등[6]은 식 (1)을 기반으로 다섯 가지 영-불 정렬모델을 제안하였다. Brown은 이들 다섯 가지 모델을 각각 모델 1, 모델 2, 모델 3, 모델 4, 모델 5로 정의하였다. 각 모델의 특성은 다음과 같다. 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.

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

여기에서 E는 영어문장을 F는 불어문장을 나타내며, m은 F의 길이, l은 E의 길이 $C_{l,m}$은 l과 m에 의해 결정되는 상수를 각각 나타낸다.

즉 모델 1은 단어간의 순서를 고려하지 않았기 때문에 "단어열(bag of words) 모델"이라 한다. 대역확률 $t(f_j|e_i)$는 *EM* 알고리즘에 의해 계산되었다.

모델 2에서는 모델 1을 확장하여 문장내의 위치정보와 문장의 길이를 고려하여 정렬을 수행하였다. 모델 2에서는 문장내의 위치와 문장의 길이에 따라 번역확률이 달라지므로, 모델 2를 '위치 및 길이 기반 모델'이라 한다. 모델 3에서는 모델 2를 기반으로 *1:n* 정렬까지 고려하여 정렬을 수행하였다. 이를 위하여 정렬된 단어사이의 거리정보를 확률식에 포함하여 사용하였다. 모델 4, 5는 모델 3의 단어단위 정렬을 구 단위 정렬로 확장한 모델이다.

Dagan 등 [7]은 [6]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[8,9].

2.2 구 기반 통계적 기계번역(구 정렬)

Och[10]는 구(phrase)간 정렬관계를 정렬 템플릿(alignment template)으로 정의하고 이를 이용한 영어-독일어간 구단위 정렬 모델을 제안하였다. 정렬 템플릿은 이중언어 코퍼스에서 원어(source language) 및 목적어(target language)의 클래스 및 문장내 위치정보에 의해 생성된 지역적 정렬 행렬(local alignment matrix)로 정의된다. Och가 사용한 정렬 템플릿은 그림 1과 같이 표현된다.

Och는 목적어의 문장을 구단위로 나눈 뒤, 각 구에 대응될 수 있는 영어의 단어들로 정렬템플릿으로 구성

표 5 조어분석된 전문분야 사전의 예

영어 용어	한글 용어	조어단위		
Achilles' tendon reflex	아킬레스 힘줄 반사	아킬레스/npp	힘줄/nc	반사/nc
apical cell	정단 세포	정단/nc	세포/nc	
crop growth rate	작물 생장율	작물/nc	생장/nc	율/xs

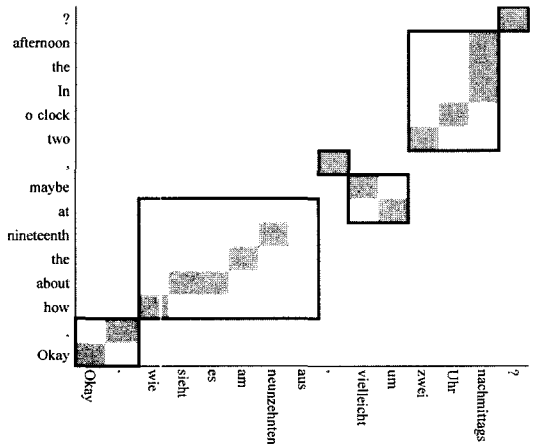

그림 1 정렬 템플릿: 실선으로 표현된 사각형은 정렬 템플릿을 나타내며, 색이 칠해진 사각형은 정렬 템플릿내에서의 단어간의 대응관계를 나타낸다.

하였다. 정렬 템플릿을 구성하는 확률은 2.1절에서 기술한 단어간 정렬 모델들을 이용하여 학습된다. Marcu와 Wong[11]은 Och의 모델과 유사한 구단위 정렬 모델을 제안하였다. [11]은 Och의 모델과 달리 대역되는 구를 구성하기 위한 확률을 구가 대응된 이중언어 코퍼스를 이용하여 학습하였다. 구 단위 정렬기법은 정렬기법 중 가장 좋은 성능을 나타내는 것으로 보고되고 있으며, 다양한 방법론이 제안되고 비교되고 있다[12-14].

2.3 기존연구의 요약

전문용어 대역쌍은 주로 명사구간의 대역관계를 나타낸다. 따라서 본 논문에서 해결하고자 하는 문제는 주어진 명사구 단위의 전문용어 대역쌍에 대하여 영어와 한국어 전문용어를 구성하는 형태소 수준의 조어단위 정렬문제이다. 따라서 2.2절에서 기술한 구정렬 모델과 2.1절에서 기술한 단어정렬모델을 구단위로 확장한 모델 4, 5보다는 모델 1, 2, 3에서 사용하는 단어 단위 정렬 방법이 보다 적합하다. 본 논문에서는 Brown의 모델 1, 2, 3을 기반으로 조어단위 정렬을 수행한다. 모델 1을 확장한 모델 2와 3을 기반으로 조어단위 정렬을 수행하였으며, *1:1* 정렬만을 고려하는 모델 2를 확장하여 모델 3과 같이 *1:n*과 *n:1*정렬까지 가능하도록 하였다. 이는 모델 3과 유사하지만 본 논문에서는 모델 3에서 사용한 거리정보를 부분적으로 사용하였다.

또한 전문용어 내의 조어단위간의 대역 특성을 정렬 모델에 반영하여 정렬을 수행한다. 기존의 단어 정렬이 양국어 대역 문장에 나타난 단어간 또는 구간의 정렬을 수행한 것에 반해, 본 논문에서는 주어진 대역 전문용어에 대하여 전문용어를 구성하는 영-한 조어 단위간의 정렬을 수행한다.

3. 영-한 조어단위정렬

3.1 영-한 조어단위 정렬 데이타

영-한 조어단위 정렬을 위한 데이타는 한국어 용어가 조어분석된 전문분야 사전을 사용하였다. 조어분석된 전문분야 사전은 물리, 화학, 생물학 분야의 사전으로 영어 용어와 영어 용어에 대역되는 한국어 용어로 구성되어 있다[15]. 한국어 용어는 조어단위로 분석되어 있으며, 각 조어단위에 대한 품사가 할당되어 있다. 표 5는 조어분석된 전문분야 사전의 예를 나타낸다. 표에서 *crop growth rate*에 대역되는 한글 용어는 '작물 생장율'이고, '작물 생장율'의 조어단위는 '작물', '생장', '율'이다. 또한 각 조어단위의 품사는 '작물/nc', '생장/nc', '율/xs'이다. 표 5에서 *npp*는 인칭대명사, *nc*는 일반명사, *xs*는 접미사를 나타낸다. 사용한 품사에 대한 정보는 부록에 자세히 기술하였다.

3.2 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, 확률 $P(A|K,E)$를 최대화하는 정렬집합 A를 찾는 문제로 정의된다. 이는 식 (3)과 같이 표현된다. 여기에서 e_i는 E의 i번째 조어단위를 나타내고, k_j는 K의 j번째 조어단위를 나타낸다. 정렬집합 A는 $A=\{a_1,....,a_t;a_p=(e_{i(p)},k_{j(p)})\}$와 같이 표현되며, 각 조어단위간의 대응관계 a_p의 집합이다.

$$A^* = \arg\max_{A} P(A \mid K, E) \tag{3}$$

3.3 확률적 모델링

본 논문에서는 영-한 전문용어 대역쌍에서 주로 나타나는 '순차정렬'과 '영(零) 정렬[2]제외'라는 두 가지 특성을 이용하였다. 두 가지 특성은 전문분야 사전에서의 영

2) 본 논문에서는 영 정렬을 조어단위 간의 *0:n* 또는 *n:0* 정렬로 정의한다.

어 조어단위와 한국어 조어단위 간의 대응관계를 수작업으로 분석하여 파악하였다.[3] 본 논문에서는 두 가지 특성을 조어단위 정렬의 제약조건으로 설정하고 조어단위 정렬을 수행한다. 사용되는 두 가지 제약조건은 다음과 같다.

제약조건 1) 부분적으로 교차정렬을 허용한다.

원어의 정렬단위 s_i, $s_j(i<j)$와 번역어 정렬단위 t_q, t_r $(q<r)$ 에 대하여, 정렬 $a_i=(s_i,t_r)$, $a_j=(s_j,t_q)$와 같이 나타날 때 이를 교차정렬이라 한다. 여기에서, i, j는 원어(source language)에서의 위치정보를, l, m은 목적어(target language)에서의 위치정보를 나타낸다. 한국어와 영어는 문장의 구조가 다르기 때문에 단어들의 정렬에 있어 교차정렬이 빈번하게 나타난다. 하지만 전문용어로 자주 나타나는 명사구의 경우 영어와 한국어 모두 수식어-피수식어 구조의 형태를 가지기 때문에 구조적 유사성을 가지며, 대부분의 경우 순차정렬의 형태로 조어단위가 정렬된다[16]. 이러한 순차정렬의 특성은 본 논문에서 사용한 실험데이타에서도 비슷한 양상을 나타내는데 약 97%가 순차정렬의 형태로 정렬된다. 하지만 예외적인 두 가지 경우에 대해서는 교차 정렬을 허용한다. 첫 번째로 영어 용어 *of*에 인해 교차정렬이 발생하는 경우가 있다. 본 논문에서는 이러한 경우 *of*를 중심으로 치환한 후 정렬을 수행한다. 예를 들어 <*clotting of blood*, '혈액 응고'>에 대하여 *of*를 중심으로 치환한 결과인 <*blood clotting*, '혈액 응고'>를 정렬 대상으로 한다. 이러한 치환에 의하여 *of*에 의해 나타나는 교차정렬은 순차정렬의 형태를 나타낸다. 두 번째로 영-한 조어단위의 개수가 같을 경우에는 교차정렬을 허용한다. 본 논문에서 사용한 데이타의 분석 결과, 그림 2와 같이 화합물을 나타내는 전문용어가 교차정렬의 형태로 정렬되는 경우가 많았으며, 영-한 조어단위 간에는 1:1 대응관계를 가지는 경우가 많았다. 이러한 조어양상을 반영하기 위하여 영-한 조어단위의 개수가 같은 경우에 한하여 교차정렬을 허용하였다. 이러한 교차정렬의 허용으로 그림 2와 같은 화합물에 대해서도 올바르게 조어단위 정렬을 수행할 수 있다. 이러한 예외적인 경우를 제외하고는 본 논문에서는 순차정렬을 가정하고 조어단위 정렬을 수행한다. 제약조건 1)에 의하여 본 논문의 조어단위 정렬문제는 부분적 교차정렬 문제로 단순화된다.

제약조건 2) 영(零) 정렬은 정렬대상에서 제외한다. 영어의 모든 조어단위가 한국어의 조어 단위로 대응된다.

정렬의 대상이 영어 전문용어와 대역어인 한국어 전문용어이기 때문에 정렬결과에서 영어의 모든 조어단위가 한국어의 조어 단위로 대응되는 것을 가정한다. 즉 정렬에서 영 정렬(*n:0*, 또는 *0:n* 정렬)이 없음을 나타낸다. 본 논문에서는 영 정렬이 나타나는 대역쌍에 대해서는 정렬을 수행하지 않는다. 실험데이타에서 1% 미만의 데이타에서 영정렬이 나타났다. 생물학 분야의 경우, 전체 대역쌍 중 약 50개 (0.8%) 정도가 해당하였다. 예를 들어, 전문용어 대역쌍 '<*Dutch elm disease*: 네덜란드+느릅나무+체관+병>'에서는 그림 3과 같이 한국어 조어단위 '체관'에 대응되는 영어 조어단위가 없기 때문에 영정렬이 존재한다고 판별된다. 본 논문에서는 영 정렬로 대응되는 대역쌍은 정렬대상에서 제외하였다.

그림 2 실험데이타에서 나타난 교차정렬 예

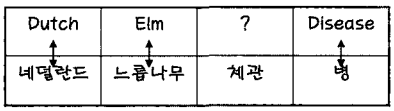

그림 3 <*Dutch elm disease*: 네덜란드+느릅나무+체관+병>에 대한 영정렬의 예

제약조건 1), 2)에 의해, 식 (3)의 조어정렬 문제는 식 (4)와 같이 표현할 수 있다. 여기에서, e_i와 k_j가 대응관계[4]에 있을 때, $a_p=(e_{i(p)},k_{j(p)})$와 같이 표현한다. 정렬집합 A는 $A=\{a_1,....,a_t;a_p=(e_{i(p)},k_{j(p)})\}$와 같이 표현된다. 제약 조건 1), 2)에 의해 가능한 조어단위 정렬의 개수는 한국어 조어단위의 개수 t와 같게 되어 $A=\{a_1,....,a_t\}$ $K=\{k_1,....,k_t\}$와 같이 표현할 수 있다. 식 (4)에서 $a(i|j,n,t)$는 위치정보를 나타낸다. 제약조건 1)에 의해 교차정렬이 허용되지 않는 조건에서 교차정렬로 나타나는 번역쌍의 부분 a_m에 대해서는 $a(i|j,n,t)$ = 0의 값을 가지며, 나머지 경우에는 1의 값을 가진다.

$$P(A|K,E)=\prod_{m=1}^{t} p(a_m|k_{j(m)},e_{i(m)})\times a(i|j,n,t) \quad (4)$$

$p(a_m|k_{j(m)},e_{i(m)})$은 한국어조어단위의 어휘정보와 품사정보를 이용하여 식 (5)로 나타내어진다. 식 (5)에서는 $k_{j(m)}=(k^w_{j(m)}, k^t_{j(m)})$로 표현된다. 여기에서 $k^w_{j(m)}$는 한국어 조어단위 $k_{j(m)}$의 어휘정보를 나타내고, $k^t_{j(m)}$는 한국

3) 본 논문에서 사용한 물리, 화학, 생물 데이타에서 교차정렬은 각각 1.3%, 0.1%, 5.65%의 비율로 나타났으며, 영정렬은 각각 0.8%, 0.2%, 0.1%의 비율로 나타났다.

4) 본 논문에서는 대응관계를 '원용(source term)의 여러 조어 단위가 대상용(target term)의 하나의 조어단위로 정렬될 때, 대상용어의 하나의 조어단위에 대응되는 원용어의 각 조어단위 간의 관계'라 정의한다. 예를 들어, 전문용어 대역쌍 〈*oocyt* 난+세에서 '*oocyt*', '*oocyt*/세포' 가대응관이다.

어 조어단위 $k_{j(m)}$의 품사정보를 나타낸다.

$$p(a_m \mid k_{j(m)}, e_{i(m)}) = p(a_m \mid k'_{j(m)}, k^w_{j(m)}, e_{i(m)}) \approx p(k'_j \mid e_i) \times p(k^w_j \mid k'_j, e_i) \quad (5)$$

3.4 조어단위 정렬을 위한 래티스(lattice)의 구성

주어진 영-한 대역쌍에 대하여 각 조어단위 간의 *1:1*, *1:n*, *n:1*정렬과 교차정렬을 위한 래티스를 구성하고 식 (5)를 이용하여 최적의 조어단위 간의 대응관계를 파악한다. *n:m* 조어단위 정렬이란 영어조어단위 *n*개에 대하여 한국어 조어단위 *m*개가 정렬되는 정렬형태로 정의된다. 예를 들어, 영어 용어 *female sex hormone*과 대응되는 한국어 용어 '자성(磁性) 호르몬'에서 *female sex*와 '자성(磁性)'이 대응되고, *hormone*과 '호르몬'이 대응된다. 이때, *female sex*와 '자성(磁性)'은 영어 조어단위 2개에 대하여 한국어 조어단위 1개가 대응되므로, 이는 *n:1* 조어단위 정렬로 파악된다. 마찬가지 방법으로 *hormone*과 '호르몬'은 1:1 조어단위 정렬로 파악된다.

본 논문에서는 주어진 영어용어와 대응되는 한국어용어에 대하여 모든 가능한 조어단위 간의 대응관계와 이들 간 연결관계를 표현하는 래티스를 생성한 후, 이들 중 가장 적합한 연결관계를 해당 대역쌍의 조어단위 정렬관계로 파악한다. 모든 가능한 대응관계와 연결관계는 부분적 교차정렬 허용(제약조건 1)과 영정렬 불가(제약조건 2)의 조건을 만족하여야 한다. 주어진 영-한 대역쌍에 대한 래티스는 수준(level)별로 가능한 대응관계를 생성한 후 이들 간의 연결관계를 파악함으로써 구성된다. 수준 *i*에서는 *i*개의 영어 조어단위와 한국어 조어단위간의 가능한 모든 대응관계를 생성한다. 그리고 주어진 영어용어에 포함된 영어 조어단위의 개수 *j*에 대하여 수준 *j*까지의 대응관계를 생성한 후, 모든 가능한 연결관계를 생성한다. 래티스의 구성은 제약조건 1)의 '부분적 교차정렬 허용'에 의해 순차정렬만을 허용하는 경우와 순차정렬과 교차정렬을 모두 허용하는 경우에 따라 그림 4와 그림 5와 같이 구성된다. 그림 4는 순차정렬만을 허용한 래티스 구성 예를 나타낸다. 세 개의 영어 조어단위로 구성된 영어 용어 *female sex hormone*에 대응하는 두 개의 조어단위로 구성된 한국어 용어 '자성 호르몬'에 대하여 그림 4와 같이 수준 3까지의 대응관계와 연결관계를 래티스로 구성한다. 주어진 대역쌍에서 영-한 조어단위의 개수가 다르기 때문에 제약조건 1)에 의해 순차정렬만을 허용한다. 수준 1에서는 *'female/자성'*와 *'hormone/호르몬'*의 두 가지 대응관계만이 생성되고, 수준 2에서는 *'female sex/자성'*, *'sex hormone/호르몬'*의 두 개의 대응관계가 생성된다. 수준 3에서는 마찬가지 방법으로 *'female sex hormone/자성'*, *'female sex hormone/호르몬'*을 생성할 수 있다. 생성된 대응관계에 대하여 가능한 모든 연결관계를 그림 4의 실선과 점선으로 표현할 수 있으며, 이들 연결관계 중 식 (5)를 이용하여 실선의 *'female sex/자성'*, *'hormone/호르몬'*을 최적의 정렬관계로 파악한다.

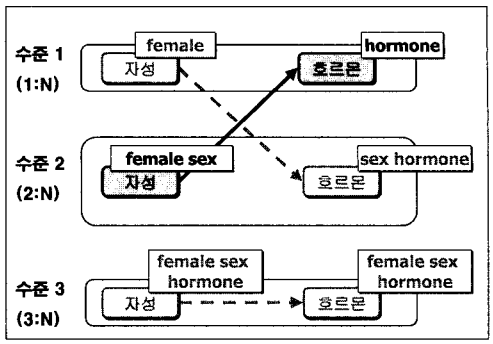

그림 4 <*female sex hormone*, '자성 호르몬'>에 대한 래티스의 구성 예

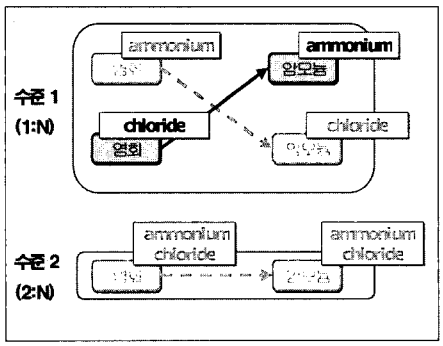

그림 5 <ammonium chloride, '염화 암모늄'>에 대한 래티스 구성 예

그림 5는 순차정렬과 교차정렬을 모두 허용한 경우의 래티스 구성 예를 나타낸다. 대역쌍<*ammonium chloride*, '염화 암모늄'>에 대하여 그림 5와 같이 수준 2까지의 대응관계로 래티스를 구성한다. 주어진 대역쌍에서 영-한 조어단위의 개수가 같기 때문에 교차정렬과 순차정렬을 모두 허용한다. 따라서 수준 1에서는 *'ammonium/암모늄'*, *'ammonium/염화'*, *'chloride/암모늄'*과 *'chloride/염화'*의 네 가지 대응관계가 생성되고 수준 2에서는 *'ammonium chloride/염화 암모늄'*의 하나의 대응관계가 생성된다. 구성된 래티스에 대하여 식 (5)를 이용하여 실선의 *'chloride/염화'*, *'ammonium/암모늄'*을 최적의 정렬관계로 파악한다.

3.4 EM 알고리즘을 이용한 파라메터 학습

식 (5)에서의 각 파라메터는 *EM(Expectation Maximization)* 알고리즘에 의해 학습된다. *EM* 알고리즘은 파라메터의 최우도추정 (Maximum Likelihood estimation)을 위한 반복적인 알고리즘이다[17,18]. 본 논문에

서의 파라메터 학습은 크게 두 단계로 구성된다. 첫 번째 단계에서는 파라메터의 반복적 학습을 위한 초기 파라메터(initial parameter)를 추정하는 단계이다. 두 번째 단계에서는 *EM* 알고리즘의 *E-step*(*Expectation step*)과 *M-step*(*Maximization step*)을 반복적으로 수행하면서 파라메터를 학습하는 단계이다. *EM* 알고리즘을 수행하기 위해서는 파라메터의 반복적 학습을 위한 시험데이타(test data)와 초기 파라메터를 결정하기 위한 학습데이타(training data) 또는 씨앗데이타(seed data)가 필요하다. 본 논문에서는 *EM* 알고리즘의 초기 파라메터를 추정하기 위한 학습데이타로 하나의 영어 조어단위로 구성된 영어용어를 포함하는 영-한 대역쌍과 하나의 한국어 조어단위로 구성된 한국어 용어를 포함하는 영-한 대역쌍을 이용하였다. 학습데이타는 그 자체로 조어단위 간의 정렬된 결과를 나타내므로 정렬의 학습데이타로 사용할 수 있다. 즉 하나의 개념단위에 대응되는 한국어 또는 영어 조어단위들의 집합으로 표현되기 때문에, 개념단위와 대응관계를 쉽게 찾을 수 있다. 학습데이타로 조어단위 간 대응관계의 집합인 $A(0)$를 구성할 수 있다. 학습데이타로 $A(0)$를 구성한 후, *EM* 알고리즘의 *M-step*을 통하여 초기 파라메터 $\theta(0)$을 계산한다. 그리고 초기 파라메터를 기반으로 *EM* 알고리즘을 반복적으로 수행하여 파라메터를 학습한다. 파라메터의 반복적 학습을 위한 시험데이타로 두 개 이상의 조어단위로 구성된 영어용어와 이에 대응되는 한국어 조어단위로 구성된 한국어용어를 이용하였다. 파라메터의 반복적 학습은 파라메터의 변화가 없을 때까지 수행한다(그림 6 참조).

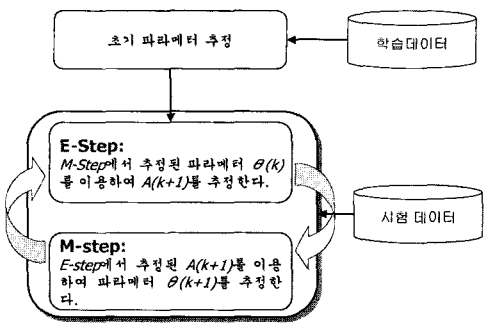

그림 6 EM 알고리즘을 이용한 파라메터 학습의 도식화

$A=\{a_1,...,a_n\}$를 정렬된 영한 조어단위 대역쌍의 집합이라고 하면, 본 논문에서 사용한 *EM* 알고리즘의 *E-Step*과 *M-Step*은 다음과 같이 표현된다.

- *E-step*: *M-Step*에서 추정된 파라메터 $\theta(k)$를 이용하여 $A(k+1)$를 추정한다.

$$A(k+1) = \arg\max_A p(A \mid E, K; \theta(k))$$

- *M-step*: *E-step*에서 추정된 $A(k+1)$를 이용하여 파라메터 $\theta(k+1)$를 추정한다.

$$\theta(k+1) = \arg\max_\theta p(\theta \mid A(k+1))$$

여기에서 파라메터는 식 (6)과 같이 정의된다. 또한 각 파라메터는 식 (7), (8)과 같이 추정된다. 식 (7), (8)은 라플라스 스무딩 기법 (Laplace smoothing method)[18]을 이용하여 영확률 (zero probability)을 방지한다.

$$\theta = \{\theta_{k^t{}_j|e_i}, \theta_{k^w{}_j|k^t{}_j,e_i}\} \tag{6}$$

$$\theta(k+1)_{k^t{}_j|e_i} = p(k^t{}_j \mid e_i; A(k+1)) = \frac{1+C(k^t{}_j, e_i; A(k+1))}{|E|+C(e_i; A(k+1))} \tag{7}$$

$$\theta(k+1)_{k^w{}_j|k^t{}_j,e_i} = p(k^w{}_j \mid k^t{}_j, e_i; A(k+1)) = \frac{1+C(k^w{}_j, k^t{}_j, e_i; A(k+1))}{|T|+|E|+C(k^t{}_j, e_i; A(k+1))} \tag{8}$$

여기에서 $C(A)$는 A의 빈도수를 나타내며, $|E|$는 전체 영어 조어단위의 개수를 $|T|$는 전체 한국어 조어단위 태그의 개수를 나타낸다.

식 (5), (6), (7), (8)에 의해서 식 (4)는 식 (9)와 같이 표현되며, 식 (9)를 이용하여 조어단위 정렬을 수행한다.

$$P(A \mid K, E) = \prod_{i=1}^{n} \prod_{j=1}^{t} \left[\theta_{k^t j,ei} \times \theta_{k^w j|k^t j,ei} \times a(i \mid j, n, t)\right] \tag{9}$$

여기에서 n은 영어 용어의 조어단위 수를, t는 한국어 용어의 조어단위 수를 각각 나타낸다.

4. 실험 및 평가

4.1 실험환경

실험을 위하여 물리, 화학, 생물학 분야의 조어분석된 영-한 대역사전을 사용하였다[15]. 사전은 영어와 한국어에 대하여 조어 분석된 결과를 포함하고 있다. 하나의 한국어 조어단위로 구성된 영-한 대역쌍과 하나의 영어 조어단위로 구성된 영-한 대역쌍을 학습데이타로 사용하고 나머지를 시험데이타로 사용하였다. 표 6은 각 분야별 학습데이타와 시험데이타의 분포를 나타낸다. 분야별로 학습데이타와 시험데이타의 비율이 다르며, 특히

표 6 실험데이타의 구성

분야	학습데이타	시험데이타	총데이타
생물	8,163	5,668	13,831
물리	2,757	8,047	10,804
화학	5,353	10,024	15,377

생물학 분야의 경우, 다른 분야에 비해 학습데이타가 시험데이타보다 많다.

평가를 위하여 두 가지 실험을 수행하였다. 첫 번째 실험은 분야별 조어단위정렬의 성능을 평가하기 위한 실험이다. 시험데이타에 대하여 정렬결과를 평가하며, 기본기법과 본 논문의 기법에 대하여 비교 평가한다. 기본기법은 한국어 조어단위에 대한 기저 명사구를 파악한 후 영어 조어단위와 한국어 기저 명사구의 개수가 같을 때, 순서대로 정렬하는 시스템이다.

두 번째 실험은 학습데이타 양에 따른 실험이다. 각 분야별로 학습데이타의 양을 변화시키면서 성능을 비교 평가한다. 학습데이타의 양은 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%로 학습데이타 수의 10%씩을 증가시키면서 평가한다.

조어정렬의 실험결과는 정확률(precision)과 재현율(recall)로 평가한다. 정확률은 전체 시험데이타의 용어 중 용어에 대한 조어단위 정렬이 올바르게 된 용어의 비율로 나타내어지며, 재현율은 전체 시험데이타의 조어단위 대역쌍 중 정렬모델이 파악한 대역쌍의 비율로 나타난다.

본 논문에서는 조어정렬에 사용한 대역모델의 성능을 평가하기 위하여 혼잡도(perplexity)를 사용하였다. 혼잡도는 단어에 대한 히스토리가 주어졌을 때, 언어모델 상에서 해당 단어의 다음에 올 수 있는 단어집합의 크기를 나타내는 척도이다. 즉 언어모델에 의해 반영되는 언어현상의 불확실성을 나타내는 척도로서 사용된다. 본 논문에서 사용한 혼잡도는 교차 엔트로피(cross entropy)의 변형으로 식 (10)의 $PP(T)$와 같이 표현된다[19].

$$PP(T) = 2^{H(T)}$$
$$H(T) = -\frac{1}{N}\log_2 P(T) = -\frac{1}{N}\sum_{i=1}^{N}\log_2 p(k_i \mid e_i) \quad (10)$$

여기에서 N은 대역관계의 수를 나타낸다.

4.2 실험결과

4.2.1 분야별 성능 평가

표 7은 분야별 영-한 조어단위 정렬 결과를 나타낸다. 실험결과에서 본 논문의 기법은 정확률에서 기본기법보다 약 30%~40%의 성능향상을 나타내며, 재현율에서도 약 16%~26%의 성능향상을 나타낸다. 또한 모든 분야에서 93% 이상의 정확률을 나타냄을 알 수 있다. 따라서 본 논문에서 제시하는 기법은 효과적으로 조어단위간 영-한 대응관계를 파악함을 알 수 있다. 그리고 본 논문에서 사용한 언어모델의 혼잡도는 화학분야가 가장 낮으며, 물리분야가 가장 높은 것으로 나타났다. 이러한 결과는 실험에 사용한 데이타의 양과 밀접한 관련성을 가지는데 실험집합의 크기가 가장 적은 물리분야의 혼잡도가 가장 높게 나타나며, 실험집합의 크기가 가장 큰 화학분야의 혼잡도가 가장 낮게 나타난다. 즉, *EM* 알고리즘에 의해 학습데이타와 시험데이타를 이용하여 비지도식으로 대역모델을 학습하기 때문에 실험집합의 크기가 클수록 혼잡도가 낮아짐을 알 수 있었다.

표 8과 9는 조어단위 정렬 결과로 나타나는 영어 조어단위와 한국어 조어단위 간의 대응쌍의 예를 분야별로 나타낸 것이다. 표 8은 하나의 영어 조어단위에 대응되는 한국어 조어단위들의 집합을 나타내고, 표 9는 하

표 7 분야별 영-한 조어단위 정렬 결과

분야	기본기법		제안기법		
	정확률	재현율	정확률	재현율	혼잡도
생물	72.65%	76.97%	94.76% (+30.43%)	91.94% (+16.28%)	103
물리	70.61%	66.03%	94.54% (+33.89%)	89.59% (+26.30%)	145
화학	66.51%	67.63%	93.47% (+40.53%)	89.13% (+24.12%)	98

표 8 분야별 영-한 조어단위 간 대응관계 (영어조어단위를 중심으로)

분야	영어조어단위	한국어조어단위
생물	abdominal	복부, 복, 배
	germinal	생식+질, 배아, 생식
	vesicle	포, 소포, 낭, 주머니
물리	resonance	공명, 껴+울림, 껴+떨기
	curvature	굽은+율, 굽음+율, 굽음, 곡률, 휨
	radiation	내+비침, 복사, 방사+선, 내+비침+선, 복사+선, 방사, 비침+선
화학	kinetic	운동, 동적, 반응+속도+론+적, 분자+운동, 반응+속도+법
	acetic	아세트산, 아세트, 초산
	excitation	들뜨기, 들뜬+상태, 들뜸, 들뜬

표 9 분야별 영-한 조어단위 간 대응관계(한국어 조어단위를 중심으로)

분야	한국어 조어단위	영어 조어단위
생물	발광	luminescent, luminous, photic
	양성	benign, bisexual, hermaphrodite, hermaphroditic, positive, proton
	교배	mating, cross, breeding, crossing
물리	축전지	acid battery, storage battery, battery
	굴절	refraction, refractive, refracted, refracting, refractor
	소리	acoustic, sound, tuning, acoustical, audio, tone, sonic
화학	브롬화	bromination, bromide, brominated
	다가	multiply charged, multivalent, ploy hydric, polyacidic, polybasic, polyfunctional, polyhydric, polyvalent
	핵	nuclear, nucleus, nuclei, nucleate, nucleic

나의 한국어 조어단위들에 대응되는 영어 조어단위의 집합을 나타낸다. 표에서 '+'는 조어단위 간의 경계를 나타낸다.

4.2.2 학습데이타 양에 따른 성능 평가

그림 7은 학습데이타 양에 따른 영-한 조어단위 정렬의 실험 결과를 나타낸다. 표 6의 각 분야별 학습데이타 중, 10%~100%로 학습데이타의 양을 변화시킬 때의 각 분야별 조어단위 정렬의 성능을 평가하였다. 실험결과에서 본 논문의 기법은 적은 양의 학습데이타만으로도 높은 성능을 나타냄을 알 수 있다. 이는 본 논문의 기법이 *EM* 알고리즘을 통한 파라메터의 반복적 학습으로 학습데이타가 부족하더라도 시험데이타로부터 파라메터를 재추정하기 때문으로 분석된다.

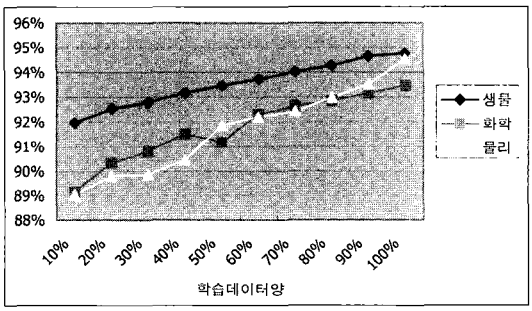

그림 7 학습데이타 양에 따른 성능

4.3 정렬 유형별 오류 분석

표 10은 생물, 화학, 물리 분야에서의 정렬 유형별 정확률과 각 정렬 유형이 전체 실험데이타에서 차지하는 비율을 나타낸다. 전체 실험데이타의 특성으로는 순차정렬로 나타나는 경우가 모든 분야에서 가장 많은 비율을 차지한다. 분야별 실험데이타의 특성으로는 생물분야의 경우 *N:1*정렬의 비율이 높으며, 화학분야의 경우 교차정렬의 비율이 높다. 또한 물리분야의 경우 순차정렬의 비율이 상대적으로 높게 나타난다. 정확률 측면에서는 순차정렬의 성능이 95~96%로 가장 높게 나타나며, 교차정렬이 32~65%의 정확률로 비교적 높게 나타나고 *N:1* 정렬이 가장 낮은 정확률을 나타낸다.

N:1 정렬에서 정확률이 낮은 이유는 실험데이타에서 영어 조어단위간의 응집성을 나타내는 데이타가 존재하지 않기 때문으로 분석되었다. 즉 영어 조어단위가 단독으로 사용되는 경우의 데이타는 많았지만 여러 개의 조어단위가 함께 사용되는 경향이 있는지에 대한 데이타의 부족으로 인하여 오류가 발생하였다. 예를 들어, 대역쌍 <*thin layer chromatography*, '박층 크로마토그래피'>에서 *thin*과 *layer*가 함께 자주 나타난다는 응집성에 대한 공기 데이타의 부족으로 인하여, 올바른 결과인 <박층, *thin layer*>, <크로마토그래피, *chromatography*>를 파악하지 못하였다. *N:1* 정렬의 오류를 줄이기 위해서는 이러한 영어 조어단위 간의 응집성을 나타내는 공기정보를 코퍼스 등을 통하여 획득하는 연구가 추가적으로 수행되어야 할 것이다.

교차정렬에 의해 나타나는 오류는 많은 경우 영-한 조어단위의 개수가 다를 경우 교차정렬을 허용하지 않았기 때문에 나타난 것으로 분석되었다. 교차정렬의 많

표 10 정렬 유형별 정확도

정렬 유형	생물		화학		물리	
	정확률	비율	정확률	비율	정확률	비율
N:1 정렬	46.28%	2.08%	36.29%	1.23%	41.13%	1.55%
교차정렬	60.00%	1.29%	65.14%	5.65%	32.14%	0.35%
순차정렬	96.27%	96.62%	95.94%	93.12%	95.60%	98.10%
전체	94.76%	100.00%	93.47%	100%	94.54%	100%

표 11 화학분야에서 교차정렬에 의한 오류의 예

영어 용어	한국어 용어
ammonium hydrogencarbonate	탄산 수소 암모늄
triethyl orthoformate	오쏘폼 산 트라이에틸
sodium hypochlorite	하이포아 염소+산 나트륨
cellulose ethanoate	에탄+산 셀룰로오스

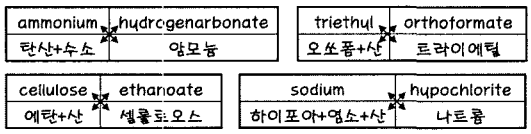

그림 8 표 11의 용어에 대한 조어단위 간 대응 관계

은 부분을 차지하는 화합물의 경우 표 11과 그림 8과 같이 영-한 조어단위의 개수가 다른 경우에도 교차정렬의 형태로 빈번하게 나타났다. 예를 들어, 두 개의 영어 조어단위와 네 개의 한국어 조어단위로 구성된 대역쌍 <*sodium hypochlorite*, '하이포아 염소+산 나트륨'>에서 *sodium*은 '나트륨'과 *hypochlorite*는 '하이포아 염소+산'과 대응관계에 있다.

5. 결 론

본 논문에서는 영-한 전문용어를 구성하는 조어단위 간의 통계적 정렬 기법을 제안하였다. 본 논문에서는 데이타 분석을 통하여 조어단위 간 정렬을 위한 제약조건을 설정하고 이를 이용하여 조어단위 간의 정렬을 위한 통계적 모델을 제안하였다. 또한 통계적 모델의 파라메터를 *EM* 알고리즘을 이용하여 추정하였다. 본 논문의 기법은 전문용어 대역쌍에서 조어단위간의 정렬 특성을 반영한 모델로서 비교적 높은 성능(평균 93% 정확률, 90%의 재현율)을 나타내었으며, 적은 학습데이타만으로도 우수한 성능을 나타내었다.

본 논문에서는 영어 단어를 하나의 개념단위로 가정하고 영-한 조어단위 정렬을 수행하였다. 하지만 "oo(egg) + cyte(cell)"로 분석되는 "oocyte"와 같이 여러 개의 조어단위로 구성되는 단어의 개념단위를 올바르게 파악하기 위해서는 영어단어에 대한 조어단위 분석이 추가적으로 필요하다.

본 논문의 기법은 새로운 영어 전문용어에 대역되는 한국어 용어에서의 한국어 조어 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. 이러한 조어 패턴과 전문용어의 변이는 영어기반의 한국어의 번역용어를 일관성 있게 제작하는데 중요한 자원으로 사용될 수 있을 것으로 기대된다.

참 고 문 헌

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

[2] 조은경, 서상규, "전문용어연구를 위한 복합용어 분석의 단위", 제 3회 전문용어언어공학심포지움, 2000.

[3] 조은경, 서상규, "전문용어의 조어 분석을 통한 개념분석", 제 4회 전문용어언어공학심포지움, 2001.

[4] 서충원, 배선미, 최기선, "조어법 정보를 이용한 전문용어의 영/한 번역 시스템 개발", 제 31회 정보과학회 춘계학술대회, 2004.

[5] 금성출판사, "금성판 뉴에이스 영-한 사전 제2판", 금성출판사, 1990.

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

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

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

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

[10] Och, F.J and Ney, H., "Statistical Machine Translation: From Single Word Models to Alignment Templates," PhD thesis, RWTH Aachen, Germany, 1999.

[11] Marcu, D. and Wong, W., "A phrase-based, joint probability model for statistical machine translation," In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002.

[12] Koehn, P. and Knight, K., "Empirical methods for compound splitting," In proceedings of the Meeting of the European Chapter of the Association of Computational Linguistics, 2003.

[13] Tillmann, C., "a projection extension algorithm for statistical machine translation," In Proceedings of Conference on Empirical Methods in Natural language Processing (EMNLP), 2003.

[14] Venugopal, A., Vogel, S., and Waibel, A., "Effective phrase translation extraction from alignment models," In proceedings of the 41st Annual Meeting of the Association of Computational Linguistics, 2003.

[15] 문화부, "전문용어 표준화를 위한 기반 조성", http://www.korterm.or.kr/ 중 자료, 2000.

[16] 이주호,최기선,이재성, "자동정렬을 통한 영한 복합어의 역어 추출", 제 12회 한글 및 한국어 정보처리 학

술발표 논문집, pp. 309-314., 2000.
[17] Demster,A.P., Laird,N.M., and Rubin,D.B., Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1), 1-38., 1977.
[18] Manning, C.D. and H. Schutze, Foundations of statistical natural language processing, MIT Press, 1999.
[19] Ney, H. Language Models, In Gibbon, D., Moore, R. & Winski, R. (eds) Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter., 1997.

부 록

한국어 조어분석에 사용한 품사 태그

	형태, 품사	표지
	기호	sy
	표지 보류	tt
	보통명사	nc
	대명사	np
명사류	고유명사	npp
	의존명사	nb
	수사	nu
	아라비아숫자포함명사	nu cc
	관형사	an
	부사	av
접사	접두사	xp
	접미사	xs
	어근적 형태소	mm
	형태, 품사	표지
	기본형	vb
용언	명사형, 명사파생접미사 결합형	vn
	관형형	va
	연결형	vc
	조사	pa
체언	형용사, 부사	af
이외의	동사	vf
외래어	전치사	pf
	접속사	cf
	접사	xf
	단일 단위 준말	표지 c
	복합 단위 준말	표지 cc
	약어	acn

오 종 훈

1998년~성균관대학교 정보공학과 졸업 (학사). 2000년~한국과학기술원 전산학과 졸업 (공학석사). 2000년~현재 한국과학기술원 전산학과 박사과정. 관심분야는 자연언어처리, 전문용어, 정보검색 등

황 금 하

1991년~중국 길림대학 물리학과 졸업 (학사). 2000년 한국과학기술원 전산학과 졸업 (공학석사). 2000년~현재 한국과학기술원 전산학과 박사과정. 1994년~1997년 중국 연변과학기술대학 전산실 (직원). 2001년~2003년 Microsoft Research, Asia (Assistant Researcher). 관심분야는 자연언어처리, 기계학습 등

최 기 선

1978년 서울대학교 수학과 졸업(학사) 1980년 한국과학기술원 전산학과 졸업 (공학석사). 1986년 한국과학기술원 전산학과 졸업(공학박사). 1985년~1986년 한국외국어대학교 전산학과 조교수. 1987년~1988년 일본 NEC C&C 정보연구소 초빙연구원. 1988년~현재 한국과학기술원 전산학과 교수 1998년~현재 한국과학기술원 전문용어언어공학연구센터 소장. 관심분야는 자연언어처리, 기계번역, 정보검색, 전문용어 등