중-한 대조분석정보를 이용한 단어정렬

리금희[0,*] 김동일[**] 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터
{ljj, dongil, jhlee}@postech.ac.kr

Word Alignment Using Chinese-Korean Linguistic Contrastive Information

Jinji Li[0,*] Dong-il Kim[**] Jong-Hyeok Lee
Department of Computer Science and Engineering,
Division of Electrical and Computer Engineering,
Pohang University of Science and Technology,
and Advanced Information Technology Research Center(AITrc)

요 약

본 논문에서는 범용 병렬코퍼스에서도 적용할 수 있는 단어정렬의 방법을 제안한다. 단어 단위로 정렬된 병렬코퍼스는 자연언어처리의 다양한 분야에 도움을 준다. 예를 들면 변환기반의 기계번역에서 변환패턴의 구축, MWTU(Multi Word Translation Unit)의 자동추출, 사전 구축, 의미 중의성 해소 등 분야에 적용된다. 중한 병렬 코퍼스의 단어정렬은 서로 다른 어족간의 관계의 규명을 포함하고 있기 때문에 본 논문에서는 통계적인 모델보다 중한 대역어 사전, 단일어 시소러스, 품사정보 및 언어학적 대조분석 정보 등 기존에 있는 리소스를 이용하여 재현율과 정확률을 높이는 방법에 대해 제시한다. 성능 평가를 위해 중앙일보에서 임의로 추출한 500개 대응문장을 이용하여 실험한 결과 82.2%의 정확률과 64.8%의 재현율을 보였다.

1. 서론

병렬 코퍼스의 중요성이 높아짐에 따라 병렬 코퍼스에 대한 연구가 점점 활성화 되어가고 있다. 병렬 코퍼스가 제공할 수 있는 정보량은 단일어 코퍼스보다 많다[4]. 또한 활용분야가 다양한 연유로 인하여 병렬 코퍼스에 대한 연구가 많이 이뤄지고 있는데 핵심은 병렬 코퍼스를 이용함에 있어서 기초 작업으로 되는 정렬방법론이다. 병렬 코퍼스의 정렬은 크게는 문서에서부터 단락, 문장, 구, 단어 등 단위로 나눌 수 있다. 단어 단위로 정렬된 병렬 코퍼스는 변환을 기반으로 한 기계번역에서 변환패턴의 구축, MWTU(Multi Word Translation Unit)의 자동 추출, 사전 구축, 의미 중의성 해소 등 자연어 처리 여러 분야에 유용하게 쓰일 수 있다.

일반적으로 중국어-한국어와 같이 서로 다른 어족에 속하는 언어들 간에는 단어단위의 정렬이 단순하지 않기 때문에, 통계적인 모델을 단독으로 적용하기 보다는 언어학적 지식과 결합하여 정렬을 하는 경우가 많다. 본 논문에서는 중한 및 한중 대역어 사전, 중국어 및 한국어 시소러스 및 부분적 언어학적 지식 등 리소스를 이용하여 단어 단위의 정렬을 하려고 한다.

단어정렬이란 병렬코퍼스에서 서로 대응되는 단어를 찾아내는 작업이다. 본 연구에서는 중국어는 단어분리(word segmentation)된 한 개의 단어를 정렬의 기본 단위로 하고 한국어의 정렬 기본단위는 한 개 어절로 한다. 그러나 정렬의 재현율과 정확률을 높이고저 중국어와 한국어의 언어학적 차이를 고려해서 전처리를 통하여, 중국어에서는 한국어의 한 개 어절로 번역되는 형태소들은 미리 결합시켜 정렬의 기본 단위로 간주하고, 한국어에서는 본용언과 보조용언은 결합시켜 한 개의 정렬단위로 정

*: 중국 길림성 연길시 연변과학기술대학교 강사.
**: 중국 길림성 연길시 연변과학기술대학교 부교수.

의한다.

본 논문의 구성은 다음과 같다. 2절에서는 단어 정렬에 관한 기존의 연구를 간략하게 소개하고, 3절에서는 본 논문이 제시한 방법론에 대해 단계별로 설명한다. 4절에서는 실험 및 결과를 제시하고 실험 결과 분석과 함께 오류 분석을 제시한다. 그리고 마지막으로 5절에서는 본 논문의 결론을 제시하고, 향후 계획에 대해서 알아본다.

2. 관련연구

병렬코퍼스에서 단어정렬은 정렬의 대상이 되는 두 언어가 속한 유형에 따라서 방법론과 정렬되는 기본단위도 현저한 차이가 있다.

영어, 불어와 같이 같은 어족에 속하고 구조가 유사한 언어 쌍에 대해서는 [8]과 같이 서로 대응되는 문장에서 대역 단어가 자주 공기한다는 특징과 위치정보를 이용하여 정렬 모델을 제안하였고, 대용량의 병렬 코퍼스를 사용하여, 모델의 파라미터를 학습시켰다. [10]에서는 중국어의 언어적 특징을 고려해 단어목록(lexical criteria)을 이용하여 중영 병렬코퍼스의 단어정렬을 진행했다. 이런 방법들의 특징은 모두 대용량의 학습 병렬 코퍼스를 이용하여 정렬모델의 파라미터를 학습시켰다는 것과 언어학적 지식을 거의 사용하지 않았다는 것이다.

통계적 방법에 의존한 이전의 단어정렬과는 달리 [9]에서는 중영 대역어 사전과 중국어 및 영어 단일언어 유의어 사전을 이용한 클래스 기반의 영어-중국어 간의 단어 정렬 기법을 제안하였으며 이런 방법을 통하여 많은 양의 병렬 코퍼스를 학습시킴에도 불구하고 얻어지는 통계적 방법의 낮은 재현율 문제를 극복하였다.

위의 논문에서는 모두 정렬은 주어진 대응 문장에서 서로 대응되는 단어나 구를 찾아내는 작업이라고 정의하고 있다.

이와는 달리 [1]에서는 정렬을 주어진 원문에 대한 번역문을 찾아내는 작업으로 단어 단위의 정렬은 정렬된 대응문장에서 의미적 유사도가 가장 큰 두개의 단어를 찾되 비슷한 유사도를 가진 후보가 하나 이상일 경우 구문적 유사도가 가장 큰 단어를 찾는 것이라고 정의하고 있다. 그리고 단어의 유사도를 찾기 위하여, 중한 대역어 사전, 유의어 사전 외에 양국어 문자, 어휘 및 구문 지식을 통계적 방법과 결합하여 사용하였고, 단어단위에서 구단위로 확장할 때는 원시 언어 구문분석기를 사용하기도 하였다.

그러나 [1]에서도 통계학적인 방법으로 인하여 많은 양의 코퍼스를 필요로 하고, 고려하는 언어학적 지식으로 인하여 다량의 리소스를 필요로 하기도 한다. 또한 54.3%의 비교적 낮은 재현율을 보였다.

본 논문에서는 중한 및 한중 대역어 사전, 중한 단일언어 시소러스, 품사정보 및 간단한 언어학적인 대조분석정보를 기반으로 하고, 범용 병렬코퍼스에도 적용할 수 있으며, 보다 높은 재현율과 정확률을 얻을 수 있는 방법을 제안하려고 한다.

3. 중한 단어 정렬

3.1 언어학적 대조분석정보 이용 단계

중국어-한국어와 같이 서로 다른 어족에 속하는 언어는 단어정렬을 할 때, 정렬하려고 하는 기본단위가 서로 다를 수 있다. 본 논문에서는 중한 단어 정렬에서의 기본단위를 다음과 같이 정의했다. 즉 중국어는 문장에서 분리된 매 단어를 정렬의 기본 단위로 하고, 한국어는 매 어절을 단어정렬의 기본 단위로 간주한다. 그러나 중국어에서 한국어의 한 개 어절로 번역되는 형태소들은 결합시켜서 한 개의 정렬 기본 단위로 하고, 한국어에서는 본용언과 보조용언은 결합하여서 단어정렬의 기본 단위로 한다. 예를 들면 [그림 1]과 같다. 이때 K1의 "돌이키"는 본용언이고 "보"는 보조용언이다.

C1: 首先/ 回□/ 一下/ □□/ 的/ 情景/ ./

K1: 당시 상황+을 돌이키+어 보+자.

[그림 1] 본용언과 보조용언의 결합 예

고립어에 속하는 중국어의 두드러진 특징중의 하나가 바로 단어구성의 간결성이다. 중국어 단어는 여러 형태소가 결합된 것이 아니라 그 자체가 하나의 형태소이다[7]. 이런 중국어 단어가 굴절어인 한국어의 한 개 어절에 대응될 때 여러 개의 형태소가 한 개 어절에 대응되는 경우가 빈번히 나타난다. 본 논문의 실험에서 사용한 중한 병렬코퍼스의 중국어 문장은 약 22.4단어로 이뤄졌고, 한국어 문장은 약 10.8어절로, 평균 중국어의 두개 단어가 한국어의 하나의 어절에 대응된다는 것을 알 수 있었다. 실험을 기반으로, 중국어와 한국어의 언어학적 대조특성을 고려하여, 가장 많이 나타나는 몇 가지 현상을 정의해 보도록 한다.

● 중국어는 양사[1]가 아주 발달한 언어지만 한국어

[1] 양사는 중국어 품사인데, 한국어의 단위성 의존명사와 같은 개념이다.

로 번역될 때 이런 양사들은 대부분이 번역되지 않는다. 때문에 중국어에서 수사 혹은 대명사는 양사와 결합하여 하나의 정렬 단위로 본다. 그러나 한국어에서 양사가 번역될 경우 수사와 단위성 의존명사를 결합하여 [그림 2]와 같이 한 개의 정렬단위로 간주한다.

```
C2: 我/ □/ 到/ 了/ 那/ 本/ 字典/ ./

K2: 나+는 그 사전+을 사+았+다.

C3: 我/ □/ 了/ 三/ 本/ 字典/ ./

K3: 나+는 사전 세 권+을 사+았+다.
```

[그림 2]수사와 양사의 결합 예

- 중국어는 명사 혹은 대명사 뒤에 방위사[2]를 사용하여 장소나 위치를 나타내는 경우가 있는데, 이때 방위사는 한국어 번역에서 나타나지 않는 경우가 대부분이다. 때문에 중국어에서 이런 구조를 가진 것은 결합하여 [그림 3]과 같이 하나의 중국어 정렬 단위로 간주한다.

```
C4: 我/ 在/ 家/ 里/ 看/ □□/ ./

K4: 나+는 집+에서 TV+를 보+았+다.
```

[그림 3]명사와 방위사의 결합 예

- 중국어에는 결과동사복합어(Resultative Verb Compounds)라는 구조가 존재하는데 크게는 두 개의 구성성분으로 나뉘어져 있다. 두 번째 구성성분은 결과보어라고 하며, 첫번째 구성성분의 동작이나 결과를 보충해 설명해주는 작용을 한다. 이때 결과보어는 한국어에서 보조용언으로 번역되거나 생략되는 경우가 많이 존재하게 된다. 때문에 결과동사복합어는 결합하여서 [그림 4]와 같이 정렬의 기본 단위로 간주한다.

```
C5: 在/ □□/ 上/ □/ 上/ 世界/ 地□/ ./

K5: 도화지+에 세계지도+를 그리+ㄴ다.
```

[그림 4]결과동사복합어의 결합 예

한국어에서 보조용언은 독자적인 의미기능을 하지 못하고 다른 동사나 형용사에 의존하여 의미를 보충해 주는 역할을 하게 된다. 그리고 보조용언은 중국어의 어떤 특정 문장 성분과 정확히 대응된다고 말할 수도 없다. 이런 특징 때문에 본용언과 보조용언을 결합시켜서 한 개의 정렬 단위로 처리한다.

그리고 단어정렬을 할 때 번역 과정에서 한국어에서 기능어로 번역되는 중국어의 허사(전치사, 조사, 어기사(□□□), 감탄사, 상태사(□□□)) 등은 정렬의 대상에서 제외시켰다. 한국어에서도 실제 정렬과정에서 유사도 계산에 사용된 어휘는 기능어를 제외한 단어집합 뿐이다.

3.2 대역어 사전과 품사정보를 이용한 정렬

대용량의 기계 가독형 사전의 사용이 가능함에 따라, 대역어 사전을 단어 정렬에 이용하는 기법이 많이 제시되고 있다. 이것은 대역어 사전이 단어와 단어간의 형태적 유사도 정보를 포함하고 있기 때문이다. 대역어 사전이 제시한 어휘와 실제 문장에 나타난 어휘간의 정확한 대응이 일어나지 않는 경우를 대비하여 Dice coefficent[5] 공식을 이용하여 형태적으로 유사도가 가장 높은 단어를 찾는 정렬을 시도한다.

한국어 단어 k_1과 k_2의 Dice coefficent를 구하는 공식은 다음과 같다.

$$Sim(k_1,k_2)=\frac{2\times|k_1 I\ k_2|}{|k_1|+|k_2|}$$

$|k_1 I\ k_2|$ = k_1과 k_2가 공통으로 가지고 있는 문자의 개수.
$|k_1|$ = k_1의 문자의 개수.
$|k_2|$ = k_2의 문자의 개수.

중국어 단어 c와 한국어 단어 k간의 유사도는 아래의 공식으로 구할 수 있다.

$$DicSim(c,k)=w*\max_{k_i\in TCset} Sim(k_i,k)$$

TCset = 대역어 사전에서 나타나는 중국어 단어 c에 대응하는 한국어 대역어 집합.
k_i = TCset의 원소.
w = 1보다 작은 상수이며, c혹은 k의 단어길이가 1 혹은 c와 k의 품사가 서로 다른 대응관계를 가질 때 적용하는 값.

[2] 방위사는 중국어 품사인데, 방향이나 위치를 나타낸다.

중국어 어휘가 한국어 어휘로 변환될 때 품사대응에 관한 일정한 규칙이 존재하기 때문에 중국어와 한국어의 품사특징을 고려해 그 대응관계를 다음 [표 1]과 같이 정했다. 중국어와 한국어의 품사분류의 차이점때문에 다대다 대응도 일어난다.

[표 1]: 중국어 품사와 한국어 품사의 대응관계

순서	중국어 품사	한국어 품사
1	성어, 습관용어, 명사, 장소사, 시간사	비서술성명사
2	성어, 습관용어, 동사	동작성명사, 지시동사
3	접속사	접속부사
4	수사	양수사, 서수사, 수관형사
5	…	…

중한 대역어 사전과 품사정보를 이용한 단어정렬에서는 DicSim(c,k) 값이 일정한 임계치보다 높은 중한 단어에 한해서 정렬을 확정 시켰다. 이 방법은 높은 정확률은 보이나 반면에 매우 낮은 재현율을 보였다. 재현율이 낮은 원인은 번역을 진행할 때 사용하는 어휘의 다양성과 두 언어의 형식, 의미, 표현 양식의 차이점 때문에 일어나는 문제라고 볼 수 있다.

3.3 시소러스를 이용한 클래스기반의 정렬

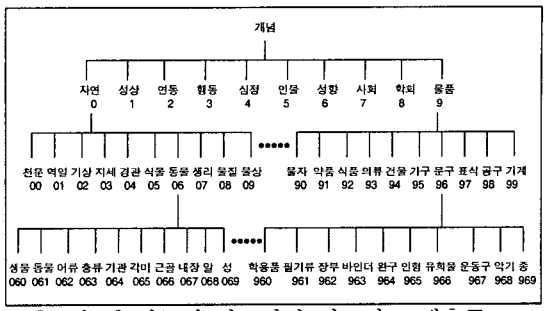
대역어 사전을 기반으로 하는 정렬에서 나타난 낮은 재현율의 문제를 극복하기 위하여 중국어와 한국어의 시소러스를 이용한 클래스 기반의 정렬을 시도하였다.

중국어 시소러스는 중국에서 많이 쓰이고 있는 중국어 유의어사전을 사용하였는데 이 사전은 12개의 대 분류, 94개의 중 분류 및 1428개의 소 분류로 이루어져 있다. 예를 들면 "文具" 란 단어는 Bp16 이란 의미코드를 가지는데, B 대분류, p 중분류 및 16 소분류에 위치하고 있다.

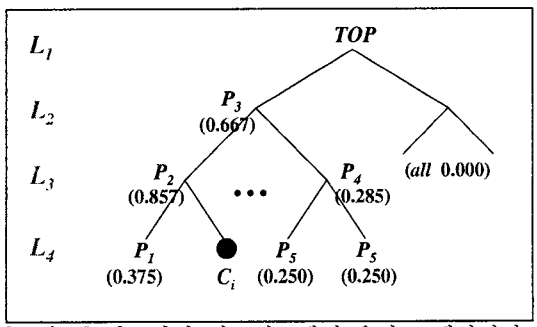
한국어 시소러스는 일본의 가도까와 시소러스의 의미분류체계를 기반으로 하였고 [그림 5]과 같다. 한국어 시소러스에서 "학용품" 이란 단어는 물품류의 문구 소분류에 속하며 "960" 란 의미코드를 가지게 된다.

한국어에서 개념노드간의 유사도는 [2]가 제시한, 다음과 같은 공식을 이용하여 구할 수 있다.

$$\mathrm{KSim}(C_i,P_j)=\frac{2\times \mathrm{level}(\mathrm{MSCA}(C_i,P_j))}{\mathrm{level}(C_i)+\mathrm{level}(P_j)}\times \mathrm{weight}$$



[그림 5] 한국어 가도까와 시소러스 계층구조



[그림 6] 가도까와 시소러스에서 유사도 계산결과

MSCA(C_i,P_j) = Most Specific Common Abstraction. 즉 Ci와 Pj의 가장 가까운 공통 상위개념.

weight = C_i가 P_j의 하위개념이면 1이고, 아니면 0.5.

level(x) = 가도까와 시소러스상에서 개념노드 x가 위치한 계층의 번호.

이 공식을 이용하여 유사도를 계산하면 [그림 6]과 같은 결과를 보인다.

가도까와 시소러스 상에서 두 한국어 개념노드의 유사도를 구하는 공식을 이용하면 중국어와 한국어 단어를 클래스로 확장하여 유사도를 계산하는 공식을 얻을 수 있다.

$$\mathrm{CKSim}(c,k)=\max_{\substack{s_c\in \mathrm{CSet}\\ s_k\in \mathrm{KSet}}} \mathrm{KSim}(S_c,S_k)$$

CSet = 중국어 단어 c의 대역어가 속해 있는 한국어 의미코드 집합.

KSet = 한국어 단어 k가 속해 있는 한국어 의미코드 집합.

S_c = CSet의 원소.

S_k = KSet의 원소.

위 공식을 이용하여 계산된 유사도는 클래스 기반의 정렬에서 첫번째 단계에서 이용되어 일정 임계치를 넘는 유사도를 가진 중한 대응단어에 대한 정렬을 확정짓는다.

다음은, 앞 단계보다 일반적인 유사도를 구하기 위하여, 서로 다른 체계를 가지는 중국어와 한국어 두 시소러스의 의미코드 간의 유사도를 계산하는 작업을 다음과 같이 진행한다.이 유사도는 중한, 한중 대역어 사전과 아래의 공식을 이용하여 계산하였다.

$$\mathrm{ConceptSim}(S_c, S_k) = \frac{|X\ I\ TY| + |Y\ I\ TX|}{|X| + |Y|}$$

S_c = 중국어 의미코드.
S_k = 한국어 의미코드.
X = 중국어 의미코드 S_c를 가지는 어휘의 집합.
Y = 한국어 의미코드 S_k를 가지는 어휘의 집합.
|X| = 집합 X에 속하는 어휘의 개수.
|Y| = 집합 Y에 속하는 어휘의 개수.
TX = X에 대응되는 한국어 대역어 집합.
TY = Y에 대응되는 중국어 대역어 집합.
|X I TY| = 집합 X와 집합 TY에 공통으로 나타나는 어휘의 개수.
|Y I TX| = 집합 Y와 집합 TX에 공통으로 나타나는 어휘의 개수.

중국어 의미코드와 한국어 의미코드의 유사도 계산 공식을 확장하면 다음과 같이 중국어 단어와 한국어 단어의 유사도 계산 공식을 얻을 수 있다.

$$\mathrm{ClassSim}(c,k) = \max_{\substack{S_c \in CSet \\ S_k \in Kset}} \mathrm{ConceptSim}(S_c, S_k)$$

CSet = 중국어 단어 c의 대역어가 속해 있는 한국어 의미코드 집합.
KSet = 한국어 단어 k가 속해 있는 한국어 의미코드 집합.
S_c = CSet의 원소.
S_k = KSet의 원소.

위 공식을 이용하여 정렬되지 않은 나머지 단어들에 대해서는 클래스 기반의 두 번째 단계의 정렬을 적용한다. 이 단계에서도 마찬가지로 일정한 임계치보다 큰 유사도를 가지는 대응 단어들은 정렬이 된다. 클래스 기반의 방법은 사전기반의 정렬에서 존재한 재현율이 낮은 문제는 어느 정도 해결되지만 정확률의 저하가 발생한다. 그러므로 클래스 기반의 정렬만 단독으로 적용하는 것은 효과적이지 못하다.

3.4 사전기반 정렬과 클래스기반 정렬의 hybrid 방법

클래스기반의 정렬을 단독으로 적용할 때 결과가 바람직하지 않기 때문에 이 방법은 사전기반 정렬과 결합한 방식이며, 사전기반 정렬의 낮은 재현율과 클래스기반의 낮은 정확률을 해결하기 위한 방법이다.

$$\mathrm{HybridSim}(c,k) = w_1 \times \mathrm{DicSim}(c,k) + w_2 \times \mathrm{CKSim}(c,k) + w_3 \times \mathrm{ClassSim}(c,k)$$

w_1, w_2 및 w_3는 각각 사전기반의 정렬에서 얻은 유사도와 클래스 기반의 정렬에서 얻은 유사도에 대한 가중치이다.

이 방법을 통하여 얻어지는 결과는 앞 두 단계에서 얻어지는 정렬결과와 충돌이 발생할 수 있다. 그러나 본 논문에서는 우선 앞 두 단계에서 정렬이 되지 않은 나머지 단어들에 대해 정렬을 진행한다.

4. 실험 및 결과 분석

본 실험은 중앙일보에서 임의로 추출한, 문장단위에서 이미 정렬된 500개의 중한 대응문장에 대하여 단어정렬을 하였다. 실험에서 84,000여 표제어의 중한 대역어 사전과 56,000여 표제어의 한중 대역어 사전, 40,000여 표제어의 중국어 유사어 시소러스와 200,000여 표제어의 한국어 가도까와 시소러스를 이용하였다. 실험에서 사용한 중국어 문장은 평균 22.4개 단어이고, 한국어 문장은 평균 10.8어절이였다. 사용한 500개의 대응문장은 품사 태거를 이용해 태깅되었다.

실험은 언어학적 대조분석 정보가 적용된 병렬코퍼스에 사전기반 정렬, 클래스기반 정렬 및 사전기반과 클래스기반 결합 방법을 차례로 적용시켰으며, 매 단계별로 유사도가 일정 임계치보다 높으면 정렬되는 방법으로 순차적인 정렬을 진행하였고 실험결과는 [표 2]에 제시했다. 또 실험결과의 분석을 위하며 사전기반 정렬, 클래스기반 정렬 방법을 각각 단독으로 적용시켜 보았고 결과는 [표 3]에 제시하였으며, 언어학적 대조분석정보 이용단계가 있을 때와 없을 때의 실험결과도 [표 4]에 제시하였다.

[실험결과]

[표 2] 단계별로 정렬을 한 결과

정렬방법	정확률(%)	재현율(%)
사전기반	95.4	25.6
클래스기반	85.1	35.2
Hybrid	82.2	64.8

[표 3] 사전기반 정렬과 클래스기반 정렬을 각각 적용한 결과

정렬방법	정확률(%)	재현율(%)
DicSim	95.4	25.6
CKSim+ClassSim	72.8	30.9

[표 4] 대조분석정보를 이용한 정렬의 결과

정렬방법	정확률(%)	재현율(%)
대조분석 정보 미적용	79.8	63.2
대조분석 정보 적용	82.2	64.8

[실험결과 분석]

실험결과로부터 볼 수 있듯이 단순히 사전에 기반한 방법은 높은 정확률은 보이나, 매우 낮은 재현율을 보인다. 앞에서도 언급했던 것처럼 이것은 중국어를 한국어로 번역할 때, 어휘선정의 다양성과 두 언어간의 표현형식의 차이점을 설명해 준다. 낮은 재현율은 클래스기반의 방법에서 어느 정도 만회할 수는 있었지만, 정확률의 저하를 일으켰다. 그러나 이 방법을 사전기반 정렬과 결합하여 사용했을 때 정확률과 재현율이 많이 높아짐을 알 수 있었다. 그리고 언어학적 대조분석 정보를 이용한 단계가 적용되었을 때 정확률과 재현율이 다소 높아짐을 보였다.

[실험오류 분석]

1. 사전 정보의 부족으로 인한 오류:
   예를 들면 사전 기반의 단어정렬에서는 사전에 기술된 대역어 정보가 아주 중요한 작용을 하지만, 사전 구축의 어려움으로 말미암아, 보통 빈도수가 높은 대역어 단어만 사용한다. 이것은 단어와 단어간의 유사도 값의 저하를 초래한다.

2. 중국어의 단어분리 오류와 중국어와 한국어 태깅오류로 인한 오류:
   중국어는 단어와 단어간에 띄어쓰기가 없는 언어이다. 또 중국어는 거의 모든 한자가 한 개의 단어를 이룰 수 있기 때문에 미등록어가 나타나면 대부분 형태소 단위로 분리시키는데, 이런 결과는 단어정렬의 오류로 연결된다. 그리고 웹에서 수집한 코퍼스이기 때문에 비문을 포함하고 있는 것이 실험오류로 나타나고 있다.

3. 중국어와 한국어가 가지고 있는 고유의 특징으로 인한 오류:
   예를 들면 중국어의 속담, 성어, 관용구 등은 한국어로 번역될 때 의역하는 경우가 많기 때문에 정확한 정렬이 이루어 지지 않는다. 그리고 중국어의 형태론적 결합방식의 다양성으로 말미암아 일어나는 오류도 있다.

이 세 가지 오류 중에서 첫 번째와 두 번째 오류는 정렬의 방법론과는 직접적인 연관이 없는 것이라고 볼 수 있다. 세 번째 유형의 오류의 발생 원인은 중국어의 형태론적 결합방식에 여러 가지가 있는데, 중첩되어 쓰이거나, 단어가 분리되어 쓰이는 경우가 있기 때문이다. 예를 들면, [그림 7]과 같다.

C6: □/ 我□/ □□/ □□/ □□ □□/ ./

K6: 우리+가 이 문제+를 좀 토론+하+게 하+여라.

C7: 我/ □/ 了/ 一下/ 他/ 的/ 玩笑/ ./

K7: 나+는 그+에게 농담+을 하+였+다.

[그림 7] 중국어 형태론적 결합방식의 예

C6에서는 "토론"이란 단어가 두 번 반복하여 쓰여서, "좀 토론하다"라는 뜻을 나타냈고, C7에서는 "농담을 하다"라는 뜻을 가진 중국어 단어가 단어의 동사 부분과 명사부분이 분리되어 나오는 현상을 보였다. 중국어는 이와 같은 형태소 결합방식이 아주 다양한 언어이다. 그러므로 세 번째 유형의 오류는, 각각의 언어에서 MWTU의 추출과 다른 언어에로의 대응에도 관련된 문제로서, 현재의 방법으로써는 해결 할 수 없지만, 단어정렬을 통하여 오류로 나타나는 문제점들을 정확히 분석하고, 기존의 단어정렬 방법과 더불어 추가적인 언어학적 지식과 구문분석 등을 통해서 정렬을 구단위로 확장하면 정렬이 가능할 것으로 예상된다.

5. 결론 및 향후 계획

본 논문에서는 범용 병렬코퍼스에서 사용할 수 있는 단어정렬의 기법을 제안하였다. 대량의 병렬 코퍼스를 구축하기 힘들고 따라서 통계적인 방법을 적용하기 힘든 상황에서, 본 논문에서는 기계번역 사전을 비롯한 기존에 구축된 다양한 언어 자원들을 이용하는 방법을 제시하였다. 그리고 정렬의 기본 단위를 새롭게 정의하고 정렬하기 전에 언어학적 대조분석 정보 이용 단계를 거쳐 단어정렬의 재현율과 정확률을 향상시켰다.

본 논문은 중앙일보에서 추출한 500개의 대응 문장에 대해 단어정렬을 하였으며 82.2%의 정확률과 64.8%의 재현율을 보였다.

단어 정렬된 병렬코퍼스는 자연언어처리의 많은 분야에 이용할 수 있다. 본 논문에서 제안한 방법과 다량의 코퍼스를 이용하는 통계적인 방법을 결합한다면 단어 정렬의 재현율과 정확률을 더 향상시킬 수 있을 것으로 예상된다. 향후에는 단어단위에서 정렬된 병렬 코퍼스를 이용하여 MWTU의 자동 추출 등에도 활용하고자 한다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

참고 문헌

[1] 황금하. 2000. *정렬을 통한 다단계 변환패턴의 자동 구축*. 석사논문, 한국과학기술원

[2] 문경희, 이종혁, 김정인, 양기주. 1998. *일-한 기계 번역 시스템: 연어 패턴을 이용한 어휘 다의성 해소*. 정보과학회논문지(B)25(8): 1270~1280.

[3] Chen, Stanley F. 1993. *Aligning sentences in bilingual corpora using lexical information.* Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pages 9~16.

[4] Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. *Two languages are more informative than one*. In Processing of the 29th Annual Meeting of the Association for Computational Linguistics, pages 130~173.

[5] Dice, L.R. 1945. *Measures of the amount of ecologic association between species.* Journal of Ecology, 26, pages 297~302.

[6] Eiichiro SUMITA. 2001. *Word-alignment using bilingual lexical resources and DP-matching.* 19th International Conference on Computer Processing of Oriental Languages, pages 263~268.

[7] Li C N., Thompson S A. 1981. *Mandarin Chinese: A functional reference grammar.* University of California Press, USA.

[8] Peter F. Brown, Stephen A. Deela Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, 19(2), pages 263~311.

[9] Sue J. Ker, Jason S. Chang. 1997. *A Class-based Approach to Word Alignment.* Computational Linguistics 1997 Volume 23, Number 2, pages 313~343.

[10] Wu, Dekai, 1994. *Aligning a parallel English-Chinese corpus statistically with lexical criteria.* Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics, pages 80~87.

[11] Wu, Dekai, 1994, *Learning an English-Chinese lexicon from a parallel corpus.* Proceedings of Association for Machine Translation in the America, pages 206~213.

[12] 王斌, □群, □祥. 1999. □英□□□□□□□□究. □算□言□文集, pages 123~128.

[13] □雅娟, □□□, 李生, □沐昀. 2001. □□□□典方法相□合的□□□料□□□□, 自然□言理解□机器翻□, pages 108~115.

[14] 梅家□. 1983. *同□□□林*. 上海□□出版社.