자동 정렬을 통한 영한 복합어의 역어 추출

이주호*, 최기선*, 이재성**
*한국과학기술원 전자전산학과, 전문용어언어공학연구센터, 첨단정보기술연구센터
**충북대학교 컴퓨터교육과
{mywork, kschoi}@world.kaist.ac.kr, jasonl@cbucc.chungbuk.ac.kr

Extraction of English-Korean Compound Noun Translation through Automatic Alignment Method

Ju-Ho Lee*, Key-Sun Choi*, Jae-Sung Lee**
*Dept. of Electrical Engineering & Computer Science / KORTERM / AITrc, KAIST
**Dept. of Computer Education, Chungbuk National University

요 약

본 논문에서는 양국어로 된 병렬 코퍼스로부터 복합어의 역어를 추출하기 위한 정렬 방법을 제시한다. 여기에서는 개념어에 대한 양국어 공기정보를 사용하여 기본 정렬을 하고, 인접한 개념어로 정렬의 단위를 확장했다. 또한 재추정 기법을 사용하여 대역 확률을 계산함으로써 보다 높은 정확률을 얻을 수 있었다. 본 논문에서 제안한 방법을 적용하여 139,265개의 영어 어절로 이루어진 우루과이 라운드 영한 병렬 코퍼스에 대해서 실험한 결과 2,290개의 대역어 쌍을 얻었고, 그 정확률은 74%였다.

1. 서론

병렬 코퍼스로부터 정렬을 통하여 여러 가지 유용한 언어학적 지식을 얻을 수 있다. 이제까지 정렬에 관한 연구는 문장 단위의 정렬에서부터 단어 단위의 정렬, 구 단위 정렬 등 여러 가지로 많이 진행되어 왔다. 일반적으로 영어-한국어 같이 문장의 구조나 단위가 다른 언어 쌍에 대해서는 정렬이 어렵다.

본 논문에서는 도메인을 제한하여 양국어 공기정보를 기초로 한 영한 복합어 정렬 방법을 제시한다. 전문분야에서는 사용하는 어휘가 제한되어 있고, 의미에 있어서 애매성이 상대적으로 적다. 따라서 비교적 쉽게 공기 정보를 이용해서 역어 정보를 추출할 수 있다. 일반적으로 전문분야에서 영어 명사구가 한자어로 번역되는 경우가 많다. 본 논문에서는 개념어에 대하여 단어 단위로 정렬을 하고, 인접한 개념어로 단위를 확장해서 보다 정확한 대역어 쌍을 얻고자 한다.

2절에서는 정렬에 관한 기존의 연구를 간단하게 알아보고, 3절에서 본 논문에서 사용하는 영한 복합어 정렬 방법을 제시한다. 4절에서는 실험 과정 및 결과를 보이고, 결과를 분석한다. 마지막으로 5절에서 본 논문의 결론을 맺는다.

2. 관련 연구

두 언어간 단위가 다른 경우에 간단한 단어 단위의 정렬은 적절하지 못하다. 더군다나 문장의 구조가 다르면 더더욱 정렬하기 힘들다[1].

일반적으로 정렬에 대한 연구는 영어-불어 같이 구조가 유사한 언어 쌍에 대해서 문장 단위[5, 7, 11] 혹은 단어나 구 단위[6, 8, 10, 15]로 활발하게 진행되어 왔다. [15]에서는 영어-불어 병렬 코퍼스로부터 간단한 명사구를 먼저 추출하고 이것들을 대상으로 정렬을 수행했다. 명사구를 추출한 규칙은 정규표현의 형태로 되어 있다. 그는 대역 확률을 Brown의 모델 1[6]을 사용하여 계산했다.

또한 영어-중국어[9, 12, 16], 영어-일본어[13, 14] 정렬에 대한 연구도 있었다. [14]에서는 일본어 코퍼스로부터 명사구를 추출하고 모든 영어 n그램에 대해서 대역 확률을 계산하여 대응되는 영어구를 찾았으며, [13]에서는 코퍼스에서 반복적으로 자주 나타난 일어, 영어 단어열에 대해서 정렬을 했다.

영어-한국어 쌍의 정렬에 대한 연구로는 [1, 2, 4]가 있다. [2]에서는 웹 상에서 번역문서 후보를 추출하여 문장에 길이에 기반하는 문장 단위 정렬을 이용하여 병렬 말뭉치를 구축하였다. [1]에서는 Brown의 모델 2[6]을 기초로 하여 기능어 정보를 사용했다. 이 방법은 모든 구에 대하여 정렬하므로 범용성을 가지지만, 파라메터 수가 많이 때문에 학습을 위해서 많은 양의 코퍼스가 필요하다. [4]에서는 확률적 음운정렬 방법을

사용해서 자동으로 외래어 사전을 추출했다.

3. 영한 복합어 정렬

본 논문에서 제안하는 전체 정렬 과정은 그림 1과 같다. 기본적으로 정렬은 양국어 공기정보를 기초로 하여 이루어지고, 인접한 개념어로 단위를 확장하며, 재추정 방법을 사용하여 정확률을 높인다. 그러면 각 부분에 대해서 자세히 알아보자.

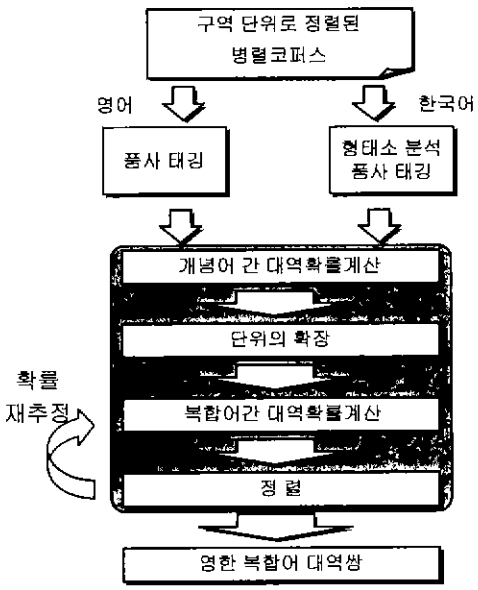

그림 1. 정렬 시스템의 구조

3.1 개념어의 추출

먼저 각 언어에 대해서 개념어만 추출하여 이것들만 정렬에 고려하도록 한다. 영어에 대해서는 명사, 동사, 형용사만 고려했고, 한국어 대해서는 오직 명사만 개념어로 선택하였다. 이는 한국어에서는 동사나 형용사가 전문용어로 사용되는 경우가 거의 없기 때문이다.

3.2 단어 대역 확률 계산

단어 대역 확률을 구하는데 있어서 기본 가정은 대응되는 양쪽 언어 문장에 동시에 같이 자주 나타나면 그 대역 확률이 높다는 것이다. 이 논문에서는 Dice 계수를 사용하여 단어간 대역 확률을 구한다. 영어 단어 E_i와 한국어 단어 K_j의 대역 확률 $C_p(E_i, K_j)$는 식(1)과 같이 구한다.

$$C_p(E_i, K_j) = \frac{2C(E_i, K_j)}{C(E_i) + C(K_j)} \quad (1)$$

식(1)에서 $C(E_i)$와 $C(K_j)$는 각각 E_i와 K_j가 코퍼스 내에서 나타난 빈도를 뜻하고, $C(E_i, K_j)$는 E_i와 K_j가 대응되는 문장에 동시에 나타난 빈도를 뜻한다. 위의 방법을 앞서 추출한 개념어에 대해서 적용하여 병렬 코퍼스의 모든 개념어 쌍에 대하여 대역 확률을 구할 수 있다.

3.3 단위의 확장

앞서 구한 대역 확률의 단위를 단어에서 인접한 단어를 포함시켜서 확장할 수 있다. 여기에 사용되는 기본 가정은 복합어를 구성하는 개념어가 원어에서나 역어에서 모두 서로 인접해서 나타난다는 것이다. 이 논문에서는 일단 1:1, 1:2, 2:1, 2:2 대응만 고려하도록 한다. 표1에 각 대응에 대한 예를 보인다.[1)]

표 1. 여러 가지 정렬의 예

대응 관계	영어	한국어
1:1	committee	위원회
1:2	counterfeit	위조 상품
2:1	(the) country (of) importation	수입국
2:2	(the) additional duty	추가 관세

일단은 단어간 대역 확률을 구했던 것과 같은 방법으로 식(2)~(4)처럼 복합어간 대역 확률을 구할 수 있다.

$$C_p(E_i, K_j K_{j+1}) = \frac{2C(E_i, K_j K_{j+1})}{C(E_i) + C(K_j K_{j+1})} \quad (2)$$

$$C_p(E_i E_{i+1}, K_j) = \frac{2C(E_i E_{i+1}, K_j)}{C(E_i E_{i+1}) + C(K_j)} \quad (3)$$

$$C_p(E_i E_{i+1}, K_j K_{j+1}) = \frac{2C(E_i E_{i+1}, K_j K_{j+1})}{C(E_i E_{i+1}) + C(K_j K_{j+1})} \quad (4)$$

식(2)~(4)에서 $E_i E_{i+1}$와 $K_j K_{j+1}$은 각각 인접한 영어, 한국어 개념어쌍을 나타낸다. 하지만 이런 식으로 모든 인접한 개념어쌍에 대해서 확장을 하면 의미없는 개념어쌍이 나올 수 있으므로 확장에 제약이 필요하다. 즉, 확장했을 때 새로 계산한 대역 확률이 확장하기 전 대역 확률들의 산술평균 혹은 최고값보다 같거나 큰 경우에만 확장을 한다. 아래에 이런 역할을 하는 제약식들을 나타낸다.

먼저 식(5)~(8)은 1:2 대응이나 2:1 대응의 경우에 사용되는 제약식을 나타낸다.

$$C_p(E_i, K_j K_{j+1}) \geq \frac{C_p(E_i, K_j) + C_p(E_i, K_{j+1})}{2} \quad (5)$$

$$C_p(E_i E_{i+1}, K_j) \geq \frac{C_p(E_i, K_j) + C_p(E_{i+1}, K_j)}{2} \quad (6)$$

$$C_p(E_i, K_j K_{j+1}) \geq \max\{C_p(E_i, K_j), C_p(E_i, K_{j+1})\} \quad (7)$$

$$C_p(E_i E_{i+1}, K_j) \geq \max\{C_p(E_i, K_j), C_p(E_{i+1}, K_j)\} \quad (8)$$

2:2 대응의 경우에 사용되는 제약식은 조금 더 복잡한데 아래의 식(9), (10)의 경우이다.

1) 개념어가 아닌 것은 괄호 안에 표기하도록 한다.

$$C_p(E_i,E_{i+1},K_j,K_{j+1}) \geq \frac{C_p(E_i,K_j)+C_p(E_{i+1},K_{j+1})}{2} \text{ and}$$
$$\geq \frac{C_p(E_i,K_{j+1})+C_p(E_{i+1},K_j)}{2} \text{ and}$$
$$\geq \frac{C_p(E_i,E_{i+1},K_j)+C_p(E_i,E_{i+1},K_{j+1})}{2} \text{ and} \quad (9)$$
$$\geq \frac{C_p(E_i,K_j,K_{j+1})+C_p(E_{i+1},K_j,K_{j+1})}{2}$$

$$C_p(E_i,E_{i+1},K_j,K_{j+1}) \geq \max\{C_p(E_i,K_j),C_p(E_i,K_{j+1}),$$
$$C_p(E_{i+1},K_j),C_p(E_{i+1},K_{j+1}),$$
$$C_p(E_i,E_{i+1}),K_j),C_p(E_i,E_{i+1},K_{j+1}),$$
$$C_p(E_i,K_j,K_{j+1}),C_p(E_{i+1},K_j,K_{j+1})\} \quad (10)$$

위에서 제시한 제약식들을 만족시키는 경우에만 정렬 단위를 확장하도록 한다. 각 대응에 대해서 좀 더 자세히 설명하겠다.

1: 2 대응의 경우

표 1의 'counterfeit'와 '위조 상품'의 경우를 생각해 보자. 여기에서 생각할 수 있는 모든 정렬의 경우는 아래와 같이 3가지이다.

counterfeit ↔ 위조 상품
counterfeit ↔ 위조
counterfeit ↔ 상품

이 경우 'counterfeit'와 '위조', 'counterfeit'와 '상품'의 대역 확률을 각각 계산한 값보다 인접한 개념어로 확장하여 'counterfeit'와 '위조 상품'의 대역 확률을 계산한 값이 더 크다. 이를 반영하는 제약식이 식(5), (7)이다.

2: 1 대응의 경우

표 1의 '(the) conuntry (of) importation'과 '수입국'의 경우를 보자. 여기에서 생각할 수 있는 모든 정렬의 경우는 아래의 3가지이다.

(the) country (of) importation ↔ 수입국
country ↔ 수입국
importation ↔ 수입국

역시 이 경우에도 각각을 나누어서 대역 확률을 계산하는 것보다 확장하여 합쳐서 (the) country (of) importation'과 '수입국'의 대역 확률을 계산하는 경우에 값이 더 크다. 이를 반영하는 제약식이 식(6), (8)이다.

2: 2 대응의 경우

표 1의 '(the) additional duty'와 '추가 관세'의 경우를 보자. 이 경우에는 앞선 두 가지 경우보다 조금 더 복잡하다. 여기에서 생각할 수 있는 모든 정렬의 경우는 아래와 같이 모두 9가지이다.

additional ↔ 추가
duty ↔ 관세
additional ↔ 관세
duty ↔ 추가
additional ↔ 추가 관세
duty ↔ 추가 관세
(the) additional duty ↔ 추가
(the) additional duty ↔ 관세
(the) additional duty ↔ 추가 관세

이 경우에도 각각을 나누어서 대역 확률을 계산하는 경우보다 '(the) additional duty'와 '추가 관세'로 확장하여 계산하는 경우 확률값이 더 크다. 이를 반영하는 제약식이 식(9), (10)이다.

3.4 문장 내에서 정렬 및 확률 재추정

앞서 구한 대역 확률을 가지고 대응되는 문장 내에서 실제로 정렬을 수행한다. 이때 대역 확률이 높은 것, 대응쌍의 단위가 큰 것이 그 우선 순위를 가진다. 이렇게 일단 정렬한 후에 정렬된 결과를 가지고 대역확률을 아래의 식(11)을 사용하여 재추정한다.

$$C'_p(E_i,K_j)=\frac{2C_a(E_i,K_j)}{C_a(E_i)+C_a(K_j)} \quad (11)$$

식(11)에서 $C_a(E_i)$와 $C_a(K_j)$는 각각 E_i와 K_j가 각각 실제 정렬에 참여한 횟수를 뜻하고, $C_a(E_i,K_j)$는 E_i와 K_j가 실제로 정렬된 횟수를 뜻한다. 앞서서 대역 확률을 구했던 식(1)에서는 대응되는 문장에 같이 나타난 경우를 모두 고려하지만 식(11)에서는 실제로 정렬된 경우만 고려하기 때문에 보다 정확한 대역 확률을 추정할 수 있다.

이렇게 수정된 대역 확률을 가지고 다시 대응되는 문장에 대해서 정렬을 수행하고, 다시 그 결과를 반영하여 새로운 대역 확률을 구해서 정렬하는 과정을 정렬의 결과가 수렴할 때까지 반복한다.

4. 실험 및 평가

4.1 실험 환경

실험에 사용된 병렬 코퍼스는 우루과이 라운드 협정문이다. 병렬 코퍼스는 전처리 단계를 거쳐서 구역 단위로 정렬된 상태이다. 한 구역은 대체로 한 문장으로 이루어져 있지만 경우에 따라서 여러 개의 문장으로 이루어지거나 문장의 일부인 경우도 있다. 사용한 병렬 코퍼스에 대한 여러 가지 통계값은 표 2와 같다.

표 2. 사용한 병렬 코퍼스에 대한 여러 가지 통계값

항목	영어	한국어
구역	4,968	4,968
어절	139,265	79,290
한 구역당 평균 어절	28.03	15.96
개념어	65,844	65,653
유일한 개념어	2,681	3,847

4.2 실험 결과 및 평가

개념어만 추출하기 위해서 먼저 각 언어에 대하여

품사 태거를 이용했다.

처음에는 양국어 공기정보를 기초로 하여 개념어 쌍에 대하여 확률을 구하고, 이를 인접한 개념어로 확장시켜 복합어 단위로 대역어 확률을 구했다. 이렇게 얻은 대역 확률을 가지고 초기 정렬을 수행하고, 정렬 결과로 다시 대역 확률을 구해서 새로 추정한 대역 확률을 사용하여 다시 정렬하는 과정을 7번 반복했다. 재추정 횟수는 실험에 의해서 결정했다. 또한 복합어로 단위를 확장할 때 제약식으로 산술평균을 이용한것과 최고값을 이용한 것을 각각 따로 실험하였다. 아래 그림은 각 경우에 대하여 추출한 유일한 대역쌍 수의 변화와 마지막에 각 대응의 분포를 나타낸 그래프이다.

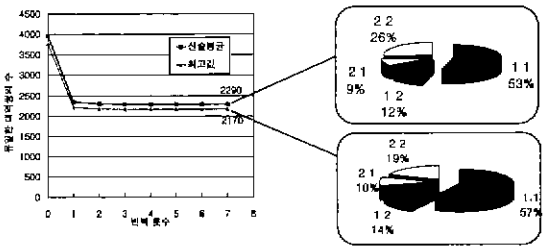

그림 2. 추출한 유일한 대역쌍 수의 변화와 각 대응의 비율

결과를 보면 산술평균을 사용했을 때나 최고값을 사용했을 때 비슷한 결과를 보이는데 첫번째 재추정할 때 유일한 대역쌍의 수가 크게 감소하고, 그 뒤로는 감소량이 줄어서 5번째 재추정 이후부터는 거의 변하지 않았다.

7번째 재추정 후 각 대응 유형의 비율도 두 경우가 비슷한 결과를 보이는데 1:1 대응이 가장 많았고, 그 뒤로 2:2 대응, 1:2 대응, 2:1 대응 순서였다. 제약식으로 산술평균을 사용한 경우보다 최고값을 사용한 경우에 추출한 유일한 대역쌍의 수가 약간 적게 나타났다.

매번 재추정할 때 정렬의 정확률은 다음의 그림 3과 같다. 역시 두 경우 모두 초기 재추정에서는 정확률이 높아지다가 그 뒤로는 크게 변하지 않았다. 정확률은 각 경우에 대하여 임의로 100개의 대역쌍을 추출하여 계산했다.

추출한 유일한 대역쌍의 개수와 정확률의 변화로 볼 때 대략적으로 5번째 재추정후 결과가 수렴한다고 볼 수 있다. 따라서 5번째부터 7번째까지만 고려해 보면 보면 전부 2,290개의 유일한 대역어쌍을 얻었고, 평균 정확률은 74%였다.

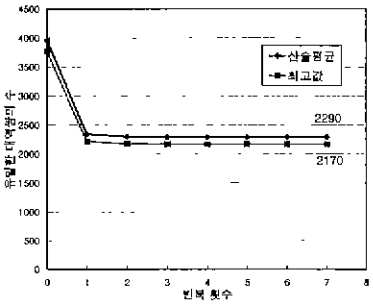

그림 3. 정렬의 정확률

4.3 결과의 예

추출한 대역쌍 중에서 몇가지 경우를 표 3에 보였다.

표 3. 추출한 대역쌍의 예

	영어	한국어
(12)	convention	협약
(13)	countermeasure	대응 조치
(14)	working party	작업반
(15)	result (of) negotiation(s)	협상 결과
(16)	result (of) negotiation(s)	협상(의) 결과
(17)	Consultation (and) Dispute	협의 (및) 분쟁
(18)	(the) custom(s) administration	세관 당국
(19)	(the) Member concern(ed)	관련 회원국
(20)	(the) period (of) (the) application	적용 기간
(21)	balance (of) payment(s)	국제 수지
(22)	(the) Director-General	사무 총장
(23)	committee	위원 회의
(24)	negative	최종 결정
(25)	agreement (and) (of) (the) multilateral	협정 (및) 다자

본 논문의 실험은 우루과이 라운드 협정문을 사용했기 때문에 그 도메인이 경제, 외교 분야로 한정되어 있다. 이런 제한된 특별한 도메인에 대해서는 일반 대역사전을 이용한 번역은 자연스럽지 않은 경우가 있다. (12)에 제시된 'convention'과 '협약'이 그 좋은 예이다. 'convention'은 일반적으로 '집회', '사회적 관습', '관례'의 뜻으로 자주 사용되지만 실험 결과에서는 외교 용어인 '협약'으로 정렬이 되었다. (13), (14), (15)는 각각 1:2, 2:1, 2:2로 정렬된 예이다. (15)와 (16)을 보면 'result (of) negotiation(s)'가 각각 '협상 결과', '협상(의) 결과'로 정렬되었다. 이는 정렬할 때 개념어만 고려했기 때문에 두 가지 경우를 같은 형태로 보고 정렬한 결과이다.

특히 2:2 대응의 결과는 다른 대응보다 여러 가지 다양한 현상을 보이는데 아래와 같이 세 가지 경우로 나누어 볼 수 있다.

순서대로 번역된 경우

복합어를 구성하는 각 단어가 순서대로 번역되어 새로운 복합어로 만들어진 경우이다. (17)을 보면 'Consultation (and) Dispute'에서 'Consultation'이 '협의'로, 'Dispute'가 '분쟁'으로 각각 번역되어 '협의 (및) 분쟁'으로 번역된 셈이다. (18)의 '(the) custom(s) administration'과 '세관 당국'도 비슷한 예이다.

이런 경우는 복합어의 구조가 병렬 관계이거나 영어쪽 구성이 '형용사+명사', '명사+명사'인 경우가 많았다.

순서가 바뀌어 번역된 경우

복합어를 구성하는 각 단어의 순서가 바뀌어 번역되어 복합어로 만들어진 경우이다. (19)를 보면 'Member'가 '회원국'으로 'concern(ed)'가 '관련'으로 번역되어 전체적으로 '(the) Member concern(ed)'가 '관련 회원국'으로 번역된 경우이다. 같은 방법으로 (20)에서 '(the) period (of) (the) application'이 '적용 기간'으로 번역되었다.

이런 경우는 영어쪽이 전치사 구로 수식되는 명사구인 경우에 자주 보였다.

새로운 단어로 번역된 경우

(21)의 'balance (of) payment(s)'가 '국제 수지'로 번역된 경우나 (22)의 '(the) Director (-) General'이 '사무 총장'으로 번역된 경우가 바로 이런 경우이다.

이런 경우는 일반 사전을 사용하여 각각을 번역하여 조합하면 어색한 번역 결과를 보이게 된다.

4.4 오류 분석

우선 전처리 단계인 형태소 분석이 잘못 된 경우 오류가 생길 수 있다. (24)의 경우를 보면 '위원회 + 의'가 '의원 + 회의'로 조사가 잘못 분석되어 결과가 잘못되었음을 알 수 있다.

또한 번역의 일관성이 결여된 경우에 결과가 잘 못 나올 수 있다. 특히 영어 약어 표기의 번역에서 일관성이 없는 경우가 많았는데 'GATT(General Agreement of Tariffs and Trade)'의 경우 '가트', 'GATT', '관세 및 무역에 관한 일반 협정' 등으로 일관성 없이 번역되어서 제대로 정렬이 되지 않았다.

통계적인 정보만 사용했으므로 아무 상관없는 두 단어가 단지 자주 같이 나타났기 때문에 정렬이 된 경우가 있다. (24)의 'negative'와 '최종 결정'을 보면 그 자체는 아무런 연관관계가 없지만 실험에서 사용된 코퍼스에서 단순히 자주 같이 사용되었기에 정렬이 되었다. 이는 단순히 양국어 공기정보만을 사용하여 정렬을 했기 때문에 생긴 오류이다.

인접한 단어로 잘못 확장해서 오류가 생긴 경우도 있었는데 (25)가 바로 그런 경우의 예이다. 이는 'agreement'와 '협정', 'multilateral'과 '다자'만 보면 각각 올바른 정렬이라고 할 수 있지만 실제 문장에서 보면 그림 4와 같이 정렬되어야 올바른 정렬이라고 할 수 있다. 이런 경우는 무조건 입접한 개념어로 확장했기 때문에 생긴 오류이다.

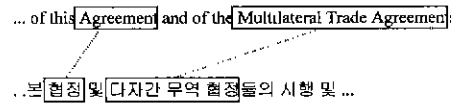

그림 4. 올바르게 확장하여 정렬한 경우

5. 결론 및 향후과제

본 논문에서는 통계 정보를 기초로 한 영한 복합어 정렬 방법을 제안했다. 도메인을 제한하여 전문분야에 대해서 정렬을 수행했으며 그 결과로 전문용어에 대한 대역어를 자동으로 구축했다. 정렬의 단위를 단어에서 복합어로 확장함으로써 복합어에 대해서 단일 단어에 대한 번역어를 조합한 것보다 더 좋은 결과를 얻을 수 있었다.

실제로 영어 139,265 어절, 한국어 79,290 어절로 구성된 우루과이 라운드 협상문에 대해서 이 방법을 적용한 결과 전부 2,290 개의 유일한 복합어 대역쌍을 얻었고, 정확률은 74%였다.

본 논문에서는 복합어의 크기를 2개의 개념어로만 제한했는데 향후 과제로서 더 큰 단위의 복합어에 대해서도 같은 방법이 그대로 사용될 수 있는지에 대한 연구가 필요하다. 또한 통계적 방법을 보완하여 언어학적 지식을 함께 사용하는 방법과 주변 문맥을 고려한 정렬도 생각해 보아야 할 것이다.

감사의 글

본 연구는 전문용어언어공학연구센터에서 수행한 과학기술부와 KISTEP의 핵심소프트웨어사업 중 "대용량 국어정보 심층처리 및 품질관리 기술개발" 과제의 일환으로 수행되었으며, 첨단정보기술연구센터를 통하여 과학재단의 지원도 받았습니다.

참고문헌

[1] 신중호. 1996. 한국어/영어 병렬 코퍼스에 대한 단어단위 및 구단위 정렬 모델. Master's thesis, 한국과학기술원.

[2] 양주일, 김선호, 송만석. 1999. 웹 문서로부터 한-영 병렬 말뭉치 자동 구축과 문장 단위 정렬. 제11회 한글 및 한국어 정보처리 학술발표 논문집, pp. 150-155.

[3] 이주호. 1999. "자동 정렬을 통한 영한 복합어의 역어 추출. Master's thesis, 한국과학기술원.

[4] 이재성. 1999. 이중언어 코퍼스로부터 외래어 표기 사전의 자동구축. 제11회 한글 및 한국어 정보처리 학술발표 논문집, pp.142-149.

[5] Brown, Peter F., Jenniger C. Lai and Robert L.

Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169-176.

[6] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. The mathematics of statistical machine translations: parameter estimation. *Computational Linguistics*, 2(19) : 263-311.

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

[8] Dagan, Ido, Kenneth W. Church and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the 29th the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1-8.

[9] Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 236-243.

[10] Gale, William A. and Kenneth W. Church. 1991a. Identifying word correspondences in parallel texts. In *Proceedings of Speech & Natural Language Workshop*, pages 177-184.

[11] Gale, William A. and Kenneth W. Church. 1991b. A program for aligning sentences in bilingual corpora, In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177-184.

[12] Ker, Sue J. and Jason S. Chang. 1997. A class-based approach to word alignment. *Computational Linguistics*, 2(23):313-343.

[13] Kitamura, Mihoko and Yuji Matsumoto. 1996. Automatic extractions of word sequence correspondences in parallel corpora. In *Proceedings of the 4th Annual Workshop on Very Large Corpora*, pages 79-87.

[14] Kumano, Akira and Hideki Hirakawa. 1994. Building an mt dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 76-81.

[15] Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17-22.

[16] Wu, Dekai and Xuanyin Xia. 1994. Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of Association for Machine Translation in the Americas*, pages 206-213.