제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

벡터를 사용한 2 단계 영한 대역어 선택

이기영, 박상규
한국전자통신연구원 언어처리연구팀
e-mail : leeky@etri.re.kr

2-Level English-Korean Target Word Selection Using Vectors

Ki-Young Lee, Sang-Kyu Park
Natural Language Processing Team, ETRI

요 약

영한 자동번역 시스템에서 대역어 선택 모듈은 어휘 변환을 수행한다. 일반적으로 영어 단어는 다양한 한국어 단어로 번역될 수 있는 의미적 모호성을 지니고 있으며, 고품질의 영한 자동번역 결과를 제공하기 위해서는, 해당 문맥에 가장 적합한 한국어 단어가 선택되어야 한다. 본 논문에서는 영어의 명사 어휘에 대하여, 벡터를 사용하는 2 단계 영한 대역어 선택 기법을 제안한다. 벡터를 사용하는 2 단계 대역어 선택 방식은 첫번째 단계에서, 원문에서 사용된 영어 명사의 의미를 결정하고, 두 번째 단계에서, 해당 의미를 지니는 유사 한국어 대역어 가운데, 생성될 한국어 문맥에 맞는 적합한 한국어 대역어를 선택한다. 또한 제안하는 방법의 타당성을 검증하기 위해 현재 우리가 개발중인 Tellus-EK 영한 자동번역 시스템에 적용한 결과를 논한다.

1. 서 론

규칙 기반 방식으로 시작한 자동번역 시스템은 현재 통계 기반, 패턴 기반, 예제 기반 등의 다양한 방식으로 개발되고 있다. 이러한 다양한 접근 방식에 따라 시스템 구조는 어느 정도의 차이는 있지만, 일반적으로 자동번역 시스템은 분석 단계, 변환 단계, 생성 단계를 포함한다.

원시 언어에 대한 형태소 분석, 구조 분석 또는 문틀을 사용한 패턴 매칭 등의 다양한 접근 방식의 분석 단계가 성공적으로 수행되었다고 하더라도 변환 단계에서 수행되는 구조 변환 과정과 어휘 변환 과정에서 원시 언어가 지니는 의미의 손실이 발생한다면 최종 번역 결과의 품질은 저하될 수 밖에 없다.

본 논문에서는 영한 자동번역에서 영어 명사 어휘를 대상으로 하는 2 단계 대역어 선택 방식을 제안한다. 제안하는 2 단계 대역어 선택 방식은 1 단계에서는, 명사가 어휘 정렬된 병렬 코퍼스로부터 추출한 공기어휘(co-occurring word)의 의미에 대한 조건부 확률값을 벡터의 요소값으로 표현한 벡터를 사용하여 영어 명사의 의미를 결정한 후, 2 단계에서는, 병렬 코퍼스의 대역부 정보를 사용하여 한국어 문맥 정보를 역시 벡터로 표현하고, 생성될 한국어 문맥에 적합한 대역어를 선택한다.

1. 1 대역어 선택의 중요성

영어 어휘는 두개 이상의 한국어로 번역될 수 있는 다의어(polysemous word)가 많다. 대표적인 영어 명사의 예는 다음과 같다.

(예 1) race/NOUN = {경주, 경쟁, 선거, 선거전, 인종}
(예 2) measure/NOUN = {조치, 측정, 법안, 법령, 기준}
(예 3) bank/NOUN = {은행, 제방}

위의 영어 어휘 이외에도 많은 다의어가 있으며, 우리가 개발중인 Tellus-EK 영한 자동번역시스템에서 사용되는 어휘 사전에서 추출한 통계 정보는 다음과 같다.

전체 명사 엔트리 (고정표현 포함)	59,448 개
2 개 이상의 대역어를 포함하는 명사 엔트리	6,624 개

표 1. Tellus-EK 사전 다의어 통계 정보

위의 표 1 에서 실제로 2 개 이상의 영어 어휘로 이루어진 고정표현을 제외하고, 현재 Tellus-EK 사전의 대역어 수가 매우 부족하다는 점을 고려하면, 실제 다의어 비율은 표 1 에서 제시한 통계보다 훨씬 크다고 할 수 있다.
또한 번역 관점에서 여러 가지 대역어를 가지는 어휘를 잘못 번역할 경우, 그 의미가 매우 우스꽝스럽거나 매우 자연스럽지 못하며, 원문의 의미가 일부 소실되거나 완전히 소실될 경우도 있다. 다음은 그러한 예를 나타낸다.

[원 문] The mainland media has ignored the mayoral **races**
[번역문] 본토 매체가 시장 **경주**를 무시하였다.

위의 문장은 실제로 자동번역 시스템에 의해서 나온 결과로서 race 의 대역어 선택에 실패한 경우에 해당하며, 단 하나의 어휘에 대한 대역어 선택이 잘못 되었다고 하더라도, 번역문 만을 본다면, 그 의미를 파악하기가 어렵다.

본 논문에서는 벡터를 이용하여 2 단계 과정을 거치는 영한 명사 대역어 선택 기법에 대하여 논하며, 본 논문의 구성은 다음과 같다. 2 장에서는 의미 모호성 해소 및 대역어 선택과 관련된 기존 연구들에 관하여 언급하고, 3 장에서는 본 논문에 제안하는 벡터를 사용하는 2 단계 영한 명사 대역어 선택 기법을 자세히 논하고, 4 장에서는 제안하는 방식을 현재 개발중인 Tellus-EK 영한 자동번역 시스템에 적용한 결과에 대하여 논하고, 마지막으로 5 장에서 결론을 맺는다.

2. 관련 연구

대역어 선택은 일차적으로 원문의 어휘가 지니는 의미적 모호성을 해소하는 과정을 포함한다. 의미 모호성 해소에 대한 많은 연구가 자연어처리의 다양한 분야, 특히, 정보검색 및 기계번역 분야에서 이루어지고 있으며, 의미 모호성 해소를 위해 사용되는 단서(clue)도 다양하다. [1],[2]는 단어가 지니는 의미적 모호성을 해소하기 위해 사용되는 다양한 단서들과 이러한 단서들을 사용하여 구현한 시스템의 성능을 비교 평가하였다. [3]은 병렬 코퍼스를 이용하여 어휘의 의미를 해당 어휘의 대역어의 집합으로 표현하고, 코퍼스 통계 정보 등을 사용하여 의미적 모호성을 해소는 기법을 제안하였다. [4]는 번역 패턴을 사용하는 한영 대역어 선택 기법을 제안하였다. [5]는 의미 태깅된 데이터로부터 추출한 공기 어휘 정보를 벡터로 구축하여, 문헌 정보 검색에서 사용하는 유사도 계산을 써서 의미 모호성을 해소하는 기법을 제안하였다. [6]은 품사, 형태소 분석 정보, 공기 정보, 동사의 하위 범주 정보 등 다양한 자질(feature)을 이용하여 의미 모호성을 해소하는 기법을 제안하였다.

단어 의미 모호성 해소에 가장 좋은 성능을 보이는 단서로는 의미 태깅된 코퍼스로부터 얻어지는 통계 정보라고 할 수 있지만, 의미 태깅이란 작업에 드는 비용 문제와 일관성이 없을 수도 있다는 문제를 생각하면 최상의 선택이라고는 할 수 없다. 이러한 이유로 인해, 일반적으로 공기 정보, 동사 하위범주 정보, 주제(topic) 정보, 도메인(domain) 정보, 의미 분류 정보 등을 주된 단서로 사용하여 의미 모호성 해소를 위한 연구가 진행되고 있다.

3. 영한 대역어 선택을 위한 2 단계 접근 방안

Tellus-EK 에서 사용했던 기존의 대역어 선택 방식은 사전에 각 영어 어휘의 대역어 뿐만 아니라 대응되는 공기 어휘들을 'COL'이라는 자질의 자질값으로 저장해두고, 입력 문장의 처리 대상 어휘 주위에 오는 어휘가 'COL' 자질의 자질값과 매칭되는지 여부에 따라 대역어를 선택하였다. 이러한 대역어 선택 방법은 커버리지 문제로 인해 좋은 성능을 낼 수 없었으며, 온-오프(on-off)식으로 대역어 선택이 결정되었기 때문에 많은 문제점을 가지고 있었다.

본 논문에서는 첫번째로, WordNet1.71 의 의미체계를 따르는 420 개의 의미코드를 설정하여 Tellus-EK 사전의 명사 어휘들에 대해서 대역어에 따른 의미코드를 부여한 후에, 각각의 명사 엔트리에 대해서 동일한 의미코드를 가지는 대역어들을 하나의 집합으로 분류하였다. 표 2 는 'race/NOUN'에 대해서 이러한 프로세스를 적용한 예를 나타낸다.

어휘	의미코드	대역어
race/NOUN	social event #1	경선, 선거전, 경주, 경쟁
	people #1	인종

표 2. 의미코드에 따른 대역어 분류 예

두 번째로, CNN 뉴스 스크립트와 국내 영어 교과서 문장으로 이루어진 약 14 만 문장의 병렬 코퍼스의 각 원문/대역문에 대해서 Tellus-EK 사전을 이용하여 명사 어휘에 대한 정렬(alignment) 작업을 수행하였다. 이렇게 구축된 정보를 주된 지식으로 사용하여 대역어 선택을 수행하였으며 자세한 내용은 다음절에서 설명한다.

3.1 대역어 선택 모듈 구성

본 논문에서 제안하는 대역어 선택 기법은 2 단계로 구성된다. 1 단계에서는 영어 입력 문장에 대해서 공기 정보를 사용하여 해당 어휘의 의미를 결정한다. 이미 기술한 바와 같이 Tellus-EK 사전의 각 영어 명사 어휘의 사전 정보는 의미코드에 따라 대역어가 분류되어 있다. 따라서 1 단계에서 해당 영어 어휘의 의미코드가 결정되면 2 단계에서 해당 의미코드를 갖는 대역어 집합 중 한국어 문맥에 가장 적합한 대역어를 선택한다. 만약 1 단계에서 결정된 의미코드에 해당하는 대역어가 하나만 존재한다면 2 단계는 거치지 않게 된다. 또한 특정 어휘의 경우, 의미코드가 하나밖에 없더라도 해당 의미코드에 속하는 대역어가 두개 이상일 경우에는 1 단계를 수행하지 않고, 2 단계만을 수행함으로써 대역어를 결정한다. 즉, 1 단계에서는 의미코드가 결정되고 2 단계에서는 해당 의미코드를 가지는 대역어들 중 가장 한국에 문맥에 적합한 대역어를 선택한다.

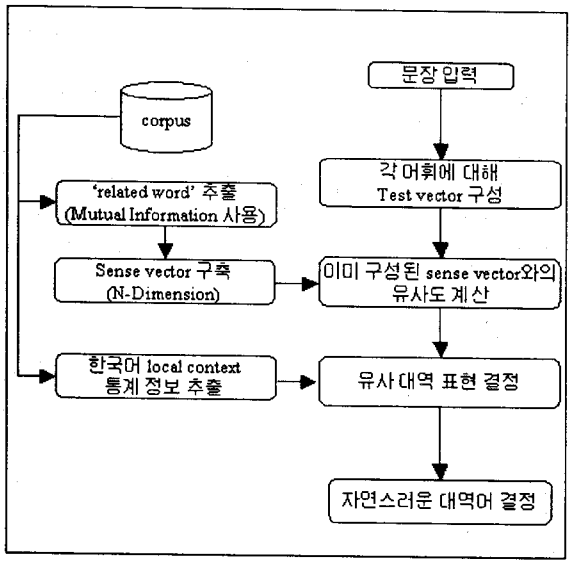

그림 1. 대역어 선택 모듈 구성도

3.2 대역어 선택 1 단계: 의미 모호성 해소

본 논문에서 제안하는 대역어 선택 기법의 1 단계에서는 의미 모호성을 해소한다. 즉, Tellus-EK 사전의 각 명사 어휘들은 대역어와 함께 WordNet1.71 의 체계에 따르는 의미코드를 가지고 있다. 이럴 경우, 각 대역어가 서로 다른 의미코드를 가질 수 있으며, 1 단계에서는 이들 가운데 하나의 의미코드를 결정한다.

1 단계에서 사용하는 의미벡터는 N 차원의 벡터로 구성되며, 그 구성요소(element)는 각 영어 명사와 공기하는 어휘들의 가중치값을 나타내며 이러한 정보는 14 만 병렬 코퍼스로부터 추출된다. 의미 벡터의 차원(dimension)은 MI(Mutual Information)에 근거하여 영어 명사 어휘와 높은 관계를 가지는 어휘들의 개수로서 정의된다. 아래의 수식은 각각 MI 와 의미벡터 SV 및 그 요소들을 나타낸다.

(수식 1) $$MI(x,y)=\frac{\Pr(x,y)}{\Pr(x)\cdot\Pr(y)}$$

(수식 2) $$SV=(w(c_1),w(c_2),w(c_3),...,w(c_n))$$

위의 의미벡터 SV 에서 C_i 는 영어 명사 어휘와 공기하는 어휘를 나타내며, 함수 $w(C_i)$는 가중치 함수(weighting function)를 나타내며, 다음과 같은 조건부 확률에 따른다.

(수식 3) $$w(c_k)=\Pr(s=s_i|w=c_k)$$ (s_i 는 임의의 sense)

즉, 의미벡터 SV 의 각 구성요소는 공기어휘에 대한 의미의 조건부 확률값을 나타내며, 그 값은 0 과 1 사이의 값을 가지며, 1 에 가까울수록 해당 공기어휘가 영어 명사 어휘의 특정 의미 결정에 매우 강력한 단서가 된다는 것을 나타낸다. 이러한 의미 벡터는 각 영어 명사 어휘에 대해 서로 다른 의미코드의 개수만큼 만들어진다. 예를 들어 앞의 표 2 의 경우 2 개의 의미 벡터가 만들어진다.

이렇게 의미 벡터가 트레이닝 단계에서 구축이 되면, 실제로 테스트 단계에서는 입력 문장의 형용사 및 명사를 대상으로 하여, 트레이닝 단계에서 구축된 의미벡터와 동일한 dimension 을 갖는 테스트 벡터가 구축된다. 테스트 벡터의 각 구성요소는 0 또는 1 의 값을 가지며, 테스트 문장에 포함된 각각의 영어 명사 어휘에 대한 공기 어휘가 해당 영어 명사 어휘의 의미벡터의 구성 요소일 경우에는 1 을 지니며, 그렇지 않을 경우에는 0 을 지닌다.

예를 들어, 'bank/NOUN'의 의미 벡터가 *(w(rain), w(commercial), w(money))*라고 가정하고, 입력 문장이 "Rain broke the bank"라고 하면 (1, 0, 0)의 테스트 벡터가 생성된다.

마지막으로 입력 문장으로부터 얻어진 테스트 벡터와 기구축된 의미벡터와의 유사도는 cosine measure 를 사용하며, 아래의 공식은 두 벡터 간의 유사도 계산 공식을 나타낸다.

(수식 4) $$sim(v,w)=\frac{\sum_{i=1}^{N}v_i w_i}{\sqrt{\sum_{i=1}^{N}v_i^2\sum_{i=1}^{N}w_i^2}}$$

3.3. 대역어 선택 2 단계: 한국어 대역어 선택

다음의 예는 영어 명사 'change/NOUN'가 한국어로 번역되는 다양한 경우를 나타낸다.

(예 4) interest rate **changes** (이자율 **변동**)
(예 5) **changes** in your cells (세포의 **변화**)
(예 6) requests for service **changes** (서비스 **변경** 요청)

위의 예에서 알 수 있듯이 'change/NOUN'의 대역어 '변동', '변화', '변경'은 단순히 영어 어휘 'change/NOUN'의 의미 결정만으로는 결정될 수 없다. 왜냐하면, 'change/NOUN'의 대역어 '변동', '변화', '변경'은 동일한 의미 코드를 가지며, 이러한 미묘한 한국어 대역어 선택은 한국어 국소 문맥 정보를 고려하여야만 가장 자연스러운 대역어를 선택할 수 있다.

본 논문에서는 이를 위해 14 만 병렬 코퍼스의 한국어 대역부만을 대상으로 하여 한국어 명사에 대한 국소 문맥 정보를 벡터로 구축하였다. 즉, 대상이 되는 한국어 명사에 대해서 공기하는 명사, 형용사에 해당하는 한국어 어휘를 조건부 확률값으로 표현한 요소값을 가지는 대역어 벡터를 이미 설명한 의미 벡터와 동일하게 구축하였다. 그리고 실제 생성 모듈로 넘어오는 대역어 리스트들을 사용하여 테스트 벡터가 만들어지며 1 단계에서와 동일한 유사도 계산 방식을 사용하여 최종적인 대역어가 선택된다.

4. 실 험

본 논문에서는 실험을 위해 CNN 웹사이트와 국내

일간지 영문판 웹사이트로부터 테스트 문장을 추출하였으며, 그 대상이 되는 영어 어휘는 'race/NOUN', 'party/NOUN', 'measure/NOUN', 'bank/NOUN', 'line/NOUN' 등 5 개의 영어 어휘에 대해 각각 40 문장씩 총 200 문장을 그 실험 대상으로 하였다.

조건 / 어휘	W: NO MI:ALL	W:NO MI:50%	W: 5 MI:ALL	W: 5 MI:50%	평균
race	73.3%	86.6%	93.3%	80%	83.3%
party	70%	55%	55%	55%	58.6%
measure	60%	65%	70%	60%	63.6%
bank	90%	85%	85%	85%	86.2%
line	85%	90%	85%	85%	86.2%
평균	75.6%	76.32%	77.65%	73%	75.6%

표 3. 대역어 선택 1 단계 정확률

표 3 에서 W 는 공기 정보를 추출할 때 적용한 윈도우 사이즈를 나타내고, W 의 값이 'NO'라는 것은 전체 문장을 대상으로 했음을 나타낸다. 또한 MI 는 MI 정보를 나타내고, MI 의 값이 'ALL'이라는 것은 추출된 공기 정보를 모두 사용했음을 나타내고, MI 의 값이 '50%'라는 것은 MI 에 의해 상위에 속한 50%만의 공기 정보를 사용했음을 나타낸다.

표 3 에서 알 수 있듯이 단어에 따라 그 정확률의 차이가 명확히 나타난다. 예를 들어, 'bank' 같은 경우는 그와 공기하는 어휘가 의미에 따라 명확히 차이가 난다는 것을 보여주며, 'measure'나 'party' 같은 경우는 그렇지 않다는 것을 보여준다. 또한 무제한 윈도우 사이즈이면서 MI 값을 제한하지 않고 가능한 공기 정보를 모두 사용할 경우에는 특히 나쁜 결과가 나오는 것을 볼 수 있다. 이러한 이유는 모호성 해소에 실질적으로 강한 단서가 되는 공기 정보 뿐만 아니라, 노이즈에 해당하는 공기 정보까지 포함되기 때문이다.

또한 위의 표에는 나타나지 않지만, 동사를 공기 정보에 포함시킬 경우, 정확률이 조금 낮아지는데, 그 이유는 목표 어휘와 정확한 구문 관계를 무시한 동사의 사용은 오히려 노이즈 역할을 하는 것으로 밝혀졌다.

결국, 의미 모호성 해소를 위해서는 윈도우 사이즈나 MI 정보 뿐만 아니라 보다 세밀한 공기 정보 추출 방안이 필수적이라고 할 수 있으며 이를 위해 구조 분석 정보의 사용을 고려하고 있다.

마지막으로 대역어 선택 2 단계에 해당하는 한국어 대역어 선택과 관련하여 실제 Tellus-EK 를 실행하여 얻은 번역문의 예는 다음과 같다.

[원 문] We do not want to join the arms **race**.
[번역문] 우리는 무기 **경쟁**에 참여하기를 원하지 않습니다.
[원 문] All **race** cars were outfitted with the GPS system.
[번역문] 모든 **경주** 자동차는 인공위성자동위치측정 시스템을 장착하였습니다.

위의 번역문에서 알 수 있듯이 'race/NOUN'의 한국어 대역어 '경쟁'과 '경주'는 Tellus-EK 영한자동번역 시스템에서 사용하는 사전에서 동일한 의미코드를 가지지만, 대역어 선택 2 단계에서 보다 자연스러운 대역어로 선택됨을 알 수 있다.

현재 대역어 선택 2 단계는 처리 대상이 되는 영어 어휘를 기준으로 공기 어휘가 추출되지 않고, 한국어 대역어만을 기준으로 하여 공기 어휘가 추출되므로 그 빈도의 편차가 존재하며, 이를 정규화하기 위한 방안이 마련중이다.

5. 결 론

본 논문에서 제안한 방법은 현재는 단순히 공기 어휘의 통계 정보만을 주된 지식으로 사용하였지만, 실험 결과에서 밝혀진 문제점을 보완하기 위해, 현재 구축중인 동사 하위범주 정보와 같은 구문 정보의 도입은 대역어 선택 모듈의 성능을 한단계 더 높여 주는데 중요한 지식이 될 수 있을 것으로 여겨진다. 또한 단순한 윈도우 사이즈가 아니라 구문 정보가 반영된 윈도우 사이즈에 대해서도 보다 나은 성능을 위해서는 고려되어야 할 요소라고 할 수 있다.

또한 제안된 방식은 병렬 코퍼스의 어휘 레벨의 정렬 정보에 크게 의존을 하므로, 보다 정확한 대역어 선택 성능을 보장하기 위해서는 대규모 병렬 코퍼스의 구축과 성능 좋은 자동 어휘 정렬 알고리즘의 개발도 필수적이라 할 수 있다.

마지막으로, 보다 자연스러운 대역어 선택이 이루어지기 위해 선행적으로 수행되어야 할 작업으로는 현재 Tellus-EK 사전의 대역어 파트가 기본적인 대역어들만으로 구성되어 있으므로, 각 사전 엔트리에 대해서 부족한 대역어의 보강 작업이 선행되어야 할 것으로 사료된다.

참고문헌

[1] Susan W. McRoy, "Using Multiple Knowledge Sources for Word Sense Discrimination", Computational Linguistics, 1992.
[2] Eneko Agirre and David Martinez, "Knowledge Sources for Word Sense Disambiguation", TSD, 2001.
[3] Hiroyuki Kaji and Yasutsugu Morimoto, "Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora", COLING, 2002.
[4] 김정재, 박준식, 최기선, "두단계 대역어 선택 방식을 이용한 구단위 패턴기반 한영 기계번역 시스템", 제 11 회 한글 및 한국어 정보처리 학술대회, 1999.
[5] Jong-Hoon Oh and Key-Sun Choi, "Word Sense Disambiguation using Static and Dynamic Sense Vectors", COLING, 2002.
[6] Hwee Tou Ng and Hian Beng Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach", ACL, 1996.