'단어-의미 의미-단어' 관계에 기반한 번역어 선택

이 현 아†

요 약

기계번역에서 올바른 번역 문장을 구성하기 위해서는 원시 문장의 의미를 올바르게 표현하면서 자연스러운 목적 문장을 구성하는 번역어를 선택해야 한다. 본 논문에서는 '단어-의미 의미-단어' 관계, 즉 원시언어의 한 단어는 하나 이상의 의미를 가지고 각 의미는 각기 다른 목적언어 단어로 표현된다는 점에 기반하여, 원시 단어의 의미 분별과 목적 단어 선택을 결합하여 번역어를 선택하는 방식을 제안한다. 기존의 번역 방식은 원시 단어에 대한 목적단어를 직접 선택하는 '단어-단어' 관계에 기반하고 있기 때문에, 원시언어를 목적 언어로 직접 대응시키기 위한 지식을 필요로 하여 지식 획득에 어려움이 있었다. 본 논문의 방식에서는 원시 단어의 의미 분별과 목적 언어의 단어 선택의 결합을 통해 번역어를 선택함으로써, 손쉽게 획득할 수 있는 원시 언어와 목적 언어 각각의 지식원에서 번역어 선택을 위한 지식을 자동으로 추출할 수 있다. 또한 원시 언어의 의미와 목적 언어의 쓰임새를 모두 반영하여 충실도와 이해도를 모두 만족시키는 보다 정확한 번역어를 선택할 수 있다.

키워드 : 기계 번역, 번역어 선택, '단어-의미 의미-단어' 관계, 자동 지식 추출, 이해도와 충실도

Translation Disambiguation Based on 'Word-to-Sense and Sense-to-Word' Relationship

Hyun Ah Lee†

ABSTRACT

To obtain a correctly translated sentence in a machine translation system, we must select target words that not only reflect an appropriate meaning in a source sentence but also make a fluent sentence in a target language. This paper points out that a source language word has various senses and each sense can be mapped into multiple target words, and proposes a new translation disambiguation method based on this 'word-to-sense and sense-to-word' relationship. In my method, target words are chosen through disambiguation of a source word sense and selection of a target word. Most of translation disambiguation methods are based on a 'word-to-word' relationship that means they translate a source word directly into a target word, so they require complicate knowledge sources that directly link a source words to target words, which are hard to obtain like bilingual aligned corpora. By combining two sub-problems for each language, knowledge for translation disambiguation can be automatically extracted from knowledge sources for each language that are easy to obtain. In addition, disambiguation results satisfy both fidelity and intelligibility because selected target words have correct meaning and generate naturally composed target sentences.

Key Words : Machine Translation, Translation Disambiguation, 'Word-To-Sense and Sense-To-Word' Relationship, Knowledge Extraction, Fidelity and Intelligibility

1. 서 론

기계번역(machine translation, MT)은 원시언어(source language)의 문서를 번역하여 목적언어(target language)의 문서를 얻는 것을 목표로 한다. 기계번역의 입력으로 들어오는 원시언어 문장의 각 단어에 대응되는 여러 목적단어 중에서 올바른 번역문장을 구성하는 단어를 선택하는 문제를 번역 애매성 해소(translation disambiguation), 또는 번역어 선택(translation selection)이라 한다.

번역어 선택은 기계번역의 성능을 좌우하는 가장 중요한 요소 중 하나이다. 예를 들어 영어 단어 'break'는 '부수다', '깨뜨리다', '(규칙을) 어기다' 등의 한국어로 번역될 수 있고, 'regulation'은 '규칙', '단속', '조절' 등으로 번역될 수 있다. 만일 원시문장 "The boy broke the regulation."에 대해서 원시언어의 어휘와 구조를 정확하게 분석하고 올바른 형태로 목적언어 문장을 생성하더라도, 번역어를 잘못 선택하여 '어기다'로 번역해야 할 'break'를 '깨뜨리다'로 번역하고 '규칙'으로 번역해야 할 'regulation'을 '단속'으로 번역하면, "그 소년은 단속을 깨뜨렸다."라는 원문의 의미에 맞지 않은 번역문을 얻게 된다.

기계번역과 번역어 선택의 성능은 충실도(fidelity)와 이해도(intelligibility)의 두 평가 기준으로 측정된다[1]. 충실도와 이해도는 각각 원시언어와 목적언어의 측면에서 번역 결과를 평가하는 척도이다. 충실도는 원시언어 문장의 정보가 번역된 문장으로 정확하고 완전하게 전달되는지를 평가한다. 이해도는 목적언어 사용자가 보았을 때 번역된 문장이 자연스럽고 유창(fluent)하게 구성되어 쉽게 이해할 수 있는지를 평가한다. 이는 기계번

† 종신회원 : 금오공과대학교 컴퓨터공학부 전임강사
논문접수 : 2005년 8월 12일, 심사완료 : 2006년 2월 6일

역이나 번역어 선택의 결과가 원시언어 문장이 가지는 의미를 제대로 나타내면서 자연스러운 목적언어 문장을 만드는 두 조건을 모두 만족시킬 때만 올바른 번역으로 평가됨을 나타낸다.

본 논문에서는 지식 획득이 쉬우면서 정확한 번역어를 선택하기 위한 새로운 방법을 제안한다. 원시언어의 한 단어는 하나 이상의 의미를 가지고 각 의미는 각기 다른 목적언어 단어로 표현되는 '단어-의미 의미-단어'관계에 기반하여, 원시단어의 의미분별과 목적단어의 선택을 결합하여 번역어를 선택한다. 번역어를 선택하기 위한 정보는 대역사전과 목적언어코퍼스에서 자동으로 추출한다. 의미분별에서는 대역사전에서 추출한 정보를 이용하고, 단어선택에서는 목적언어 공기빈도를 사용한다. 단어선택에서는 올바른 의미를 가지는 단어들에 대해서만 목적언어 정보를 활용함으로써 이해도가 높은 문장을 만들면서도 정확한 의미를 가지는 즉 충실도 높은 번역어를 선택할 수 있다.

2. 기존 연구

기계번역 시스템은 언어학적 지식에 기반하는가의 여부에 따라 크게 변환기반 방식과 중간언어기반 방식, 확률기반 방식과 예제기반 방식으로 나뉜다[2]. 변환기반 방식과 중간언어기반 방식은 규칙이나 지식을 이용하는 방식으로 주로 사람이 수동으로 작성한 규칙을 이용한다. 확률기반 방식은 코퍼스에서 추출한 통계정보를 이용하고, 예제기반 방식에서는 잘 구축된 예제와의 유사도에 기반하여 번역문장을 얻는다.

규칙에 기반한 역어선택 방식에서는 원시단어가 각 목적단어로 번역될 때의 원시단어의 품사나 구문 정보, 문맥 정보가 번역어 선택을 위한 규칙의 조건으로 기술된다. 규칙기반이나 지식기반 방식은 잘 기술된 사전이나 규칙을 이용하므로 비교적 정확한 번역을 얻을 수 있는 장점은 있지만, 번역을 위한 지식을 수동으로 구축해야 하므로 역어선택 지식을 획득하기 어렵다. 올바른 번역을 얻기 위해서는 다양한 번역 현상을 일관성 있게 처리할 수 있는 구조화된 지식이나 규칙이 필요하지만, 예측 불가능한 수많은 번역 현상을 조건간의 충돌 없이 일관성 있게 처리할 수 있는 규칙을 작성하기는 매우 어렵다. 또한, 원시언어의 모든 번역어 각각을 선택하기 위한 조건을 기술해야 하므로 사전이나 규칙의 수가 매우 커질 수 있다.

근래에 들어 대용량의 언어지식이 전자화된 형태로 구축되면서 하나의 문서를 두 언어로 기술한 병렬코퍼스(bilingual corpus)나 연관되지 않은 두 언어의 문서를 가진 양언어 코퍼스(non-parallel corpus, dual corpus)를 이용한 통계적 접근 방법들이 기계번역과 역어선택에서 등장하였다. 병렬코퍼스는 번역을 위한 지식을 자동으로 추출하기 좋은 지식원이지만, 번역 지식을 용이하게 획득하기 위한 대안이 되기는 힘들다. Koehn and Knight[3]는 병렬코퍼스 사용의 문제점을 지적하고, 양 언어의 단일언어코퍼스와 대역사전을 이용한 역어선택 방식이 병렬코퍼스만을 이용한 방식과 유사한 결과를 낼 수 있음을 보였다. 목적언어코퍼스를 이용한 방식[4]이나 양언어 코퍼스에 기반한 방식[3]에서는 대역사전을 이용하여 원시 단어에 대응되는 목적단어를 얻고, 번역 문장을 구성하게 될 목적 단어 간의 연관도를 목적언어 코퍼스에서 추출한 통계값에 기반하여 계산하여 가장 자연스러운 목적문장을 생성하는 번역어를 선택한다. 양언어 코퍼스 기반 방식은 목적언어 정보에만 의존하지 않도록 원시언어 코퍼스에서의 단어간 연관도를 목적언어 통계값에 반영하여 번역어를 선택한다.

3. 번역어간의 '단어-의미 의미-단어' 대응관계

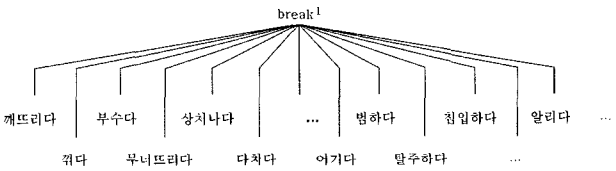
본 장에서는 '단어-단어' 대응관계에 기반한 기존의 역어선택 방식의 문제점을 살펴보고, 본 논문에서 제안하는 '단어-의미 의미-단어' 대응관계에 기반한 번역방식에 대해서 설명한다. 그리고, '단어-의미 의미-단어' 대응관계를 반영하기 위한 번역어 선택 방식으로 의미분별과 단어선택의 결합을 제안한다.

3.1 '단어-단어' 대응관계

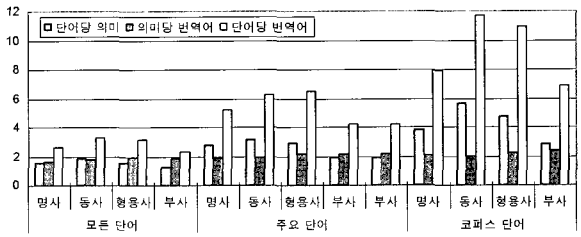
앞 장에서 나열한 기존의 역어선택 방식에서는 원시언어의 단어를 목적언어의 단어로 직접 번역하는 방식을 사용한다. 이러한 직접 대응관계는 '단어-단어' 대응관계로 표현될 수 있다. (그림 1)은 기존의 번역어 선택 방식에서 이용하던 '단어-단어' 대응관계를 영어단어 'break'와 그에 대한 한국어 번역어의 대응관계를 예제로 표현한다.

'단어-단어' 대응관계에 기반한 번역어 선택 방식은 지식 획득이 어렵고, 잘못된 번역어를 선택할 가능성이 높다. '단어-단어' 번역을 하기 위해서는 원시단어에 대응되는 모든 목적단어 각각을 선택하기 위한 지식이 필요하다. 이러한 지식을 사람이 수작업으로 구축하기 위해서는 많은 시간과 노력이 필요하고, 자동으로 추출하기 위해서는 병렬코퍼스나 의미태그가 부착된 코퍼스가 필요하다. 일부 연구에서는 지식 구축의 어려움을 줄이기 위해서 자주 이용되는 번역어들만을 목적단어로 취급하여 번역어를 선택하였다. 하지만, 이와 같이 제한된 수의 목적단어만을 번역어로 한정하면 얻어지는 목적언어 문장이 부자연스럽거나 원시문장의 의도를 정확하게 전달하지 못 한다.

'단어-단어' 대응관계는 문제 복잡도를 높인다. (그림 2)는 영영한 사전[5]에서 추출한 품사별 단어와 의미, 번역어의 분포를 보인다. 그림에서 단어당 의미는 한 단어가 평균적으로 가지는 의미의 개수를, 의미당 번역어는 하나의 의미가 가지는 평균 번



(그림 1) 영어 동사break와 한국어 번역어 간의 '단어-단어' 대응관계



(그림 2) 영영한 사전에서 추출한 단어-의미-번역어 분포

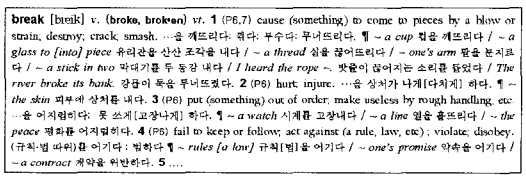
역어 개수를, 단어당 번역어는 한 단어가 가지는 평균 번역어 개수를 보인다. 그림의 왼쪽의 모든 단어는 사전의 모든 단어의 분포를 보이고, 중간의 주요 단어는 사전에서 중요하거나 자주 쓰이는 것으로 표시된 단어에 대한 분포를 보인다. 오른쪽의 코퍼스 단어는 영어 코퍼스의 문장들에서 나타나는 단어에 대한 분포를 보인다. 주요 단어는 하나의 단어가 평균 2.82개의 의미를 가지고 하나의 의미는 1.96개의 한국어에 대응되었으며 하나의 단어는 5.55개의 한국어 단어에 대응되었다. 영어 코퍼스에서는 평균 단어당 의미는 4.2개, 단어당 번역어는 8.6개로, 사전에서 추출한 값에 비하여 대응 개수가 더 많았다. 이에서 실제 언어 환경에서 자주 쓰이는 단어일수록 단어당 번역어 비율이 크다는 것을 알 수 있다.

'단어-단어' 대응관계에 기반한 역어선택은 원시단어를 직접 목적단어, 즉 번역어로 대응시키므로 그 문제 복잡도가 단어당 번역어에 비례하고, 지식 추출의 복잡도나 어려움도 마찬가지로 단어당 번역어 개수에 비례한다.

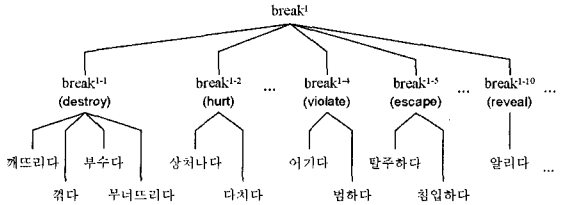
3.2 '단어-의미 의미-단어' 대응관계

단어의 의미 중의성은 다의적(polysemous) 중의성과 동음이의적(homograph) 중의성으로 나눌 수 있다. 두 종류의 의미적 중의성은 사전의 구성에도 반영된다. 일반적인 대역 사전(bilingual dictionary)은 하나의 단어가 가지는 동음이의적 중의성은 각기 다른 표제어로 구분하고, 다의적 중의성은 하나의 표제어 안에서 순서를 매겨 나타낸다. (그림 3)은 영영한 사전의 동사 'break'에 대한 내용의 일부이다[5]. 사전에서는 영어 동사 'break'가 가지는 다의적 중의성을 나열하고, 각각의 의미에 대응되는 한국어 번역어와 예문들을 제시한다.

이처럼 하나의 단어는 하나 이상의 의미를 가지고, 각각의 의미는 하나 이상의 다른 언어의 단어에 대응될 수 있다. 본 논문에서는 이러한 원시언어 단어와 원시단어의 의미, 목적단어와의 관계를 '단어-의미 의미-단어' 대응관계로 표현한다. 영어 동사 'break'와 한국어 번역어 간의 대응관계를 그림으로 나타내면 (그림 4)와 같다. 예를 들어, 사전에서 단어 'break'의 첫번째 항목 $break^1$은 여러 개의 의미를 가진다. 첫번째 의미 $break^{1-1}$은 부수다(destroy)의 의미이고, 이 의미는 한국어 '깨뜨리다', '꺾다', '부수다', '무너뜨리다' 등으로 번역될 수 있다. 네번째 의미 $break^{1-4}$는 위반하다(violate)의 의미이고, 이는 한국어 '어기다', '범하다' 등으로 번역될 수 있다. 이러한 '단어-의미 의미-단어' 대응관계는 대부분 언어의 단어들에 적용된다.


break [breik] *v.* **(broke, broken)** *vt.* **1** (P6,7) cause (something) to come to pieces by a blow or strain, destroy; crack, smash. …을 깨뜨리다; 꺾다; 부수다; 무너뜨리다. ¶ *~ a cup* 컵을 깨뜨리다 / *~ a glass to [into] piece* 유리잔을 산산 조각을 내다 / *~ a thread* 실을 끊어뜨리다 / *~ one's arm* 팔을 분지르다 / *~ a stick in two* 막대기를 두 동강 내다 / *I heard the rope ~.* 밧줄이 끊어지는 소리를 들었다 / *The river broke its bank.* 강물이 둑을 무너뜨렸다. **2** (P6) hurt; injure. …을 상처가 나게[다치게] 하다. ¶ *~ the skin* 피부에 상처를 내다. **3** (P6) put (something) out of order; make useless by rough handling, etc …을 어지럽히다; 못 쓰게[고장나게] 하다. ¶ *~ a watch* 시계를 고장내다 / *~ a line* 열을 흩뜨리다 / *~ the peace* 평화를 어지럽히다. **4** (P8) fail to keep or follow; act against (a rule, law, etc); violate; disobey. (규칙·법 따위)를 어기다; 범하다 ¶ *~ rules [a law]* 규칙[법]을 어기다 / *~ one's promise* 약속을 어기다 / *~ a contract* 계약을 위반하다. **5** ....

(그림 3) 영영한 사전의 동사 'break'에 대한 내용의 일부



(그림 4) 영어 동사 'break'와 한국어 번역어 간의 '단어-의미 의미-단어' 대응관계

원시언어 단어와 목적언어 단어는 '단어-의미 의미-단어' 대응관계를 가지고, 원시단어의 의미에 맞지 않은 번역어를 선택하면 틀린 번역 문장을 얻고 의미는 맞지만 목적문장에 어울리지 않은 번역어를 선택하면 부자연스러운 번역문을 생성한다. 예를 들어, 아래의 예문 (1)에서 볼 수 있듯이, 'break a regulation'에서 'break'의 번역어로 '부수다'를 선택하면 'break'의 원시문장에서의 의미인 위반하다(violate)에 맞지 않으므로 틀린 번역문을 얻게 된다. 의미에 맞는 '어기다'나 '범하다'를 번역어로 선택하면 올바른 번역문을 얻을 수 있다.

(1) They broke the regulation.
⇒ (X) 그들은 규칙을 부수었다.
⇒ (O) 그들은 규칙을 어겼다.
⇒ (O) 그들은 규칙을 범했다.

하지만, 원시문장의 의미에 맞는 단어를 선택하는 것이 항상 정확한 번역문을 만들지는 않는다. 한국어에서 '깨뜨리다'는 단단한 물체를 조각나게 하는 의미로 쓰이고, '꺾다'는 긴 물체를 휘어 부러뜨릴 때 이용된다. 따라서, 'break this window'를 번역할 때에 'break'의 번역어로 '꺾다'를 선택하면 'break'의 의미 부수다(destroy)에는 맞지만 부자연스럽고 이해하기 어려운 한국어 문장을 생성하지만, '깨다' 또는 '깨뜨리다'를 번역어로 선택하면 원문의 의미에 맞으면서 자연스러운 한국어 번역 문장을 얻을 수 있다.

(2) He broke this window.
⇒ (O) 그가 이 창문을 깨뜨렸다.
⇒ (?) 그가 이 창문을 꺾었다.
⇒ (X) 그가 이 창문을 어기었다.

예제에서 살펴본 바와 같이, 정확한 번역어를 선택하기 위해서는 원시언어 문장에서의 의미에 맞으면서 목적언어 문장을 자연스럽게 만드는 번역어를 선택하여야 한다. 본 논문에서는 이 점에 기반하여 원시단어의 의미분별과 목적단어의 선택을 결합하여 번역어를 선택할 것을 제안한다.

3.3 의미분별과 단어선택을 결합한 역어선택

기존의 역어선택에 대한 연구에서는 원시단어와 목적단어의 직접 대응관계 즉 '단어-단어'관계에서 번역어를 선택하는 방식을 취한다. 이 경우 다양한 번역어들 각각을 선택하기 위한 조건이나 지식이 필요하므로 번역 지식 획득이 어렵다. 또한 선택 대상이 많으므로 문제복잡도가 높아지고 역어선택의 정확률이 떨어질 가능성이 높다.

본 논문에서는 '단어-의미 의미-단어' 대응관계에 기반하여 의미분별과 단어선택을 결합하여 번역어를 선택할 것을 제안한다. 역어선택은 원시문장의 각 단어의 의미를 분별하고, 분별된 의미에 대응되는 번역어 중에서 가장 적합한 단어를 선택하는 과정으로 수행된다. 의미분별의 척도로 의미선호도(sense preference)를 사용하고, 단어선택은 단어확률(word probability)을 척도로 이용한다. 의미선호도와 단어확률을 결합하여 번역어를 선택한다.

의미분별에서는 대역사전에 나타난 원시단어의 의미 분류를 의미의 단위로 사용한다. 의미분별에서는 의미선호도를 이용한다. 의미선호도는 대역사전의 번역 예문을 이용하여 구한다. (그림 3)의 사전에서 볼 수 있듯이, 'break'가 부수다(destory)의 의미로 쓰이는 경우에는 예문에 'cup', 'piece', 'thread'와 같은 단어들이 발생하고, 위반하다(violate)의 의미로 쓰이는 경우에는 'promise'나 'contract'와 같은 단어가 발생한다. 이처럼 사전의 예문에 나타난 단어들은 단어의 의미를 분별하기 위한 문맥정보나 공기정보로 활용될 수 있으므로 이를 이용하여 의미선호도를 계산한다.

원시 단어의 의미는 여러 개의 목적 단어로 표현되고 같은 의미를 나타내는 목적 단어 집합을 구성한다. 단어확률은 한 의미에 대응되는 목적단어 집합의 각 단어가 번역어로 사용될 확률을 나타낸다. 예문 (2)에서 보았듯이, '깨뜨리다', '부수다', '꺾다', '무너뜨리다'는 영어 단어 'break'의 하나의 의미를 표현하지만, 한국어에서는 그 쓰임새가 다르다. 'window'의 번역어 '창문'과 '깨뜨리다'는 한국어 코퍼스에서 자주 공기하여 발생하지만, '창문'과 '꺾다'는 공기하여 발생하는 경우가 거의 없다. 이처럼 번역 문장의 자연스러움이나 올바른 번역 문장을 만드는가의 여부는 목적단어간의 공기 빈도를 이용하여 판단할 수 있고, 본 논문에서는 이를 단어확률로 나타낸다.

마지막으로 의미선호도와 단어확률을 결합하여 어떤 단어를 번역어로 선택할지 결정한다. 의미분별과 단어선택의 척도를 결합하여 선택된 번역어는 정확한 의미를 가지면서도 자연스러운 목적문장을 만들 수 있다.

의미분별과 단어선택을 결합하여 번역어를 선택하는 방법은 기존의 방식들에 비하여 문제 복잡도를 낮출 수 있으며, 획득하기 쉬운 지식을 사용하면서도 의미적으로 적합하면서 자연스러운 번역어를 선택할 수 있다.

- **문제 복합도의 감소** : (그림 1)에서 볼 수 있듯이, 단어당 번역어 수는 단어당 의미 수에 비해 큰 값을 가지고, 그 값은 실제 문서에서 자주 나타나는 단어일수록 높아진다. 기존의 역어선택에 대한 연구들을 원시단어를 직접 목적단어로 대응시킨다. 따라서 그 문제 복잡도가 단어당 번역어에 비례하고, 문제 복잡도가 높은 만큼 번역어선택의 정확률도 떨어지게 된다. 하지만, 본 논문의 방식은 의미분별과 단어선택을 통해 번역어를 선택하므로 그 복잡도가 단어당 의미 또는 의미당 번역어에 비례하고, 따라서 기존의 방식에 비해 문제의 복잡도가 낮다.
- **용이한 지식 획득** : 역어선택은 원시언어를 목적언어에 대응시키는 문제이므로, 단일언어를 대상으로 하는 문제에 비해 지식 획득이 더 어렵다. 기존의 역어선택 방식은 원시단어를 목적단어로 직접 대응시키기 때문에, 번역을 위한 지식을 자동으로 추출하기 위해서는 양 언어가 연결된 형태의 복잡한 지식원, 즉 수동으로 구축한 복잡한 지식 베이스나 병렬코퍼스와 같이 획득하기 어려운 지식원을 요구한다. 하지만, 원시언어의 의미분별과 목적언어의 단어선택으로 번역어를 선택하면 양 언어가 연결된 지식원으로는 각 의미에 대응되는 번역어 집합 정보만 필요하고, 원시언어의 의미분별을 위한 지식과 목적언어 단어선택을 위한 지식은 원시언어와 목적언어 각각에 대한 개별 지식원에서 획득할 수 있다. 본 논문에서는 의미에 대한 번역어 대응관계와 원시언어 의미분별 지식은 대역 사전에서 추출하고, 목적언어 단어를 선택하기 위한 지식은 목적언어 단일언어 코퍼스에서 추출한다.
- **강건한 역어선택** : 기존의 역어선택에 대한 연구는 역어선택을 의미분별 문제로 취급하거나 통계 정보에 의존하여 번역어를 선택한다. 통계정보를 활용하는 코퍼스 기반 방식에서는 단어 의미에 대한 정보 없이 단어열의 통계적 분포만을 이용하므로 의미에 맞지 않은 번역어를 선택하기 쉽다. 역어선택을 의미분별 문제로 보는 경우에는 다양한 번역어 각각을 각기 다른 의미로 보고, 이를 선택하기 위한 조건이나 정보를 필요로 하므로 지식 획득이 어렵거나 자료 부족 문제가 발생할 수 있다. 본 논문에서는 의미분별 결과에 기반하여 단어를 선택하므로 의미에 맞지 않는 번역어를 선택할 가능성이 적어, 충실도(fidelity)가 높은 즉 강건한 결과를 보장한다.
- 자연스러운 목적문장 생성 : 올바른 의미에 대한 번역어들이 항상 바른 번역 문장을 만들지는 않는다. 'break this window'에서 'break'의 의미가 destroy이고 '꺾다'는 destroy의 의미에 대응되는 단어이지만, 이 문장을 '창문을 꺾다'로 번역하면 원문의 의미를 이해하기 어렵다. 본 논문에서는 목적단어의 공기빈도를 이용하여 분별된 원시단어의 의미에 대응하는 단어 중에서 가장 적절한 단어를 번역어로 선택한다. 따라서, 올바른 의미를 가지면서 자연스러운 목적문장을 구성하는, 즉 이해도(*intelligibility*)가 높은 번역어를 선택할 수 있다.

4. 역어 선택 모델

본 논문에서는 원시단어의 의미분별과 목적단어의 선택의 과정을 통해 번역어를 선택한다. 의미분별과 단어선택에서는 그 척도로 의미선호도와 단어확률을 이용한다. 의미 분별은 대역사전에서 추출한 정보를 이용하고, 단어 확률은 목적언어코퍼스에서 추출한 정보를 사용한다. 두 값의 결합치가 가장 큰 목적 단어가 번역어로 선택된다.

4.1 의미분별을 위한 의미 선호도의 계산

본 논문에서는 입력 문장의 단어가 사전에서 추출한 의미 분별 정보와 얼마나 일치하는지를 의미선호도로 정의하고, 이 값을 이용하여 원시단어의 의미를 분별한다. 예를 들어 (그림 3)의 사전에서 'break'가 부수다(destroy)의 의미로 쓰이는 경우에는 주변에 'cup'이나 'bank'와 같은 단어가 발생하고, 어기다(violate)의 의미로 쓰이는 경우에는 문맥에 'contract'나 'promise'와 같은 단어를 가진다는 정보를 예문에서 추출할 수 있다. 예를 들어 'break the agreement'라는 입력 문장에서 'break'의 문맥에 발생하는 'agreement'는 'cup'이나 'bank'보다 'contract'나 'promise'와 더 유사하므로 'break'의 의미가 어기다(violate)임을 알 수 있다.

의미선호도(sense preference, *spf*)는 입력 문장의 단어가 사전에서 추출한 각 의미분별 정보와 얼마나 일치하는지를 나타낸다. 문장에서의 원시단어가 가지는 품사에 대응 가능한 품사를 가지는 표제어의 사전 정보를 추출한다. 의미선호도는 의미 문맥 정보를 기본으로 아래의 식 (1)을 통해서 계산한다. 의미분별 대상이 되는 원시단어를 s_i라 하고, s_i의 k번째 의미를 s_i^k 라 하면, 입력 문장에서의 s_i^k 의 의미선호도는 $spf(s_i^k)$로 표시한다. 식에서 집합 $EX_{s_i^k}$ 는 의미 s_i^k 의 예문에서 추출한 모든 실질어를

가지고, $|EX_{s_i^k}|$는 집합 $EX_{s_i^k}$가 가지는 단어의 개수를 표시한다. 식에서는 문장에 나타나는 모든 실질어($s_j \in SNT$, 단 $s_j \neq s_i$)에 대해서 예문에 있는 모든 단어($w_e \in EX_{s_i^k}$) 중에서 가장 유사한 것들의 거리를 합산하여 문맥에 의한 기본 의미선호도를 계산한다. 단어간의 유사도 $sim(s_j, w_e)$의 계산은 WordNet을 이용하는 Rigau et al. [6]의 방법을 사용한다. 식 (2)는 각 의미의 정규화된 의미선호도를 구한다.

$$spf_{context}(s_i^k) = \frac{\sum_{s_j \in SNT} \max_{w_e \in EX_{s_i^k}} sim(s_j, w_e)}{|EX_{s_i^k}|} \quad (1)$$

$$spf_{norm}(s_i^k) = \frac{spf_{context}(s_i^k)}{\sum_x spf_{context}(s_i^x)} \quad (2)$$

만일 원시단어의 모든 의미에 대해서 사전에 의미분별을 위한 예문 정보가 존재하지 않는다면 모든 의미에 대해서 $spf_{context}(s_i^k)=1$을 적용하여 정규화된 의미선호도를 구한다.

4.2 단어 선택을 위한 단어 확률의 계산

단어확률은 같은 의미를 가지는 번역어의 집합 중에서 어떤 단어가 가장 자연스러운 번역 문장을 구성하는지를 표현한다. n개의 실질어로 구성된 원시문장 "$s_1\ s_2 \ldots\ s_n$"에서, i번째 단어의 k번째 의미를 s_i^k라 하고, s_i^k에 대응되는 번역어 집합을 $T(s_i^k)$라 하자. t_{i1}^k는 $T(s_i^k)$의 첫번째 요소라고 하자. 단어 s_i와 동일 문장에 나타나는 단어의 집합 $\Theta(s_i)$의 한 원소 (s_j, m)는 t_{j1}, t_{j2}, ... , t_{jm}의 m개의 번역어를 가지는 원시단어 s_j를 표시한다 ($t_{jp} \in T(s_j)$, 단 $j \neq i$). s_j에 대응되는 목적단어 t_{jp}와 s_i^k에 대응되는 목적단어 t_{iq}^k가 목적언어코퍼스에서 동일한 문장에 나타나는 횟수를 $f(t_{iq}^k, t_{jp})$라고 하자. 이 때, 단어확률 $wp(t_{iq}^k)$는 단어 집합 $T(s_i^k)$ 중에서 t_{iq}^k를 선택할 확률을 의미한다.

$$n(t_{iq}^k) = \sum_{(s_j,m)\in\Theta(s_i)} \sum_{p=1}^{m} \frac{f(t_{iq}^k, t_{jp})}{f(t_{iq}^k) + f(t_{jp})} \quad (3)$$

$$wp(t_{iq}^k) = \hat{p}_{word}(t_{iq}^k) = \frac{n(t_{iq}^k)}{\sum_x n(t_{ix}^k)} \quad (4)$$

'break'의 예에서, $n(t_{iq}^k)$은 각 번역어 t_{iq}^k, 즉 '깨뜨리다', '어기다' 등이 목적문장에서 입력 문장의 주변 단어의 번역어들과 발생하는 빈도를 의미한다. 'break the agreement'의 경우에 'agreement'는 '동의', '일치', '협정', '계약'의 네 개의 번역어를 가지므로, (agreement, 4)$\in\Theta(break)$를 구성한다. t_j^p는 '동의', '협정' 등에 대응되므로, f(깨뜨리다, 동의), ... , f(어기다, 계약) 등의 공기빈도가 이용된다. 공기빈도를 각 단어의 발생빈도 $f(t_{iq}^k)$와 $f(t_{jp})$의 합으로 나누어, 발생빈도가 높은 단어를 선호하는 것을 막는다. 만일, 식 (3)의 모든 $n(t_{iq}^k)$가 0이라면, 단어 자체의 발생 빈도를 이용한 아래의 식을 이용하여 평탄화(smoothing)를 수행한다.

$$\text{if } \sum_y n(t_{iy}^k) = 0,\quad n(t_{iq}^k) = \sigma \cdot \frac{f(t_{iq}^k)}{\sum_p f(t_{ip}^k)} \quad (5)$$

4.3 의미선호도와 단어확률의 결합을 통한 번역어 선택

번역어 선택을 위해서는 원시 단어의 의미 분별 척도인 의미선호도와 각 의미에 대응되는 목적단어 선택 척도인 단어확률을 곱한 값을 이용한다. 단어확률 $wp(t_{iq}^k)$는 집합 $T(s_i^k)$ 내에서의 확률이므로, $T(s_i^k)$이 원소를 하나만 가지면 $wp(t_{iq}^k)$=1이 되고 원소를 여러 개 가질 수록 0에 가까운 값을 가지게 된다. 즉, 의미에 대응되는 번역어의 수가 작은 의미를 선호하게 되기 때문에 하나의 의미에 대응되는 단어의 단어확률의 최대값이 1이 되도록 $\max_j(wp(t_{ij}^k))$으로 $wp(t_{iq}^k)$를 나누어 준다.

$$target\ word = \operatorname*{argmax}_{t_{iq}^k} spf_{norm}(s_i^k) \cdot \frac{wp(t_{iq}^k)}{\max_j(wp(t_{ij}^k))} \quad (6)$$

5. 실험 및 평가

본 논문에서 제안하는 역어선택 방식의 평가를 위한 실험을 수행하였다. 평가에서는 객관적인 평가 결과를 얻기 위해서 사람의 수작업을 배제하고 병렬코퍼스와 사전을 이용하여 자동으로 추출한 평가집합에 대해서 자동으로 역어선택 결과를 평가하였다.

5.1 실험 환경

기계가독형 대역사전은 엣센스 영영한 사전에서 추출하였다 [5]. 사전은 약 4만3천개의 엔트리로 구성되어 있고, 약 8만개의 의미가 정의되어 있으며 유일한 단어의 수는 3만4천여개이다. 목적언어에서의 공기빈도는 세종코퍼스와 KAIST 코퍼스, 연세 코퍼스에서 약 63만쌍의 공기빈도를 추출하였다.

역어선택을 위한 실험집합은 Koehn and Knight[3]의 실험 집합 추출 방식을 사용하였다. 이 방식에서는 원시문장의 단어가 하나 이상의 목적문장의 단어에 대응되는 경우는 드물다고 가정하고, 대역사전을 이용하여 정렬된 문장에서의 번역어 대응관계를 얻는다. 원시문장의 각 단어에 대응되는 번역어들을 대역사전에서 찾고, 만일 정렬된 목적문장에 대응되는 번역어를 가지지 않는 원시단어가 하나라도 있으면 그 원시문장은 평가 대상에서 제외한다. 추출한 정렬코퍼스를 이용하면, 사람의 개입 없이 역어선택의 정확도를 측정할 수 있다. (그림 5)는 평가 집합 중 한 문장의 예를 보인다. 한국어 문장의 실질어를 기준으로 두번째 실질어의 '구성'이 영어단어 'structure'에 대응됨을 보이고, 이는 'structure'의 첫번째 엔트리인 명사의 두번째 의미에 대응됨을 나타낸다.

영-한 정렬된 코퍼스는 한영 사전의 예문, 소설, 고등학생을 위한 교과서 등에서 얻었다. 번역은 명사, 동사, 형용사, 부사를 대상으로 한다. 총 3,678개의 문장에서 11,042개의 단어에 대한 번역을 수행하였다.

#EXAMPLE	The sentence structure is awkward.
>EXAMPLE	The/DT sentence/NN structure/NN is/VBZ awkward/JJ ./.
#K_EXAMPLE	문장의 구성이 어색하다
>K_EXAMPLE	문장/ncn+의/jcm 구성/ncpa+이/jcs 어색/ncps+하/xsm+다/ef
>KE_MAP	#&sentence/1/n/1#&structure/1/n/2#&awkward/ 1/adj/1

(그림 5) 자동 평가를 위한 영-한 대응된 평가 문장의 예

5.2 평가

아래의 <표 1>은 의미선호도의 정확도를 보인다. 가장 높은

선호도를 가지는 의미에 대응되는 번역어 중의 하나가 정렬된 목적문장에 발생하면 정확하게 의미가 분별된 것으로 보고 의미 분별 결과를 평가하였다. 대역사전에서 추출한 예문 정보만을 이용하여 명사의 경우에는 약 60%의 정확도를, 동사에 대해서는 46% 수준의 정확도를 얻었다. 이러한 결과는 명사의 의미는 넓은 문맥에 의존하고 동사의 의미는 구문 관계로 연결된 문맥 단어에 의존적이라는 Towell and Voorhees [7]의 연구결과에 부합한다.

〈표 1〉 의미선호도에 의한 의미 분별 정확도

명사	동사	형용사	부사	전체
60.71%	46.54%	58.23%	57.09%	55.92%

아래의 <표 2>는 번역어 선택의 정확도를 보인다. 번역어 선택에 대한 실험은 베이스라인(baseline)에 대한 실험과 단어확률만을 이용한 번역어 선택에 대한 실험, 의미선호도와 단어확률의 조합을 통한 실험으로 구성된다. 역어선택의 베이스라인으로는 무작위선택과 사전에 정의된 첫번째 의미의 첫번째 번역어를 선택한 경우, 그리고 번역어 중에서 가장 자주 발생하는 단어를 선택한 경우의 세가지를 사용하였다. 사전에서 첫번째로 정의된 의미가 가장 자주 이용되는 의미이고 번역어정의문의 처음에 나타나는 단어가 가장 기본이 되는 번역어이지만, 이를 선택한 경우의 정확률은 무작위 선택에 비해 그다지 높지 않았다. 그에 비하여 자주 발생하는 단어를 무조건 선택하는 경우의 정확률이 높아, 단어의 발생 빈도가 번역에서도 좋은 정보가 되는 것으로 나타났다.

〈표 2〉 번역어 선택의 정확도

	명사	동사	형용사	부사	전체
무작위	11.22%	4.24%	9.61%	10.82%	7.41%
첫번째 의미의 첫번째 번역어	14.57%	14.94%	8.28%	4.98%	13.00%
가장 자주 발생하는 번역어	48.30%	34.11%	37.21%	25.78%	41.01%
단어확률 기반	51.00%	31.25%	45.88%	37.59%	43.77%
의미선호도-단어확률 기반	51.48%	36.83%	47.97%	38.54%	45.93%

단어확률을 사용하는 역어선택 방식은 원시언어가 가지는 의미에 상관없이 목적단어간의 공기빈도만을 이용하여 번역어를 선택한 경우의 정확도이다. 이는 기존의 목적언어코퍼스에 기반한 '단어-단어' 관계에 기반한 역어선택 방식의 결과에 해당한다. 표의 마지막 행인 의미선호도-단어확률 기반 선택은 식(5)에 의해 의미선호도와 단어확률을 결합한 값을 이용하여 번역어를 선택한 결과의 정확도이다. 표에서 볼 수 있듯이 간단한 사전 예문만을 이용하여 의미 분별 단계를 거치고 목적언어 통계정보를 활용하여 '단어-의미 의미-단어' 관계에 기반하여 번역어를 선택한 방식이 '단어-단어'기반 방식보다 좋은 결과를 보였다.

자동 평가 방식에서는 평가에 사용된 번역 문장 쌍에 나타나는 단어와 동일한 단어를 선택하여야 정확한 번역으로 평가한다. "It is a quality frequently found in Korean works of art." 에 대응된 번역문장이 "그것은 한국의 미술 작품에서 빈번하게 발생하는 특징이다."으로 되어 있을 때, 'art'의 번역어로 '예술'을, 'frequently'의 번역어로 '자주'를 선택해도 맞는 결과이지만, 자동 평가 방식은 정렬 문장의 단어와 똑같지 않은 경우에는 올바른 번역어가 선택되어도 실패한 것으로 평가하기 때문에 두 개의 번역어 선택을 모두 틀린 결과로 평가한다. 이처럼 자동 평가 방식에서는 역어 선택의 결과가 저평가되기 쉽다. 평가 대상 중 200여개의 문장을 수작업으로 평가한 결과에서는 단어확률 기반 선택의 정확도가 약 52%, 의미선호도-단어확률 기반 선택이 약 63%의 정확도를 보였다. 오류가 발생한 경우에 대한 결과 분석에서는 사전 예문의 부족 등으로 인하여 의미분석 결과가 맞지 않아 틀린 번역어를 선택한 경우가 가장 많은 부분을 차지하였고, 사전에서의 번역어 추출과정이나 번역 집합 추출과정에서 단어 대응에서 실패한 경우도 나타났다.

6. 결 론

하나의 단어는 하나 이상의 의미를 가지고 각 의미는 하나 이상의 단어로 대응될 수 있으므로, 원시단어와 목적단어 간에는 '단어-의미 의미-단어' 대응관계가 있다. 기존의 번역어선택에 대한 연구는 원시단어를 직접 목적단어로 번역하는 '단어-단어' 직접 대응관계를 이용하기 때문에 수동으로 기술한 규칙이나 병렬코퍼스와 같이 사용하기 어려운 지식을 필요로 하였다. 이를 해결하기 위해서 목적언어에서 추출한 정보만을 이용하는 방식이 제안되었으나 이 방식은 원시언어에서의 의미를 반영하지 못하므로 틀린 번역어를 얻기 쉽다.

본 논문에서는 '단어-의미 의미-단어' 대응관계에 기반하여 원시언어의 의미분별과 목적언어의 단어선택의 과정을 통해서 번역어를 선택할 것을 제안하였다. 원시 언어의 의미를 반영하면서 자연스러운 목적 언어를 구성하는 단어를 선택하므로 충실도와 이해도를 모두 만족시킬 수 있는 방식으로, 병렬 코퍼스를 이용한 자동 평가에서 '단어-단어' 기반 방식보다 좋은 결과를 보였다.

본 논문에서는 사전의 번역 예문만을 이용한 의미 분별 방식을 사용하였으나, 사전에 나타나는 다양한 정보들을 활용하여 의미 분별의 정확도를 높이는 연구를 진행할 예정이다. 또한 의미 분별과 단어 선택의 결합하는 다양한 방식에 대한 시도도 필요할 것으로 본다.

참 고 문 헌

[1] John White and Teri O'Connell, "Machine Translation Evaluation", Tutorial, the Second Conference of Association for Machine Translation in the Americas, 1996.

[2] Bonnie J. Dorr and Pamela W. Jordan and John W. Benoit, "A Survey of Current Research in Machine Translation", 1-68, Advances in Computers, M. Zelkowitz (Ed). Academic Press, London, 1999.

[3] Philipp Koehn and Kevin Knight, "Knowledge Sources for Word-Level Translation Models", Proceedings, Empirical Methods in Natural Language Processing, 2001.

[4] Ido Dagan and Alon Itai, "Word Sense Disambiguation Using a Second Language Monoligual Corpus", 563-596, Vol.20, No.4, Computational Linguistics, 1994.

[5] 엣센스 영영한 사전, 민중서림, 1995.

[6] German Rigau and Jordi Atserias and Eneko Agirre, "Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation", Proceeding of Annual Meeting of the Association for Computational Linguistics, 1997.

[7] Geoffrey Towell and Ellen M. Voorhees, "Disambiguating Highly Ambiguous Words", Computational Linguistics, Vol. 24, No.1, pp.125-147, 1998.

이 현 아

e-mail : halee@kumoh.ac.kr

1996년 연세대학교 컴퓨터과학과(학사)

1998년 한국과학기술원 전산학과(공학석사)

2004년 한국과학기술원 전산학과(공학박사)

2000년~2004년 ㈜ 다음소프트

2004년~현재 금오공과대학교 컴퓨터공학부 전임강사

관심분야 : 자연언어처리, 정보검색, 지식공학, 기계번역