韓國 컴퓨터情報學會 論文誌
第11卷 第5號, 2006. 11.

2006-11-5-1-9

다양한 지식을 사용한 영한 기계번역에서의 대역어 선택

이 기 영*, 김 한 우**

Target Word Selection for English-Korean Machine Translation System using Multiple Knowledge

Lee Ki Young*, Kim Han Woo**

요 약

일반적으로 영어를 한국어로 번역할 때, 대부분의 영어 명사 어휘들은 해당 어휘가 사용되는 문맥에 따라 다양한 한국어 명사로 번역될 수 있다. 따라서 영어 원문이 갖는 의미를 손실 없이 번역문으로 전달하기 위해서는 문맥에 맞는 올바른 한국어 대역어를 선택할 수 있어야 한다. 본 논문에서는 동사구패턴, 공기 정보에 기반한 의미벡터, 공기 품사 정보 및 한국어 문맥 통계 정보 등의 다양한 지식을 사용하여 영어 명사 어휘의 대역어를 올바로 선택하는 방안을 제공한다. 동사구 패턴은 사전과 코퍼스를 사용하여 구축되었으며, 의미 벡터는 영어 어휘가 특정 한국어 어휘로 번역될 때 공기하는 정보들의 조건부 확률을 나타낸다. 한국어 문맥 통계 정보는 한국어 코퍼스로부터 추출된 N-그램 정보를 나타내며, 품사 공기 정보는 대역어 선택 모호성을 지니는 영어 어휘와 통계적으로 깊은 관련성을 지니는 품사를 나타낸다. 마지막으로 본 논문에서 제안한 대역어 선택 모호성 해소 방안을 평가하기 위한 실험을 수행하였으며, 실험 결과, 제안하는 방법이 기존의 방법보다 성능이 좋다는 것을 확인할 수 있었다.

Abstract

Target word selection is one of the most important and difficult tasks in English-Korean Machine Translation. It effects on the translation accuracy of machine translation systems. In this paper, we present a new approach to select Korean target word for an English noun with translation ambiguities using multiple knowledge such as verb frame patterns, sense vectors based on collocations, statistical Korean local context information and co-occurring POS information. Verb frame patterns constructed with dictionary and corpus play an important role in resolving the sparseness problem of collocation data. Sense vectors are a set of collocation data when an English word having target selection ambiguities is to be translated to specific Korean target word. Statistical Korean local context Information is an N-gram information generated using Korean

• 제1저자 : 이기영
• 접수일 : 2006.10.20, 심사일 : 2006.11.05, 심사완료일 : 2006.11.12
* 한양대학교 컴퓨터공학과 ** 한양대학교 컴퓨터공학과 교수

corpus. The co-occurring POS information is a statistically significant POS clue which appears with ambiguous word. The experiment showed promising results for diverse sentences from web documents.

▸ Keyword : 기계번역(machine translation), 대역어 선택(target word selection), 단어 의미 모호성 해소(word sense disambiguation)

I. 서 론

컴퓨터의 등장과 함께 컴퓨터를 이용하여 인간이 사용하는 다양한 언어를 번역하고자 하는 기계번역에 대한 연구가 계속되어 왔다. 초기에는 규칙 기반 방식으로 개발되기 시작한 기계번역 시스템은 현재 통계 기반(Statistics-Based MT), 패턴 기반(Pattern-Based MT), 예제 기반(Example-Based MT) 등의 다양한 방법론을 사용하여 개발되고 있다. 이렇게 기계번역 시스템을 구분하는 주된 차이는 일반적으로 번역을 위해 사용하는 주된 지식이 무엇인가에 따른 것이다. 이렇듯 기계번역을 위해 사용하는 주된 지식의 차이에 따라 기계번역을 위한 다양한 개발 방법론이 사용될 수 있지만, 이러한 범주들을 포괄하여 대부분의 기계번역 시스템은 원시 문장 분석 단계, 원시 문장-목표 문장 변환 단계, 목표 문장 생성 단계를 포함한다. 원시 문장 분석 단계에서는 형태소 분석(morphological analysis) 기능과 구조 분석(syntactic analysis) 기능을 수행한다. 원시 문장 분석 단계에서는 입력된 문장을 구성하는 각 어휘에 대한 형태론적 분석을 수행하여 활용된 형태의 표층 구조로부터 어휘의 원형 등을 복원하고, 전체 문장의 문법적 구조를 밝혀낸다. 변환 단계는 원시 문장을 목표 문장으로 번역하기 위한 과정의 중간 단계로서, 어순 등을 고려하여 원시 언어와 목표 언어 간의 구조 변환과 어휘 변환을 수행한다. 마지막으로 생성 단계는 변환 단계의 결과를 목적 언어의 문법, 활용에 맞도록 최종 번역 결과를 생성한다.

기계번역 시스템을 구성하는 이러한 세 가지 단계들 가운데, 변환 단계는 원시 문장이 가지는 의미를 손실 없이 목표 문장으로 전달하는 매우 중요한 역할을 수행한다. 즉, 변환 단계에서 수행되는 구조 변환 과정과 어휘 변환 과정에서 원시 문장이 지니는 의미의 손실이 발생한다면 최종 번역문의 표층적 표현은 잘못된 부분이 없다고 하더라도 실제 원시 문장이 지니고 있는 의미를 전달하는데 오류가 발생하게 되어, 전체적인 번역 품질은 떨어진다. 특히 자주 사용되는 고빈도 어휘일수록 이러한 대역어 선택 모호성은 증가한다는 특성을 지니고 있기 때문에, 영한 기계번역 시스템의 성능 개선을 위해서 대역어 선택은 매우 중요한 분야라고 할 수 있다. 본 논문에서는 영한 기계번역에 있어서 대역어 선택 모호성을 지니는 영어 명사 어휘를 대상으로 다양한 지식을 사용하여 의미 모호성 해소 및 대역어 선택 과정을 포함하는 효과적인 대역어 선택 방안을 제시한다. 본 논문의 구성은 다음과 같다. 2절에서는 대역어 선택에 대한 관련연구를 설명한다. 3절에서는 기계번역에서의 대역어 선택 모호성에 대해서 설명하며, 4절과 5절 및 6절에서는 다양한 지식을 사용하여 대역어 선택 모호성을 해소하는 방안을 제시하며, 7절에서는 본 논문에서 제안하는 방식에 대한 실험 결과를 논한다. 마지막으로 결론과 향후연구 방향을 8절에서 제시한다.

II. 관련 연구

대역어 선택은 기계 번역에서 원문이 담고 있는 의미를 번역문으로 정확히 전달하기 위해서 필수적인 기능이다. 기계 번역에서 대역어 선택이 일차적으로 원문의 어휘가 지니는 의미적 모호성을 해소하는 과정을 포함하기 때문에, 기존의 연구로서 정보 검색 등의 분야에서 행해져온 의미 모호성 해소에 관한 연구도 전체적인 대역어 선택의 관점에서 함께 살펴보는 것이 타당하다고 할 수 있다.

[1], [2], [3]과 [4]는 의미 모호성 해소에 사용되는 유용한 지식들을 분류하였고, 또한 각각의 지식들이 어느 정도 유용한 지에 대한 연구를 수행하였다. 어휘가 갖는 의미적 모호성을 해소하는데 유용한 지식으로는 품사, 형태소 정보, 공기 정보(collocation), 의미 정보, 문법 정보, 의미의 사용 빈도 정보, 주제 정보(domain)

및 화용 정보(pragmatics) 등 매우 다양하다. 이러한 종류의 지식들은 단독으로 사용될 수도 있으며, 서로 조합되어 함께 사용될 수도 있다. 동시에, 코퍼스로부터 추출된 통계 정보를 사용하여 의미 모호성을 해소하려는 시도가 [5], [6], [7] 등에서 있었다.

일반적으로, 의미 모호성을 해소하는데 가장 유용한 지식으로 의미 태깅된 대규모 코퍼스로부터 얻어진 통계 정보를 생각할 수 있다. 하지만, 일관성 있는 의미 태깅 코퍼스를 구축하는 것은 비용과 시간 면에서 매우 어려운 작업이기 때문에, 이러한 방법 또한 현실적인 면과는 거리가 멀다. 결국 이러한 이유로 인해 많은 연구자들은 보다 현실적이고 지식 습득이 보다 쉬운 지식들을 사용하여 의미 모호성을 해소하려는 많은 노력을 하고 있다 [8], [9], [10], [11].

III. 대역어 선택 모호성 해소의 중요성

영한 번역에서, 일반적으로 품사를 막론하고 많은 영어 어휘들이 문맥에 따라 다양한 한국어 어휘로 번역될 수 있다. 즉, 사전에 등재된 각 영어 어휘들에 대해서 가능한 한국어 대역어 정보가 등록되어 있다고 할 때, 이러한 후보 대역어들은 각 영어 어휘가 사용된 문맥에 따라 적절하게 선택되어 번역되어야 문장 전체가 갖는 의미가 손실 없이 번역문으로 전달된다. 다음의 예문들은 이러한 경우를 나타낸다.

(E3.1) A missile defense system (NMD) could lead to a possible arms *race*.
(K3.1) 미사일 방어 체제가 가능한 무기 *경쟁*으로 이끌 수 있다.

(E3.2) Bradford has a history of *race* related violence.
(K3.2) 브래드포드는 *인종*과 관련된 폭동의 역사가 있다.

예문 (E3.1)과 (E3.2)에서 사용된 'race'는 {경주, 경쟁, 선거, 인종} 등의 대역어를 가지고 있으며, 올바른 번역은 사용된 문맥에 따라 다르다고 할 수 있다.

우리는 이렇게 2가지 이상의 대역어가 사전에 등록되어 있어서 번역시 가능한 대역어 후보들 가운데 하나를 선택해야 하는 영어 어휘들에 대해서 대역어 선택 모호성을 지녔다고 정의한다. 특히, 본 논문에서는 영어 명사 어휘들을 대상으로 하여 기계번역에서의 대역어 선택 방안에 대해서 자세한 설명을 한다.

표 1. 대역어 선택 모호성 어휘(명사) 통계
Table 1. Statistics about English polysemous words

전체 명사 어휘 수 (숙어 제외)	31,836
대역어 선택 모호성을 지닌 어휘 수	10,493

표 1은 현재 영한 기계번역용 기계 가독형 사전(machine readable dictionary)에서 영어 명사 어휘를 대상으로 그 대역어 개수가 2개 이상인 어휘들에 대한 통계 정보를 나타낸다. 즉, 전체 명사 어휘의 약 3분의 1 정도가 대역어 선택 모호성을 지니며, 실제로, 고빈도 어휘일수록 후보 대역어의 개수는 더 많다고 할 때, 영한 기계번역에서 대역어 선택이 전체 번역률에 얼마나 큰 영향을 주는지 알 수 있다.

표 2는 사전에서 대역어 선택 모호성을 지니는 대표적인 어휘들에 대한 대역어와 각 대역어에 할당된 의미 코드를 보여준다. 본 논문에서는 워드넷 1.71 (WordNet 1.71)에 기반을 둔 1,163개의 의미 코드가 사용되었다 [12].

표 2. 한국어 대역어 및 의미 코드 예
Table 2. Korean target words and semantic codes

영어 어휘	의미 코드	한국어 대역어
race	contest#1	경주, 경쟁, 선거
	race#2	인종
president	communicator#1	대통령
	leader#1	회장, 위원장
life	person#1	생명
	being#1	삶, 생활, 인생
	life#5	수명
	living_thing#1	생물

표 2에 보이는 영어 명사 어휘 'race'는 4개의 대역어 후보들을 가지며, 이중 {경주, 경쟁, 선거}에는 의미 코드 'contest#1'이 할당되고, {인종}에는 의미 코드 'race#2'가 할당되어 있음을 알 수 있다.

본 논문에서는 원시 언어 모델과 목표 언어 모델을 모두 고려하여 대역어 선택을 수행하는 방안을 제안한다. 즉, 표 2의 3개 어휘 모두 2개 이상의 의미 코드가 할당되어 있다. 이런 경우에 하나의 의미 코드를 결정하는 것은 원시 언어 모델에 기반한다. 원시 언어 모델에 의해 의미 코드가 결정된 후, 해당 의미 코드를 가지는 대역어가 2개 이상인 경우, 이들 중에서 최종 한국어 대역어를 선택하는 것은 목표 언어 모델에 기반한다.

IV. 대역어 선택 방안의 2단계 구성

그림 1은 본 논문에서 제안하는 대역어 선택 방식을 전체적으로 보여주는 개념도이다. 제안하는 대역어 선택 방식은 전체적으로 2단계로 구성되어 있으며, 각 단계에서 사용되는 지식은 서로 다르다. 1단계에서는 의미 모호성을 해소하여 어휘가 가진 의미 코드를 결정하고, 2단계에서는 결정된 의미 코드를 공유하는 대역어 후보들 가운데 가장 적합한 대역어를 결정한다. 그림 1을 보다 자세히 설명하면 다음과 같다. 우선, 공기 품사 정보를 적용 대상 어휘에 대해서는 공기 품사 정보가 적용되어 가장 먼저 대역어가 결정된다. 공기 품사 정보는 적용되는 어휘의 수는 많지 않지만, 일단 적용되면, 매우 정확하게 대역어를 선택하는데 도움이 된다. 1단계에서 동사구 패턴이 문장을 구성하는 각 단문에 대해서 적용된다. 이때, 각 단문의 논항 자리의 헤드 명사는 동사구 패턴에 의해 의미가 결정된다. 그리고 헤드 이외의 명사나 동사구 패턴이 매칭되지 않는 경우에는 의미 벡터를 사용하여 의미 모호성을 해소하고 해당 어휘의 의미 코드가 결정된다. 2단계에서는 1단계에서 결정된 의미 코드를 공유하는 한국어 대역어의 개수가 2개 이상인 경우, 한국어 문맥 통계 정보를 활용하여 최종 한국어 대역어를 결정한다.

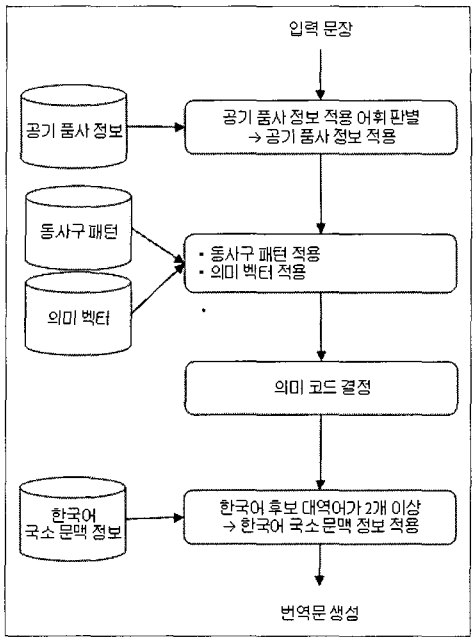

그림 1. 대역어 선택 과정
Fig. 1 Overview of the target word selection

V. 제 1단계: 의미 모호성 해소

5.1 동사구 패턴의 사용

5.1.1 동사구 패턴의 정의

언어학적 의미에서 격틀(case frame)은 동사와 동사가 갖는 필수격에 해당하는 논항 정보로 구성된다. 본 논문에서는 격틀이 갖는 문법적이면서 동시에 의미적인 성질을 대역어 선택에 이용하고자 한다. 동사구 패턴은 이러한 목적을 위해서 도입되었으며, 격틀과 매우 유사한 형태를 지닌다. 동사구 패턴은 기본적으로 원시 언어 표현과 해당 대역 표현인 목표 언어 표현을 모두 가진다. 이러한 특징 때문에 동사구 패턴은 입력 문장의 문법적 구조를 밝히는 구조 분석에서도 사용되며, 동사구 단위의 번역을 위해 구조 변환에서도 사용된다. 동사구 패턴은 동사가 갖는 필수격에 해당하는 논항 정보를 포함하며, 해당 논항의 자리에는 가능한 의미 코드가 위치한다. 이때의 의미 코드는 앞 절에서 기술한 바와 같이, 워드넷 1.71에 기반한 의미 코드가 사용된다.

S=(person#1) abandon:v O=(conveyance#3) =〉 S! 가 O! 를 포기하!

S=(person#1) abandon:v O=(female#2) =〉 S! 가 O! 를 버리!

S=(conveyance#3) abandon:vn =〉 S! 가 방치되!

S=(all) abandon:vn PREP=by PO=(all) =〉 S! 가 PO! 에_의해 폐지되!

S=(person#1) abandon:v O=(unit#3) =〉 S! 가 O! 를 떠나!

S=(idea#1) abandon:v O=(person#1) PREP=to PO=(change#1) =〉 S! 가 O! 를 PO로 몰!

그림 2. 동사 'abandon'에 대한 동사구 패턴 예
Fig. 2 Verb frame patterns for verb 'abandon'

동사구 패턴에서 사용되는 의미 코드는 동사구 패턴의 적용 가능성과 적용되었을 때의 정확률에 매우 중요한 영향을 미친다. 의미 코드의 수가 너무 많은 경우와 너무 적은 경우 모두 적용 가능성 및 정확률에 나쁘게 작용한다. 특히 이러한 의미 코드의 적정 개수에 대해서는 수학적 적정값이 존재하지 않으므로 지속적인 실험과 관찰을 통해서 적절한 수의 의미 코드가 결정될 수 있다.

동사구 패턴은 명사와 동사의 의미를 결정하는데 있어서 매우 중요한 역할을 수행한다. 그림 2는 영어 동사 'abandon'에 대한 동사구 패턴의 예를 보인다. 그림 2에서 'conveyance#3', 'female#2' 등은 동사 'abandon'의 논항 자리를 채우는 의미 코드를 나타낸다.

현재 사용 중인 대부분의 동사구 패턴들은 수동으로 구축되었으며, 현재까지 구축되어 사용 중인 동사구 패턴의 수는 약 4만7천개이다. 현재까지 구축된 동사구 패턴은 그 효용성을 위해 고빈도 동사를 중심으로 구축되었다.

5.1.2 동사구 패턴의 적용

동사구 패턴은 앞서 기술한 격틀과 같이, 동사의 의미를 결정하는 동시에 해당 동사의 논항의 자리에 올 수 있는 명사의 의미적 속성을 제약한다. 동사구 패턴은 의미 모호성 해소를 위한 일종의 문법적 클루(clue)라고 생각할 수 있다.

본 논문에서는 대역어 선택 모호성을 지니는 명사의 의미 결정을 위해 가장 먼저 적용하는 지식으로 동사구 패턴을 사용한다. 동사구 패턴으로 의미 모호성을 해결하는 대상은 입력 문장을 구성하는 각 절에서 동사의 논항 자리를 차지하는 헤드 명사로 제한한다. 이와 같은 이유는 동사구 패턴 자체가 갖는 제약이기도 하다.

입력 문장에 대한 구조 분석이 끝난 후, 구조 분석기는 분석된 각 절에 대해서 적용 가능한 동사구 패턴 후보들을 변환기에 넘긴다. 이 과정에서 변환기가 받는 동사구 패턴 후보들은 구조 분석기에 의해서 분석된 입력 문장 특히 절에서의 동사구 구조와 매칭되는 것들이다. 변환기는 입력 문장을 구성하는 각 절들의 동사구 구조에서 해당 동사의 논항에 위치한 헤드 명사의 의미 코드와 후보 동사구 패턴들의 각 논항 자리에 위치한 의미 코드를 비교하여 최적의 동사구 패턴을 선택한다. 이렇게 동사구 패턴이 선택됨과 동시에 절에서 사용된 동사의 의미와 해당 동사의 논항에 위치한 명사의 의미가 결정된다. 그림 3은 임의의 입력 문장에 대해 동사구 패턴을 적용하는 과정을 나타낸다. 그림 3에서 의미 모호성 어휘 'point'에 대한 의미 코드가 'meaning#1'으로 결정되는 이유는 동사구 패턴 'S=(person#1) absorb:v O=(meaning#1) ==〉 S!가 O!를 이해하!' 를 제외한 나머지 2개의 동사구 패턴은 'point'가 갖는 의미 코드를 목적어 논항에 가지고 있지 않기 때문에, 상기의 동사구 패턴이 적용되며, 이때 'point'의 의미 코드도 'meaning#1'으로 함께 결정된다. 즉, 동사구 패턴을 적용함으로써 동사의 대역어와 논항에 위치한 명사의 의미 코드가 동시에 정해진다.

사전 내용,

<point/NOUN>
..... [의미 코드: point#2], [대역어: 지점], ...
..... [의미 코드: characteristic#2], [대역어: 핵심], ...
..... [의미 코드: meaning#1], [대역어: 요점], ...

동사구 패턴 DB 내용,

<absorb/VERB>
S=(market#1) absorb:v O=(container#1) ==> S! 가 O! 를 소화하!
S=(person#1) absorb:v O=(loss#2) ==> S! 가 O! 를 부담하!
S=(person#1) absorb:v O=(meaning#1) ==> S! 가 O! 를 이해하!

입력 문장: I absorbed the full point of a his remark.

논항 자리의 의미 코드를 비교,

<S=(person#1) absorb:v O=(meaning#1) ==> S! 가 O! 를 이해하!>

→ 입력 문장과 매칭되는 동사구 패턴으로 선택

최종적으로,
후보 의미 코드 가운데, 'meaning#1' 가 선택

그림 3. 동사구 패턴 적용 예
Fig. 3 Example of applying verb frame pattern

5.2 의미 벡터의 사용

5.2.1 의미의 정의

본 논문에서는 의미 벡터를 적용하는 관점에서, 명사 어휘가 갖는 의미를 재정의 하였다. 즉, 영어 명사 어휘가 갖는 의미를 기계 번역을 위한 기계 가독형 사전 상에서 같은 의미 코드를 공유하는 대역어들의 집합으로 정의하였다. 여기서, 의미 벡터를 사용하기 위해서 의미를 다시 정의하는 이유는 문장 정렬이 이루어진 병렬 코퍼스를 사용하여 영어 어휘와 한국어 대역어 간의 관계를 이용하기 때문이다. 즉, 이미 앞에서 예를 든 표 2에서 {경주, 경쟁, 선거}, {회장, 위원장}, {삶, 생활, 인생} 등은 각각 하나의 의미 코드를 공유하고, 이러한 유사 대역어들의 집합을 하나의 의미로서 정의한다.

5.2.2 의미 벡터의 구축

의미 벡터는 N-차원의 벡터이며, 각 어휘에 따라 서로 다르다. 의미 벡터의 각 요소는 대역어 선택 모호성을 지니는 어휘와 함께 나타나는 공기 어휘들에 대한 가중치 값이다. 이러한 의미 벡터는 문장간 정렬(sentence alignment)이 이루어진 영한 병렬 코퍼스를 사용하여 만들어졌으며, 사용된 영한 병렬 코퍼스는 뉴스 도메인에 속하는 30만 문장으로 구성되어 있다. 그림 4는 의미 벡터를 구축하는 프로세스를 보인다.

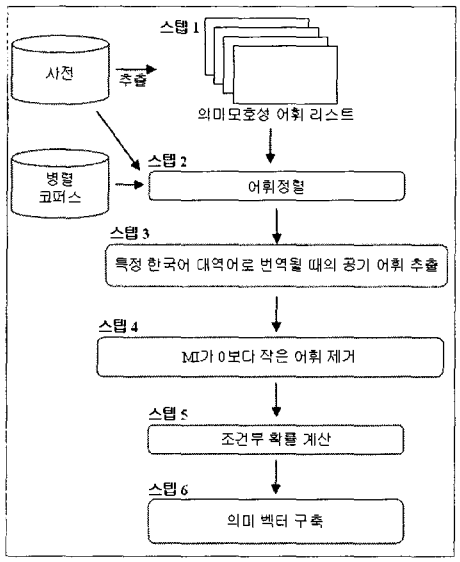

그림 4. 의미 벡터 구축 과정
Fig. 4 Construction flow of sense vectors

영한 병렬 코퍼스를 사용하여 대역어 선택 모호성을 갖는 영어 명사 어휘에 대한 의미 벡터를 구축하는 스텝은 다음과 같다:

- 스텝 1 사전의 각 명사 어휘를 대상으로 하나 이상의 의미 코드를 갖는 명사 어휘들을 추출하여 의미 벡터에 의해 의미가 결정될 목표 어휘로서 추출한다. 이렇게 추출된 명사 어휘들은 의미 벡터를 사용하여 의미 모호성을 해소하는 대상 어휘가 된다.

- 스텝 2 스텝 1에서 얻어진 영어 명사 어휘들에 대해서 영한 병렬 코퍼스에서의 어휘 정렬을 수행한다. 이때 어휘 정렬을 위한 지식으로는 기계 가독형 사전을 사용한다.

- 스텝 3 스텝 2에서의 어휘 정렬이 수행되면, 해당 문장에서 대역어 선택 모호성을 지니는 어휘와 함께 나타나는 공기 어휘들을 추출한다. 즉, 어휘 정렬에 의해 모호성 어휘가 특정 한국어 대역어로 매핑될 때

모호성 어휘와 함께 문장에 나타나는 공기 어휘들이 추출되며, 이렇게 추출된 공기 어휘들은 모호성 어휘가 한국어 대역어로 번역될 때의 클루로서 사용된다.

- 스텝 4 스텝 3에서 추출된 공기 어휘들 가운데, 대역어 선택 모호성을 지니는 어휘와의 상호 정보(MI: Mutual Information)가 0보다 큰 공기 어휘들만이 통계적으로 의미 있는 데이터로 간주하여 필터링을 수행한다.

- 스텝 5 스텝 4에서 추출된 공기 어휘들에 대해서, 대역어 선택 모호성을 지니는 어휘가 해당 공기 어휘와 함께 나타날 때, 특정 대역어로 번역되는 조건부 확률 값을 구한다.

- 스텝 6 의미 벡터는 상기의 과정들로 구해진 공기 어휘들에 대한 조건부 확률 값들로 구성된다.

의미 벡터의 차원은 상호 정보가 0보다 큰 공기 어휘들의 개수로서 정의된다. 두 어휘(대역어 선택 모호성을 지닌 어휘와 그와 공기하는 어휘) 간의 관련성에 대한 척도로서 상호 정보를 사용한다. 상호 정보는 아래의 수식 (4.1)과 같이 정의된다.

$$MI(x,y)=\log\frac{\Pr(x,y)}{\Pr(x)\cdot\Pr(y)} \quad \cdots\cdots (4.1)$$

위의 수식 (4.1)에서 x와 y는 각각 대역어 선택 모호성을 지니는 어휘와 그와 공기하는 어휘를 나타낸다. 또한, 상호 정보가 0보다 큰 경우를 통계적으로 의미있다고 판단하고, 공기 어휘로 사용할 경우에 대한 임계치로서 0보다 큰 경우만을 대상으로 하였다.

다음 수식 (4.2)는 의미 벡터를 정의한다.

$$SV=(w(c_1),w(c_2),w(c_3),...,w(c_n)) \quad \cdots\cdots (4.2)$$

위의 수식에서 w(ck)는 공기하는 어휘 Ck에 대한 가중치 함수이다. 여기서 w(ck)는 다음과 같이 수식 (4.3)에 의해 계산된다.

$$w(c_k)=\Pr(s=s_i|w=c_k) \quad \cdots\cdots (4.3)$$

여기서 si는 대역어 선택 모호성을 지니는 어휘의 i번째 의미(같은 의미 코드를 공유하는 대역어들의 집합)이다. 가중치 함수 w(ck)가 '1'인 경우는 공기 어휘 Ck가 대역어 선택 모호성 어휘와 함께 나타날 때, 특정 의미로 사용될 확률이 '1'이라는 것을 의미한다.

5.2.3 의미 벡터의 적용

테스트 단계에서는 대역어 선택 모호성을 지니는 어휘에 대한 의미 벡터와 동일한 차원의 테스트 벡터가 생성된다. 테스트 벡터의 각 요소의 값은 0 또는 1의 값을 가진다. 텍스트 벡터의 요소값이 0이라는 것은 해당 공기 어휘가 문맥에서 발견되지 않는다는 것을 의미하고, 1은 해당 공기 어휘가 문맥에서 발견된다는 것을 의미한다.

이렇게 입력 문장으로부터 구축된 생성된 테스트 벡터는 대역어 선택 모호성을 지니는 각 의미 벡터와의 유사도가 계산되며, 유사도 계산을 위해 사용되는 수식은 코사인 메져를 사용하며, 수식 (4.4)와 같다. 그림 5는 의미 벡터가 적용되는 예를 보여준다.

$$sim(v,w)=\frac{\sum_{i=1}^{N}v_i w_i}{\sqrt{\sum_{i=1}^{N}v_i^2\sum_{i=1}^{N}w_i^2}} \quad \cdots\cdots (4.4)$$

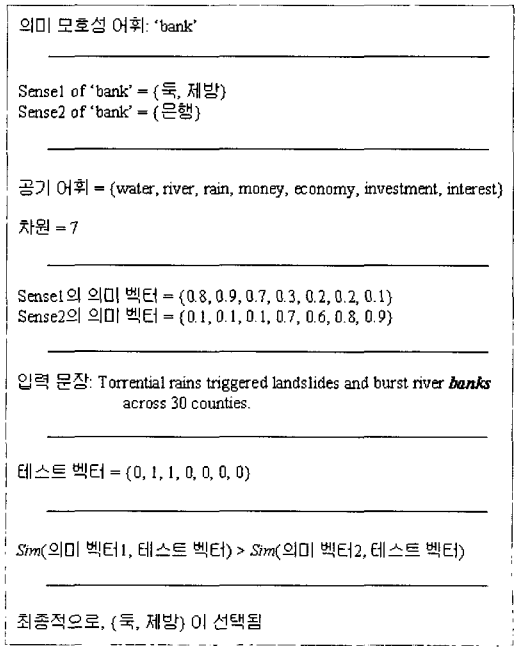

그림 5. 의미 벡터의 적용 예
Fig. 5 Example of applying sense vectors

VI. 제 2단계: 한국어 최적 대역어 결정

6.1 한국어 문맥 통계 정보의 사용

한국어 문맥에 대한 통계적 정보는 약 30만개의 문장을 포함하는 한국어 코퍼스를 사용하여 구축된 N-그램 정보이다. 한국어 문맥 통계 정보는 어떤 영어 어휘가 같은 의미 코드를 공유하는 한국어 대역어를 2개 이상 가지고 있을 때, 이중 최적의 하나의 대역어를 선택하기 위해 사용된다. 아래의 예문은 한국어 코퍼스에서 추출된 한국어 문맥 통계 정보를 사용하여, 최적의 한국어 대역어를 선택하는 것이 필요한 이유를 보여준다. 영어 어휘 'change'의 경우, {변동, 변화, 변경} 과 같은 대역어가 사전에 등록되어 있다. 이러한 세 가지 대역어는 모두 같은 의미 코드를 공유하므로, 이미 앞에서 기술한 방법으로는 이러한 3가지 대역어 가운데 하나를 선택하는 것이 불가능하다. 즉, 이러한 경우에는 목표 언어 모델을 고려하여 최종 대역어를 선택하게 된다.

(E5.1) Interest rate *changes* => (K5.1) 이자율 *변동*

(E5.2) *changes* in your cells => (K5.2) 세포의 *변화*

(E5.3) requests for services *changes* => (K5.3) 서비스 *변경* 요청

본 논문에서는 이와 같이 동일한 의미 코드를 공유하는 의미적으로 매우 유사한 대역어들 가운데 문맥에 맞는 최적의 대역어를 선택하기 위해 한국어 코퍼스로부터 추출된 N-그램 통계 정보를 사용한다. N-그램 통계 정보를 사용하여 최종 한국어 대역어를 선택하는 것은 변환 단계에서 각 어휘들의 의미가 결정된 가장 마지막 단계에서 수행된다. 즉, 변환 단계가 종료된 후, 한국어 생성 단계에서 최종 한국어 대역어가 생성될 한국어 문맥을 고려하여 결정된다. 결국 한국어 문맥 통계 정보는 목표 언어 관점에서 번역문의 자연스러움에 기여한다.

6.2 공기 품사 정보의 사용

공기 품사 정보는 본 논문에서 제안하는 한국어 대역어 선택 기법에서 가장 먼저 적용되는 지식이지만, 의미 모호성 해소 단계를 거치지 않고, 한국어 대역어를 결정한다는 점에서 제 2단계(한국어 대역어 결정 단계)로 분류하였다.

표 3. 영어 명사 'degree'의 의미(WordNet1.71 기반)
Table 3. Senses of 'degree'(based on WordNet)

Sense 1	A position on a scale of intensity or amount or quality
Sense 2	A specific identifiable position in a continuum or series or especially in a process
Sense 3	An award conferred by a college or university signifying that the recipient has satisfactorily completed course of study.
Sense 4	A unit of temperature on a specified scale
Sense 5	A measure for arcs and angles
Sense 6	The highest power of a term or variable
Sense 7	The seriousness of something

앞 절에서 언급했듯이, 동사구 패턴과 의미 벡터는 영어 어휘가 갖는 의미 모호성을 해소하기 위한 가장 기본적인 지식이다. 또한 많은 기존의 연구자들도 공기 정보와 문법 정보가 의미 모호성 해소에 매우 효과적이라고 밝혔다. 하지만, 연구실이 아닌 실제 분야에서는 이러한 단편적인 지식에만 의존해서 의미 모호성을 해소하고 이를 통해 상용화까지 가능한 번역 시스템을 개발하는 데는 한계가 있다. 즉, 사용 가능한 다양한 정보를 사용하여 의미 모호성 해소 및 대역어 선택에 접근할 필요가 있다. 예를 들어, 영어 명사 어휘 'degree'는 7가지 의미를 지닌다. 표 3은 'degree'가 갖는 가능한 의미를 보여준다.

'degree'가 'Sense 4'와 'Sense 5'의 의미로 사용될 때, 그 의미를 결정하는 가장 강력한 단서는 앞에서 언급한 공기 정보와 같은 문맥 데이터나 동사구 패턴과 같은 문법적 정보가 아니라 바로 그 앞에 위치하는 어휘의 품사 정보이다. 즉, 'degree'는 품사 'NUM(수사)'과 매우 밀접한 관련이 있다. 즉, 'degree'의 앞 위치에 숫자 관련 어휘가 올 경우, 'degree'의 의미는 'Sense 4' 이거나 'Sense 5'일 확률이 매우 높다고 할 수 있다. 결국 이러한 사실은 특정 어휘의 경우, 그와 공기하는 품사

정보도 그 어휘의 대역어를 결정하는데 매우 중요한 역할을 한다는 것을 알 수 있다. 표 4는 이러한 종류의 어휘의 예를 나타낸다. 따라서 일부 어휘에 대해서는 그와 공기하는 품사 정보가 대역어 선택에 매우 효과적이며 실질적인 기계번역 성능을 높이는데 있어서 매우 유용한 지식으로 사용될 수 있다.

표 4. 공기 품사에 의해 대역어가 결정되는 어휘들
Table 4. Ambiguous words with co-occurring POS

어휘	대역어 후보	NUM과 사용될 때의 대역어
degree	{정도, 도, 학위, 등급}	도
feet	{발, 피트}	피트
year	{년, 시대, 시간}	년
day	{낮, 하루, 일, 날, 시대}	일
yard	{마당, 야드, 우리}	야드
generation	{세대, 생성}	세대

VII. 실 험

7.1 실험 방법

총 47,531개의 동사구 패턴이 구축되었으며, 이들 모두가 의미 모호성 해소를 위해 사용되었다. 의미 벡터를 구축하기 위해서, 뉴스 분야의 30만 문장으로 구성된 영한 병렬 코퍼스가 사용되었다. 이 과정에서 공기 어휘를 추출하는 윈도우 사이즈는 한 문장으로 정의하였으며, "밀접한" 관계로 여겨지는 공기 어휘의 측정 단위로서 MI가 '0' 이상인 공기 어휘만을 대상으로 하였다. 한국어 문맥 통계 정보는 앞에서 언급한 병렬 코퍼스의 한국어 부분만을 사용하여 구축되었으며, 이때 윈도우 사이즈는 앞/뒤 3으로 제한하였다. 또한, 가능한 공기 품사에 대해서는 'NUM'으로 제한하였다. 현재는 'NUM' 이외의 품사로 확장하는 것에 대한 연구도 진행중에 있다.

본 논문에서 제안하는 대역어 선택 방식을 평가하기 위하여, 대역어 선택 모호성을 지니는 70개의 어휘를 선별하여 실험을 수행하였다. 실험을 위한 테스트 문장은 단어당 50문장씩 총 3,500문장으로 구성되었다.

7.2 실험 결과

본 논문에서는 제안하는 방법의 성능에 대해서 주관적 평가를 수행하였다. 그 이유는 기계번역 시스템이 갖는 특성과 사용하는 사전 및 의미 코드 등이 서로 다르기 때문에 객관적 평가가 어렵기 때문이다.

표 5는 실험 결과를 보인다. 베이스라인은 기본적으로 가장 자주 사용되는 대역어가 선택되는 경우를 의미한다. 표 5를 보면, 일반적으로 다양한 지식을 사용하는 것이 대역어 선택에 있어서 유리하다는 것을 알 수 있다. 물론, 각 지식들이 어떤 영향을 미쳤는지는 해당 어휘에 따라 다르다.

실험 결과, 의미 벡터가 가장 강력한 대역어 선택 지식으로 생각되지만, 의미 벡터가 제한된 사이즈의 병렬 코퍼스로부터 구축되었기 때문에 데이터 부족 문제(data sparseness problem)가 발생한다. 동사구 패턴은 의미적 관계를 지닌 일종의 문법적 정보이다. 동사구 패턴에 의한 대역어 선택 성능의 개선은 얼마나 많은 패턴이 구축되었으며, 동사구 패턴에서 사용되는 의미 코드의 개수가 어느 정도 적당한가에 따라 다르다. 동사구 패턴 역시 데이터 부족 문제가 발생하지만, 일단 입력 문장과 매칭되어 적용될 경우, 그 정확률은 매우 높다. 한국어 문맥 통계 정보는 의미 벡터나 동사구 패턴에 의해 의미가 결정된 후에, 해당 의미를 갖는 한국어 대역어가 2개 이상인 경우에만 최종 한국어 대역어를 선택하기 위해서 사용된다. 즉, 한국어 문맥 통계 정보는 목표 언어 모델에 기반하여 최종 한국어 대역어를 선택하도록 해준다. 공기 품사는 매우 적은 수의 어휘에 적용될 수 있었지만, 그 성능은 매우 높았다.

표 5. 대역어 선택 결과
Table 5: Results of target word selection

사용된 지식	평균 정확률
베이스라인	61.09 %
의미 벡터	66.36 %
의미 벡터 + 동사구 패턴	69.93 %
의미 벡터 + 동사구 패턴 + 한국어 문맥 통계 정보	71.15 %
의미 벡터 + 동사구 패턴 + 한국어 문맥 통계 정보 + 공기 품사	72.84 %

표 6은 각 지식이 서로 다른 영어 어휘에 대해서 대역어 선택 성능에 어떠한 영향을 주는가에 대한 실험 결과를 나타낸다. 'account'의 경우에, 의미 벡터에 의한 성능 개선은 그렇게 높지 않다. 그 이유는 'account'와 같은 어휘는 다양한 한국어 대역어를 가지고 있으며, 이들 대역어 역시 서로 다른 의미 코드를 지니고 있고,. 이렇게 서로 다른 의미로 사용될 때, 실제로 그것들을 구별해줄만한 식별력 있는 공기 어휘들이 없기 때문이다. 이러한 경우에는 오히려 동사구 패턴이 보다 좋은 지식으로 사용될 수 있다. 또한, 공기 품사의 경우도 'point'와 같은 어휘에 대해서 그 대역어를 결정하는데 매우 유용한 지식임을 알 수 있다.

그림 6은 각각의 대역어 선택을 위한 지식들에 의한 성능 개선의 차이를 나타낸다. 그림 6에서 (a), (b), (c), (d)는 각각 의미 벡터, 동사구 패턴, 한국어 문맥 통계 정보, 공기 품사에 의한 대역어 선택 성능의 개선 정도를 나타낸다. 재현률과 정확률의 관점에서 의미 벡터가 가장 강력한 대역어 선택을 위한 지식임을 알 수 있다. 하지만, 동사구 패턴의 경우도, 지속적인 구축 작업에 의해 대역어 선택에 있어서 많은 기여를 할 수 있을 것으로 기대한다.

표 6. 각 지식에 의한 성능 변화
Table 6. Change of performance by each knowledge

영어 어휘	베이스 라인	(a)	(b)	(c)	(d)
account	60 %	60 %	68 %	68 %	68 %
bank	60 %	76 %	78 %	78 %	78 %
operation	58 %	76 %	88 %	88 %	88 %
point	54 %	60 %	60 %	60 %	82 %

(a): 의미 벡터
(b): 의미 벡터 + 동사구 패턴
(c): 의미 벡터 + 동사구 패턴 + 한국어 문맥 통계정보
(d): 의미 벡터 + 동사구 패턴 + 한국어 문맥 통계정보 + 공기 품사

VIII. 결 론

본 논문에서는 영한 기계번역에서 영어 명사 어휘에 대해서 대역어 선택 모호성을 해소하고 문맥에 맞는 한국어 대역어를 선택하는 방안을 제안하였다. 다양한 지식들이 서로 다른 지식 소스로부터 추출되어 실험되었다. 또한 대역어 선택 모호성을 해소하기 위해 어느 한 종류의 지식만을 사용하는 것보다 나은 성능을 보였다. 향후 연구 과제로 보완이 필요한 부분이 도출되었으며, 다음과 같다.

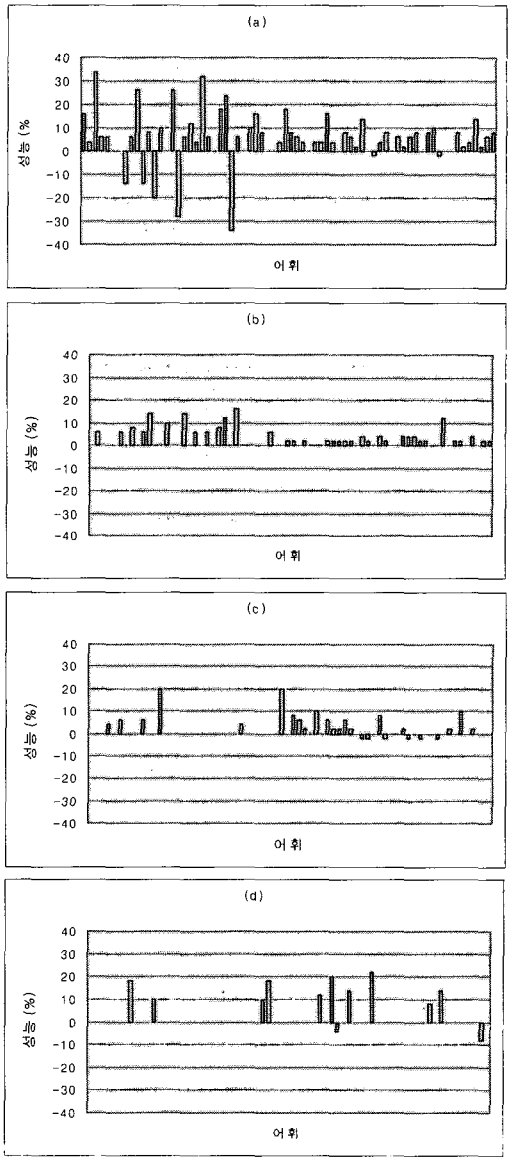

그림 6. 각 지식에 의한 성능 개선 차이
Fig. 6 The difference of improvement by each knowledge

- 최적의 의미 코드 개수에 대한 문제가 제기되었다. 만약 의미 코드의 수가 너무 많으면 커버리지에 문제가 발생하며, 또한 너무 작은 경우에는 식별력에 문제가 발생한다. 이를 해결하기 위해 많은 시행 착오가 필요하다고 생각된다.

- 의미 벡터 적용시 발생하는 데이터 부족 문제를 효과적으로 대처하는 방안이 필요하다. 현재 문장 단위의 정렬이 이루어지지 않은 대규모의 비교 가능 코퍼스(comparable corpus)로부터 의미 벡터를 구축하는 연구가 진행 중이다 [13].

- 동사구 패턴의 커버리지를 향상시키기 위해 보다 많은 동사구 패턴의 구축이 필요하다. 이를 위해 동사구 패턴의 반자동 확장에 대한 연구도 필요하다 [14].

- 다양한 지식들이 함께 사용될 때, 이러한 각각의 지식들에 대해서 서로 다른 가중치를 주는 방법도 대역어 선택 모호성 해소의 성능을 향상시키는데 도움을 줄 수 있다.

- 대역어 선택 모호성을 해소하는데 도움을 줄 수 있는 또 다른 종류의 지식들을 찾아내는 것도 중요하다.

참고문헌

[1] Eneko Agirre and David Martinez, "Knowledge Sources for Word Sense Disambiguation," *TSD*, 2001.

[2] Susan W. McRoy, "Using Multiple Knowledge Sources for Word Sense Discrimination," *Computational Linguistics*, 1992.

[3] Hirst, G, *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press. 1987.

[4] Hwee Tou Ng and Hian Beng Lee, "Integrating Multiple Knowledge Soruces to Disambiguate Word Sense: An Exemplar-Based Approach," *Proceedings of the ACL*, 1996.

[5] Yarowsky, David, "Word Sense Disambiguation Using Statistical Models of Roget's categories trained on large corpora," *COLING*, 1992.

[6] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "Word Sense Disambiguation using statistical methods," *ACL*, 1991.

[7] Yarowsky, David, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *ACL*, 1995.

[8] Rada Mihalcea, "Bootstrapping Large Sense Tagged Corpora," *Proceedings of the 3rd International Conference on Languages Resources and Evaluations LREC 2002*, 2002.

[9] Kiyoaki Shirai and Tsunekazu Yagi, "Learning a Robust Word Sense Disambiguation Model using Hypernyms in Definition Sentences," *COLING 2004*, 2004.

[10] Masaki Murate, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma and Hitoshi Isahara, "Japanese word sense disambiguation using the simple bayes and support vector machine methods," *Proceedings of the SENSEEVAL-2*, 2001.

[11] Hiroya Takamura, Hiroyasu Yamada, Taku Kudoh, Kaoru Yamamoto and Yuji Matsumoto, "Ensembling based on feature space restructuring with application to WSD," *NLPRS 2001*, 2001.

[12] http://wordnet.princeton.edu/

[13] Kumiko TANAKA and Hideya IWASAKI, "Extraction of Lexical Translations from Non-aligned Corpora," *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

[14] Hong et al., "Semi-Automatic Construction of Korean-Chinese Verb Patterns," *COLING 2004 Workshop on Multilingual Linguistic Resources*, 2004.

저 자 소 개

이 기 영

2000년 2월 : 한양대학교 컴퓨터공학과 박사 수료

현재: 한양대학교 공학기술연구소 연구원

김 한 우

한양대학교 컴퓨터공학과 교수