한국산업정보학회논문지 제5권 제2호 (2000.6)

영한번역 시스템에서 연어 사용에 의한 실용적인 대역어 선택

(Practical Target Word Selection Using Collocation in English to Korean Machine Translation)

김 성 묵*
(Seong-Mook Kim)

요 약 기계번역시스템에서 번역의 우수성은 중의성이 심한 동사의 대역어 선택에 좌우된다. 동사의 의미 분별은 함께 어울려 사용되는 연어들에 의해 해소될 수 있지만, 이러한 연어들을 획득하기에는 많은 어려움과 비용의 문제가 발생한다. 이에 따라 기존의 많은 연구 중에서 실용성을 검토해 볼 필요가 있다. 본 논문에서는 영한번역시스템의 성능 향상을 위해 기존에 획득된 연어에 최소한의 명사 의미자질을 구축하여 계산한 의미거리(Semantic Distance)에 의한 실용적인 대역어 선택 방법을 기술하고자 한다.

Abstract The quality of English to Korean Machine Translation depends on how well it deals with target word selection of verbs containing enormous ambiguity. Verb sense disambiguation can be done by using collocation, but the construction of verb collocations costs a lot of efforts and expenses. So, existing methods should be examined in the pratical view points. This paper describes the practical method of target word selection using existing collocation and semantic distance computed from minimum semantic features of nouns.

1. 서 론

문장을 이루는 핵심 요소는 동사이고 이러한 동사는 여느 품사보다 다양한 의미를 지니고 있으며 번역에서는 그 대역어가 다양하게 나타나는 특징을 보여 준다.[1] 90년대 초에 개발되어 나온 초기 번역 시스템들은 단어의 일대일 번역을 하여 이러한 동사의 대역어 선택 문제를 전혀 고려하지 못했다. 그 후 90년대 중반 이후 번역시스템에서 의미의 중의성이 있는 동사의 대역어를 잘 선택하는 것이 번역 시스템의 성능에 직결되고 있음을 인식한 후, 이에 대한 연구가 활발히 진행되고 있다. 동사의 대역어 선택은 다양한 언어 지식을 요구하는데 그 중 구조적인 구문 지식과 서로 어울려 사용되는 연어들의 지식이다. 즉, 아래의 (1)과 같이 동사의 대역어는 구조 상 목적어를 가지고 있는지 없는지, 그 주어나 목적어가 어떠한 명사인지에 따라 결정이 된다.

(1) a. Birds fly. - 날다
b. Time flies. - 흐르다
c. Children is flying kites. - 날리다
d. He flew the Stars and Stripes - 게양하다

이러한 지식들을 획득할 수 있는 자원은 기존에 존재하는 사전을 생각할 수 있다. 사전에는 언어처리에 필요한 어휘의 형태 정보, 구문정보, 의미정보 등이 망라되어 있고 다양한 대역어와 연어에 대한 정보가 잘 표현되고 있으므로 기계번역 시스템을 시작하려는 단계에서는 쉽게 이용할 수 있는 장점이 있다. 90년대 이후 컴퓨터의 급속한 발전에 따라 대량의 자료를 처리할 수 있게 됨에 따라 근래에 발달한 대량의 코퍼스를 이용한 방법도 가능하다. 인터넷과 전자 문서의 발달 등으로 코퍼스는 손쉽게 구할 수 있게 되었으며 현재의 언어 생활에 밀접하고 빈도를 갖는 유용한 정보를 제공한다. 따라서 코퍼스를 통계적으로 처리

* 한국아이비엠 소프트웨어 연구소

하여 대역어 선택에 필요한 연어정보를 얻어내려는 연구 노력도 진행되고 있다. 그렇지만, 이러한 연구 방법들이 실용적인 측면에서 검토되어야 할 필요가 있다. 실제 언어에 적용되기 위해서는 광범위한 무작위 언어를 대상으로 삼아야 하므로 실험적인 수준을 넘어 실용적인 방법을 전제로 해야 하고 개발시의 경제적 비용 문제를 고려해야 한다. 비용은 시간과 노력으로 도출되고 그에 따른 효과와도 상관이 있으므로 많은 방법론 중에서 적절한 연구 방법을 찾는 것이 중요한 일이다. 본 논문에서는 동사의 대역어 선택에 중요한 역할을 하는 연어에 대한 지식을 얻어 내는 기존의 방법론을 고찰해 봄으로써 실용적이고 경제적인 관점에서 동사의 대역어 선택 방법을 제시하려고 한다.

2. 기존 방법에 대한 고찰

지금까지 연구되고 개발된 동사의 대역어 선택 문제를 타동사와 목적어인 연어의 예로 국한하여 간략히 핵심만 고찰해 봄으로써 그 장단점을 파악하고 개발 비용의 측면에서 가장 효율적인 방법론을 도출하고자 한다.

2.1 연어리스트에 의한 대역어 선택

기존의 많은 시스템들이 일대일 단어 번역을 극복하기 위해서 연어를 사용하여 동사의 대역어를 결정하였다. 연어는 두 언어 사이의 표현 차이를 극복하기 위해 도입된 것으로 다음과 같이 쉽게 연어를 사전에 나열하여 동사의 대역어 선택을 할 수 있다.[5]

(2) give
vt(주다)
* 전하다 (information knowledge)
* 베풀다(favor kindness)
* 열다(party)

위의 예에서 각 연어는 괄호에 리스트 형태로 표현되고 있다. 즉, give의 목적어로 information과 knowledge를 가지면 대역어가 '전하다'로, kindness를 가지면 '베풀다'로, party를 가지면 '열다'로 선택된다. 연어 리스트의 정보로 대역어를 선택하지 못했을 때는 디폴트로 '주다'로 선택된다. 이 방식에는 이러한 구조정보 및 연어 정보를 일반사전 및 관용어 사전이나 능숙한 영어 사용자의 경험을 이용하여 수작업으로 구축하게 된다. 어느 정도의 연어를 모으는 데에는 쉽게 가능하지만 모든 연어를 얻어내는 것은 언어의 무한성을 생각할 때 가능하지 않으며 근본적으로 지식획득(knowledge acquisition)의 병목현상(bottleneck)이 발생한다. 비용 측면에서도 초기에는 저비용으로 시작할 수 있지만 오랜 시간을 작업하는 가운데 고비용이 발생하게 된다. 시스템 측면에서는 연어 리스트의 증가에 따라 사전이 커지면 관리하기가 어려워지고 내용의 일관성이 떨어지고 중복 오류 등이 나타나는 등 시스템이 저하될 문제가 발생한다. 또, 적용할 도메인이 바뀜에 따라 대응력도 한계가 있으며 연어 리스트에 나타나지 않은 경우 전부 디폴트의 대역어를 선택하게 된다는 단점이 있다.

2.2 의미자질, 사전정의나 개념을 이용한 지식 기반형 대역어 선택

의미자질은 많은 구체적인 어휘를 추상화한 표현의 속성이다. 의미자질은 존재론적 공리 체계로 언어를 포함한 모든 현상을 인식할 수 있는 기저의 역할을 한다. 이런 이론적 체계에 의한 추상화된 의미자질을 설정하여 대역어 선택을 하게 되면 연어의 나열에서 오는 지식획득의 병목현상을 줄일 수 있는 가능성이 된다. 예를 들어, 의미자질에 의한 동사의 의미 분별을 통해 대역어를 선정하는 방법을 살펴 보자. 영어의 take를 살펴보면 다음과 같다.

(3) take 타다 [+운송수단]
잡다 [+물건]
데리고 가다 [+사람]
맡다 [+책임]
포획하다 [+동물]

take의 대역어는 목적어로 나오는 명사의 의미자질이 어느 것을 포함하느냐에 따라 결정된다. 이에 따라, 명사 의미 사전을 구축하거나 이미 구축되어진 유의어 사전인 시소러스나 동의어 집합으로 표현된 워드넷[7]을 이용하여 대역어 선택 기준을 작성해야 한다. 의미자질만으로 대역어 선택이 가능하지만 각종 명사에 해당 의미자질을 부여하는 일이 필요하게 된다. 이것은 구체어인 연어를 모으는 일만큼 어려운 일에 속하게 된다. 실제 무한 문장을 다루다 보면 많은 의미자질을 수작업에 의해 기입해야 하며 이러한 일은 전문 지식가를 통해서 가능한 일이다. 또한, 의미자질은 서로 세분화되다 보면 일관성이 결여되고 그 체계가 복잡하게 되어 관리가 어려워진다.

각 동사의 의미는 사전에 정의 되어 있는 사전의 정의와 예문이 보여주는 지식을 이용하여 대역어 선택을 하려는 입장이다. 예를 들어, '타다'의 의미를 지닌 take를 사전에 나오는 정의와 예문을 살펴보면 다음과 같이 나타난다.

(4) take : To use (something) as a means of conveyance or transportation.
Will you take a bus or taxi ? [11]
If you take a car, train, bus or plane, you use it to go from one place to another. She took the train to New York. [8]

사전의 정의나 예문에 사용된 명사를 부류개념이나 의미표지어로 간주하고 주어진 명사가 이 단어의 유사성을 통해 대역어 선택을 이루는 방법이다. 예를 들어, 예문에 나오는 car, tarin, bus, plane, taxi 이라는 어휘를 의미표지어로 삼아 주어진 예문에 나온 어휘가 이에 속하는지 비교하고 아니면 해당 어휘의 의미가 정의에 주어진 conveyance나 transportation과 같은 부류 개념과 유사성을 찾아가는 것이다.[2][10] 즉, 사전에 정의나 예제로 나와 있는 문장을 구문분석을 통해 관련 지식을 자동으로 획득할 수 있다. 이 방법은 오랜시간에 걸쳐 농축된 조직화된 지식을 이용하고 있는 장점이 있다. 그러나 사전에 사용된 어휘가 너무 제한적이고 너무 짧은 경우가 많아 무제한의 어휘를 다루는 실제 문장에 적용하기는 한계를 드러낸다.

2.3 코퍼스로부터 자동 습득된 통계 정보 이용

코퍼스를 이용하면 대량의 수집이 가능하므로 가장 빈번히 쓰이는 언어 지식을 추출할 수 있고 다양한 자료로부터 지식이 획득되므로 수작업에서 오는 지식획득 병목현상을 피할 수 있으며 자료의 검증 없이 직관적이고 이론적으로 생기는 문제를 극복할 수가 있어 현재 많은 연구가 이루어지고 있다.[3][6] 그러나 코퍼스는 양적인 문제에서는 만족을 할 수 있으나 질적인 문제에서 여전히 문제점을 가지고 있다. 양질의 코퍼스를 구하기는 쉽지 않으며 원시 코퍼스를 검토하여 깨끗한 코퍼스로 만드는 일이 비용을 발생시킨다. 특히, 단일언어의 코퍼스는 쉽게 구해지지만 양언어 코퍼스는 상대적으로 구하기가 어렵다. 번역에서의 대역어의 선택 문제는 단일 코퍼스로부터 얻는 지식으로는 한계가 있다. wear의 예를 살펴 보면, 하나의 의미에도 다음과 같이 다양한 대역어가 존재함을 알 수 있다.

(5) wear 1. To carry or have on the person as covering, adornment, or protection: wearing a jacket; must wear a seat belt.
1. When you wear clothes, shoes, or jewelry, you have them on your body. He was wearing a brown uniform.

(6) 의미 wear1의 대역어

wear(OBJ shoes boots socks stocking slipper) -> 신다
wear(OBJ glove mitten gauntlet ring) -> 끼다
waer(OBJ belt band sash) ->매다
waer(OBJ charm) ->지니다
wear(OBJ veil mask glasses hat cap bonnet sunglasses monocle) -> 쓰다
wear(OBJ watch bracelet) -> 차다
wear(OBJ badge) -> 달다
wear(OBJ cross necklace rivier choker collar) -> 걸다

하나의 의미에도 대역어가 달라지는 것을 단일 코퍼스에서는 직접 파악하기 어려우므로 [4]에서는 새로운 메타 의미를 설정하고 대역의 메타 어휘를 얻어서 연어 관계를 이용하여 대역어를 선택하는 방법을 제시하고 있다. 결국 사람의 개입을 요구하므로 쉽게 개발 비용의 측면에서장점이 되기 어렵다. 또한 자주 사용하지 않으나 매우 중요한 정보를 찾기가 어려운 자료 부족 문제를 야기하므로 완전한 해결이 아니다.

3. 실용적인 대역어 선택 방법

3.1 코퍼스에 나타난 연어의 특징

일상 언어 생활에 나타나는 코퍼스에서 타동사의 목적어로 등장하는 연어의 특징을 살펴 보면 다음과 같다. 동사의 의미는 연어에 따라 설정이 된다. 즉, 동사의 목적어로 오는 단어들의 의미적 분류에 따라 동사의 의미가 세분화된다. 동사의 의미가 동일한 연어는 의미적으로 응집된 의미장(semantic field)을 이루고 있다.

그런데, 대역어는 이런 동사의 세분된 의미와는 다르게 다양하게 나타난다. 같은 유사어나 동의적 관계를 보이는 연어에도 대역어가 다르게 나타난다. 동사의 의미와 대역어와의 관계를 표현하면 다음과 같다.

(7)	동사	의미1	대역어1
			대역어2
			대역어3
		의미2	대역어4
			대역어5

또한, 의미적으로 연결이 안되는 연어들도 대역어가 동일하게 나타나는 경우가 존재한다. 동사의 의미가 세분화되어도 대역어가 같은 경우에는 연어들이 의미적으로 연결되지 않는다. .

(8)	동사	의미1	대역어1
		의미2	대역어1

즉, 대역어는 연어들의 의미적 분류에 일관되게 선택되는 것이 아니므로 사람의 개입 없이 코퍼스만으로부터 자동으로 지식을 얻기가 어려움을 보여주고 있다. 이외에도 실제 코퍼스에서 연어에 대한 지식 획득이 어려운 문법적인 현상은 다음과 같다.

(9) a. 수동구문으로 나타난다
Blue is much worn at present.
b. 양화사의 피수식어로 나타난다.
Do not take a lot of this medicine.
c. 대명사화
Do you have Aspirin? I want to take some.
d. 준동사에 의한 수식
Did you decide the plane to take to New York?
e. 관계문의 선행사로
The diamond ring which she wore in her finger

이상의 코퍼스에 나타난 연어들과 대역어의 관계를 살펴 볼 때, 이러한 지식들은 코퍼스에서 자동으로 획득되기에는 난점이 많이 있고 전문 지식과 판단력을 가진 사람의 개입을 요한다.

3.2 사전을 이용한 의미 예문 수집

초기 비용을 많이 들이지 않고 단기간 내에 연어 리스트를 모으고 동사 대역어를 선택하는 방법은 역시 사전의 지식을 이용하는 방법이다. 사전은 일반적으로 정의 문장에 사용된 어휘가 제한적이라는 단점이 있지만 코빌드 사전은 사전 편집자의 직관 이외에 대량의 코퍼스를 이용하고 있다. 코빌드 사전에는 실제 생활에서 자주 사용되는 문장을 중심으로 구성되고 동사의 의미가 연어를 중심으로 나타나 있다. 또한, 대역어 사전에 주어진 예문을 통해 하나의 의미에 대응하는 여러 대역어를 수집할 수 있다. 연어 수집에 일차적으로 고려하는 것은 대역어 수집이다. 대역어 선택은 영어 어휘의 의미 분별 이상의 문제이며 동사의 대역어는 연어의 대역어에 연결성을 지니고 있다. 대역어 중심으로 연어를 설정하는 방법으로 사전 작업자들은 사전에 사용된 예문과 대응하는 번역문을 입력하고 함께 쓰이는 연어들을 나열하는 참여하는 사전 작업자는 일반 영어지식을 갖고 있거나 생각할 수 있다. 또한 코퍼스를 통해 얻어진 문장들을 이용해 순서나 빈도를 결정할 수 있다. 이렇게 작성된 동사의 대역어 수는 통상 동사의 의미수의 2.5배에 달하는 36000개가 작성되었다. 따라서, 수집된 연어리스트와 주어진 동사의 목적어를 단순 문자열 검사를 통해 동사의 대역어를 쉽게 선택할 수 있게 되었다.

3.3 의미거리(semantic distance)에 의한 대역어 선택

연어 리스트를 모두 다 수집하는 것은 계속 반복되는 일인 만큼 수집된 연어 리스트만으로 리스트에 주어지지 않은 경우의 대역어 선택 문제를 해결할 대안이 필요하다. 기본적인 생각은 자연언어의 특성인 어휘의미의 추상성을 이용하여 연어의 단순 문자열 검사대신 목적어와 연어리스트와의 의미거리를 통해 대역어를 선택하는 것이다. 그렇게 하기 위해서는 명사 사전에 의미자질을 추가하는 작업이 필요하고 의미자질을 설정하는 기준이 필요하다. 일반 명사의 어휘 수는 대략 5만여개에 이르므로 단시간에 사전 작업자가 쉽게 작업하기 위해서는 복잡하지 않고 소수의 의미자질을 준비하는 것이 비용절감에 절대적이다. 한편으로는 의미자질이 의미 응집성을 보여 주어 대역어 선택에 변별이 가능한 정도의 의미자질 설정이 필요하다. 이상의 원칙을 준수하여 다음의 의미자질을 설정하였다. 의미 자질에는 대분류로 생물, 부분, 자연, 인공, 추상, 시간, 단위, 행동, 속성, 장소, 사건의 11개가 사용되고 각각의 세분화에 의해 69개의 의미자질을 개발했다. 이 방법의 장점은 변별력이 높은 적은 수의 의미 자질만을 이용하므로 사전의 확장시에도 용이하다는 점이다. 예를 들어 bus는 다음과 같은 의미 자질을 갖고 있다.

(10) bus : #인공물 #차량

명사 사전에 의미자질이 준비되었을 때 연어의 단순 문자열 검사가 실패한 경우 대역어를 선택하는 방법을 살펴보면 다음과 같다. 연어 리스트에 나타나 있지 않은 목적어가 주어지면 바로 디폴트 의미를 갖지 않고 연어리스트의 전체 의미에 어느 정도 의미적으로 관련성이 있는지 의미거리를 계산하여 관련된 연어리스트를 구해서 대역어를 선택한다. 즉, 연어리스트의 의미밀도(sense densty) 벡터와 주어진 단어의 의미 벡터를 계산하고 그 내적을 통한 거리를 수치화하는 방법이다. 의미거리는 대역어를 선택해야 할 주어진 동사의 목적어인 명사가 가지고 있는 의미자질의 벡터를 w라 하고, 동사의 연어 리스트의 i번째 어휘의 벡터를 ci라 하면 두 의미 벡터간의 내적인 의미거리 di

가 다음과 같이 구해진다.

(11)

$$d^i = s_U^i \cdot w_U^i$$

$$s_U = s/\|s\| \quad s = \sum_i c_U^i \quad w_U = w/\|w\| \quad c_U^i = c^i/\|c^i\|$$

연어리스트가 여러 개가 있으므로 이중 가장 의미거리가 최대인 것으로 결정된다. 연어리스트의 의미밀도는 연어리스트에 나열된 명사들의 각각의 의미벡터의 총합의 평균 값을 의미한다. 이런 계산 결과로 나온 의미거리를 통해 실험을 해 보면 의미자질의 변별력이 상대적으로 약하고 분포가 밀접되어 있으면 원하지 않은 대역어를 선택하게 된다. 따라서 예제 문장의 튜닝과정을 통해 유의미한 임계치를 설정하여 임계치 이하의 의미거리는 무시하고 의미거리의 허용 최대치와 최소치를 결정하는 휴리스틱한 조절이 필요로 한다.

(12)

$$d_s^i = \begin{cases} \text{if } d^i < \theta_d, 0 \\ \text{otherwise}, d_s^{\min} + \left(\left(d_s^{\max} - d_x^{\min} \right) \cdot \frac{d^i - \theta_d}{1 - \theta_d} \right) \end{cases}$$

4. 실험 및 남은 문제

이상에서 구축된 연어리스트와 의미거리를 통한 대역어 선택을 500 문장에 대해 실험을 하였다. 실험 결과 기본 연어 리스트 만으로 대역어가 선택된 경우는 63%인 316문장이었으며 의미거리를 통해 대역어가 선택된 경우가 31%인 155문장이었으며 그 밖의 경우는 디폴트 값으로 선택되어 졌다. 의미거리를 통해 대역어가 선택되어진 경우의 95%가 정확도를 보여 주고 있었으며 디폴트 값으로 선택되어진 경우의 75%는 대역어가 잘못 선택된 것이었다. 이것은 임계치의 조정과 사전의 의미자질의 미비로 인한 것이 원인이었다. 보다 정확도를 높이기 위해서는 많은 양의 코퍼스를 이용한 튜닝을 통해 가중치의 조절과 의미자질의 개선이 필요하다. 또한 복합 명사로 쓰이는 경우 자동으로 의미자질 할당을 부여하는 작업이 필요하다.

5. 결 론

본 논문은 영한번역시스템의 성능 향상을 위해 사전에서 수작업으로 획득된 연어에 최소한의 명사 의미자질을 구축하여 의미자질의 벡터를 이용한 의미거리에 의한 실용적인 대역어 선택 방법을 제시하였다. 연어리스트는 정의가 실제 생활에서 사용된 문장으로 이루어진 코빌드 사전과 대역어 사전의 예문을 이용하여 3만개의 연어를 작성하였고 69개의 의미자질을 설정하여 주어진 어휘의 의미자질 벡터 값과 연어리스트의 의미밀도벡터 값의 내적인 의미거리를 계산하는 방법으로 대역어를 선택함으로써 기존의 디폴트 값으로만 대역어가 선택되어지는 경우의 80%의 대역어 선택이 가능하게 되었다.

참 고 문 헌

[1] 김성묵. 1993. 기계번역에서의 다의어 동사 처리 연구. 제27회 어학연구. 서울대학교

[2] 김성묵. 1998. 독일어 명사합성어의 기계분석. 서울대학교

[3] 이승우. 1999. 국소 문맥과 공기정보를 이용한 비교사 방식의 명사의미 중의성 해소. 포 항공과대학교

[4] 이현아. 1997. 영-한 변환 시스템에서 한국어 구문 공기 정보를 이용한 역어 선택. 한국 과학기술원.

[5] 이호석, 김영택. 1993. 영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전. 한국정 보과학회 논문지 Vol. 20. No.07

[6] 조정미. 1998. 코퍼스와 사전을 이용한 동사 의미 분별. 한국과학기술원.

[7] Christiane Fellbaum. 1998. WordNet. The MIT Press.

[8] HarperCollins Publisher. 1997 COLLINS COBUILD new Student's Dictionary

[9] Kim Seong Mook. 1999. IBM English to Korean Lexical DB Guide(MS)

[10] Karen Jensen, George Heidorn and Stephen Richrdson. 1993. Natural Language Processing: The PLNLP approach. Kluwer Academic Publishers.

[11] Microsoft Corporation. 1997. Microsoft Bookshelf 2.0. Software.

김 성 묵

1989 - 현재 : 한국아이비엠 소프트웨어 연구소 연구원
1998 : 서울대학교 독어독문학과 문학박사
1988 : 서울대학교 독어독문학과 문학석사