독일언어문학 제70집(2015.12). 1-21

한-독 기계번역을 위한 한국어 영형목적어 처리 연구*

박아름 · 홍문표**(성균관대)

Ⅰ. 들어가는 말

한국어는 소위 Pro-삭제 언어 Pro-Drop Sprache로서 특정 대명사가 생략될 수 있는데, 이러한 대명사를 종종 영형대명사 Null-Pronomen라고 부른다. 이러한 종류의 대명사는 일본어, 스페인어 또는 이탈리아어에도 등장한다. 스페인어나 이탈리아어에서는 주어 위치에서 영형대명사가 나타나는 반면, 한국어에서는 주어 위치뿐만 아니라 목적어 위치에서도 영형대명사가 등장한다. 한국어에서는 영형주어 Null-Subjekt가 가장 빈번하게 나타나는 조응 표현이다. 홍민표(2000)에서는 한국어 대화체 텍스트에 등장하는 대명사 중 영형주어의 비율이 약 57%라고 보고했다.

영형목적어 Null-Objekt는 한국어 대화체 텍스트에서 두 번째로 빈번하게 등장하는 영형대명사이다. 홍민표(2000)에 따르면 영형목적어는 약 12%의 비율로 등장한다. 영형목적어가 자주 사용됨에도 불구하고 기존 연구에서는 대부분 영형주어에 초점을 맞추어 연구를 진행해 왔다. Ryu(2001)은 한국어 문어체 텍스트와 대화체 텍스트에 등장하는 영형대명사의 비율을 비교하여, 한국어 문어체 텍스트에서는 영형대명사가 거의 나타나지 않는다고 보고했다. 따라서 본 연구에서는 한국어 대화체 텍스트에 나타나는 영형목적어를 다룰 것이며, 영형목적어 해소를 위한 언어학적인 단서들을 찾아낼 것이다.

기계번역 maschinelle Übersetzung의 관점에서 한국어 영형목적어의 해소는 매우 중요한 문제 중 하나인데, 특히 한국어를 출발언어, 독일어를 목표언어로 하는 기계번역 시스템에서는 한국어의 영형목적어가 반드시 복원되어야만 한다.

---

* 본 연구는 지식경제부의 지식경제 기술혁신사업의 일환(10041807)으로 수행되었습니다.

** 교신저자

왜냐하면 한국어에서 생략된 목적어가 독일어에서는 반드시 명시적인 목적어로 번역돼야 하기 때문이다. 현재 상용화 된 대부분의 기계번역 시스템은 대용어 해소와 같이 문장 층위를 넘어서는 현상에 대해서는 처리할 수 없다. 이러한 문제점을 설명하기 위해 다음의 예문[1]을 살펴보자.

표 1 : 한국어→독일어 언어쌍에 대한 기계번역 결과 예시

	화자	한국어 원문	독일어 번역 결과 1 (Google)	독일어 번역 결과 2 (Systran)
(1)	A	영수증ø 필요하세요?	Haben Sie eine Quittung benötigen?	Der Empfang ist notwendig?
(2)	B	네, 필요합니다.	Ja, es ist notwendig.	Ja ist notwendig.
(3)	A	그러면 식사 후에 øi 드리도록 하겠습니다.	Dann werde ich Ihnen nach dem Essen.	Dann nach Mahlzeit zwecks geben.

예문 (1)~(3)은 화자 A와 B의 대화인데, A가 발화한 두 번째 문장인 (3)에 목적어가 생략되어 있으며, 이는 'ø'라는 기호로 표시되어 있다. 이 영형목적어의 선행사는 A가 발화한 첫 번째 문장 (1)에 등장하는 '영수증'이다.

이 한국어 예문에 대해 통계기반 기계번역 시스템인 구글 Google 번역기와 규칙기반 기계번역 시스템인 시스트란 Systran 번역기를 사용하여 독일어 번역을 수행하였다. 예문 (3)에 제시된 두 개의 독일어 번역 결과를 살펴보면, 한국어 원문에 목적어가 생략되어 있기 때문에 독일어 번역 결과 또한 목적어가 생성되지 않았다. 이는 앞서 말한 바와 같이 현재 상용화 된 기계번역 시스템의 한계를 보여주는 것으로서, 기계번역 시스템의 종류에 상관없이 독일어 목적어가 생성되지 않아서 번역 결과로 독일어 비문이 나오게 되었다. 따라서 한국어 영형목적어는 한-독 기계번역의 관점에서 반드시 복원되어야 하는 현상이다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 한국어 영형목적어 현상에 대해 더 자세히 살펴볼 것이다. Ⅲ장에서는 영형대명사 해소에 관한 기존 연구들과

1) 본 논문에 제시된 모든 예문과 실험에 사용된 문장은 ETRI(한국전자통신연구원)에서 제공해 준 관광 분야 코퍼스이다.

그들의 한계점에 대해 다룬다. 한국어 영형목적어 해소를 위해 Ⅳ장에서는 본 연구에서 제안한 기계학습 방법을 소개할 것이며, 이를 위해 8개의 기계학습 자질들을 도입할 것이다. 또한 한국어 영형목적어 해소를 위한 기계학습 방법의 효과를 평가할 것이다. Ⅴ장에서는 한국어 영형목적어를 복원하기 전과 복원한 후의 독일어 번역률을 평가하여 비교할 것이다. 마지막으로 Ⅵ장에서는 연구결과를 정리하고 향후 연구방향을 제시할 것이다.

Ⅱ. 한국어 영형목적어 현상

한국어에 등장하는 영형목적어를 복원하기 위해 센터링 이론이 사용될 수 있다. 센터링 이론에서는 담화 층위에서 현가성이 높은 요소를 초점 Fokus (vgl. Sidner 1979) 또는 센터 center (vgl. Grosz et al. 1983)라고 부른다. 한 문장과 관련하여 가장 중심이 되는 개체가 (영형) 대명사로 표현되는 경향이 있다는 특성을 활용하여 (영형) 대명사의 선행사를 확인할 수 있다. 기존 연구들에서도 센터링 이론을 대용어 해소에 적용하고자 하는 시도들이 존재한다 (최재웅/이민행 1999; 홍민표 2000; 홍문표 2001). 센터링 이론에서는 각 발화에 실현된 개체들의 집합을 그 발화의 전향적 센터 forward-looking center라고 부르고, 한국어와 관련된 기존 연구들에서는 이러한 개체들의 현가성 salient에 따라 문법 기능에 따른 전향적 센터 순위 forward-looking center ranking를 제안하였다. 본 연구에서는 홍문표 (2011)의 논의에 따라 한국어의 전향적 센터 순위를 다음과 같다고 가정한다:

• 한국어의 전향적 센터 순위 (홍문표 2011)
주제격 > 주어격 > 목적격 > 부사격 > 기타

앞서 제시된 전향적 센터 순위를 고려해 보면 영형목적어는 담화 내에서 가장 현가성이 높은 개체인 토픽 Topik이라고 이해될 수 있다. 한국어에서 문장의 토픽은 주제 표지 '은', '는'과 초점사 '도', '만'을 통해 나타나기 때문에 문장 내에서 토픽인 개체를 쉽게 판별할 수 있다. 따라서 만약 담화 내의 선행사 후보

중 주제 표지와 초점사를 지니는 후보가 있다면 이것이 영형목적어의 선행사일 확률이 높다. 이와 관련된 예문을 살펴보자.

표 2 : 전향적 센터 순위와 관련된 대화문 예시

화자	한국어 대화문
A	저기 큰 다리$_1$가 육교$_2$인가요?
B	네, 엘리베이터$_3$도 있고요.
A	노인분들$_4$과 장애우$_5$를 위한 거겠군요?
B	네, 일반인$_6$이 ø 타지 않습니다.

표2의 대화문에서 생략된 목적어는 기호 'ø'로 표시되어 있다. 이 예시에서는 총 6개의 선행사 후보가 존재한다: 1.다리 2.육교 3.엘리베이터 4.노인분들 5.장애우 6.일반인. 첫 번째, 여섯 번째 후보는 주어 위치에 등장하며, 두 번째 후보는 보어 위치, 네 번째, 다섯 번째 후보는 목적어 위치, 그리고 세 번째 후보가 초점사 '도'를 지니고 있다. 앞서 살펴보았듯이 토픽이 전향적 센터 순위에서 가장 높은 순위를 차지하고 있기 때문에 세 번째 후보인 '엘리베이터'가 영형목적어의 선행사일 확률이 높으며, 실제로 이 예시에서 '엘리베이터'가 정답 선행사이다. 표2의 예시에서 살펴보았듯이, 선행사 후보의 통사적 기능이 한국어 영형목적어를 해소하는데 중요한 정보이기 때문에, 우리는 이 정보를 한국어 영형목적어를 해소하는데 활용할 것이다.

특성-공유 제약 Property-sharing constraints도 한국어 영형목적어 해소를 위해 활용될 수 있다. Kameyama(1986)은 일본어의 영형대명사를 해소하기 위해 특성-공유 제약을 제안하였다. 이 제약에 따르면 영형대명사가 주어 위치에 나타나면, 선행사도 주어일 확률이 높으며, 영형대명사가 목적어 위치에 나타나면, 선행사도 목적어일 확률이 높다. 일본어와 한국어는 언어학적 특성을 많이 공유하기 때문에, 이 제약을 한국어 영형목적어를 해소할 때에도 적용할 수 있을 것이다. 표3은 특성-공유 제약을 한국어 영형목적어 해소를 위해서도 사용할 수 있다는 것을 보여주는 예시이다.

표 3 : 특성-공유 제약과 관련된 대화문 예시

화자	한국어 대화문
A	동물들$_1$이 모두 밖$_2$에 나오지 않는군요.
B	날씨$_3$가 추워져서 밖$_4$에 잘 나오지 않습니다.
A	아기 동물들$_5$을 보러 왔는데 잘 보이지가 않네요.
B	오후$_6$에는 ø 보실 수 있을 겁니다.

이 예시에서 영형목적어에 대해 6개의 선행사 후보가 존재한다: 1.동물들 2.밖 3.날씨 4.밖 5.아기 동물들 6.오후. 선행사 후보들 중에 정답 선행사인 '아기 동물들'보다 전향적 센터 순위가 더 높은 후보들이 있음에도 불구하고, 정답 선행사는 목적어 위치에 있는 '아기 동물들'이다. '아기 동물들'이 목적어 위치에 등장하였고, 이것이 영형목적어의 선행사가 되었으므로, 이는 특성-공유 제약을 보여주는 하나의 예시라고 할 수 있다. 따라서 영형 목적어와 선행사 후보의 통사적 기능이 같은지의 여부 또한 한국어 영형목적어 해소를 위해 사용될 수 있다.

영형목적어의 술어와 선행사 후보의 술어가 의미적인 관련성이 있다는 것 또한 한국어 영형목적어의 한 가지 특징이라고 할 수 있다. 영형목적어와 선행사 후보 간의 술어가 의미적으로 관련성이 있다면 해당 선행사 후보가 영형목적어의 정답 선행사로 선호된다.

표 4 : 술어의 의미적 관련성에 관한 대화문 예시 1

화자	한국어 대화문
A	여보세요, 객실 담당자$_1$ 부탁합니다.
B	네, 말씀하세요.
A	객실$_2$에서 인터넷$_3$을 사용하려면 어떻게 해야 하나요?
B	인터넷 카드$_4$를 구입하셨습니까?
A	아니요, 어디서 ø 구입할 수 있죠?

표4는 영형목적어와 선행사 후보의 술어 사이의 의미를 사용하는 것이 한국어 영형목적어 해소를 위해 중요하다는 것을 보여주는 예시이다. 이 예시에서는 영형목적어에 대해 4개의 선행사 후보가 존재한다: 1.객실 담당자 2.객실 3.인터넷 4.인

터넷 카드. 4개의 선행사 후보 중 두 번째 선행사 후보를 제외한 후보들이 모두 목적어 위치에 등장하는데, 이러한 경우 이들의 통사적 기능이 모두 같기 때문에 이들 사이의 순위를 매길 수는 없다. 이 때 선행사 후보 중 '인터넷 카드'와 영형목적어의 술어가 '구입하다'로 같기 때문에, 이들 술어의 의미적 관계로 인해 '인터넷 카드'가 정답 선행사가 된다. 선행사 후보와 영형목적어의 술어가 동의어 관계에 있지 않더라도, 술어들이 유의어 관계 또는 반의어 관계에 있는 것 또한 한국어 영형목적어의 선행사를 선택하는데 도움이 되는 정보가 될 수 있다.

표 5 : 술어의 의미적 관련성에 관한 대화문 예시 2

화자	한국어 대화문
A	무엇을 도와드릴까요?
B	호텔 전용 리무진$_1$을 이용할 수 있나요?
A	네, 물론입니다.
A	이용 요금$_2$은 다소 비싼데 괜찮으신가요?
B	괜찮습니다.
A	네, 며칠 동안 ø 사용하실 건가요?

표5에서 영형목적어에 대한 선행사 후보는 '호텔 전용 리무진'과 '이용 요금'이다. '이용 요금'에 주제 표지가 부착되어 있기 때문에 '호텔 전용 리무진'보다 전향적 센터 순위가 높음에도 불구하고 '호텔 전용 리무진'이 정답 선행사가 된다. 그 이유를 '호텔 전용 리무진'과 영형목적어의 술어가 각각 '이용하다' 그리고 '사용하다'이고, 이들이 유의어 관계에 있기 때문이라고 추정할 수 있다.

표 6 : 술어의 의미적 관련성에 관한 대화문 예시 3

화자	한국어 대화문
A	죄송합니다, 박물관$_1$에서는 휴대전화$_2$를 꺼 주셔야 합니다.
B	그런가요, 알겠습니다.
A	관람$_3$이 끝난 후에는 ø 키셔도 됩니다.

이 예시에서는 영형목적어에 대해 세 개의 선행사 후보가 존재한다. 이 선행사 후보들 중 두 번째 선행사 후보인 '휴대전화'가 영형목적어의 정답 선행사이다. '휴

대전화'의 술어는 '끄다'이며, 영형목적어의 술어는 '켜다'인데, 이들 술어가 반의어 관계에 있으므로, 이 정보는 선행사 후보들의 통사적 기능보다 더 중요한 역할을 한다. 이것이 바로 우리가 한국어 영형목적어 해소를 위해 선행사 후보와 영형목적어의 술어가 지니는 의미적 관계를 활용하고자 하는 이유이다.

영어의 WordNet과 유사하게 술어들이 지니는 의미에 대한 정보를 제공해주는 한국어 사전이 존재한다. 세종 전자 사전[2]과 KorLex[3]는 술어들 간의 의미 관계에 대한 정보를 추출할 수 있는 한국어 사전의 예시이다. KorLex는 WordNet을 한국어로 번역한 사전이며, 세종 전자 사전은 동의어와 반의어와 같은 단어의 의미 관계에 대한 정보를 포함하고 있기 때문에 이 사전들을 활용하면 선행사 후보와 영형목적어의 술어가 지니는 의미를 자동으로 추출하여 비교할 수 있다.

Ⅲ. 영형대명사 해소에 관한 기존연구

영형목적어와 같은 영형대명사를 해소하기 위한 연구들은 일본어 (vgl. Nakaiwa/Ikehara 1995; Nakaiwa/Shirai 1996) 그리고 스페인어 (vgl. Park/Hong 2014; Palomar et al. 2001; Ferrádez/Peral 2000)와 같은 언어들과 관련하여 진행되어 왔다. 이러한 연구들은 대용어 해소와 관련된 연구에 기반을 두고 있는데, 대용어 해소와 관련된 연구들은 영어를 중심으로 1970년 이후에 수행되어 왔다. 다양한 언어에 나타나는 대용어 현상들은 언어에 상관없이 유사한 전략을 사용하여 해소되어 왔다. 가장 대표적인 방법은 언어학적 정보를 사용하는 것이다. 관련 연구들에서는 제약 Constraint과 선호도 Präferenz를 구분한다 (Baldwin 1997; Lappin/Leass 1994; Carbonell/Brown 1988).

제약은 선행사 후보들 중 특정한 후보들을 제거하는 역할을 하는데, 이는 절대적인 기준으로 간주된다. 선호도는 휴리스틱 규칙의 형태로서 선행사 후보들에 제약을 적용한 후 남아 있는 후보들 사이의 순위를 결정해주는 역할을 한다. Nakaiwa/Shirai(1996)은 일본어의 영형대명사를 해소하기 위해 격, 양상 표현, 동

2) https://ithub.korean.go.kr/
3) http://klpl.re.pusan.ac.kr/

사의 의미 속성 그리고 접속사와 같은 의미 그리고 화용론적인 제약을 제안하였다. 하지만 이들이 제안한 제약은 주로 일본어 영형주어 해소에 초점을 맞추었으며, 문어체 코퍼스를 대상으로 하므로, 이들의 접근법을 본 연구에서 다루는 대화체 문장에 등장하는 영형목적어 해소에 적용하는 것은 어려움이 있다.

센터링 이론(Grosz et al. 1995)은 휴리스틱 규칙을 사용하는 접근법 중 하나인데, 이 이론에서는 화자가 집중하는 대상인 센터 또는 초점이 담화의 국부적 응집력 lokale Kohärenz을 유지하기 위해 결정적인 역할을 한다고 주장한다. 따라서 한 발화 내에서 언급된 특정한 개체들이 다른 개체들보다 더 중요하다고 주장하였으며, 이러한 특성이 대용어의 선행사를 결정하기 위해 적용되어 왔다. Walker et al.(1994)는 일본어 영형대명사 해소를 위해 센터링 모델을 사용했으며, Roh/Lee(2003)은 추론 비용 inference cost을 고려한 비용 기반 센터링 모델 Cost-based Centering Model을 사용하여 영형대명사를 해소하는 알고리즘을 제안하였다. 특정 담화에서 가장 현가성이 높은 요소가 영형 대명사로 실현될 확률이 높다고 알려져 있기 때문에 우리는 이러한 정보 또한 기계학습의 자질로 사용하고자 하였다.

현재의 대용어 해소 방법론들은 대부분 형태-통사적 정보 또는 피상적인 의미 분석을 활용한 제약과 선호도를 사용한다. 하지만 이러한 방법들은 결정론적 방법론 deterministischer Ansatz으로서 특정한 조건에서는 항상 같은 출력이 나오도록 한다. 하지만 특정 조건이 적용되더라도 출력이 다르게 나오는 경우도 존재하며, 아예 조건이 적용될 수 없는 경우도 존재한다. 예를 들어, 선행사 후보가 두 개 이상 남아 있는 경우 조건에 맞으나 선택된 출력이 정답이 아닌 경우도 있으며, 아예 조건이 적용되지 않아 두 개 이상의 선행사 후보가 남는 경우도 존재할 수 있다.

따라서 대용어 해소를 위해 이미 텍스트 논조자동분석 등의 자연언어처리 학습에 널리 사용되고 있는 기계학습과 같은 비결정론적 방법론 nicht-deterministischer Ansatz이 사용될 수 있다 (vgl. Connolly et al. 1994; Paul et al. 1999; 홍문표 2014). 기계학습 방법론에서는 주어진 데이터로부터 학습을 하여 데이터에서 가장 가능성이 높은 후보를 추천하는 방법이므로, 결정론적 방법론의 한계를 극복할 수 있다는 장점이 있다.

Park/Hong(2014)는 스페인어 영형주어를 해소하기 위해 휴리스틱 규칙과 기계학습 방법론을 통합시킨 하이브리드 방법론을 제안하였다. 스페인어 영형주어는 동사의 어미로부터 복원 가능한 경우가 존재하는데, 이를 위해 우선 동사의 형태론적 굴절 정보를 사용하여 영형주어를 복원한다. 그 이후 동사의 정보로부터 선행사가 결정되지 않는 모호한 경우에 대해 기계학습을 적용하여 하나의 선행사를 선택한다. 이 연구와는 달리 본 연구에서는 한국어 영형목적어를 다루며, 한국어에서는 동사의 형태론적 정보를 사용하여 영형목적어를 복원할 수 없기 때문에 이들의 방법론을 본 연구에 적용하기에는 어려움이 따른다. 이러한 이유로 본 연구에서는 한국어 영형목적어를 해소하기 위해 기계학습 방법론만을 채택한다.

Ⅳ. 제안하는 방법론

Ⅳ.1. 기계학습 방법론을 위한 자질 제안

본 연구에서는 한국어에 나타나는 영형목적어 현상을 다루기 위해 기계학습 방법을 사용한다. 우리는 한국어 대화체에 등장하는 영형목적어의 특성을 고려하여 기계학습을 위한 총 8개의 기계학습 자질을 제안하였으며, 표7은 각 자질과 자질에 대한 값을 설명한다.

표 7 : 기계학습을 위해 제안한 8개의 자질

	자질	값
f1	선행사 후보의 통사적 기능	*top*, *sub*, *obj*, *adv*, *comp*, *poss*
f2	통사적 기능의 일치 여부	*para*, *diff*
f3	술어들의 의미적 관계	*sim*, *same*, *oppo*, *diff*, *loc*[4)]
f4	문장 거리	*loc*, *0*, … *n*
f5	영형목적어 화자 기준 문장 거리	*-n* … *0* … *n*
f6	선행사 후보가 핵인지의 여부	*head*, *not*
f7	가장 현가성이 높은 요소인지 여부	*1,0*
f8	공지시 관계인지 여부	*yes*, *no*

f1은 선행사 후보의 통사적 기능을 나타내는 것으로, 선행사 후보의 통사적 기능의 순위가 높을수록 정답 선행사가 될 확률이 높아지는 경향을 반영하기 위해 상정한 자질이다. 만약 한 개체가 *top*이라는 값을 할당받으면 다른 선행사 후보들보다 정답 선행사가 될 확률이 높아진다. 한국어에서는 주제 표지 '은', '는' 그리고 초점사 '도', '만'을 통해 해당 개체가 주제격인지를 쉽게 파악할 수 있다. 또한 f2에서 영형목적어와 선행사 후보의 통사적 기능을 비교하여 값을 할당하는데, f1의 값이 f2의 통사적 기능 일치 여부를 계산하는데 사용될 뿐만 아니라 f7의 가장 현가성 높은 개체를 찾아내는 데에도 도움이 된다.

f2는 선행사 후보가 영형목적어와 통사적 기능이 같은지의 여부를 나타내는 것이다. 만약 선행사 후보와 영형목적어의 통사적 기능이 다르다면 *diff*라는 값을 할당받게 된다. 이는 Ⅱ장에서 살펴본 자질-공유 제약을 반영한 것으로서 영형목적어와 통사적 기능이 같은 선행사 후보가 있다면, 해당 선행사 후보가 정답 선행사가 될 확률이 높은 경향을 반영하도록 설정한 자질이다.

f3은 선행사 후보와 영형목적어의 술어의 의미적 관계[5]에 대한 값을 할당해 줄 수 있는 자질이다. 한국어에서는 선행사와 영형목적어의 술어가 동의어, 반의어 등과 같이 의미적으로 특정한 관계에 있는 경우들이 존재하므로, 이러한 현상을 반영한 자질을 상정하였다. 따라서 f3의 값을 통해 이러한 경향성을 나타내줄 수 있다.

f4는 영형목적어와 선행사 후보의 문장 거리에 관한 것이다. 영형목적어와 선행사 후보가 같은 절에 나타날 경우 선행사 후보는 *loc*라는 값을 지니게 되며, 같은 절은 아니지만 같은 문장에 나타나는 경우 선행사 후보가 *0*이라는 값을 지니게 된다. f4의 값이 더 커질수록 선행사 후보와 영형목적어의 거리가 멀다는 것을 나타낸다. 이 자질은 영형목적어와 거리가 가까운 선행사 후보일수록 정답

4) 만약 선행사 후보와 영형목적어가 같은 절 내에 등장하면 loc라는 값을 할당받는다.

5) 익명의 심사자가 지적한 바와 같이 술어의 의미적 관계를 파악하기 위해서는 해당 술어의 의미 모호성 해소 단계가 필요하다. 그러나 본 연구에서는 영형 목적어를 해소하기 이전에 이미 술어의 의미 모호성이 해결된 것으로 가정하여 이 과정에 대해 언급하지 않았음을 밝힌다. 논문의 질적 향상을 위해 큰 도움을 주신 익명의 심사자께 감사드린다.

선행사가 될 확률이 높다는 기존 연구들의 결과를 반영하는 것이다.

본 연구에서는 대화체 문장을 다루기 때문에 문어체에 초점을 맞춘 기존 연구들의 방법론을 그대로 적용하기에는 어려움이 있다. 이러한 이유로 우리는 대화체 문장의 특성을 반영한 f5를 도입하였다. f5는 영형목적어의 화자를 기준으로 영형목적어와 선행사 후보의 문장 거리를 계산하는 자질이다. 우리는 영형목적어의 화자를 기준으로 문장 거리를 계산하는 것이 기존 연구에서 대용어 해소를 위해 문장 거리에 대한 정보를 활용하는 원래의 목적을 더 잘 반영할 수 있다고 가정하였다. f4와 달리 f5는 영형목적어와 선행사 후보의 화자가 다르면 음수를 값으로 할당받을 수 있다.

f6은 선행사 후보가 자신이 속한 구에서 핵 Kopf인지의 여부에 관한 것이다. 예를 들어, 명사구인 선행사 후보가 자신이 속한 명사구에서 핵이라면 해당 명사구가 핵이 아닌 후보 선행사보다 정답 선행사일 확률이 높아지는 경향을 반영하기 위해 이 자질을 설정하였다.

f7은 센터링 이론의 틀을 기계학습 자질의 형태로 제안한 것이다. 기존 연구들에서는 문맥으로부터 추론 가능한 현가성이 높은 개체가 자주 생략된다고 주장해 왔다(Walker et al. 1994; Iida 1998; Hong 2000). 따라서 우리는 한국어의 전향적 센터 순위를 고려하여 특정 영형목적어에 대한 선행사 후보들 중 가장 현가성이 높은 선행사 후보에 대해 *1*이라는 값을 할당하도록 값을 설정하였다.

f8은 실제로 영형목적어와 특정 선행사 후보가 공지시 Koreferenz 관계에 있는지에 대한 정답을 제시해주는 것이다. 만약 특정 선행사 후보가 해당 영형목적어의 정답 선행사라면 *yes*라는 값을 할당받고, 만약 정답 선행사가 아니라면 *no*라는 값을 할당받는다.

Ⅳ.2. 실험

본 연구에서 제안한 기계학습 자질을 사용하여 한국어 영형목적어 해소의 성능을 평가하기 위해 실험을 수행하였다. 실험을 위해 ETRI에서 제공해 준 관광 분야의 대회체 문장들 중 영형목적어가 포함된 문장들[6]을 추출하였으며, 총

6) 실험에 사용된 문장 예시는 부록에 제시되어 있다.

1123개의 공지시 쌍이 추출되었다; 이 중 308개는 실제 공지시 관계에 있는 긍정쌍이며, 824개는 공지시 관계가 아닌 부정쌍이다.

기계학습 방법의 효과를 평가하기 위해 웨카 WEKA 3.6.10 버전을 사용하였으며, 기계학습 알고리즘으로는 'SVM (Support Vector Machine)'을 선택하였다. SVM 알고리즘은 다양한 자연언어처리에서 좋은 성능을 보인다는 것이 증명되어 왔다(vgl. Kudo/Matsumoto 2001; Isozaki/Kazawa 2002). 평가 방식으로는 '10-fold' 평가 방식을 활용하여 실험 결과를 얻었는데, 이 방식은 예를 들어, 1000문장의 코퍼스가 존재하면 이를 10개의 세트로 나누어 그 중 9개 세트인 900문장에 대해 기계학습을 수행하고, 나머지 1개 세트인 100문장에 대한 성능평가를 수행하는 방식이다. 이 과정을 모든 세트에 대해 반복한 후 평가 결과를 모두 합산하여 평균을 구하게 된다.

본 연구에서 제안한 8개의 기계학습 자질을 사용하여 한국어 영형목적어 해소를 위한 실험을 한 결과 73.37%의 정확도 Accuracy가 측정되었다. 이는 전체 코퍼스에서 기계학습을 통해 정확하게 복원된 한국어 영형목적어의 비율을 의미한다. 본 연구에서 측정된 정확도를 기존 연구들의 결과와 직접적으로 비교하는 것은 어려움이 있는데, 그 이유는 기존 연구들에서는 대화체가 아닌 문어체 텍스트를 연구 대상으로 삼고 있기 때문이다. 따라서 우리는 실험 결과의 비교를 위해 홍문표(2011)에서 제안한 한국어 전향적 센터 순위에 따라 담화 내 가장 현가성이 높은 선행사 후보를 정답 선행사로 결정하는 베이스라인을 설정하였다. 표8에서 볼 수 있듯이 본 연구의 방법론을 사용하면 베이스라인 대비 정확도가 61.71% 향상된다.

표 8 : 실험 결과

	베이스라인	제안한 방법론	비고
정확도 Accuracy	11.66%	73.37%	61.71% 상승

웨카 시스템에서 'InfoGainAttribute Evaluator'를 활용하면 기계학습을 위해 사용된 자질 중 영형목적어 복원을 위해 큰 영향력을 지닌 자질의 순위가 결과로 나오는데, 다음 표는 자질들의 영향력에 따른 순위를 보여준다.

표 9 : 자질들의 순위

순위	자질	
1	f4	문장 거리
2	f3	술어들의 의미적 관계
3	f7	가장 현가성이 높은 요소인지 여부
4	f5	영형목적어 화자 기준 문장 거리
5	f1	선행사 후보의 통사적 기능
6	f2	통사적 기능의 일치 여부
7	f6	선행사 후보가 핵인지의 여부

표9에서 볼 수 있듯이 1위를 차지한 자질은 f4였고, 이는 영형목적어와 선행사 후보의 문장 거리에 관한 것이다. 문장 거리는 대용어 해소와 관련된 기존 연구에서 자주 사용되는 정보이다. 왜냐하면 영형목적어와 거리가 가까운 선행사 후보가 정답 선행사가 될 확률이 높기 때문이다. McEnercy et al.(1997)은 대명사와 그들의 선행사의 거리와 관련하여 연구를 수행하였고, 연구 결과 대명사의 선행사가 그 분포에 있어 명확한 패턴이 있다는 것을 증명하였다. 문장 거리에 대한 자질이 한국어 영형목적어 해소를 위해 가장 영향력이 높은 자질이라는 결과는 바로 이러한 문장 거리의 중요성을 증명하는 것이라고 볼 수 있다.

f4 다음으로는 영형목적어와 선행사 후보의 술어의 의미 관계를 나타내는 자질인 f3이 두 번째로 높은 순위를 차지했다. 이전 연구들에서는 대용어 해소에 대한 의미적 제약으로 하위 범주화와 같은 정보를 사용해 왔다. 예를 들어, ‘eat'의 주어가 생략되어 있으면, 선행사는 사람 또는 동물이라는 의미를 지닐 것이라는 제약과 같은 것이다. 다른 연구들의 의미적 제약과는 달리 본 연구에서 제안한 f3은 영형대명사 해소에 처음 제안되는 자질이다. 이 자질이 두 번째 순위를 차지했기 때문에 한국어 영형목적어 해소를 위해 술어의 의미적 관계가 매우 중요한 역할을 한다고 주장할 수 있다.

표9에 제시된 자질들의 순위를 통해 우리는 센터링 이론이 한국어 영형목적어 해소를 위해서도 중요하다는 것을 확인할 수 있었다. 센터링 이론에 따르면 가장 현가성이 높은 선행사 후보가 영형 대명사로 실현되는 경향이 있다. f7은 이러한 특성을 반영하였는데, 제안된 자질 중 이 자질이 한국어 영형목적어 해

소를 위해 3번째로 영향력이 큰 자질이라는 결과가 나왔다. 따라서 센터링 이론에서 제안한 주장이 한국어 영형목적어 해소를 위해서도 중요하다는 것이 증명된 것이라 볼 수 있다.

Ⅴ. 독일어 번역률 평가

본 연구에서 제안하는 방법이 한-독 기계번역 결과에 미치는 영향을 알아보기 위해 한국어 영형목적어 복원 전/후의 독일어 번역 결과를 비교하였다. 독일어 번역은 구글 번역기를 사용하여 얻게 된 결과이다. ETRI에서 제공한 총 180문장의 관광 분야 대화체 문장을 대상으로 번역 점수를 평가하였으며, 번역 점수 평가 기준은 표10과 같다. 이 평가 기준을 바탕으로 영형목적어를 수동으로 복원하기 전과 후의 번역점수를 평가한 후 번역률을 환산하였다.

표 10 : 번역 점수 평가 기준

점수	기준
4점	원문이 정보의 손실없이 완벽하게 번역됨
3점	번역상 약간의 어색함은 있으나 정보가 거의 완벽하게 전달됨
2점	의미가 구 Phrase 단위로 부분적으로 전달됨
1점	의미가 단어레벨에서만 전달됨
0점	번역 실패
참고	각 점수가 나타내는 기준을 고려해봤을 때 중간 수준이라고 판단되면 0.5점 단위로 점수를 부여한다.

번역률 평가 결과 한국어 영형목적어 복원 전의 독일어 번역률은 50.83%였으며, 영형목적어를 복원한 후의 독일어 번역률은 65.76%로 번역률이 14.93% 정도 향상됨을 알 수 있었다.

표 11 : 영형목적어 복원 전/후 번역률 비교

	영형목적어 복원 전	영형목적어 복원 후	비고
번역률	50.83%	65.76%	14.93% 상승

한국어 영형목적어를 복원하여 독일어 번역 점수가 높아지는 문장의 예시는 표12, 표13과 같다.

표 12 : 한국어 영형목적어 복원 전/후 독일어 번역 결과 예시 1

선행 문장	A: 편의점에서 간편히 식사하실래요?				
	B: 네?				
	B: 편의점이요?				
	영형목적어 복원 전	독일어 번역 결과		영형목적어 복원 후	독일어 번역 결과
(4)	A:네, 많은 사람이 ø 이용합니다.	Ja, viele Menschen mit.	(4)'	네, 많은 사람이 편의점을 이용합니다.	Ja, viele Menschen werden die Convenience-Store zu verwenden.

표 13 : 한국어 영형목적어 복원 전/후 독일어 번역 결과 예시 2

선행 문장	A: 메인으로 가는 버스 맞나요?				
	B: 네, 맞습니다.				
	A: 버스 안에서 기다려도 될까요?				
	B: 그러시죠.				
	A: 감사합니다.				
	B: 이제 출발하겠습니다.				
	A: 실례합니다만, 에어콘을 꺼주시겠어요?				
	영형목적어 복원 전	독일어 번역 결과		영형목적어 복원 후	독일어 번역 결과
(5)	B: 네, ø 꺼드릴게요.	Ja, ich werde gehen.	(5)'	네, 에어컨을 꺼드릴게요.	Ja, ich werde schalten Sie die Klimaanlage.

독일어 번역결과를 얻기 위해 구글 번역기를 활용하였는데, 구글 번역기는 통계기반 번역 방식을 사용한다. 통계 기반의 기계번역은 2개 국어 말뭉치를 통계적으로 분석한 결과를 학습하여 번역하는 방식이다. 특히 구축된 언어 모델을 통해 특정 단어가 번역되면 그 다음 단어로 나올 수 있는 것을 확률적으로 계산하여 가장 확률이 높은 단어를 번역 결과로 생성한다.

표 12와 표 13에서 제시된 독일어 번역 결과를 살펴보면 한국어 영형목적어가 복원됨으로서 (4)'처럼 한국어 영형목적어가 복원되기 이전에는 생성되지 않았던 동사가 생성되기도 하며, (5)'처럼 목적어에 따른 정확한 동사가 결과로 나오기도 한다. 이 예시들에서도 알 수 있듯이 한국어에서는 단순히 생략된 목적어를 복원하였지만 독일어 번역 결과에서는 목적어를 복원하기 이전보다 훨씬 의미 전달이 잘 되는 독일어 문장이 생성된 것을 볼 수 있다. 따라서 한국어 영형목적어를 복원하는 것이 실제로 한-독 기계번역 결과의 번역률을 향상시키는데 중요한 역할을 한다는 것을 알 수 있다.

Ⅵ. 맺는말

본 연구에서는 한국어 대화체 문장에서 빈번하게 등장하는 영형목적어를 복원하기 위해 기계학습 방법을 제안하였다. 한국어 영형목적어는 한국어를 출발언어로 하고 목적어가 명시적으로 생성되어야 하는 독일어와 같은 언어를 목표언어로 하는 기계번역 시스템에서 필수적으로 복원되어야만 하는 현상이었다. 본 연구에서는 한국어 대화체에 등장하는 영형목적어의 특성을 반영하여 총 8개의 기계학습 자질들을 제안하였다. 우리가 제안한 자질들을 사용하여 한국어 영형목적어 해소의 성능을 평가하기 위해 실험을 한 결과 73.37%의 정확도가 측정되었으며, 이는 베이스라인 대비 정확도가 61.71% 향상된 수치였다. 한국어 영형목적어를 복원하는 것이 독일어 번역 결과에 어떠한 영향을 미치는지 알아보기 위해 한국어 영형목적어 복원 전/후의 번역률을 비교한 결과 영형목적어를 복원한 이후의 번역률이 65.76%로 14.93% 상승된 것을 알 수 있었다. 현재 향후 실험을 위해 더 큰 규모의 코퍼스를 수집 중이며, 본 논문에서 제안한 방법론을 다시 적용한 후 번역률을 평가해 보고자 한다.

참고문헌

최재웅 & 이민행(1999) :『초점- 형식의미론과 한국어 기술』. 한신문화사.

홍문표(2011) :「한-독 대화체 기계번역을 위한 주어생략현상의 처리방안」. 『독어학』 24집, 417-439.

홍문표(2014) :「바이그램을 활용한 텍스트 논조자동분석」. 『독일언어문학』, 65집, 27-46.

홍민표(2000) :「센터링 이론과 대화체에서의 논항 생략 현상」. 인지과학, 11(1)집, 9-24.

Baldwin, B.(1997): CogNIAC: high precision coreference with limited knowledge and linguistic resources. In: Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts. Association for Computational Linguistics, 38-45.

Carbonell, J. G., & Brown, R. D.(1988): Anaphora resolution: a multi-strategy approach. In: Proceedings of Proceedings of the 12th conference on Computational linguistics, 1. Association for Computational Linguistics, 96-101.

Connolly, D., Burger, J. D., & Day, D. S.(1997): A machine learning approach to anaphoric reference. In: New Methods in Language Processing, 133-144.

Ferráandez, A., & Peral, J.(2000): A computational approach to zero-pronouns in Spanish. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 166-172.

Grosz, B. J., Weinstein, S., & Joshi, A. K.(1995): Centering: A framework for modeling the local coherence of discourse. In: Computational linguistics, 21(2), 203-225.

Hong, M.(2002): A review on zero anaphora resolution theories in Korean. In: Studies in Modern Grammar, 29, 167-186.

Iida, M.(1998): Discourse coherence and shifting centers in Japanese texts. In: Centering theory in discourse, 161-180.

Isozaki, H., & Kazawa, H.(2000): Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th international conference on Computational linguistics, Volume 1. Association for Computational Linguistics, 168-184.

Kameyama, M.(1986): A property-sharing constraint in centering. In: Proceedings of the 24th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 200-206.

Kim, L. K.(2010): Korean Honorific Agreement too Guides Null Argument Resolution: Evidence from an Offline Study. University of Pennsylvania Working Papers in Linguistics, 16(1), 12. 101-108.

Kudo, T., & Matsumoto, Y.(2001): Chunking with support vector machines. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics, 1-8.

Lappin, S., & Leass, H. J.(1994): An algorithm for pronominal anaphora resolution. In: Computational linguistics, 20(4), 535-561.

Lee, D.(2002): Discourse Representation Methods and Korean Dialogue. In: Proceedings of the 2002 Winter Linguistic Society of Korea Conference. Seoul National University, 88-104.

Okumura, M., & Tamura, K.(1996): Zero pronoun resolution in Japanese discourse based on centering theory. In: Proceedings of the 16th conference on Computational linguistics, Volume 2. Association for Computational Linguistics, 871-876.

McEnery et al.(1997): Corpus annotation and reference resolution. In: Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, 67-74.

Nakaiwa, H., & Shirai, S.(1996): Anaphora resolution of Japanese zero pronouns with deictic reference. In: Proceedings of the 16th conference on Computational linguistics, Volume 2., Association for Computational

Linguistics, 812-817.

Palomar, M., Ferráandez, A., Moreno, L., Martíinez-Barco, P., Peral, J., Saiz-Noeda, M., & Muñnoz, R.(2001): An algorithm for anaphora resolution in Spanish texts. In: Computational Linguistics, 27(4), 545-567.

Park, A., & Hong, M.(2014): Hybrid Approach to Zero Subject Resolution for multilingual MT-Spanish-to-Korean Cases. In: Proceedings of the 28th Pacific Asia Conference On Language Information and Computing. 254-261.

Paul, M., Yamamoto, K., & Sumita, E.(1999): Corpus-based anaphora resolution towards antecedent preference. In: Proceedings of the Workshop on Coreference and its Applications. Association for Computational Linguistics, 47-52.

Roh, J. E., & Lee, J. H.(2003): An empirical study for generating zero pronoun in Korean based on Cost-based Centering Model. In: Proceedings of Australasian Language Technology Association, 90-97.

Ryu, B. R.(2001): Centering and zero anaphora in the Korean discourse. Seoul National University, MS Thesis.

Walker, M., Cote, S., & Iida, M.(1994): Japanese discourse and the process of centering. In: Proceedings of Computational linguistics, 20(2), 193-232.

Zusammenfassung

Objektellipse im Koreanischen und deren Behandlung für die maschinelle Übersetzung ins Deutsche

Park, Arum · Hong, Munpyo (Sungkyunkwan Uni)

In der vorliegenden Arbeit wird das Objekt-Ellipse Phänomen in der

maschinellen Übersetzung des Koreanischen ins Deutsche behandelt. In einer sogenannten Pro-Drop Sprache, wie im Koreanischen, wird ein Objekt oft ausgelassen. Bei der Übersetzung aus dem Koreanischen in das Deutsche müssen die ausgelassenen Objekte explizit ausgedrückt werden.

In den meisten bisherigen Ansätzen handelt es sich um eine deterministische Methode. Hier spielen empirische Regeln eine entscheidende Rolle, um ein ausgelassenes Objekt wieder zu finden. Zu den empirischen Regeln gehören u.a. morphologische Informationen, semantische Eigenschaften bestimmter Verben und Adjektive und Informationen aus Kontexten. Eins von den größten Problemen der Ansätze ist, dass es Fälle gibt, wo die Resolution der Anaphern nicht deterministisch erfolgt.

Um das Problem zu lösen, wird hier ein neuer Ansatz vorgestellt. Der neue Ansatz stützt sich auf maschinelles Lernen. Zu diesem Zweck werden insgesamt 8 Merkmale mit Rücksicht auf die Eigenschaft des koreanischen Zero-Objekt vorgeschlagen.

Das Experiment zeigte, dass unser Ansatz im Vergleich zu der Basislinie die Genauigkeit der Abstad von 11.66% auf 73.37% um 61.71% erhöhen kann.

핵심어 : 대용어해소 Anaphernresolution, 영형대명사 Zero-Pronomen,
영형목적어 Zero-Objekt, 기계번역 maschinelle Übersetzung,
기계학습 maschinelles Lernen

필자 E-mail : remin2@skku.edu, skkhmp@skku.edu

논문투고일 : 2015. 10. 15 / 심사일 : 2015. 11. 15 / 게재확정일 : 2015. 12. 5

〈부록: 실험에 사용된 코퍼스 예시〉

실험 문장 예시 1

화자	한국어 대화문
A	지금 여행$_1$중인데 여권$_2$을 잃어버렸어요.
B	우선 여기 민원신청서$_3$를 작성해 주세요.
A	예.
B	어떻게 ø 분실하셨나요?
A	ø 도둑맞은 것 같아요.

실험 문장 예시 2

화자	한국어 대화문
A	음식 주문$_1$을 어떻게 하는 거죠?
B	저 기계$_2$에서 메뉴$_3$를 선택한 후 식권$_4$을 뽑으세요.
A	아침 식사$_5$는 11시까지만 되는 건가요?
B	네, 지금 정확히 11시니까 ø 원하신다면 ø 해 드릴게요.
A	감사합니다.

실험 문장 예시 3

화자	한국어 대화문
A	죄송한데 여기 화장실$_1$을 이용할 수 있을까요?
B	저쪽의 통로$_2$로 쭉 가시면 있습니다
A	네.
B	무슨 일 있으신가요?
A	화장실$_3$에 휴지$_4$가 없어서요.
B	잠시만요.
B	제가 ø 꺼내 드릴게요.
A	알겠습니다.