한-독 대화체 기계번역을 위한 주어생략현상의 처리방안*

홍문표(성균관대)

1. 서론

한국어와 독일어간의 기계번역에 대한 산업적 활용도는 아직 영어와 한국어 또는 영어와 다른 언어들 간의 기계번역만큼 높지는 않지만, 꾸준히 그 수요와 필요성이 높아져가고 있다. 특히 기술분야 및 특허분야의 한국어 문서에 대한 독일어로의 번역수요가 급증하고 있기 때문에 한-독 기술문서 기계번역 시스템의 개발에 대한 관심은 날로 높아져가고 있다.

이와 더불어 한국어와 독일어간의 기계번역에 대한 산업적 수요가 발생하고 있는 분야는 대화체 혹은 구어체 문장에 대한 기계번역이다. 수년전부터 아시아지역을 중심으로 불었던 한류열풍이 이제는 유럽 대륙에까지 상륙하여 독일과 프랑스를 중심으로 유럽 내의 타 국가에 번지고 있다고 한다. 한국 문화, 특히 한국대중문화에 관심이 많은 독일인들은 한국에 관한 정보를 실시간으로 접하고자 하며, 비슷한 관심사를 가진 한국인들과 온라인으로 대화를 나누기를 원한다. 이러한 목적 등으로 독일인들 및 한국인들은 구글사 Google Inc. 등에서 제공하는 실시간 기계번역서비스를 활용하고 있지만 특히 대화체 문장에 대한 낮은 번역수준으로 인해 이 번역결과를 실제로 사용하는 경우는 매우 드물다.

영어와 독일어 혹은 영어와 스페인어 등과 같은 많은 다른 언어쌍에 대해서는 뛰어난 성능을 보이는 이 시스템이 한국어와 독일어의 대화체 번역에 대해서는 유독 조악한 번역결과를 보이는 이유는 여러 가지를 들 수 있다. 우선 한국어와 독일어로 동시에 작성된 문서가 극히 제한되어 있기 때문이다. 한국어와 독일어로 동시에 작성된 문서의 양이 영어와 독일어로 동시에 작성된 문서의 양보다 적을 것임은 자명하다. 한-독 병렬언어 말뭉치가 부족하면

* 본 연구는 지식경제부의 지식경제 기술혁신사업의 일환(2009-S-034-01)으로 수행되었습니다.

통계기반 기계번역시스템의 성공적인 적용은 매우 어렵다.

또 다른 이유는 번역시스템이 문어체 텍스트 geschriebene Texte에 특화되어 있기 때문이다. 문어체 텍스트는 대화체 텍스트 gesprochene Texte에 비하여 철자오류나, 문법오류, 혹은 생략현상 등이 드물기 때문에 비교적 정형화되어 있다고 볼 수 있다. 그러나 이에 비해 대화체 텍스트는 위에 언급한 여러 오류 및 대화체만의 여러 가지 언어현상들을 포함하고 있기 때문에, 번역시스템이 처리하기에 더 어려운 점이 있다.

본 연구에서는 한국어 대화체 문장을 독일어로 기계번역할 때 번역성능을 저하시키는 이유 중의 대표적인 현상인 주어생략현상 Subjekt-Ellipse을 다룬다. 생략현상은 기존의 언어학 연구에서 의미론, 화용론적인 시각으로 접근되었다. 이러한 순수 이론언어학적 기반의 방법론들은 자연언어처리 분야에 직접 적용하기에는 많은 지식 혹은 리소스를 필요로 하기 때문에, 실시간 번역을 요구하는 상황에 적합하지 않은 측면이 있다. 따라서 본 연구에서는 주어생략현상의 처리를 위해 비교적 간단한 언어학적 단서를 활용한 방법론을 제안하고자 한다. 이 방법론은 문장의 형태소분석 morphologische Analyse과 품사태깅 POS-Tagging정보만을 활용하므로, 구조분석, 의미분석을 모두 수행해야 하는 기존의 방법론에 비해 적용하기가 쉽고, 적은 양의 리소스를 필요로 하는 장점이 있다.

논문의 2장에서는 한국어를 출발어 Quellsprache로 하는 기계번역에서 나타나는 주어생략현상에 대해 살펴본다. 3장에서는 주어생략현상을 다루기 위한 기존 연구에 대해 논의한다. 여기서는 일-영 기계번역을 위한 Nariyama et al. (2002)의 연구와 센터링 이론 Centering Theory 관련 연구를 중심으로 논의를 진행할 것이다. 4장에서는 2, 3장의 논의를 바탕으로 한-독 대화체 기계번역을 위한 주어생략처리 방법론을 제안한다. 5장에서는 제안한 방법론의 타당성 검증을 위한 실험을 수행하고, 그 결과를 소개한다. 6장은 본 연구의 결과를 정리하고, 향후 연구방향에 대해 논의한다.

2. 기계번역에서의 주어생략현상

한국어와 일본어 등과 같은 주제지향언어 topik-orientierte Sprache에서 독일어와 영어와 같은 주어지향언어 Subjekt-orientierte Sprache로의 기계번역에서 발생하는 가장 큰 문제 중의 하나는 주제지향언어에서 빈번히 발생하는 주어생략현상이다. 주제지향언어에서 문법적 주어 grammatisches Subjekt는 문맥이나 상황 또는 세상지식 등을 통해 유추할 수 있는 경우 흔히 생략될 수 있기 때문에, 반드시 문법적 주어가 문장 내에 등장해야 하는 독일어와 같은 주어지향언어로의 번역시에는 생략된 주어를 복원하는 과정이 필요하다. 특히 주어의 생략현상은 채팅, 메신져 등과 같은 대화체 텍스트에서 더 빈번히 일어나므로 대화체 기계번역시스템의 개발을 위해서는 이 문제의 해결이 필수적이라고 할 수 있다.

일반적으로 관련 연구에서 대명사가 실현되어야 할 자리에 실현되지 않고 생략되어 있는 것을 영형대명사 Zero Pronoun이라고 부르는데, 본 연구에서 다루는 주어생략현상은 영형대명사의 한 부분이라고 볼 수 있다.

(1) A: 너 무슨 일 있니? B: 아뇨, 그냥 ∅ 너무 피곤해요
(2) 형i은 정신차리라고 혼내지는 못할망정 은새가 묻는 말에 ∅i 대꾸하고 실실 쪼개면서 웃어주고!
(3) 실은 어제 저i 아내j와 다퉜어요. ∅i 아무리 말을 걸어도 ∅j 아무 대답 하지 않더라고요...

(1)~(3)은 본 연구를 위해 수집한 한국어 대화체 코퍼스에서 발췌한 예문들이다. 예문에서 '∅'기호는 생략된 주어의 위치를 나타낸다. 대화체 문장에서는 위의 예문들과 같이 주어가 명시적으로 실현되지 않는 경우가 빈번하다. 그 이유 중의 하나는 (1)의 예문에서와 같이 발화를 주고 받는 대화상황에서 발화 순서에 기반하여 명시되지 않은 주어가 무엇인지가 명확하기 때문이다. 또 다른 이유는 (2)번 예문에서처럼 명시되지 않은 주어가 동일 문장 내에 이미 등장하였거나 혹은 (3)번 예문에서와 같이 담화구조상에서 선행 문장 등에 등장하였기 때문이다.

Nariyama et al.(2002)은 일-영 기계번역에서 나타나는 주어생략현상을 생략된 주어의 선행사를 어떻게 복원할 수 있냐에 따라 다음의 세 가지 유형으로 구분하였다.

i) 문장 내에 생략된 주어의 선행사가 있는 경우 (Intra-Sentential Zero Pronoun)
ii) 담화 내에 생략된 주어의 선행사가 있는 경우 (Inter-Sentential Zero Pronoun)
iii) i)과 ii)의 경우가 아니면서 언어외적 요인을 통해 생략된 주어의 선행사를 파악할 수 있는 경우 (Extra-Sentential Zero Pronoun)

이와 같은 분류는 한국어에도 동일하게 적용될 수 있으므로, 본 연구에서는 위의 분류를 그대로 따르고자 한다. 문장 내에 생략된 주어의 선행사가 있는 경우 (앞으로 'Intra-ZP'로 표기함)는 주로 연결어미 등으로 연결된 복문의 경우 앞선 문장 등에 나타난 주어가 'Intra-ZP'의 선행사 역할을 할 수 있는 경우이다. 다음 예문 (4)는 'Intra-ZP'의 예를 보여준다.

(4) 나i도 공부하다가 ∅i 쉬다가 ∅i 공부하고 있어.

이 문장은 3개의 단문이 연결어미 '-다가'로 연결되어 있고, 첫 번째 문장의 주어 '나'가 두 번째 단문 및 세 번째 단문의 주어로 해석될 수 있다.

두 번째 유형의 주어생략현상은 동일 문장 내가 아니라 담화 내에서 선행사를 찾을 수 있는 경우이다 (앞으로 'Inter-ZP'로 표기함). (3)의 예문은 'Inter-ZP' 현상을 보여준다. (3)에서 생략된 주어의 선행사들은 담화를 구성하는 첫 번째 문장 ('실은 어제 저 아내와 다퉜어요')에서 찾을 수 있다.

마지막으로 세 번째 유형은 생략된 주어의 선행사를 담화 내에서 찾을 수 있는 것이 아니라, 언어외적인 요인들, 예를 들어 세상지식이나 시각정보 등을 통해 파악할 수 있는 경우이다 (앞으로 'Extra-ZP'로 표기함). (5)의 예문이 이에 해당한다.

(5) ∅ 점심 같이 먹어요

(5)의 예문에서 생략된 주어의 선행사는 상황에 따라 1인칭, 2인칭, 3인칭으로 모두 해석이 가능하다. 이러한 'Extra-ZP'의 선행사는 언어외적 지식을 동원해야 파악할 수 있는 경우가 많다.

본 연구를 위해 수집한 총 715문장의 한국어 대화체 코퍼스를 분석한 결과 총 359의 문장에서 적어도 하나의 영형주어대명사가 발견되었다. 이는 메신져 텍스트 및 드라마 대본과 같은 구어체 텍스트에서는 거의 두 문장 중 한 문장에서 하나 이상의 영형주어대명사가 사용된다는 것을 의미한다. 359개의 문장에서 사용된 영형주어대명사의 개수는 총 435개로서 전체 문장의 개수를 고려해보면 문장당 0.61개의 영형주어대명사가 사용되었다. 또한 영형주어대명사가 사용된 문장만을 고려할 경우 문장당 1.21개의 영형주어대명사가 사용됨을 알 수 있었다.

표 1에서 보는 바와 같이 이를 문장의 유형별로 구분하면, 메신져 문장의 경우 총 358문장중 155문장에서 영형주어대명사가 사용되었고 (43%), 드라마 대본의 경우는 204문장에서 영형주어대명사가 사용되었다 (57%). 이는 아마도 드라마 대본의 경우 시각정보가 많이 활용되므로, 메신져 문장보다는 주어가 생략되는 경우가 많을 것이기 때문으로 예상된다.

<표 1> 구어체 코퍼스에 등장하는 영형주어대명사의 비율

	전체문장	메신져 문장	드라마 대본
코퍼스 문장수	715	358	357
ZP 등장문장수	359 (50%)	155 (43%)	204 (57%)
ZP 개수	435	191	244
ZP개수 1[1)]	1.21	1.23	1.2
ZP개수 2[2)]	0.61	0.53	0.68

435개의 영형주어대명사를 앞서 소개한 영형주어대명사의 유형에 따라 분류해보면 다음의 표 2와 같다. 전체 문장에 나타나는 총 435개의 영형주어대명사 가운데 81.14%나 'Extra-ZP'로 분류되었으며, 15.4%가 'Inter-ZP', 3.44%가 'Intra-ZP'로 분류되었다. 이러한 비율은 메신져 문장과 드라마 대본에서도

1) ZP 개수 1 = 전체 ZP 개수/ZP가 등장하는 문장수
2) ZP 개수 2 = 전체 ZP 개수/전체 문장수

크게 다르지 않다. 다만 드라마 대본의 경우 메신져 문장보다는 ‘Inter-ZP’가 다소 많이 등장하는 것으로 조사되었다.

<표 2> 영형주어대명사 유형별 개수 및 비율

	전체문장	메신져 문장	드라마 대본
ZP 개수	435	191	244
Extra-ZP	353 (81.14%)	162 (84.41%)	191 (78.28%)
Inter-ZP	67 (15.4%)	19 (9.94%)	48 (19.67%)
Intra-ZP	15 (3.44%)	10 (5.23%)	5 (2.05%)

위와 같은 코퍼스 분석결과는 주어생략현상 문제의 해결방안을 고안하는데 여러 가지를 시사한다. 첫째, 센터링 이론과 같은 주어생략현상을 다루기 위한 기존의 방안만으로는 대화체에 절대적으로 많이 등장하는 ‘Extra-ZP’ 현상을 해결할 수 없으므로 새로운 해결방안이 필요하다. 둘째, ‘Extra-ZP’는 생략된 주어의 선행사를 문맥이나 상황정보, 세상지식 등과 같이 언어외적인 요인을 통해 파악할 수 있으므로, 이를 기계번역을 위해 현실적으로 이용할 수 있는 방안이 필요하다. 셋째, ‘Inter-ZP’ 현상의 해결을 위해 센터링 이론을 도입할 수 있으나, 기존의 방법론을 대화체 기계번역이라는 목적에 부합하도록 수정할 필요가 있다.

3. 관련연구

주제지향적 언어로부터 주어지향적 언어로의 기계번역에서 흔히 발생하는 주어생략현상은 주로 일본어-영어 기계번역시스템의 개발을 위해 다루어졌다. Nakaiwa et al.(1995)은 ‘Extra-ZP’의 해결을 위해 의미론적, 화용론적 제약을 사용하였다. 특히 일본어의 양상조동사 기능을 하는 어미 및 동사의 의미적 속성과 접속사의 의미를 활용하는 방안을 제안하였다. 이들은 특히 동사의 의미적 속성을 주어복원을 위해 활용하였는데, 이를 위해 동사의 의미를 97개로 나누었다. 이들의 연구는 본 연구를 위해서도 많은 부분 비판적으로 수용되나, 이들이 주로 다룬 텍스트 유형은 신문기사 등과 같은 문어체 텍스트이므

로, 본 연구에서 제안하는 구어체의 처리방안과는 약간의 차이가 있다.

'Inter-ZP'의 처리를 위한 대부분의 연구는 센터링 이론을 중심으로 이루어졌다. Okumura & Tamura(1996)는 일본어의 'Inter-ZP'를 처리하기 위해 센터링 이론을 약간 수정하여 제안하였다. 이들의 연구는 영어를 중심으로 한 기존의 센터링 이론에서 제안한 현가성 체계에 일본어의 특징을 반영하여 'TOPIC'과 'EMPATHY'를 추가했다는 특징이 있다. 이들이 제안한 알고리즘의 정확률은 71.3%에서 최대 78.3%이다.

Nakaiwa et al.(1995)에서는 접속사 정보를 활용하여 'Inter-ZP'를 처리하는 방법론도 소개하였다. 이들의 실험 결과에 따르면 선행절과 후행절의 주어공유와 관련된 접속사의 유형을 분류하여 주어를 복원하는 방법이 순수 센터링 이론에 기반한 방법론보다 오히려 더 우수한 성능을 보인다고 한다.

한국어 구어체 기계번역의 측면에서 센터링 이론을 주어생략 현상에 적용한 연구로는 차건회 외(1997)를 들 수 있다. 이들은 한국어 구어체에 나타나는 주어생략현상을 센터링 이론을 적용하여 해결하려 했다. 이들의 연구에서는 선호센터에 대한 계산을 베이지안 Bayesian 수식을 활용하여 수행하였다. 그러나 이 방안을 실시간으로 진행되는 구어체 기계번역시스템에 적용하기에는 무리가 있다. 또한 연구의 대상이 단문에 한정된다는 단점을 갖고 있다.

4. 한-독 기계번역을 위한 주어생략 처리방안

여기서는 생략된 주어의 유형 중 가장 큰 비율을 차지하는 'Extra-ZP'를 처리하는 방안과 그 다음 비중을 차지하는 'Inter-ZP' 및 'Intra-ZP'를 처리하는 방안을 소개한다.

4.1. 'Extra-ZP' 처리방안

Nariyama et al.(1995)이 지적한 바와 같이 'Extra-ZP'의 처리를 위해 반드시 언어외적인 요인만이 활용될 수 있는 것은 아니다. 많은 경우 언어외적 요인 이외에도 한국어의 종결어미 정보와 보조용언을 통한 양상 Modalität이나

동작태 Aktionsarten 정보 등을 생략된 주어의 선행사를 찾는데 활용할 수 있다.

한국어의 '~ㅂ니다', '~어/아요?', '~세요!' 등과 같은 종결어미는 평서문, 의문문, 명령문 등과 같은 문장의 유형을 결정하는 역할을 한다. 문장 유형이 명령문인 경우, 생략된 주어는 (6)과 같이 'Sie'나 'du'로 복원될 수 있지만, 평서문인 경우에는 (7)과 같이 'ich'를 포함한 모든 인칭대명사가 가능할 것이다. 따라서 평서문 종결어미 하나만으로는 생략된 주어를 추정하여 복원하기가 어렵다.

(6) ∅ 빨리 와 ! ⇔ (Du) Komm schnell her !
(7) ∅ 열심히 공부하고 있습니다 ⇔ Ich/Er/Sie/Es/Wir/Sie lern(en) fleißig

그러나 종결어미가 양상 정보를 나타내는 보조용언 등과 결합될 경우 생략된 주어를 복원하는데 결정적인 역할을 할 수 있다. 예를 들어 평서형 종결어미 '~ㄴ다'에 '의도'의 의미를 나타내는 보조용언 '~ㄹ까 하', '~기로 하', '~고자 하' 등이 결합되면 생략된 주어가 'ich'일 가능성이 매우 높다 (8).

(8) (나는/*너는/*그는/*그들은) 그냥 집에 갈까 해 ⇔ Vielleicht gehe ich einfach nach Hause

이와 같이 종결어미가 보조용언과 결합하여 문장유형을 결정하고 주어를 복원하는데 역할을 할 수 있는 예는 종결어미 '~어/아'의 경우에도 찾아볼 수 있다. 종결어미 '~어/아'는 의문형 종결어미 또는 명령형 종결어미로 사용될 수 있으므로 (9)에서 보는 바와 같이 문장부호를 사용하지 않을 경우 그 쓰임새가 모호하다.

(9) 점심 같이 먹어 ⇔ Gehen wir zusammen Mittag essen / Gehst du zusammen Mittag essen?

그러나 이와 같은 모호한 종결어미도 특정 보조용언과 결합될 경우 문장유

형의 모호성이 해소된다. 위의 종결어미 '~어/아'도 보조용언 '~어/아 버리'와 결합하면 그 쓰임새가 명령형 종결어미로만 한정될 가능성이 높다 (10).

(10) 가버려 ⇔ Geh weg / *Gehst du

본 연구에서는 이와 같이 종결어미와 보조용언의 결합형태가 생략된 주어를 복원하는데 결정적인 역할을 하기 때문에, 'Extra-ZP'의 복원을 위해 이 정보를 활용하고자 한다.

이 외에도 정진우/박종철(2009)의 연구는 '~어/아요'와 같이 평서형, 의문형, 명령형, 청유형 등으로 모두 사용이 가능한 모호한 종결어미의 모호성 해소방안을 다루고 있다. 정진우/박종철(2009)의 연구결과를 문장유형을 결정하는데 적극적으로 도입하여 생략된 주어를 복원한다면 보다 좋은 결과를 얻을 수 있을 것으로 기대된다. 그러나 본 연구의 주요 목적은 종결어미의 모호성을 해소하는 방안이 아니므로, 이에 대한 논의는 차후 연구로 미루기로 한다.

이은희(2009)와 유혜령(2010)은 보조용언과 종결어미의 결합형을 일종의 복합형 종결어미로 정의하고, 그 기능 및 의미에 따라 분류하였다. 이 중 화자의 의도, 계획, 의지, 추측, 판단, 의심, 아쉬움, 놀람, 약속, 다짐 등과 같은 의미를 나타내는 보조용언과 종결어미의 결합형은 구어체 문장에서 주어가 1인칭일 경우 주로 사용된다. 따라서 생략된 주어의 선행사는 'ich'로 복원하는 것이 타당하다. (표3 참조)

그 밖에 국어학에서 '해체', '하세요체', '해요체', '해라체', '해체' 등으로 분류되는 명령어미는 중의성이 없으므로 해당 문장을 명령문으로 간주하고 주어를 복원하지 않는다. 그 밖에 '~어/아 주겠니?', '~어/아 주실래요?', '~어/아 주시지 않을래요?' 등과 같은 의문형 복합종결어미는 형태상으로는 의문형이지만 명령문의 역할을 하므로 주어를 'du'나 'Sie'로 복원을 할 수도 있고 안할 수도 있다.

<표 3> 생략된 주어를 1인칭 대명사로 복원할 수 있는 '보조용언+종결어미' 유형

의미	보조용언 + 종결어미
의도, 계획, 의지	~ㄹ까 하다, ~기로 하다, ~ㄹ까 보다
추측	~ㄹ까 싶다, ~ㄴ가 하다, ~려니 하다, ~려니 싶다
판단	~ㄴ가 보다
의심	~냐 싶다, ~랴 싶다, ~지 싶다
아쉬움	~었어야 하다
약속, 다짐	~ㄹ께요

마지막으로 청유형 종결어미로 분류되어 주어가 'wir'로 복원되는 종결어미로는 '~자', '~세', '~ㅂ시다' 등이 있다.

'Extra-ZP'의 문제를 해결하는 또 하나의 방안은 형용사의 의미적 속성을 활용하는 것이다. 한국어의 형용사 중 주체의 심리상태를 나타내는 형용사는 주어가 1인칭일 때만 사용이 가능하다 (11~13).

(11) 나는 기쁘다 :: 그는 *기쁘다 / 기뻐한다
(12) 나는 반갑다 :: 그는 *반갑다 / 그는 반가워한다
(13) 나는 부럽다 :: 그는 *부럽다 / 그는 부러워한다

이와 관련하여 정연주(2010)의 연구에서는 형용사의 종류를 주관형용사 subjektive Adjektive와 객관형용사 objektive Adjektive로 나누었다. 이 연구에서는 '기쁘다', '무섭다', '부럽다' 등과 같은 주관형용사는 어미 '~어/아 하'와 결합하여 '기뻐하다', '무서워하다', '부러워하다' 등과 같은 객관형용사로 바뀌며, 이 경우 객관형용사가 서술하는 대상은 1인칭이 아닌, 2인칭과 3인칭일 수 있다고 본다. 주관형용사는 객관형용사와는 달리 화자의 심리적 상태를 기술하므로, 주관형용사가 사용된 문장의 주어는 1인칭 화자로 보는 것이 타당할 것이다.[3] 따라서 이와 같은 주관형용사가 사용되고 주어가 생략된 문장의 경우는 생략된 주어를 'ich'로 복원하는 것이 맞다. 이러한 주관형용사에는 '거슬리다', '기쁘다', '두렵다', '밉다', '반갑다', '벅차다', '부끄럽다', '부럽

3) 이원경(2006)의 연구에서도 이와 같은 주장은 뒷받침된다.

다', '싫다', '애달프다', '애타다', '즐겁다' 등이 있다.

그 외에도 'Extra-ZP'의 처리를 위해 높임 선어말 어미를 활용할 수 있다. 주체높임 선어말 어미 '~시'는 정상적인 대화상황에서 1인칭 대명사에 부착될 수 없다. 따라서 용언에 주체높임 선어말 어미가 사용된 경우, 생략된 주어의 후보로 1인칭 대명사는 제외된다.

(14) 아버지가 오신다
(15) ?내가 오신다

4.2. 'Inter-ZP' 처리방안

'Inter-ZP'는 선행사를 담화 내에서 찾을 수 있는 영형대명사를 말한다. 자연언어처리분야에서 이와 관련된 대다수의 연구는 Grosz & Sidner(1986)의 센터링 이론 Centering Theory을 기반으로 하여 이루어졌다. 센터링 이론에서는 담화의 구조가 언어구조, 의도구조, 초점구조와 같은 세 개의 층위로 이루어졌다고 본다. 이 중 초점구조가 영형대명사의 선행사를 탐색하는데 중요한 역할을 한다.

초점구조를 설명하는데 가장 중요한 요소는 센터 Center이다. 센터는 담화 속에 등장하는 논항 역할의 명사구들로 볼 수 있다. 이 명사구들은 잠재적으로 후속 담화에서 대명사로 지시될 수 있다. 명사구가 담화 속에 처음 사용되는 순간 그 명사구는 후속 대명사의 잠재적인 선행사가 될 수 있으며, 이러한 명사구들을 전향적 센터 forward-looking center라고 부른다. 이 중 문법기능 grammatische Funktion이나 주제화 Topikalisierung 등에 의해 표시되는 가장 현가성 Saliency이 높은 센터를 선호되는 센터 preferred center라고 하고, 다음 발화에서 대명사 혹은 영형대명사로 지시되는 센터를 후향적 센터 backward-looking center라고 부른다. 담화는 일반적으로 이 현가성이 가장 높은 센터에 대해 진행되며 이는 대명사로 나타나는 경우가 많다. 현가성을 결정하는 요인은 독일어와 영어 등의 경우는 주어, 목적어 등과 같은 문법기능의 위계성이며, 한국어나 일본어의 경우는 주제성이 문법기능에 우선하는 것으로 알려져 있다.[4] 차건회 외(1997)의 연구에서는 한국어 텍스트내에서 현가성을 결정하

기 위해 베이지안 Bayesian 확률공식을 사용하기도 하였다.

담화가 진행되면서 이 후향적 센터는 그대로 유지되거나 바뀔 수가 있다. 후향적 센터가 그대로 유지되는 전환 Transition 방식에는 '지속 Continuing'과 '전이 Retaining'가 있다. '지속'과 '전이'는 해당 문장의 후향적 센터가 앞 문장의 후향적 센터와 동일하다는 공통점이 있으나, 후향적 센터가 해당 문장에서 가장 현가성이 높은 센터('지속')이냐 아니냐('전이')는 점에서 차이가 있다.

(17) 나i 어제 여자친구 만났어 (fc: 나, 여자친구)
∅i 여자친구한테 점심도 사주고 영화도 보여줬지 (bc=pc=나) {지속}[5)]

(18) 나 어제 여자친구i 만났어 (fc: 나, 여자친구)
내가 ∅i 엄청 보고 싶었거든 (bc=여자친구≠pc) {전이}

해당 문장의 후향적 센터가 앞선 문장의 후향적 센터와 다를 수도 있다. 일반적으로 이러한 경우는 담화가 진행되면서 담화의 주제가 바뀌는 경우로 이해할 수 있는데, 이 경우 새로운 후향적 센터가 해당 문장에서 가장 현가성이 높은 요소일 경우는 '약전환 Smooth Transition'이라 부르고, 새 후향적 센터가 해당 문장에서 가장 현가성이 높지 않은 경우는 '강전환 Rough Transition'이라 부른다.

(19) 나i 어제 여자친구i 만났어 (fc: 나, 여자친구)
∅i 나한테 그만 만나자고 하더라 (bc=여자친구=pc) {약전환}

(20) 나 어제 여자친구i 만났어 (fc: 나, 여자친구)
철수도 우연히 그저께 ∅i 만났단다 (bc=여자친구≠pc) {강전환}

센터링 이론의 언어학적 설명력에도 불구하고, 기계번역에 센터링 이론을

---

4) 이익환/이민행(1999), 홍민표(2000), Lee/Lee(2000)
5) fc: forward-looking center
bc: backward-looking center
pc: preferred center

그대로 적용하여 영형대명사의 선행사를 찾아내는 일은 쉽지 않다. 그 첫째 이유로는 메모리 문제를 들 수 있다. 대부분의 기계번역시스템은 문장단위의 번역을 수행한다. 즉, 한 문장에 대한 번역작업이 완료되면 그 문장에 대한 정보는 메모리에서 사라지고 새로운 문장의 번역작업이 착수된다. 만약 혹시 등장할지 모르는 영형대명사의 선행사 탐색을 위해 이전 문장의 초점구조 정보를 폐기하지 않고 저장하고 있으면, 문장이 누적될수록 메모리 부담이 커지고 그만큼 처리속도에도 문제가 있다. 결국 메모리 문제로 인해 센터링 이론을 적용하기 위해서는 해당 문장의 앞선 몇 문장까지의 정보를 저장해야하는지 결정해야할 필요가 있다.

둘째, 대화체나 구어체 번역은 대개의 경우 실시간으로 수행되어야 한다. 그만큼 컴퓨터의 연산속도도 구어체 기계번역의 매우 중요한 요소이다. 센터링 이론을 그대로 구어체 기계번역에 수용할 경우 현가성이 가장 높은 센터를 문법지식이나 통계지식을 활용하여 계산해내야 하는 부담과 현가성이 높은 센터를 계산해내었더라도, 담화의 전환방식이 무엇인지를 계산해내야 하는 부담이 있다. 또한 기존의 센터링 이론 연구는 단문으로 구성된 텍스트를 대상으로 하였기 때문에, 실제 발화에서 많이 등장하는 복문을 처리하기에 적합하지 않은 면도 있다.

이와 관련하여 Okumura/Tamura(1996)는 일-영 기계번역에서 생략된 일본어 영형대명사의 선행사를 찾는 알고리즘을 제안하였다. 이들의 연구에 따르면 일본어의 경우 영형대명사의 선행사는 조사 대상 코퍼스의 95.1%의 경우 단문 기준으로 선행 2문장 내에서 찾을 수 있다. 또한 선행 4문장까지 고려하면 점차 성능이 향상되나, 그 이상을 넘어서까지 선행사를 탐색하게 되면 오히려 성능이 떨어지는 것으로 보고되었다. 또한 이들은 복문의 경우 단문단위로 분절하여 센터링 이론을 처리하였다.

본 연구에서는 영형대명사에 'Extra-ZP'와 관련된 규칙이 적용되지 않는 경우, 'Inter-ZP'일 가능성으로 보고 단순화된 센터링 이론을 적용하였다. 본 연구에서 제안하는 'Inter-ZP'처리 알고리즘은 다음과 같다.

'ZP'가 발견될 경우
i) 종결어미/동사 의미속성 등과 같은 'Extra-ZP'처리 규칙을 적용함

ii) i)이 적용되지 않을 경우 'ZP'가 등장한 Si의 바로 전 문장 Si-1의 담화지시체 (주제격, 주격, 목적격, 부사격 명사구)를 모두 추출한다
iii) Si-1에 명사구가 없을 경우 Si-2, Si-3, Si-4의 순서로 명사구를 탐색한다
iv) 현가성이 가장 높은 명사구를 ZP의 선행사로 결정한다
iv) 복문의 경우 단문 단위로 나누어 처리한다

위와 같이 단순화된 센터링 이론이 어느 정도의 정확률을 갖는지 알아보기 위해 간단한 실험을 수행하였다. 'Inter-ZP'가 총 46개 출현하는 46개의 문장에 대해 위의 알고리즘을 적용한 결과, 30개의 문장에서 선행사를 정확하게 찾아내어 65.2%의 정확률을 보였다. 총 16개의 오류는 10개의 오류는 선행사가 선행 4문장의 범위를 넘어서 등장하는 경우였고, 6개의 오류는 선행사가 앞선 4문장의 범위에서 등장하기는 하지만, 가장 현가성이 높은 성분이 아닌 경우, 즉, 강전환이나 약전환에 속하는 경우였다. 약 65%의 정확률은 일-영 기계번역과 관련된 연구에서 보고된 70% 정도의 정확률에 비하면 약간 떨어지는 수치이기는 하지만, 매우 간단한 처리만으로도 어느 정도 정확하고 빠르게 선행사를 찾아낼 수 있다는 점에서 실시간으로 진행되는 대화체 기계번역에 적용해볼만한 것으로 판단된다. 그러나 향후에는 고정된 현가성의 적용이 아닌 좀 더 다이내믹한 현가성의 계산 방안을 고려해야 할 것으로 보인다.

4.3. 'Intra-ZP' 처리방안

'Intra-ZP'는 선행사를 문장내의 다른 부분, 즉, 복문의 경우 앞선 단문에서 찾을 수 있는 영형대명사를 말한다. 본 연구에서는 'Intra-ZP'만을 위한 별도의 처리방안을 제안하는 것이 아니라, 복문도 단문으로 나누어 앞서 소개한 'Extra-ZP', 'Inter-ZP'처리 방안을 동일하게 적용한다.

추가적으로 우리는 본 연구에서 한국어 연결어미를 그 기능에 따라 세 가지 그룹으로 분류한 후, 주어를 공유하는 속성이 있는 연결어미는 이 특징을 활용하도록 한다. 첫 번째 유형의 연결어미는 선행절과 후행절의 주어가 일치하는 경향이 많은 연결어미이다. 여기에는 다음과 같은 연결어미가 속한다.

Klasse A
'~고', '~면서', '~며', '~려고', '~려면', '~느라고/느라', '~어서'

이 유형에 속하는 대표적인 연결어미 '~고'는 박소영(2000)의 연구에 따르면 '동작지속'과 '수단/방법'의 의미로 사용될 때 선행절과 후행절의 주어가 동일한 것이 일반적이다.

(21) 철수i는 새로 산 책가방을 메고 ∅i 학교에 갈 것이다
(22) 잠시 후 기차i가 석탄을 싣고 ∅i 도착한다.

'~고'가 '계기나열'의 의미로 사용될 때는 선행절과 후행절의 주어가 동일해야 하는 제약은 없다.

(23) 도둑이 도망가고 경찰이 집에 도착했다

그러나 이 경우는 선행절과 후행절의 주어가 명시되는 것이 일반적이므로, 우리는 연결어미 '~고'가 사용되고 후행절의 주어가 생략되었다면 선행절의 주어와 후행절의 주어가 동일한 것으로 볼 수 있을 것이다.
'~면서'의 경우도 '~고'의 경우와 유사하다. '~면서'가 '동시, 대립'의 의미로 사용될 경우는 선행절과 후행절의 주어가 동일한 것이 일반적이다.

(24) 영희i가 밥을 먹으면서 ∅i 신문을 본다
(25) 순희i가 턱을 괴며 ∅i 생각에 잠겼다

그러나 '~면서'가 '계기'의 의미로 사용되면, 선행절과 후행절의 주어가 동일해야 하는 제약은 없다.

(26) 시간이 지나면서, 비바람이 거세졌다.

이 경우도 '~고'가 '계기나열'의 의미로 사용될 때와 마찬가지로, '계기'의

의미로 사용되면 선행절과 후행절의 주어가 명시되는 것이 일반적이기 때문에, '~면서'가 사용된 복문에서 영형대명사가 사용되면 그 선행사는 선행절의 주어로 보아도 타당할 것이다.

이렇게 A부류에 속하는 연결어미들은 선행절의 주어와 생략된 후행절의 주어를 동일시하여 처리한다. 이와는 달리 B부류에 속하는 연결어미들은 선행절과 후행절의 주어에 대한 특별한 제약이 존재하지 않는다. 따라서 이러한 연결어미로 연결된 복문의 경우는 단문단위로 나누어 'Inter-ZP'의 처리를 위해 제안한 센터링 이론을 적용하게 된다.

Klasse B
'~길래', '~자', '~자마자'

연결어미의 마지막 부류는 주어가 항상 1인칭인 경우이다. 여기에는 유일하게 '~었더니'가 속한다.

Klasse C
'~었더니'

'~었더니'는 주어가 항상 1인칭 화자이어야 한다는 제약이 있다. 따라서 '~었더니'로 연결되는 복문에서 생략된 주어는 항상 '나', 즉 'ich'이다.

(27) 어제 동생이 무슨 숙제를 하나 ∅ 보았더니 그림일기였다
(28) 오랜만에 ∅ 어제 집에 갔더니 부모님이 반가워하셨다

4.4. 생략된 주어복원 알고리즘

우리는 4.1에서 4.3까지 생략된 주어대명사의 유형에 따라 주어를 복원하는 방법에 대해 살펴보았다. 여기서는 이 방법들을 모두 종합하여 기계번역 도중 생략된 주어가 발견될 경우 이의 순차적인 처리 프로세스에 대해 언급하도록 한다.[6]

6) 기계번역 등과 같은 자연언어처리분야에서 문장성분의 생략여부를 정확하게 파악

먼저 생략된 주어가 있을 경우, 번역시스템은 이 영형대명사가 'Extra-ZP' 인지 'Inter-ZP'인지 'Intra-ZP'인지 알 수 없다. 따라서 먼저 4.1에서 제안한 'Extra-ZP'의 해결방안을 먼저 적용한다. 즉, '보조용언+연결어미' 정보와 감정형용사 정보 등을 적용하여 생략된 주어를 복원한다('Extra-ZP' 처리방안). 만약 이 방법이 적용되지 못하면, 다음 단계로 넘어간다.

만약 해당 문장이 복문이라면 복문을 구성하는 연결어미의 유형을 파악한다. 연결어미가 4.3에서 제안한 A부류나 C부류에 속하면 연결어미의 주어공유 관련 속성을 적용하여 생략된 주어를 복원한다('Intra-ZP' 처리방안). 만약 연결어미가 A나 C부류에도 속하지 않는다면, 제안한 단순화된 센터링 이론을 적용한다('Inter-ZP' 처리방안). 해당문장이 복문이 아니라면 다음 단계로 넘어간다.

입력문이 단문인 경우 'Extra-ZP'의 해결을 위한 규칙이 적용되지 않는다면, 단순화된 센터링 이론을 적용하여 생략된 주어를 복원한다. 이 모든 규칙이 적용되지 않는다면 평서문의 경우 'ich', 의문문의 경우 'Sie', 명령문인 경우 'Sie'로 주어를 생성한다.

다음 장에서는 이 연구에서 제안하는 방법론의 성능평가 결과를 소개한다.

5. 실험 및 결과분석

본 연구에서 제안하는 주어생략 처리방안의 성능평가를 위해 실험을 수행하였다. 실험 대상 문장은 한국 드라마 대본이었으며, 총 300문장을 '시크릿 가든', '장난스런 키스'라는 드라마의 대본에서 발췌하였다. 이 300문장에는 총 207개의 영형주어대명사가 사용되었고, 2명의 대학원생이 공동으로 생략된 주어를 복원하여 정답셋으로 사용하였다.

본 실험에서는 첫째, 문장유형만으로 주어를 복원하는 디폴트 방안[7], 둘째,

하는 것 또한 생략된 성분을 복원하는 것 못지 않게 어려운 작업이다. 여기서는 입력문에서 주어가 생략된 것이 정확하게 파악된 것으로 가정하고 논의를 진행하도록 한다.

7) 본 연구에서 활용한 문장유형 분류방식은 단순히 문장부호만을 활용하는 것이다.

‘Extra-ZP’의 처리방안 중 ‘보조용언+종결어미’ 정보를 활용하며, 디폴트 방안을 백업으로 활용하는 방안, 셋째, ‘보조용언+종결어미’ 정보와 ‘감정형용사’ 정보를 활용하며, 디폴트 방안을 백업으로 활용하는 방안, 넷째 ‘보조용언+종결어미’ 정보와 ‘감정형용사’ 정보 및 접속사 정보를 활용하며, 디폴트 방안을 백업으로 활용하는 방안, 마지막으로 본 연구에서 제안하는 최종 방안인 ‘보조용언+종결어미’ 정보와 ‘감정형용사’ 정보 및 접속사 정보와 센터링 이론을 적용하며, 디폴트 방안을 백업으로 활용하는 방안의 성능을 서로 비교하였다. 각 방안별로 단순히 문장유형만을 고려하여 생략된 주어를 복원하는 방안인 디폴트 방안을 백업으로 활용한 이유는 제안한 규칙의 커버리지가 떨어져 재현률 Recall이 낮을 수 있는 문제를 해결하기 위해서이다.

각 방안별 성능은 정확률 Precision, 재현률 Recall 그리고 정확률과 재현률을 모두 고려한 ‘F-measure’를 활용하여 평가되었다.

- 정확률 = (올바른 영형 주어대명사 / 복원한 영형 주어대명사) * 100
- 재현률 = (복원한 영형 주어대명사 / 전체 영형 주어대명사) * 100
- F-measure = 2*Precision*Recall/(Precision+Recall)

성능 평가에서 우리는 표4와 같은 결과를 얻었다. 평가의 베이스라인으로 활용된 디폴트 주어복원 방안은 총 74.47%의 재현률을 보였으나, 정확률 측면에서는 46.41%에 그쳤다. 여기에 보조용언을 포함한 종결어미 정보를 반영할 경우 재현률은 77.29%로 베이스라인 대비 약 2.9% 상승하며 정확률은 약 16.1%가 상승하였다. 여기에 감정형용사 정보를 추가적으로 반영할 경우, 재현률과 정확률 측면에서 각각 약 0.5%와 1.5%의 상승효과를 보였다. ‘Intra-ZP’의 처리를 위해 연결어미 정보를 추가적으로 고려한 경우 약 6.3% 정도의 재현률 상승효과와 1% 가량의 정확률 상승효과를 얻었다. 최종적으로 본 연구에서 제안한 방법을 모두 고려한 경우, 베이스라인 대비 정확률은 약 27%, 재현률은 11%의 상승효과가 있었으며, ‘F-measure’의 경우에도 약 0.21

마침표로 끝나는 문장은 평서문으로 분류하고 주어를 ‘ich’로 복원하며, 물음표로 끝나는 문장은 의문문으로 분류하며, 주어를 ‘Sie’로 복원하고, 느낌표로 끝나는 문장은 명령문으로 분류하여 주어를 ‘Sie’로 복원하였다.

의 상승효과가 있었다.

<표 4> 주어복원 성능평가 결과

	디폴트[8]	종결어미+디폴트	종결어미+감정형용사+디폴트	종결어미+감정형용사+연결어미+디폴트	종결어미+감정형용사+연결어미+센터링 이론+디폴트
정확률	46.41%	62.50%	63.98%	64.94%	73.29%
재현률	74.40%	77.29%	77.78%	84.06%	85.02%
F-measure	0.572	0.691	0.702	0.733	0.787

결과를 분석해보면 디폴트 주어생성방안과 비교하여 종결어미와 보조용언 정보를 추가하였을 때의 정확률 상승폭이 가장 컸으며, 감정형용사와 연결어미 정보는 정확률 향상의 측면에서는 그리 큰 기여를 하지는 못하는 것으로 나타났다. 그러나 종결어미 정보, 감정형용사 정보, 연결어미 정보를 활용한 규칙이 적용되지 못한 경우에 디폴트 방안을 적용하는 대신 센터링 이론을 적용한 결과 재현률 자체는 큰 변화가 없었지만 디폴트 대비로는 정확률이 약 27% 상승하였으며, 종결어미와 감정형용사, 연결어미 및 디폴트를 활용한 방법론과 비교하여서도 약 8% 이상의 정확률 향상효과를 거두었다.

본 실험을 통해 현재 한국어를 출발어로 하는 한-독 혹은 한-영 대화체 기계번역 시스템에서 대부분 활용하고 있는 문장유형 기반의 영형주어대명사 복원 방안은 종결어미와 보조용언 정보만을 활용하더라도 정확률을 대폭 향상시킬 수 있으며, 센터링 이론까지 적용할 경우 정확률 및 재현률을 크게 향상시켜서 궁극적으로는 번역률의 향상에 기여할 수 있음을 보았다.

6. 결론

본 연구에서는 한국어를 출발어로 하여 독일어나 영어와 같은 주어지향 언

8) 평가의 베이스라인을 의미함

어로 기계번역을 수행할 때 큰 문제가 되는 주어생략현상을 다루었다. 일반적으로 언어학에서 문장성분의 생략현상은 담화구조 또는 정보구조와 같은 담화의 글로벌 구조에 대한 접근 없이는 설명이 어려운 것으로 알려져 있다. 그러나 기계번역 분야에서는 이와 같은 담화구조나 정보구조에 대한 접근이 어렵기 때문에 최소한의 리소스만을 사용하여 생략현상을 다루는 것이 불가피하다.

우리는 본 연구에서 주어생략현상을 생략된 주어의 선행사를 어디서 찾을 수 있느냐에 따라 'Extra-ZP', 'Inter-ZP', 'Intra-ZP로 나눌 수 있음을 보였다. 생략현상에 대한 기존의 연구는 주로 'Inter-ZP'현상의 처리에 집중되었는데, 코퍼스 분석결과 드라마 대본과 같은 대화체 텍스트에서는 오히려 'Extra-ZP'가 훨씬 빈번히 나타나기 때문에, 이 유형의 영형주어대명사 처리가 더 중요함을 보았다. 이를 위해 보조용언과 종결어미 정보를 활용한 주어복원 방안을 제안하였다. 보조용언과 종결어미 정보는 형태소 분석과 품사태깅 단계에서 획득할 수 있는 정보로서, 담화구조나 정보구조 등의 정보에 비해 매우 간단한 분석만으로도 얻을 수 있기 때문에 본 연구에서 제안한 방법론은 적용하기가 매우 간단하다는 장점이 있다. 그 밖에 일부 감정형용사는 의미적 특성상 항상 1인칭 주어만을 취한다는 점을 고려한 주어복원 방안도 제안하였다.

'Extra-ZP'의 처리를 위해서는 기존의 센터링 이론을 단순화한 방법론을 제안하였다. 기존의 센터링 이론을 그대로 기계번역 시스템에 적용하기는 컴퓨터 프로세싱 측면에서 메모리 문제 등과 같은 어려운 점이 있다. 이의 보완을 위해 본 연구에서는 앞선 4개의 문장에 대한 정보만을 저장하여 선행사를 탐색하는 방법론을 제안하였다. 그러나 선호되는 센터를 결정할 때 무조건 현가성이 가장 높은 문장성분으로 결정하는 방법은 향후 좀 더 개선해야할 내용이다. 그 밖에 'Intra-ZP'의 처리를 위해서 한국어의 연결어미를 3가지로 구분하였으며, 연결어미의 특성에 따라 선행절의 주어를 후행절의 주어와 동일시할 수 있음도 보였다.

본 연구결과의 실용화를 위해 좀 더 개선해야할 부분은 앞서 언급한 센터링 이론의 적용시 선호되는 센터를 상황에 맞게 계산하는 방법을 찾아내는 것이다. 또한 통계기반 한-독 기계번역시스템에 본 연구에서 제안한 방법론을 어떻게 통합하여 적용할 것인가도 중요한 향후 연구테마 중 하나일 것이다.

참고문헌

박소영 (2000): 양태의 연결어미 '-고'에 대한 연구. 한국언어학회,『언어학』26, 167-198.

박재연 (2003): 국어 양태의 화·청자 지향성과 주어 지향성. 국어학회,『국어학』41, 249-275.

안주호 (2006): 현대국어 '싶다' 구문의 문법적 특징과 형성과정. 한국어의미학회,『한국어의 미학』20, 371-391.

유혜령 (2010): 국어의 형태,통사적 공손 표지에 대한 연구. 청람어문교육학회,『청람어문교육』41, 377-409.

이명희 (2010): 한국어와 중국어의 요청 화행 대조 연구. 이중언어학회,『이중언어학』42, 103-134.

이원경 (2006): 감정동사의 분류와 특성분석.『담화와 인지』13-1, 163-182.

이은희 (2009): 한국어 교육을 위한 명령형 어미 연구 -사용 실태 분석을 통한 교육 내용 구성 관점에서-. 청람어문교육학회,『청람어문교육』40, 단일호, 71-95.

이익환/이민행 (1999): 한국어 대화에서의 대명사의 선행사 탐색. 제11회 한글 및 한국어 정보처리 학술대회 학술지, 382-388.

정연주(2010): "-어 하-"와 통합하는 객관형용사의 의미 특성.『의미학』33, 297-320.

정진우/박종철 (2009): 형태소 분석을 통한 한국어 문장 유형 자동 분류. 한국언어정보학회,『언어와정보』13-2, 59-97.

차건회/송도규/박재득 (1997): 한국어 대용과 생략 해결을 위한 센터링 이론의 적용. 제9회 한글 및 한국어 정보처리 학술대회 학술지, 347-352.

한송화 (2000): 한국어 보조용언의 상적 기능과 양태기능, 화행적 기능에 대한 연구 :'하다'를 중심으로. 국제한국어교육학회,『한국어교육』11-2, 189-209.

홍민표 (2000): 센터링 이론과 대화체에서의 논항 생략 현상. 한국인지과학회,『인지과학』11, 9-24.

홍윤기/서희정 (2009): 종결어미-용언 긴밀구성의 정도성 연구. 한국어교육학회,『한국어교육학』128, 525-554.

Grosz, B./Sidner L. (1986): Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12, 175-204.

Kameyama, M. (1986): A Property-Sharing Constraint in Centering. *Proc. of the 24th Annual Meeting of the Association for Computational Linguistics*,

200-206.

Lee, I.-H./Lee, M.-H. (2000): Anaphora Resolution and Discourse Structure: A Controlled Information Packaging Approach. *Language and Information 4-1*, 67-82.

Nakaiwa, H./Yokoo, A./Ikehara, S. (1994): A System of Verbal Semantic Attributes Focused on the Syntactic Correspondence between Japanese and English. *Proc. of COLING '94*, 672-678.

Nakaiwa, H./Ikehara, S. (1995) Intrasentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. *Proc. of The 6th TMI*, 96-105.

Nariyama, S. (2002): Grammar for ellipsis resolution in Japanese. *Proceedings of the 9th International Conference on Theoretical and Methodological Issuesin Machine Translation*, 135-145.

Okumura, M./Tamura, K. (1996): Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory. *Proceeding of the 16th conference on Computational linguistics(COLING) - Volume 2*.

Walker, M. A./Iida, M./Cote, S. (1994) Japanese Discourse and the Process of Centering. *Computational Lintuistics, Volume 20, Number 2*, 193-232.

Zusammenfassung

Subjektellipse im gesprochenen Koreanischen und deren Behandlung für die maschinelle Übersetzung ins Deutsche

Hong, Munpyo (Sungkyunkwan Univ.)

In der vorliegenden Arbeit wurde das Subjekt-Ellipse Phänomen in der maschinellen Übersetzung des Koreanischen ins Deutsche behandelt. Bei der Übersetzung von einer sogenannten topik-orientierten Sprache wie dem Koreanischen in eine sogenannte subjekt-orientierte Sprache wie das Deutsche werden die Satzteile, insbesondere die Subjekte oft ausgelassen. In den meisten bisherigen Untersuchungen über das Problem wurden die Versuche unternommen, die z.B. die globale Information eines Textes wie die Diskursstruktur heranziehen. Der bekannteste Versuch in dieser Kategorie ist die

‘Centering Theorie’ von Grosz&Sidner(1986). Diese Theorie ist zwar geeignet für die linguistische Erklärung des Phänomens aber kann nur einen Teil des Phänomens erklären.

In der Korpus-Analyse wurde herausgestellt, dass das meiste Subjekt-Ellipse Phänomen in dem gesprochenen Koreanischen das sogenannte ‘Extra-ZP’ aufweist. Unter dem ‘Extra-ZP’ versteht man ein Zero-Pronomen, dessen Antezedenz man ausserhalb des Diskurses finden oder vermuten kann. Die ‘Centering-Theorie’ kann dieses Phänomen nicht erklären, da die Antezedenz nicht im Diskurs oder im Text vorkommt. Dafür stellten wir eine alternative Methode vor, die sich hauptsächlich auf die morphologische Information des Hauptverbs des Inputsatzes stützt. Dazu können die Semantik der bestimmten Adjektive zur Resolution des ausgelassenen Pronomens dienen.

Für die Behandlung des ‘Inter-ZP’ schlugen wir eine vereinfachte ‘Centering-Theorie’ vor. Um das Memory-Problem bei der Anwendung der ‘Centering-Theorie’ zu lösen, wurde die Anzahl des vorherigen Satzes, in dem die Antezedenz gesucht werden soll, auf 4 begrenzt.

Das Experiment zeigte, dass unsere Methode im Vergleich zu den bisherigen Methoden einfach anzuwenden ist und relativ hohe Korrektheit aufweist.

[검색어] 주어생략, 대화체 자동번역, 한독 기계번역, 센터링 이론
Subjekt-Ellipse, Maschinelle Übersetzung der gesprochenen Sprache, Koreanisch-Deutsch maschinelle Übersetzung, Centering-Theorie

홍문표 110-745 서울시 종로구 명륜동 3가 53
성균관대학교 인문대학 독어독문학과
skkhmp@skku.edu

논문접수일: 2011.10.30
논문심사일: 2011.11.30
게재확정일: 2011.12.12