한-독 기계번역을 위한 한국어 영형 주어 처리연구*

박 아 름, 홍 문 표** (성균관대)

I. 서론

대용어 Anaphern 현상은 자연언어처리에서 다루기 어려운 현상 중 하나이다. Hirst(1981)에 따르면 의미해석을 위해 다른 언어표현의 해석을 필요로 하는 특정 개체를 대용어 Anapher, 그리고 그 개체가 지시하는 것을 선행사 Antezedenz라고 칭한다.

(1) [Maria]$_i$ aß [eine Banane]$_j$, weil sie$_i$ hungrig war.

예문 (1)에서 독일어 대용어 'sie'가 지시하는 개체는 앞 절에 등장하는 명사구 'Maria'와 'Banane' 중 'Maria'이므로 'Maria'가 'sie'의 선행사이다. 이렇듯한 대명사의 선행사를 결정하는 과정을 대용어 해소 Anapherresolution라고 부르는데, 이러한 대용어 해소는 기계 번역 machinelle Übersetzung과 같은 자연언어처리 응용 분야에서 매우 중요한 과제이다.

기계 번역의 관점에서 출발언어 Quellsprache의 대명사를 목표언어 Zielsprache로 올바르게 생성하는 것은 아주 중요하다(vgl. Mitkov & Schmidt 1998). 특히 독일어처럼 대명사에 성이 표시되는 언어가 목표언어일 경우, 특정 대명사의 선행사가 무엇인지에 따라 독일어 대명사의 성이 달라진다.

Hutchins & Somers(1992)에서 가져온 영어를 출발언어로 그리고 독일어를 목표언어로 할 때의 예시를 살펴보자.

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

(2) [The monkey]$_i$ ate the banana because it$_i$ was hungry.
(3) The monkey ate [the banana]$_i$ because it$_i$ was ripe.

위의 각 문장에서 영어 'it'이라는 대명사가 지시하는 것이 다르다. 문장 (2)에서는 'it'이 '원숭이'를 가리키며, 문장 (3)에서는 '바나나'를 지시한다. 만약 이 문장들을 독일어로 번역한다면, 독일어에서는 대명사가 선행사의 성을 취하기 때문에 다음과 같은 대명사로 번역될 것이다.

(2') er (선행사가 남성 명사인 'der Affe')
(3') sie (선행사가 여성 명사인 'die Banane')

대부분의 기계 번역 시스템은 이러한 대용어 해소를 다루지 않으며 대용어 해소와 같이 문장 층위를 넘어서서 처리해야 하는 현상에 대해서는 기계 번역 시스템의 종류에 상관없이 올바른 번역 결과를 얻을 수 없다. 예시 (4)와 (5)는 규칙기반 기계 번역 시스템인 시스트란 Systran 번역기와 통계기반 기계 번역 시스템인 구글 Google 번역기를 사용하여 예시 (2)와 (3)을 자동 번역한 결과이다.

	영어 문장	독일어 번역 결과 1 (Systran)	독일어 번역 결과 2 (Google)
(4)	The monkey ate the banana because it was hungry.	Der Affe aß die Banane, weil sie Hunger hatte.	Der Affe die Banane gegessen, denn es war hungrig.
(5)	The monkey ate the banana because it was ripe.	Der Affe aß die Banane, weil sie reif war.	Der Affe die Banane gegessen, denn es war reif.

<표 1> 영어→독일어 언어쌍에 대한 기계 번역 시스템 번역 결과 예시

(4)번과 (5)번 예시를 살펴보면 규칙기반 기계 번역 시스템에서는 영어 대용어 'it'을 모두 독일어 여성 대명사 'sie'로 생성하였고, 통계기반 기계 번역 시스템에서는 이를 독일어 중성 대명사 'es'로 생성하였다. 이는 앞서 말한 현재 상용화 되어 있는 기계 번역시스템의 한계를 보여주는 것으로서, 기계 번역 시

스템의 종류에 상관없이 영어의 대명사에 대한 독일어 대명사의 기본값을 설정해놓고 이를 그대로 생성하는 것으로 보여진다.

본 논문에서는 한국어를 출발언어로 하고 독일어를 목표언어로 하는 기계번역 시스템에서 반드시 처리되어야만 하는 한국어 영형 주어 Null-Subjekt를 해소하는 방법을 제안하고자 한다. 앞서 살펴본 영어를 출발언어로 하고 독일어를 목표언어로 하는 경우와는 달리, 한국어를 출발언어로 하는 경우 문제가 되는 대용어 현상은 영형 주어이다. 영형 주어는 영형 대명사와 관련이 있는 현상인데 영형 대명사란 생략된 논항의 통사적 범주가 대명사라는 문법적 요소인 것을 의미하며 이러한 영형 대명사가 주어 위치에 등장할 때 이를 영형 주어라고 부른다. 한국어나 일본어와 같은 주제 지향 언어 Topik-orientierte Sprache에서는 주어 대명사가 종종 생략되지만 한국어의 영형 주어는 독일어처럼 선행사의 성이나 수와 같은 정보를 포함하고 있지 않기 때문에 독일어 대용어 해소와는 다른 방식의 접근법이 필요하다. 한국어 영형 주어 해소를 위해 본 논문에서는 그 동안 다양한 자연언어처리 응용 분야에서 분류 문제를 성공적으로 해결하는데 사용돼 온 기계 학습 maschinelles Lernen방법을 사용하고자 한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 한→독 기계 번역 시스템에서 문제가 되는 한국어 영형 주어 현상에 대해 자세히 살펴볼 것이다. Ⅲ장에서는 대용어 해소에 관한 기존 연구를 소개하면서 본 연구와의 차이점을 설명하고자 한다. Ⅳ장에서는 기계 학습을 위해 본 논문에서 제안한 12개의 자질에 대해 상세히 소개한다. Ⅴ장에서는 Ⅳ장에서 제안한 자질을 사용한 기계학습 방법이 한국어 영형 주어를 해소하는데 얼마나 효과적인지를 알아보기 위해 수행한 실험과 그 결과에 대해 제시한다. 마지막으로 Ⅵ장에서는 결론을 내리고 향후 연구에 대해 언급할 것이다.

Ⅱ. 한국어 영형주어현상

한국어나 일본어와 같은 주제 지향 언어 Topik-orientierte Sprache에서 영어, 독일어와 같은 주어 지향 언어 Subjekt-orientierte Sprache로의 번역 시 영형 주

어와 같은 영형 대명사 현상이 문제가 된다. 왜냐하면 한국어에서 주어가 생략되었더라도 독일어에서는 이를 필수적으로 복원해야 하기 때문이다.

	한국어 문장	독일어 문장
(6)	마리아는 배가 고팠기 때문에 * 바나나를 먹었다.	[Maria]$_i$ aß [eine Banane]$_j$, weil sie$_i$ hungrig war.
(7)	원숭이는 배가 고팠기 때문에 * 바나나를 먹었다.	[Der Affe]$_i$ aß [eine Banane]$_j$, weil er$_i$ hungrig war.

<표 2> 한국어→독일어 기계 번역시 문제가 되는 한국어 영형 주어 현상 예시

예문 (6)과 (7)을 살펴보면 한국어에서 이유를 나타내는 부사절에 등장한 주어가 주절에서는 생략되어 있으며, 이는 '*'로 표시되어 있다. 한국어 문장을 독일어로 번역할 때에는 한국어의 생략된 영형 주어가 반드시 생성되어야 한다. 이 때 독일어 선행사의 성과 수에 따라 적절한 대명사가 선택되어야 한다. 만약 한국어의 영형 주어가 복원되지 않는다면 독일어 번역 결과는 비문이 될 것이다. 따라서 한국어 영형 주어의 해소가 한국어를 출발 언어로 독일어를 목표 언어로 하는 기계 번역 시스템에서 필수적이다.

본 연구에서는 한국어 대화체 문장을 연구 대상으로 삼는데, 노대규(1996)에 따르면 문어체 문장과 대화체 문장을 비교할 때, 주어와 목적어의 생략 현상, 주제화 현상 및 이동현상이 대화체에서 더 빈번히 일어난다고 한다. 그럼에도 불구하고 한국어 대화체 문장에서 한국어 영형 주어를 복원할 수 있는 정보들이 종종 문장 내에 등장한다.

우선 한국어의 문장 유형을 나타내는 동사의 형태론적 정보가 한국어 영형 주어 해소를 위해 사용될 수 있다. 예를 들어 명령문, 청유문과 같은 한국어의 문장 유형을 나타내는 동사의 종결어미가 존재하는데, 이 정보를 활용하면 한국어 문장에서 생략된 주어를 복원할 수 있다. 명령문의 경우 '해체', '하세요체', '해요체', '해라체', '해체' 등과 같은 종결어미가 사용되며 문장 내에 이러한 어미가 등장하면 명령의 화행을 수행하므로 생략된 주어는 '당신' 또는 '너'이다.

(8) * 비행기 번호를 *말씀해주세요.*
(9) * 조심히 *들어가십시오.*

예문 (8)과 (9)는 본 논문에서 구축한 코퍼스의 일부인데, 명령문 어미가 사용되었으므로, 명령문 어미를 통해 생략된 주어가 '당신'이라는 것을 알 수 있다.

이렇게 동사의 종결어미에 따라 문장의 유형이 결정되고, 표면적인 정보를 통해 영형 주어를 복원할 수 있지만 문장의 유형과 그 문장의 화행이 항상 일치하는 것은 아니다. 특정 의문형 종결어미들은 의문문이지만 명령의 화행을 수행한다. 예를 들어, '~어/아 주겠니(줄래)?', '~어/아 주실래요?', '~어/아 주시지 않을래요?' 등의 의문형 종결어미들이 실제로 요구하는 화행은 명령이기 때문에 앞서 본 명령형 종결어미와 마찬가지로 생략된 한국어 주어는 예문 대화 상황 내의 화자인 '당신' 또는 '너'이다. 예문 (10)과 (11)에서는 화자 존칭 어미 '시'가 등장하므로 생략된 한국어 주어는 '당신'이다.

(10) * 캬라멜 마끼아또 한 잔 *주실래요*?
(11) * 샘플은 선크림으로 많이 좀 *챙겨주실래요*?

동사의 종결어미 외에도 한국어 문장에 등장하는 주관형용사가 문장의 생략된 주어에 대한 단서를 제공한다. 주관형용사란 주어의 의미역이 경험주인 형용사 부류인데, 여기에는 경험주의 심리 상태, 감각 상태 그리고 경험주의 판단을 나타내는 형용사가 포함된다. 이러한 이유로 주관형용사가 등장하는 문장이 평서문일 때에 주어가 화자(1인칭)와 일치해야 한다는 제약이 있다. 예를 들어, '즐겁다', '기쁘다'와 같은 소위 심리 형용사가 평서문에 등장하는 경우 예문 (12)와 (13)에서 알 수 있듯이 생략된 주어가 화자인 '나'일 수밖에 없다.

(12) * 다시 만나게 되어 * 정말 *기뻐*.
(13) * 이런 말을 해서 * *미안해*.

주관형용사에 '~어/아 하'와 같은 어미가 부착되면 주관형용사가 객관형용사로 바뀌게 되는데, 정연주(2010)에 따르면 이러한 경우 생략된 주어가 1인칭이 아닌 2인칭 또는 3인칭이 된다는 제약이 생긴다. 따라서 예문 (14)와 (15)처럼 주관형용사에 특정 어미가 부착되어 생성된 객관형용사가 문장 내에 등장한다면 이 또한 한국어 영형 주어를 복원하는데 도움이 되는 정보가 될 수 있다. 이 예문들에서 생략된 주어는 '나' 외의 주어일 것이다.

(14) * 엄청 *기뻐했어.*
(15) * 그 소식을 듣고 * 너무 *슬퍼했어.*

본 논문에서는 한국어 대화체 문장에 등장하는 다양한 현상들을 고려하여 한국어 영형 주어를 복원할 수 있는 정보들을 기계 학습을 위한 자질로 삼는다. 그 외에도 센터링 이론과 같은 담화 응집성과 관련된 정보를 반영한 자질도 제안할 것이다. 이와 관련된 자세한 내용은 Ⅳ장에서 다루도록 하겠다.

Ⅲ. 기존 연구

영형 주어와 같은 영형 대명사를 해소하기 위한 연구들은 한국어와 유사한 현상이 많이 존재하는 일본어를 중심으로 진행되어 왔다(vgl. Nakaiwa & Ikehara 1995; Nakaiwa & Shirai 1996). 이러한 연구들은 모두 대용어 해소와 관련한 연구에 기반을 하고 있는데, 대용어 해소에 대한 연구들은 1970년대부터 영어를 중심으로 수행되었다. 대용어 현상을 처리하는데 있어 언어에 상관없이 유사한 전략이 사용되어 왔다. 대표적인 방법론은 문장에 드러나는 언어학적인 정보를 활용하는 것으로 관련 연구들에서는 제약 Constraint과 선호도 Präferenz를 사용한다(vgl. Baldwin 1997; Lappin & Leass 1994; Carbonell & Brown 1988).

제약은 가능한 선행사 후보들을 제거하는 역할을 하는데 이는 절대적인 것으로 간주되고, 선호도는 상대적인 기준으로서 대부분 휴리스틱 규칙의 형태로 제안된다. 우선 선행사 후보들에 제약이 적용된 이후 남아있는 선행사 후보들이 존재하면 선호도는 이들 사이의 우선순위를 결정해주는 역할을 한다.

일본어 대용어 해소 모델에 대한 연구들은 이러한 제약과 선호도를 사용하며, 특히 일본어의 양상 표현, 동사의 의미 자질 또는 접속사를 제약으로 사용한다(vgl. Nakaiwa & Ikehara 1995; Nakaiwa & Shirai 1996). 본 연구에서도 이 연구들을 비판 수용하였으나 이들의 연구는 문어체 코퍼스를 대상으로 하며, 또한 기계 학습 방법을 사용하지 않기 때문에 본 연구의 제안과 차이가 있다.

이 외에도 대용어의 선행사를 확인하는데 주로 사용되는 것은 담화 구조이다. 특히 담화의 국부적 응집력 lokale Kohärenz을 유지하기 위해서는 담화 내

에서 화자가 집중하는 대상인 센터 center 또는 초점 Fokus이 결정적인 역할을 하는데, 센터링 이론 centering theory은 이러한 국부적인 응집력이 어떻게 유지되는지 분석하기 위해 제안된 이론이다(vgl. Grosz et al. 1983; 1995). Okumura & Tamura(1996)는 일본어 영형 대명사의 해소를 위해 센터링 이론을 적용하였으며, Kameyama(1986)와 Walker et al.(1994)은 일본어를 위한 전향적 센터의 순위를 제안했다.

홍문표(2011)에서는 센터링 이론을 단순화한 알고리즘을 제안하여 선행사가 담화 내 존재할 때의 영형 주어를 해소하고자 하였다. 이 연구에서는 Okumura & Tamura(1996)의 제안에 따라 선행사 후보 검색 범위를 영형 주어가 있는 문장(S_i)을 선행하는 앞의 4문장(S_i-1, S_i-2, S_i-3, S_i-4)까지로 지정하였다. 그 후 지정한 범위 내에서 담화지시체(주제격, 주격, 목적격, 부사격 명사구)를 추출한 후 그 후보들 중 현가성 salient이 가장 높은 명사구를 선행사로 결정하였다. 홍문표(2011)에서는 이 알고리즘을 적용하여 65.2%의 정확률을 보고하였다. 그러나 모든 담화에서 현가성이 가장 높은 개체가 선행사가 되는 것은 아니기 때문에 이 알고리즘을 그대로 적용하기에는 한계가 있다.

현재의 대용어 처리 방법론들은 대부분 앞서 본 제약과 선호도 휴리스틱 규칙을 사용하며 형태-구문론적 정보 또는 의미 분석을 사용한다. 대용어 해소를 위한 이러한 방법들은 결정론적 방법론 deterministischer Ansatz으로서 특정한 입력이 주어진 조건에 맞으면 항상 같은 출력이 나오도록 한다. 그러나 결정론적 방법은 주어진 조건이 적용이 안 되는 경우, 조건이 적용되었으나 선행사 후보가 두 개 이상 남는 경우 그리고 조건에 맞으나 제안된 출력이 정답이 아닌 경우에 적절한 선행사를 선택할 수 없다는 문제가 있다.

따라서 대용어 해소를 위해 비결정론적 방법론 nicht-deterministischer Ansatz인 기계 학습 방법을 사용하기도 한다(vgl. Connoly et al. 1994; Kazuhide & Eiichiro 1998). 기계학습은 문맥에 따라 확률적으로 가장 가능성이 높은 후보를 추천하는 방법이므로 제약규칙과 선호도 규칙의 결정론적 방법론의 한계를 극복 가능하다. 본 연구에서도 기존 연구의 한계점을 극복하기 위해 기계 학습 방법론을 도입하여 한국어 대화체에 등장하는 영형 주어의 선행사를 결정하고자 한다.

Ⅳ. 기계 학습을 위한 자질 제안

앞서 Ⅱ장에서 살펴보았듯이 한국어의 영형 주어를 복원하기 위한 여러 가지 단서가 문장 또는 텍스트 내에 존재한다. 우리는 한국어 대화체의 특성을 고려하여 기계 학습을 위한 총 12개의 자질을 제안하였으며 제안한 자질은 표 3과 같다. 각 자질을 제안한 이유는 이후에 차례대로 설명하도록 하겠다.

feature	value (binary)	내용
f1	0/1	명령형 종결어미 등장 유무
f2	0/1	의문형 복합종결어미 등장 유무
f3	0/1	청유형 종결어미 등장 유무
f4	0/1	주관형용사 등장 유무 (평서문)
f5	0/1	객관형용사 등장 유무 (평서문)
f6	0/1	복합형 종결어미 등장 유무 (평서문)
f7	0/1	높임 선어말 어미 '~시' 등장 유무
f8	0/1	주관형용사 + '~십니까?'/ '~세요?'
f9	0/1	주제격 선행사(~은/는/도/만)가 윈도우 범위 4문장 이내 등장 유무
f10	0/1	주격 선행사(~이/가)가 윈도우 범위 4문장 이내 등장 유무
f11	0/1	목적격 선행사(~을/를/에게)가 윈도우 범위 4문장 이내 등장 유무
f12	0/1	부사격 선행사(~에서/으로 등)가 윈도우 범위 4문장 이내 등장 유무

<표 3> 기계 학습을 위해 제안한 12개의 자질

우선 한국어의 문장 유형을 나타내는 동사의 형태론적 정보가 영형 주어를 복원할 수 있는데 도움을 줄 수 있기 때문에, 명령형 종결어미('해체', '하세요체', '해요체', '해라체', '해체' 등) 그리고 청유형 종결어미('~자' '~세', '~ㅂ시다')를 반영한 자질 f1과 f3을 설정하였다. 의문형 종결어미 중 명령의 화행을 수행하는 '~어/아 주겠니(줄래)?', '~어/아 주실래요?', '~어/아 주시지 않을래요?' 등의 형태를 지니는 동사의 경우 생략된 주어가 명령형 종결어미와 같기 때문에 이러한 특성 또한 자질 f2에 반영하였다.

문장에 등장하는 주관형용사 또한 문장에 생략된 주어에 대한 단서를 제공하는데, 주관형용사는 경험주의 심리 상태, 감각 상태 그리고 경험주의 판단을 나타낸다. 이러한 형용사가 평서문에 등장할 경우 생략된 주어는 '나'가 된다. 따라서 본 연구에서는 문장이 평서문일 때 문장 내 주관형용사의 등장여부를 자질 f4로 결정하였다.

유현경(1998)에 따르면 주관형용사는 크게 심리형용사, 감각형용사, 판단형용사가 있는데 본 연구에서는 유현경(1998)에서 제안하는 형용사 리스트를 사용하며 이와 관련된 형용사 리스트는 부록에 제시하였다. 만약 주관형용사에 '~어/아 하'와 같은 어미가 부착되어 객관형용사가 되는 경우 생략된 주어가 2인칭 또는 3인칭이 된다는 제약이 발생하기 때문에 우리는 이러한 제약에 따라 주관형용사에 어미 '~어/아 하'가 부착된 형태를 객관형용사로 보고 평서문 내 객관형용사의 등장유무를 자질 f5로 결정하였다.

동사의 종결어미에 따라 문장의 유형이 결정되기 때문에 앞서 이와 관련된 자질 f1, f2, f3를 제안하였다. 특정 종결어미의 경우 여러 개의 문장 유형에서 사용 가능하여 중의성이 존재하는 경우가 있다. '~어/아'는 중의성이 존재하는 종결어미의 예시이다. 예를 들어, 종결어미 '어'가 등장하는 '*밥 먹어*' 라는 문장은 평서문, 의문문, 명령문으로 모두 사용할 수 있다. 그러나 종결어미와 보조용언이 결합하는 복합형 종결어미의 경우 한국어 영형 주어 복원을 할 때 도움이 될 수 있다. 특히 화자의 의도, 계획, 추측, 판단, 의심, 아쉬움, 다짐과 같은 의미를 지니는 복합형 종결어미가 문장 내에 등장했을 때에는 생략된 주어를 화자인 '나'로 복원하는 것이 가능하다. 따라서 우리는 복합형 종결어미의 리스트를 표 4와 같이 제시하며 이러한 정보를 반영하는 자질 f6을 설정하였다.

의미	보조용언+종결어미
의도, 계획, 의지	~ㄹ까 하다, ~기로 하다, ~ㄹ까 보다, ~려고 하다, ~고 싶다, ~겠다
추측	~ㄹ까 싶다, ~ㄴ가 하다, ~려니 하다, ~려니 싶다
판단	~ㄴ가 보다
의심	~냐 싶다, ~랴 싶다, ~지 싶다
아쉬움	~었어야 하다
약속, 다짐	~ㄹ께(게)요

<표 4> 영형 주어 해소를 위한 복합형 종결어미 리스트

한국어의 존칭 표지 '~시' 또한 생략된 주어의 정답이 아닌 것에 대한 제약을 줄 수 있는 정보이다. Lee et al.(1994)은 존칭 표지 '~시'가 청자 논항 복원에 어느 정도 기여를 한다고 보고했다. Lee et al.(1994)에서는 대부분의 대화에서 상대방인 청자에 대한 존중이 언어에 직접적으로 반영되어 있으며, 청자 논항이 생략되어 있다 하더라도 동사의 어미가 존칭형태로 표현되기 때문에 이러한 형태론적 정보가 청자 논항의 복원에 큰 기여를 할 수 있다고 주장한다. 우리는 이러한 이유로 문장 내 높임 선어말 어미 '~시'의 등장 유무를 f7로 결정하였다.

목정수(2003)는 높임 선어말 어미 '~시'와 관련하여 '~십니까?', '~세요?'가 주관동사[1]와 결합하면, 주어는 2인칭 대명사인 '당신'으로 선택된다고 주장하였다. 이에 대한 근거로 종결어미 '~습니까?'가 가리키는 인칭성은 굴절 차원의 2인칭성이지만, [+존대]라는 자질값이 포함되어 있고 선어말어미 '시'가 행위자 인칭 존대와 관련이 있으므로, '십니까?/세요?'는 잠재적으로 3인칭화된 2인칭성으로 보아야 한다는 점을 제시하였다. 주관동사(주관형용사)는 주어의 의미역을 경험주로 제약하기 때문에 결과적으로 '~십니까/~세요'와 주관동사가 함께 등장하는 경우에 생략된 주어는 2인칭 대명사라는 제약이 발생하게 된다. 따라서 우리는 종결어미 '~십니까/~세요'가 주관동사와 결합하는 구문의 등장 유무를 자질 f8로 결정하였다.

본 연구에서 제안한 자질 f1에서 f8까지는 문장 단위에 대한 자질이었는데, 이 외에도 담화 층위의 정보가 한국어 영형 대명사의 정답을 찾는데 단서를 제공해줄 수 있다. 앞서 본 형태 그리고 구문론적인 정보가 문장 단위에 존재하지 않는 경우, 특히 문맥 내 영형 주어의 선행사 후보들이 많아서 그 중 하나의 선행사를 택해야 할 때는 그 후보들 중 가장 현가성이 높은 요소를 선택해야 한다.

센터링 이론에서 이러한 현가성이 높은 요소를 초점 Fokus(vgl. Sidner 1979) 또는 센터 center(vgl. Grosz et al. 1983)라고 부른다. 따라서 영형 주어를 위한 여러 선행사 후보들이 경쟁할 때의 대용어 해소는 문장들의 센터/초점을 추적하는 작업으로 전환되며 이를 위해서 이미 다양한 방법론이 제안돼 왔

1) 목정수(2003)에서 사용되는 '주관동사'라는 용언의 분류는 심리형용사와 같은 주관형용사를 의미한다.

다(vgl. Brennan et al. 1987; Dahl & Ball 1990; Mitkov 1994; Sidner 1986; Stys & Zemke 1995).

센터링 이론을 통해 영어 대명사의 선행사를 결정하고자 하는 기존 연구들이 존재한다(vgl. Brennen et. al 1987; Kameyama 1997). 일본어와 한국어를 대상으로도 이러한 센터링 이론을 영형 대명사에 적용하고자 하는 연구들이 진행되어 왔다(vgl. Walker et al. 1990a; 1990b; 홍민표 2000).

Kameyama(1994)는 영어에 대해 문법기능순위 Grammatical Function Order: GF Order를 정하였고, 이 순위에 따라 최상위 요소로 실현된 개체가 출력 초점 요소 중에서 가장 현가성이 높은 것이라고 주장하였다. 이 연구에서 주장한 영어의 문법기능순위는 다음과 같다.

- 영어의 문법기능순위(GF Order):
  주어>목적어>목적어2>기타

Walker et al.(1990a, b)은 이를 일본어에 적용하여 일본어의 문법기능순위를 다음과 같이 결정하였다.

- 일본어의 문법기능순위(GF Order):
  주제(Topic) > 공감도(Empathy) > 주어 > 간접목적어 > 직접목적어> 기타

한국어에 대해서도 이러한 시도를 한 연구들이 있고(vgl. 최재웅 & 이민행 1999; 홍민표 2000; 홍문표 2011), 각 연구마다 문법기능순위를 다르게 정의내렸지만 본 연구에서는 홍문표(2011)의 논의에 따라 한국어의 문법기능순위를 다음과 같이 정하였다.

- 한국어의 문법기능순위(GF Order):
  주제격 > 주어격 > 목적격 > 부사격 > 기타

이러한 담화 응집성에 관련된 정보를 기계학습의 자질로 사용하기 위해서는 영형 주어가 등장하는 앞 문장들의 몇 번째까지 선행사 후보를 탐색할 것인지를 설정하는 것이 중요하다. 본 연구에서는 영형 주어가 등장한 문장(S_i) 앞의

4번째 문장(S_i-1, S_i-2, S_i-3, S_i-4)까지를 선행사 후보를 검색하는 범위로 지정하였다. 홍문표(2011)에서 제시한 문법기능순위에 따라 바로 앞 문장(S_i-1)부터 시작하여 앞의 4번째 문장(S_i-4)에 주제격 명사구가 등장하는지의 여부를 f9, 주어격 명사구가 등장하는지의 여부를 f10, 목적격 명사구가 등장하는지의 여부를 f11, 부사격 명사구가 등장하는지의 여부를 f12로 정의 내렸다.

Ⅴ. 실험

본 연구에서 제안한 자질들을 사용하여 한국어 영형 주어 해소의 성능을 평가하기 위해 실험을 수행하였다. 실험을 위해 영형 주어가 포함된 총 1,000문장의 한국어 대화체 문장[2]을 수집하였는데, 이는 관광 분야의 대화이다. 생략된 한국어 주어에 대한 독일어 주어 정답셋을 구축하기 위해 2명의 대학원생이 작업을 수행하였다. 이 때 선행사에 따라 독일어 주어의 성이 바뀌게 되는 모든 경우를 'der/die/das' 주어 유형으로 설정하였다.

주어	빈도수
ich	472
Sie	472
der/die/das	53
wir	3

<표 5> 코퍼스 내 독일어 정답 주어 분포

코퍼스 내 독일어 정답주어의 분포를 조사한 결과, 총 4개의 주어 유형이 등장하였다. 우선 가장 높은 빈도로 등장하는 독일어 주어는 대화 내 화자인 'ich'와 청자인 'Sie'[3]였으며, 그 다음으로 'der/die/das' 유형 그리고 'wir'가 등장하였다. 각 주어 유형에 대한 빈도수는 표 5와 같다.

2) 실험코퍼스는 ETRI에서 제공한 관광분야 대화코퍼스이며, 이 문장들은 관광 분야 통·번역 시스템을 개발하기 위해 구축되었다.

3) 독일어에서 청자는 'Sie'(당신), 'du'(너)가 모두 가능하지만 본 연구에서는 데이터 빈약성 data sparseness 문제를 방지하기 위해 대표형을 'Sie'라고 결정하였다.

본 논문에서는 웨카 WEKA 3.6.10 버전을 사용하여 한국어 영형 주어 해소를 위해 기계학습 방법을 적용할 때의 효과를 평가했다. 기계학습 알고리즘으로는 'SVM (Support Vector Machine)'을 선택하였는데, 이 알고리즘은 자연언어처리의 다양한 작업에서 좋은 성능을 보인다고 알려져 있다(vgl. Kudo & Matsumoto 2001; Isozaki & Kazawa 2002). 또한 '10-fold' 평가 방식을 활용하여 실험 결과를 얻었는데, 이 평가 방식은 예를 들어, 1,000문장의 코퍼스가 존재하면 이를 10개의 세트로 나누어 그 중 9개 세트인 900문장에 대해 기계학습을 수행하고, 나머지 100문장에 대한 성능평가를 수행한다. 이 과정을 모든 세트에 대해 반복 수행한 후 각 평가결과를 모두 합산하여 평균을 구하게 된다.

본 연구에서 제안한 12개의 자질을 모두 사용하여 생략된 한국어 영형 주어에 대한 독일어 주어를 찾기 위해 기계학습을 한 결과 89.3%의 정확도 Accuracy가 측정되었다. 이는 전체 코퍼스에서 기계학습을 통해 정확하게 선택된 독일어 주어의 비율을 의미한다. 홍문표(2011)에서는 한국어 대화체에 등장하는 영형 주어를 처리하여 독일어 주어를 복원하기 위한 다양한 휴리스틱 규칙을 제안하였는데, 이 때의 정확도는 73.29%였다. 물론 홍문표(2011)의 연구에서는 한국어 대화체 문장 중에서도 메신저 문장과 드라마 대본을 코퍼스로 사용하였으며, 기계학습을 사용하지 않았다는 점에서 본 연구와 직접적인 비교는 어려우나 본 연구의 방법론을 사용하면 정확도가 약 16% 이상 향상되었다.

	베이스라인	제안한 방법론	비고
정확도 Accuracy	73.29%	89.3%	약 16% 상승

<표 6> 베이스라인과 제안한 방법론의 정확도 비교

독일어의 각 정답 주어의 유형별로 측정된 정확률 Precision, 재현률 Recall 그리고 이 두 가지 수치를 가중치에 따라 평균을 낸 'F-measure'에 대한 결과값은 표 7과 같다.

주어	Precision(%)	Recall(%)	F-measure(%)
ich	0.857	0.939	0.896
Sie	0.961	0.89	0.924
der/die/das	0.652	0.566	0.606
wir	0	0	0

<표 7> 독일어 정답 주어별 정확률, 재현률, F-measure

실험 결과 정답 주어 'Sie'의 'F-measure'값이 0.924%로 가장 높게 측정되었다. 그 다음으로는 'ich'가 0.896%, 'der/die/das' 유형이 0.606%였다. 우선 정답 주어 'Sie'와 'ich'는 전체 코퍼스 내에서 둘 다 '472'회 등장하였으므로 이들이 다른 주어 유형에 비해 기계학습을 하기에 상대적으로 충분한 수치였기 때문에 'F-measure'값이 높게 측정된 것으로 추정할 수 있다. 정답 주어 'wir'의 'F-measure'값이 0%인 이유는 전체 코퍼스 내 'wir'가 정답으로 3번밖에 등장하지 않아 다른 정답 주어 유형에 비해 이 클래스에 대해 기계학습을 하는 것이 용이하지 않았을 것으로 예상한다.

웨카 시스템에서 'InfoGainAtrribute Evaluator'를 활용하면 기계학습을 위해 사용되었던 총 12개의 자질 중 독일어 정답 주어 선택을 위해 큰 영향력을 지녔던 자질의 순위를 알 수 있는데 이에 대한 결과는 다음과 같다.

순위	자질
1	f7
2	f6
3	f1
4	f9
5	f10
6	f2
7	f11
8	f4
9	f3
10	f8

11	f5
12	f12

<표 8> 자질 내 순위

1위를 차지한 자질은 f7이었고, 이는 '높임 선어말 어미 '~시' 등장 유무'에 관한 것이다. 일반적으로는 한국어 문장에서 높임 선어말 어미 '~시'가 등장한다고 해서 생략된 주어를 하나의 특정한 주어로 제한할 수 없다. 하지만 코퍼스 내에서 자질 f7이 값을 '1'로 지니는 경우, 즉 높임 선어말 어미 '~시'가 등장하는 경우는 총 432회였으며, 이 때 'Sie'가 정답 주어 클래스로 413회 등장하였다. 따라서 자질 f7이 선택되면 'Sie'라는 주어가 정답으로 학습될 확률이 높다고 추정할 수 있다. 또한 이 중에서는 자질 f7과 명령형 종결어미 등장 유무와 관련된 자질 f1이 함께 '1'의 값을 지니는 경우가 존재했으므로 이 두 자질의 조합이 함께 학습되었을 확률도 있다.

앞서 코퍼스 분석에서 이미 제시했듯이 실험 대화체 코퍼스 내에서 화자인 'ich'와 청자인 'Sie'가 정답 주어로 등장하는 비율이 다른 정답 주어에 비해서 높았다. 일반적으로 화자가 본인에 대해 말할 때 높임 선어말 어미 '~시'를 사용하지는 않으며 대화 상대방인 청자 'Sie'에 대해 높임말을 사용하기 때문에 이 실험 결과는 대화체 문장의 일반적인 경향성을 반영했다고 볼 수 있다.

독일어 정답 주어 선택을 위해 영향력을 크게 미친 두 번째 자질은 'f6'으로 이는 '복합형 종결어미 등장 유무'에 관한 것이다. 본 논문에서는 화자의 의도, 계획, 추측, 판단, 의심, 아쉬움, 다짐과 같은 의미를 지니는 복합형 종결어미 리스트를 제안하였는데, 이를 통해 생략된 주어를 화자인 '나'로 복원하는데 도움을 줄 수 있도록 자질 'f6'을 설정하였다. 전체 코퍼스에서 'f6'이 값을 지니는 경우는 195회였는데, 이 중 'ich'가 정답 주어인 경우가 188회로 약 96.4%의 비율이었다. 따라서 복합형 종결어미가 등장하면 독일어 정답 주어로 화자인 'ich'가 선택될 확률이 높도록 기계학습 되었을 것이라고 추정한다. 앞서 1위를 차지했던 자질 'f7'과 2위를 차지했던 자질 'f6' 모두 형태소 분석단계에서 쉽게 파악될 수 있으므로 이를 구현하기 용이하다는 장점이 있다.

이 자질들 외에도 주목해봐야 할 것은 4위에 오른 자질 'f9'이다. 1위부터 3위까지는 문장 단위에서 반영되는 자질이라면, 자질 'f9'는 담화 단위에서 적용되는 자질로 이는 주제격 선행사가 선행하는 네 개의 문장에 등장하는지의

여부와 관련된 것이다. 자질 ‘f9’가 독일어 정답 주어를 선택하는데 영향력을 미친 자질 순위에서 4위인 것은 앞서 한국어의 문법기능순위를 ‘주제격 > 주어격 > 목적격 > 부사격 > 기타’ 순으로 설정한 것과 관련이 있다. 즉 가장 높은 문법기능순위를 지닌 ‘주제격’ 선행사가 한국어 영형주어를 해소하는데 도움이 되었다는 것을 의미한다. 따라서 문법기능순위에서 ‘주제격’이라는 최상위 요소로 실현된 선행사가 정답 주어로서 가장 현가성이 높다는 주장을 뒷받침할 수 있다고 볼 수 있다.

본 연구에서 제안하는 방법이 한-독 번역결과에 어떠한 영향을 주는지 알아보기 위해 한국어 영형 주어 복원 전/후의 독일어 번역결과를 비교해 보았다. 번역률 향상을 알아보기 위해 총 120문장을 대상으로 하였으며 이는 관광 분야의 대화체 문장이었다. 번역률 평가를 위한 기준은 표 9와 같으며 영형 주어 복원 전/후의 번역률을 수동으로 평가하였다.

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.

<표 9> 번역률 평가 기준

한국어 영형 주어 복원 전의 독일어 번역률을 평가한 결과 약 59.68%의 번역률이 측정되었으며, 한국어 영형 주어를 복원한 후의 독일어 번역률은 약 63.85%로 번역률이 4% 정도 향상됨을 알 수 있었다.

	영형 주어 복원 전	영형 주어 복원 후	비고
번역률	59.68%	63.85%	약 4% 상승

<표 10> 한국어 영형 주어 복원 전/후의 번역률 비교

한국어 영형 주어를 복원함으로서 독일어 번역 결과의 번역률이 높아지는 문장들을 예시로 들면 표 11과 같다.

	영형 주어 복원 전 한국어 문장	독일어 번역 결과 (Google)		영형 주어 복원 후 한국어 문장	독일어 번역 결과 (Google)
(16)	상점은 몇 시에 열리나요? * 오후 8시에 열리네요.	Ich öffne am 20.00.	(16‘)	상점은 몇 시에 열리나요? 그것은 오후 8시에 열리네요.	Es ist um 8:00 Uhr geöffnet.
(17)	* 제 동생이 탄 비행기의 도착 시각을 좀 알고 싶은데요.	Mein Bruder möchte die Ankunfts-zeit der Flug-aufnahmen kennen.	(17‘)	제가 제 동생이 탄 비행기의 도착 시각을 좀 알고 싶은데요.	Ich möchte die Ankunfts-zeit Flug meines Bruders kennen.

<표 11> 한국어 영형 주어 복원 전/후 독일어 번역 결과 예시

독일어 번역 결과는 구글 번역기를 활용하였는데 구글 번역기는 통계기반 번역 방식을 사용하는 시스템이다. 통계기반 번역은 서로 다른 두 언어의 대역 코퍼스 bilingual Corpus를 통계적 방법으로 분석한 결과를 학습하여 번역하는 방식이다. 통계기반 번역기는 언어 모델을 통해 특정 단어가 번역되면 그 단어 다음에 나올 수 있는 단어를 확률적으로 계산하게 된다.

표 11에서 한국어 번역 결과를 살펴보면 구글 번역기의 이러한 특성으로 인해 한국어 영형 주어가 복원된 경우 독일어 주어가 복원됨에 따라 독일어 번역문 전체 결과가 훨씬 더 좋아지는 것을 살펴볼 수 있다. 예문 (17)의 생략된 주어가 ‘나’이기 때문에 예문 (17‘)에서는 주어를 복원하여 기계번역을 하였다. 독일어 번역 결과를 비교해보면 영형 주어 복원 전에 등장하지 않았던 독일어 주어 ’ich’만 독일어 번역 결과에 추가되는 것이 아니라 ‘ich’가 번역됨으로서 해당 단어 뒤에 올 수 있는 단어들이 확률적으로 계산되어 연쇄적으로 새로운 단어들이 번역된 결과를 볼 수 있다. 따라서 한국어 영형 주어를 복원하는 것이 한-독 기계번역 결과의 번역률을 향상시키는데 중요한 역할을 한다.

Ⅵ. 결론

본 논문에서는 한국어 대화체 문장에서 빈번하게 등장하는 한국어 영형 주어 복원을 위해 기계학습 방법론을 적용하였다. 한국어 영형 주어는 한국어와 같은 주제 지향 언어에서 독일어와 같은 주어 지향 언어로 번역할 때 반드시 해소되어야만 하는 현상이다. 따라서 한국어 영형 주어 해소는 한국어를 출발 언어로 하고 독일어를 목표 언어로 하는 기계 번역 시스템에서 필수적이다.

이를 위해 우리는 기계 학습을 위한 총 12개의 자질을 제안하였는데, 이는 문장 내 등장하는 동사의 형태론적 정보를 반영하고(f1~f3, f6~f8), 주관 그리고 객관 형용사의 등장 유무와 관련된 자질(f4, f5)에 관한 것이었으며 또한 센터링 이론을 반영하여 담화 내의 정보와 관련된 자질(f9~f12)도 제안하였다.

우리가 제안한 자질들을 사용하여 한국어 영형 주어 해소의 성능을 평가하기 위해 실험을 한 결과 89.3%의 정확도가 측정되었으며, 이는 베이스라인에 비해 약 16%가 향상된 결과였다.

현재 구축한 코퍼스에는 각 정답 주어 유형의 빈도수에 차이가 있다. 실험 결과 코퍼스 내 빈도수가 높은 정답 주어들이 기계학습 방법론에서 정확도가 높게 분류됐다. 따라서 향후 모든 정답 주어 유형에 대한 빈도수를 유사하게 조정하고 더 큰 규모의 코퍼스를 수집하여 본 논문에서 제안한 방법론을 다시 적용해보고자 한다.

참고문헌

노대규(1996): 한국어의 입말과 글말, 국학자료원.

목정수(2003): 한국어 문법론. 월인.

유현경(1998): 국어형용사연구. 한국문화사.

정연주(2010): "-어 하-" 와 통합하는 객관형용사의 의미 특성. 한국어의미학, 33, 297-319.

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

홍문표(2011): 한-독 대화체 기계번역을 위한 주어생략현상의 처리방안. 독어학, 24, 417-439.

홍민표(2000): 센터링 이론과 대화체에서의 논항 생략 현상. 인지과학, 11(1), 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.

Brennan, S. E., Friedman, M. W., & Pollard, C. J.(1987): A centering approach to pronouns. In Proceedings of the 25th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 155-162.

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

Connolly, D., Burger, J. D., & Day, D. S.(1997): Machine learning approach to anaphoric reference. International Conference on New Methods in Language Processing, UMIST Manchester, United Kingdom, 33-144.

Dahl, D. A., & Ball, C. N.(1990): Reference resolution in PUNDIT. Chapter 8 in Logic and Logic Grammars for Language Processing, P. Saint-Dizier and S. Szpakowicz, Ed., Ellis Horwood, New York, 168-184.

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

Grosz, B.J., Joshi, A.K., & Weinstein, S.(1983): Providing a unified account of definite noun phrases in discourse. In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, 44-50.

Hirst, G.(1981): Anaphora in Natural Language Understanding: A Survey. Lecture Notes in Computer Science. Springer Verlag.

Hutchins, W. J., & Somers, H. L.(1992): An introduction to machine translation (Vol. 362). London: Academic Press.

Isozaki, H., & Kazawa, H.(2002): 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.

Kameyama, M.(1994): Indefeasible semantics and defeasible pragmatics. Technical Note 544, SRI International. A shorter version to appear in Kanazawa Makoto,

Christopher Pinon, and Henriette de Swart, editors, Quantifiers, Deduction, and Context. CSLI, Stanford.

Kameyama, M.(1997): Intrasentential centering: A case study. In Proceedings of the Workshop on Centering Theory in Naturally Occurring Discourse, Institute for Research in Cognitive Science, University of Pennsylvania. 89–112.

Kazuhide, Y., & Eiichiro, S.(1998): Feasibility study for ellipsis resolution in dialogues by machine-learning technique. In Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1428-1435.

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. Computational linguistics, 20(4), 535-561.

Lee, S. H., Byron, D., & Gegg-Harrison, W.(1994): Annotations for Zero Pronoun Resolution in Korean Using the Penn Korean Treebank. In The 3rd Workshop on Treebanks and Linguistic Theories (FLT 2004), Tübingen, Germany, 75-88.

Manabu, O., & Kouji, T.(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.

Mitkov, R.(1994): An integrated model for anaphora resolution. In Proceedings of the 15th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1170-1176.

Mitkov, R., & Schmidt, P.(1998): On the complexity of pronominal anaphora resolution in machine translation. In Mart´ın-Vide, C. (Ed.), Mathematical and computational analysis of natural language. John Benjamins Publishers, Amsterdam, 207-222.

Nakaiwa, H., & Ikehara, S.(1995): Intrasentential resolution of Japanese zero pronouns in a Machine Translation system using semantic and pragmatic constraints. In Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95), 96-105.

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.

Sidner, C. L.(1979): Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. PhD Thesis, Massachusetts Institute of Technology.

Sidner, C.(1986): Focusing in the comprehension of definite anaphora. In Readings in Natural Language Processing. Morgan Kaufmann Publishers Inc., 363-394.

Stys, M. E. & Zemke, S. S.(1995): Incorporating discourse aspects in English-polishMT: Towards robust implementation, in 'Recent Advances in NLP', Velingrad/BG.

Walker, M., & Whittaker, S.(1990b): Mixed initiative in dialogue: An investigation into discourse segmentation. In Proceedings of the 28th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 70-78.

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

Walker, M., Iida, M., & Cote, S.(1990a): Centering in Japanese discourse. In COLING90: In Proceedings of the 13th International Conference on Computational Linguistics, Helsinki, 1-8.

Zusammenfassung

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

PARK Arum, HONG Munpyo (Sungkyunkwan Uni)

In der vorliegenden Arbeit wird das Subjekt-Ellipse Phänomen in der maschinellen Übersetzung des Koreanischen ins Deutsche behandelt. In einer sogenannten topik-orientierten Sprache wie im Koreanischen wird ein Subjekt oft ausgelassen. Bei der Übersetzung aus dem Koreanischen in eine sogenannte subjekt-orientierte Sprache wie das Deutsche müssen die ausgelassenen Subjekte 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 Subjekt 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 12 Merkmale über die morphologischen Informationen der bestimmten Verben und Adjektive, semantische Eigenschaften besimmter Prädikate sowie Diskursinformationen vorgeschlagen.

Das Experiment zeigte, dass unser Ansatz im Vergleich zu den bisherigen Methoden die Genauigkeit der Anaphernresolution von 73.29% auf 89.3% um 16% erhöhen kann.

주제어: 대용어해소, 영형대명사, 영형주어, 기계번역, 기계학습
Schlüsselbegriffe: Anaphernresolution, Zero-Pronomen, Zero-Subjekt, maschinelle Übersetzung, maschinelles Lernen

필자 이메일 주소: remin2@skku.edu, skkhmp@skku.edu
논문투고일: 2015.01.03 | 논문심사일: 2015.01.22 | 게재확정일: 2015.02.15

<부록>

대상 심리형용사의 목록(114개)	원인 심리형용사의 목록(141개)	감각형용사의 목록(126개)	판단형용사의 목록(12개)
가소롭다	감격스럽다	가뜬하다	관계없다
가증스럽다	감사하다	가렵다	괜찮다
가엾다	거북살스럽다	가물가물하다	그만이다
감감하다	거북스럽다	가볍다	끄떡없다
감개무량하다	거북하다	가뿐하다	되잖다
같잖다	겸연쩍다	가쁘다	마땅하다
거추장스럽다	겁다	간지럽다	무방하다
걱정스럽다	고맙다	간질간질하다	상관없다
경이롭다	고생스럽다	갑갑하다	소용없다
고깝다	고소하다	개운하다	시원찮다
고단하다	곤혹스럽다	거뜬하다	싸다
고달프다	괘씸하다	거북하다	좋다
고되다	구차스럽다	고단하다	
고통스럽다	권태롭다	고달프다	
궁금하다	괴롭다	고되다	
궁하다	근심스럽다	고프다	
귀엽다	기껍다	곤하다	
귀찮다	기막히다	곱다	
그립다	기쁘다	궁금하다	
기특하다	긴가민가하다	굴뚝같다	
껄끄럽다	까마득하다	근지럽다	
꼴사납다	꺼림칙하다	근질근질하다	
끔찍스럽다	꺼림하다	급하다	
낯설다	께름직하다	깔깔하다	
낯익다	난감하다	껄껄하다	
다행스럽다	난처하다	나른하다	
다행하다	남부끄럽다	노곤하다	
달갑다	남부럽다	느끼하다	
답답하다	낭패스럽다	답답하다	
대견스럽다	낯간지럽다	더부룩하다	
대견하다	낯뜨겁다	덥다	
대수롭다	노엽다	든든하다	
더럽다	놀랍다	따끔하다	
덧없다	눈물겹다	떨떠름하다	

대상 심리형용사의 목록(114개)	원인 심리형용사의 목록(141개)	감각형용사의 목록(126개)	판단형용사의 목록(12개)
두렵다	당혹하다	뜨끈하다	
든든하다	당황하다	뜨끔하다	
딱하다	따분하다	띵하다	
마땅찮다	떠름하다	마렵다	
막연하다	뜨악하다	마르다	
만만하다	마땅하다	막막하다	
만족스럽다	마뜩찮다	맛없다	
못마땅하다	마뜩하다	맛있다	
못미덥다	마지못하다	맨숭맨숭하다	
무섭다	망신스럽다	맹맹하다	
무엇하다	망연하다	먹먹하다	
문제없다	머쓱하다	멍멍하다	
미덥다	멋적다	멍하다	
미쁘다	면목없다	메스껍다	
미안스럽다	무료하다	메슥메슥하다	
미안하다	무색하다	메시껍다	
믿음직스럽다	무안스럽다	면구스럽다	
밉다	무안하다	목마르다	
밉살맞다	무참하다	무겁다	
밉살스럽다	미심쩍다	뭉클하다	
반갑다	미안쩍다	배고프다	
번거롭다	민망스럽다	배부르다	
부끄럽다	민망하다	부르다	
부담스럽다	바쁘다	불편하다	
부럽다	벙벙하다	뻐근하다	
불쌍하다	분통하다	뻑뻑하다	
뿌듯하다	분하다	뻑적지근하다	
사랑스럽다	불만족스럽다	뻣뻣하다	
새롭다	불안스럽다	뻥하다	
새삼스럽다	불안하다	뿌듯하다	
생경스럽다	불쾌하다	산란하다	
생소하다	불행하다	시다	
석연하다	비통하다	시리다	
설다	뼈아프다	시원하다	
성가시다	뼈저리다	시큰하다	
수상쩍다	서글프다	심란하다	
수상하다	서럽다	싸하다	
신기롭다	서운하다	쓰라리다	

대상 심리형용사의 목록(114개)	원인 심리형용사의 목록(141개)	감각형용사의 목록(126개)	판단형용사의 목록(12개)
신기하다	섧다	쓰리다	
싫다	섬뜩하다	씁쓉하다	
아깝다	섭섭하다	아뜩하다	
아니꼽다	속상하다	아른아른하다	
아리송하다	속시원하다	아리다	
아쉽다	솔깃하다	아릿하다	
안쓰럽다	송구스럽다	아찔하다	
애석하다	송구하다	아프다	
야속하다	수치스럽다	알알하다	
얄밉다	스스럽다	어수선하다	
어렵다	슬프다	어지럽다	
역하다	시들하다	어찔하다	
염려스럽다	시원섭섭하다	어떨떨하다	
염치없다	실망스럽다	얼얼하다	
예쁘다	심드렁하다	울울하다	
우습다	심심하다	울적하다	
원망스럽다	쑥스럽다	저리다	
의문스럽다	쓸쓸하다	저릿하다	
의심스럽다	씁쓰레하다	조마조마하다	
의심쩍다	씁쓰름하다	졸립다	
의아하다	아득하다	짜릿하다	
의외롭다	아연하다	짠하다	
자랑스럽다	아찔하다	쩌릿하다	
저주스럽다	안타깝다	쭈뼛하다	
절실하다	암담하다	찌르르하다	
정겹다	애달프다	찌릿하다	
조심스럽다	애통하다	찌뿌드드하다	
좋다	어리둥절하다	착찹하다	
주체스럽다	어색하다	처량하다	
지긋지긋하다	어이없다	출출하다	
짐스럽다	어처구니없다	춥다	
징그럽다	억울하다	침침하다	
짠하다	언짢다	칼칼하다	
창피스럽다	역겹다	캄캄하다	
창피하다	열없다	컬컬하다	
측은하다	열적다	텁텁하다	
탐탁하다	영광스럽다	팍팍하다	
한스럽다	외롭다	편안하다	

대상 심리형용사의 목록(114개)	원인 심리형용사의 목록(141개)	감각형용사의 목록(126개)	판단형용사의 목록(12개)
한심스럽다	우울하다	편찮다	
한탄스럽다	원통하다	편편하다	
혐오스럽다	유감스럽다	편하다	
힘겹다	유쾌하다	평강하다	
	의아스럽다	평안하다	
	이상스럽다	피곤하다	
	일없다	피로하다	
	재미없다	허전하다	
	재미있다	허탈하다	
	죄송스럽다	허하다	
	죄송하다	헛헛하다	
	죄스럽다	혼란스럽다	
	적적하다	혼란하다	
	즐겁다	홀가분하다	
	지겹다	후련하다	
	지루하다	흐릿하다	
	짜증스럽다		
	찜찜하다		
	참담하다		
	창연하다		
	터무니없다		
	통쾌하다		
	행복스럽다		
	행복하다		
	황송스럽다		
	황송하다		
	황홀하다		
	흐뭇하다		
	흡족하다		
	흥겹다		
	흥미롭다		

주관 형용사 리스트 (vgl. 유현경 1998)