한국어-영어 대화체 번역시스템을 위한 영형 대명사 해소

박아름, 지은별, 홍문표
성균관대학교 독어독문학과
{remin2,byul0907,skkhmp}@skku.edu

Zero Pronoun Resolution for Korean-English Spoken Language MT

Arum Park, Eun-Byul Ji & Munpyo Hong
Sungkyunkwan University, 53 Myeongnyun-dong 3-ga, Seoul, Korea

요 약

이 논문은 한-영 대화체 번역 시스템에서 영형 대명사 해소를 위한 새로운 방법론을 제시하였다. 영형 대명사는 문맥, 상황, 세상 지식으로부터 추론될 수 있는 문장에서 생략된 요소이다. 이 논문은 특히 주어-대명사 생략 현상에 대해 다루고 있는데, 그 이유는 드라마 대본이나 인스턴트 메신저 채팅과 같은 한국어 대화체에서는 매우 일반적인 현상이기 때문이다. 이 논문에서 우리는 많은 양의 지식을 요구하지 않는 간단한 방법론을 제시하였다. 평가결과 우리의 방법은 0.79의 F-measure 스코어를 달성하였고, 전체번역률의 측면에서는 약 4.1% 정도의 향상효과가 있었다.

주제어: 영형 대명사, 한국어 대화체, 주어 생략 현상

1. 서론(1)

한국어가 유럽 언어와 구분될 수 있는 주요한 언어학적 속성 중 하나는 바로 주제 지향적이라는 것이다. 이런 이유로 한국어에서는 주어가 생략되는 경우가 매우 일반적이며, 특히 한국어 대화체에서는 주어 생략이 더 빈번하다.

생략된 주어를 비롯한 생략된 대명사는 일반적으로 '영형 대명사(ZP: Zero Pronoun)'라고 불린다. 영형 대명사와 관련한 기존 연구들은 대부분 일본어를 다루고 있으며, 영형 대명사의 선행사를 찾기 위해 센터링 이론, 화행 이론과 같은 언어학적 이론을 사용한다 cf. Walker et al. (1994)[1], Kameyama (1996)[2], Lee et al. (1997)[3], Nariyama (2002)[4]. 영형 대명사 해소에 대한 대표적인 연구라고 할 수 있는 Nakaiwa et al. (1995)[5]에서는 일본어에서의 extra-sentential ZP의 선행사를 결정하기 위해 양태 표현, 동사의 의미적 속성, 접속사와 같은 의미 · 화용론적 제약을 사용하였다. 이 제약들은 intra-sentential ZP 해소에서도 적용되었다 cf. Nakaiwa and Ikehara (1995)[6].

비록 이런 연구들이 영형 대명사 해소에 관해 세련된 방법을 제시하지만, 사실상 대화체를 다루는 자동번역 시스템에는 적합하지 않다. 왜냐하면 해당 방법들은 영형 대명사 해소를 위해 일반적으로 엄청난 양의 지식 베이스를 필요로 하기 때문이다.

본 연구에서는 한국어-영어 대화체 자동번역 시스템에서 영형 대명사 해소를 할 수 있는 새로운 방법을 제시한다. 제안된 방법들은 무거운 지식 베이스를 필요로 하지 않기 때문에, 대화체 자동번역 시스템에서 구현되기에 적합하다.

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

2. 방법론

한국어 대화체에서는 주어생략현상이 빈번하게 일어나는데 이는 매우 다루기 어렵기 때문에 번역률을 크게 저하시키는 요인 중 하나이다. 2.1에서는 한국어 대화체에서 생략된 주어들의 언어학적 특성을 분석할 것이다. 이 결과를 토대로 2.2에서는 생략된 주어를 복원하기 위한 언어학적 제약들을 제안할 것이다. 해당 제약들은 번역 시스템의 프로세싱 부담을 낮출 수 있는 가벼운 리소스들이기 때문에 이를 한국어-영어 통번역시스템에 적용하기에 적합하다.

2.1 한국어 대화체 문장에서 생략된 주어의 특성

한국어 대화체에서 생략된 주어에 대해 3가지 유형으로 나눌 수 있다:

(1) 같은 문장 내에 생략된 주어의 선행사가 존재하는 유형 (intra-sentential ZP), (2) 생략된 주어의 선행사가 텍스트 내에 존재하는 유형 (inter-sentential ZP), (3) 생략된 주어가 지시적이고 담화 내에 존재하지 않는 유형 (extra-sentential ZP). 각 유형에 대한 한국어 예문은 표 1과 같다.(2)

(2) 본 연구는 영형 대명사 중에서도 주어생략현상에 관한 것이므로, 예문에서 주어가 생략된 자리를 'ZP'로 표시할 것이다.

표 1: 한국어 대화체에서 유형별 생략된 주어의 예문

ZP의 유형	한 국 어 예 문
(1) intra-sentential	A: 너 이리 와서 ZP 이것 좀 봐. *ne ili wase ZP* ikes com pwa.(3) you here come and ZP this Ø look-at '(You) come here and look at this.'
(2) inter-sentential	A1: 너 원래 이렇게 버릇없었어? ne wenlay ikehkey pelusepsesse? you always like this rude 'Are you always rude like this?'
	A2: ZP 왜 이렇게 변했니? *ZP way ilehkey pyenjayssni*? ZP why like this have changed 'Why have you so changed like this?'
(3) extra-sentential	A: ZP 오늘 기분이 훨씬 좋아. *ZP onul kipuni hwelssin coha.* ZP today feel a lot better 'I feel a lot better today.'

생략된 주어가 유형별로 어느 정도 빈도로 나오는지 알아보기 위해 한국어 대화체 총 715문장을 분석하였다. 이 문장들은 메신저 대화 358문장, 드라마 대본 357문장으로 구성되어 있다.

Nakaiwa et al.(1995)에서는 72%의 intra-sentnetial ZP (총 515개의 영형대명사 중 140개), 73%의 extra-sentential ZP (총 515개의 영형대명사 중 375개)가 등장했다는 결과가 나왔다. 하지만 표 2를 보면 본 연구 결과는 Nakaiwa et al.(1995)과 크게 다르다는 것을 알 수 있다.

표 2: 한국어 대화체 코퍼스에서의 ZP 유형별 출현 빈도수

	총	메신저 대화	드라마 대본
생략된 주어의 수	435	191	244
extra-sentential	353 (81.14%)	162 (84.41%)	191 (78.28%)
inter-sentential	67 (15.40%)	19 (9.94%)	48 (19.67%)
intra-sentential	15 (3.44%)	10 (5.23%)	5 (2.05%)

한국어 대화체 문장에서는 전체 문장에서 생략된 주어 435개 중 353개(81.14%)가 extra-sentential ZP였고, 67개(15.40%)는 inter-sentential ZP, intra-sentential ZP는 15개(3.44%)정도에 지나지 않는다는 사실을 알 수 있다. 즉, 한국어 대화체에서는 extra-sentential ZP가 약 8%이상 많았고, intra-sentential ZP는 매우 낮은 비율을 차지하고 있었다. 이는 일본어를 대상으로 한 실험에서는 뉴스 기사라는 문어체 코퍼스를 사용하였고, 본 연구에서는 메신저 대화나 드라마 대본과 같은 대화체를 대상으로 실험을 하였기 때문이라고 예상된다.

앞선 분석 결과를 토대로 한국어 대화체 번역의 주어복원을 위해서는 extra-sentential ZP에 대한 처리가 가장 중요함을 알 수 있다. 하지만 이 유형의 ZP는 문장 내 혹은 텍스트 내의 다른 문장에 선행사가 존재하지 않고, 화행이나 세상에 대한 지식과 같은 언어외적인 단서에 의존해야 하기 때문에 해소되기 어렵다. 하지만 실제 경험적 데이터에 의하면 한국어 대화체에서의 extra-sentential ZP를 해소하기 위한 언어학적 단서들이 존재한다는 것을 알 수 있다.

2.2 의미, 화용론적 제약을 사용한 한국어의 생략된 주어 복원

본 논문에서는 한국어 대화체에서 생략된 주어를 복원하기 위한 3가지 유형의 언어학적 속성을 제안한다: 1) 종결어미, 2) 감정 상태 동사, 3) 접속사.

먼저 여기에서는 주어를 제약할 수 있는 총 122개의 종결어미 리스트를 제안한다. 이 중 주어가 화자인 'I'가 될 수 있는 종결어미는 52개, 청자 'you'일 수 있는 65개, 주어가 'we'가 될 수 있는 3개이다. 해당 종결어미들은 종결어미나 보조용언에 화자의 의도나 추측과 같은 양상정보를 포함하고 있거나 명령문과 청유문과 같이 문형을 하나로 결정할 수 있기 때문에 생략된 주어의 선행사를 제약할 수 있다. 예를 들어 '-ㄹ까 하다'와 '-기로 하겠다'는 화자의 의도를 나타내고, '-으려니 싶다'와 '-으려니 하다'는 화자의 추측을 나타내기 때문에 생략되어 있는 주어는 'I'로 복원할 수 있다. '-(아/어) 주겠니?'나 -(아/어) 주실래요?'는 의문문이나 명령문을 나타내기 때문에 주어를 'you'로 복원할 수 있으며, '-자', '-ㅂ시다'는 중의성의 없는 청유형 종결어미이므로 주어를 'we'로 복원할 수 있다.

두 번째로 총 65개의 감정 상태 동사(4) 리스트를 제안한다. 감정 상태 동사의 화자는 내부의 감정을 스스로 경험하는 경험자일 수밖에 없기 때문에 주어는 무조건 'I'가 된다. '기쁘다', '반갑다', '설레다', '밉다' 등이 바로 여기에서 제안하는 감정 상태 동사의 예이다.

표 3: 감정 상태 동사가 포함된 한국어 대화체 예문

감정 상태 동사 예문	한국어 대응 영어
(4) 'ZP 기쁘다.'	I-be-glad.
	*He-be-glad.
(5) 'ZP 진짜 설렌다.'	I-be-really-excited.
	*He-be-really-excited.

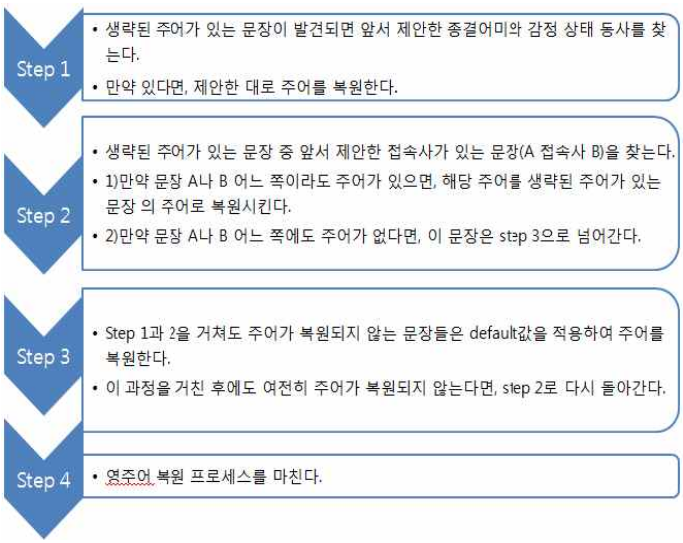
(3) 한글에 대한 전사는 예일 로마법을 따른다.

(4) 한국어에서 감정을 표현하는 동사는 크게 감정 반응 동사와 감정 상태 동사로 나눌 수 있다. 감정 상태 동사와 달리 감정 반응 동사가 표현하는 감정은 외부에서도 쉽게 관찰할 수 있기 때문에 주어의 지시사가 'I'이외의 3인칭 주어가 올 수 있다. 또한 감정 반응 동사는 대개 감정 상태 동사의 어간에 '-하'가 붙어서 생성될 수 있다. '기뻐하다', '반가워하다' 그리고 '미워하다' 등이 감정 반응 동사의 예이다.

세 번째로 한국어 구어체 문장에서 생략된 주어를 복원하기 위해 우리는 7개의 한국어 접속사를 제안한다. 복문의 경우 한국어의 특정 접속사에 따라 복문 내의 단문들이 주어를 공유할 수 있다. 총 7개의 접속사 중 6개의 접속사 '-고', '-(으)면서/-(으)며', '-(으)려고', '-(으)려면', '-느라고/-느라', '-어서'는 복문 A 접속사 B에서 A와 B의 주어가 같을 확률이 매우 높기 때문에, A와 B의 주어를 일치시킬 수 있다. 나머지 1개의 접속사 '-었더니'는 주어가 1인칭인 문장에서만 사용할 수 있고, A의 주어는 언제나 말하는 사람, 즉 화자이어야만 한다는 제약이 있다. 다음 장에서는 본 논문에서 지금까지 제안한 제약들을 기반으로 알고리즘을 제안한다.

3. 알고리즘(5)

한국어에서 생략된 주어를 해소하기 위한 과정은 다음과 같다.



4. 평가

본 논문에서 제안한 방법론을 평가하기 위해 한국 드라마 대본에서 추출한 1,013 문장을 대상으로 평가를 수행하였다. 생략된 주어 복원의 정확성을 측정하는 데에는 다음의 3가지 평가 요소를 사용하였다. Recall은 제안된 제약을 통해 찾을 수 있는 생략된 주어의 수의 비율을 나타낸 것이고, Precision은 제안된 제약을 통해 정확하게 해소된 생략된 주어의 수의 비율을 말한다. F-measure는 Recall과 Precision 이 두 가지를 반영하고 다음 공식에 의해 계산된다.

F-measure=2*Precision*Recall/(Precision+Recall)

또한, 다음과 같이 4가지 조건에 따른 생략된 주어 해소의 정확성을 조사하였다: 1) 베이스라인으로서의 default값, 2) 종결 어미+default값, 3) 종결 어미+감정 동사+default값, 4) 종결 어미+ 감정 동사+접속사+default값

표 4. 생략된 주어 해소에 대한 정확성(accuracy)

	1)	2)	3)	4)
Precision	52.12	68.96	69.53	70.88
Recall	81.15	83.44	83.93	89.51
F-measure	0.635	0.755	0.761	0.791

제안된 방법론들과의 비교를 위해 문장 유형을 바탕으로 한 default값을 베이스라인으로 사용하였다. 베이스라인의 precision은 52.12%, recall은 81.15%로 나타났다.

모든 제약들은 precision과 recall 두 가지 모두를 향상시킬 수 있었다. 종결 어미를 적용했을 때는 precision의 비율이 16% 이상, recall이 약 2% 향상되었고, 여기에 감정 동사를 더하면, precision은 69.53%가 되고, recall 역시 default값과 비교했을 때 2% 올랐다.

마지막으로 단계적으로 접속사까지 모든 제약들을 적용했을 때 (종결 어미+감정 동사+접속사+default값), precision은 약 19%, recall은 약 8% 향상되어 나타났다. 결론적으로 접속사에 대한 제약이 생략된 주어를 해소하는데 있어서 매우 효과적임을 알 수 있다.

이 결과가 번역률 향상에 있어서 얼마나 효과적인지를 알아보기 위해서 0-4점 사이의 등급으로 구성되어 있는 번역률 평가 방법을 사용하였다.

표 5: 자동 번역 시스템 평가 기준

(5) 본 연구에서는 주어가 생략된 문장과 그 위치가 이미 파악되었다고 가정한다. 본 연구에서는 default값을 문장의 기능(sentential function)에 따라 설정하였다. 문장 유형은 단순하게 구두법에 의해 분류하였다: 마침표가 있으면 평서문, 물음표가 있으면 의문문 그리고 느낌표가 있으면 명령문이라고 간주하였다. 평서문에서 주어가 생략된 경우 그 선행사를 'I'로 설정하였고, 의문문이나 명령문의 경우 생략된 주어의 선행사를 'You'로 설정하였다.

점수	평가 기준
4	원어문의 의미가 그대로 전달된 경우
3.5	원문의 문장전체가 잘 분석되어 문장의 전체적인 의미의 골격이 전달되지만 동사를 제외한 1-2단어의 대역어가 잘못된 경우
3	원문의 문장전체가 잘 분석되어 문장의 전체적인 의미의 골격이 전달되지만 여러 단어의 대역어가 잘못된 경우
2.5	원문의 문장 전체의 분석은 실패했으나, 하나 이상의 동사구가 잘 분석되고 정확히 번역되어 부분적으로 문장의 의미가 전달된 경우
2	원문의 문장 전체의 분석은 실패하여 전체적인 문장의 의미를 파악하기 어려우나, 하나 이상의 명사구가 잘 분석되고 정확히 번역됨.
1	원문의 문장 전체의 분석은 실패하여 전체적인 문장의 의미를 파악하기 어려우나, 문장 중에 하나 이상의 단어 또는 한 개의 명사구라도 정확히 번역된 경우
0	원문이 번역문에 그대로 출력됨

총 1,013문장 중 생략된 주어가 610개이므로 한 문장 당 평균 0.6개의 주어가 생략됨을 알 수 있다. 또한 모든 제약을 적용했을 때 베이스라인과 비교하여 주어 복원 정확률이 약 19% 향상되었다. 위의 표 5를 참조하여 주어를 틀리게 복원하는 것이 2점이고, 주어를 맞게 복원하는 것이 3점이라고 가정하면, 베이스라인과 비교하여 약 2.8%의 번역률 향상을 가져온다고 할 수 있다. 또한 주어를 맞게 복원한 결과를 3.5점으로 가정하면 약 4.1% 정도의 번역률 향상을 기대할 수 있다. 이 번역률 향상 수치가 낮아 보일 수 있으나 구어체 자동 번역에서 특히 주어가 잘 복원되어 번역률이 향상되는 경우는 문장의 기본 골격이 전달되는 것이므로 실제로 사용자의 체감 번역률 향상은 훨씬 높다고 할 수 있다.

5. 결론

본 연구에서는 한-영 대화체 기계번역 시스템에서 나타나는 주어생략 현상의 해소를 위한 언어학적 제약들을 제시하였고, 또한 생략된 주어의 선행사를 찾기 위한 4가지 단계의 알고리즘도 제안하였다. 본 연구의 방법론에서는 extra-sentential, inter-sentential, intra-sentential ZP를 구분할 필요 없이 파서(parser)가 문장에서 생략된 주어를 발견하면, 제안된 알고리즘을 단계적으로 적용하면 된다. 또한 방법론 평가에서는 precision, recall, F-measure의 수치가 기본 베이스라인에 비해 각각 약 18%, 8%, 16% 향상되었음을 보여주었다. 약 18%의 정확률 향상은 구어체 번역에서 약 4.1%의 번역률 향상을 의미하며, 이는 문장의 기본 골격이 잘 번역되는 것이므로 사용자의 체감 번역률 향상은 훨씬 높아진다.

참고문헌

[1] Walker M.A., Iida M., and Cote S. Japanese Discourse and the Process of Centering, Computational Linguistics, Volume 20, Number 2:193-232, 1994.

[2] Kameyama M. A Property-Sharing Constraint in Centering, Proc.of the 24th Annual Meeting of the Association for Computational Linguistics:200-206, 1986.

[3] Lee, J.W., G.C. Kim, J.Y Seo. A Dialogue Analysis Model with Statistical Speech Act Processing for Dialogue Machine Translation:10-15, 1997.

[4] Nariyama S. Grammar for ellipsis resolution in Japanese. In Proceedings of the 9th International Conferenceon Theoretical and Methodological Issuesin Machine Translation:135-145, 2002.

[5] Nakaiwa H., Ikehara S. Intra-sentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints. In Proc.of The 6th TMI:96-105, 1995.

[6] Nakaiwa H., Shirai S., Ikehara S. and Kawaok T. Extra-sentential Resolution of Japanese Zero Pronouns using Semantic and Pragmatic Constraints. In Proc.ofAAAI:99-105, 1995.