C-STAR 인터체인지 포멧을 이용한 다국어 대화체 번역시스템의 성능

최운천, 박준, 양재우
한국전자통신연구원 교환전송기술연구소
대전시 유성구 가정동 161 우: 305-350,
ucchoi@etri.re.kr

Performance of Multi-Lingual Spoken Language Translation System using C-STAR Interchange Format

Un-Cheon Choi, Jun Park, Jae-Woo Yang
ETRI-Switching & Transmission Technology Laboratory

요약

ETRI 통신단말연구부에서는 1999년 7월 22일에 C-STAR 회원국이 함께 참여하는 국제간 음성언어번역 시스템 공동 시연을 가졌다. 이 논문은 다국어 대화체 번역 시스템인 음성언어번역 시스템의 국제간 공동 시연에 사용된 한국어 번역 시스템의 성능에 대해 기술한다. 번역 시스템의 성능은 전사문장을 이용한 영어, 일본어, 한국어의 번역 결과와 음성인식 결과를 이용한 각 언어의 번역 결과를 평가하여 얻었다. 그리고 세부 시스템의 성능을 알아보기 위해 음성인식의 결과로부터 C-STAR IF(interchange format)까지의 해석 시스템과 C-STAR IF로부터 한국어, 영어, 일본어로 생성해 내는 생성 시스템의 성능으로 나누어서 평가한다.

1. 개요

미국의 카네기멜론 대학과 일본의 ATR 및 한국의 전자통신연구원 등이 가입한 C-STAR (Consortium for Speech Translation Advanced Research)[1]에서는 C-STAR-IF (interchange format)[2]를 이용하여 1999년 7월 22일에 국제간 음성언어번역 시스템 공동 시연을 하였다[3]. C-STAR IF는 일종의 중간언어(interlingua) 표현이다[1,2]. 각 회원국들은 자국의 음성인식부터 C-STAR IF의 생성과, C-STAR IF로부터 자국의 언어 생성만을 책임지기로 약속하였다.

이번 국제간 음성언어번역 시스템 공동 시연에서는 ETRI가 손님이 되어 미국, 일본, 프랑스의 여행사와 여행일정을 논의하고, 호텔이나 비행기 예약 등을 수행하였다. 또한 프랑스의 CLIPS가 손님 역할을 할 때, ETRI가 한국측 여행사가 되어 대전 여행에 대해 상담해 주고, 호텔 등을 예약해 주었다. 일본 ATR, 미국 카네기멜론 대학, 독일 칼슈르에 대학이 손님 역할을 한 경우엔 인사 세션에 참가하여 다국어 번역이 동시에 가능함을 보였다.

이 논문에서는 전사문장을 이용한 영어, 일본어, 한국어의 번역 결과와 음성인식 결과를 이용한 각 언어의 번역 결과를 평가하여 얻은 번역 시스템의 성능에 대해 기술한다. 그리고, 한국어 음성인식의 결과로부터 C-STAR IF를 생성해 내는 한국어 해석 시스템[1]과 C-STAR IF로부터 한국어, 영어, 일본어를 생성해 내는 생성시스템[1]의 성능도 세부 시스템 자체의 성능을 알아보기 위해 평가하였다. ETRI 측에서는 한국어 생성만을 하면 되지만, 외국 기관과의 연동이 이루어지지 않을 경우나, 자체 번역 시스템의 데모를 위해 영어와 일본어 생성 시스템도 구축하였다[1]. 그래서 영어, 일본어 생성 시스템의 성능도 함께 기술한다.

2. 번역 시스템의 성능

번역 시스템의 성능은 전사문장에 대한 의미 전달률과 음성인식 결과에 대한 의미 전달률로 나타낸다. 두 가지 모두 해석문법 작성에 사용되지 않은 데이터를 사용하여 평가하였다. 해석문법 작성에 사용된 데이터는 160 대화이고, 평가에 사용된 데이터는 20대화이다. 모두 여행계획 영역에 대한 것으로 여행 계획이나 안내에 대해 손님과 여행사측이 서로 주고 받는 대화를 모아서 전사한 것이다.

[1]C-STAR에 대한 보다 자세한 내용은 CSTAR 홈페이지 http://www.is.cs.cmu.edu/cstar/ 참조.

평가의 단위는 SDU(semantic dialog unit)[2]를 사용하였다. 이것은 하나의 C-STAR IF로 표현할 수 있는 단어의 열로, 한 단어가 될 수도 있고, 하나의 구나 문장이 될 수도 있다. 때로는 두 개의 문장이 하나의 SDU를 이루기도 한다. 평가에 사용된 20대화에는 총 753개의 SDU가 있다.

평가의 척도로 사용된 의미 전달률은 화자의 의도가 어느 정도 상대방에게 전달되었는지를 나타내는 것으로 다음의 4가지로 구분하였다. 첫째로 의미전달이 완벽하고, 생성문 자체에 전혀 오류가 없는 경우(perfect), 둘째로 의미전달이 충분히 가능하며 약간의 구문이나 형태소 오류가 생성문에 있는 경우(good), 셋째로 의미전달에서 부분적인 오류가 있지만 발화의 취지는 전달된 경우(okay), 그리고 의미전달에 실패한 경우(bad)이다. 번역결과가 없을 경우는 무조건 bad로 처리하였다. 이 중에서 bad를 제외한 경우를 의미전달이 성공한 것으로 보았다.

2.1 전사문장을 이용한 평가

전사문장을 이용한 평가는 음성인식의 결과가 100%인 경우와 동일한 것으로 녹음된 음성을 그대로 받아 적은 것이다. 대화 내용을 전사한 것이기 때문에 간투사나, 머뭇거림, 같은 표현의 반복 등도 포함되어 있다. 평가 결과는 표1과 같다. 세 개의 언어가 거의 비슷한 결과를 얻었지만, 그 중 영어가 조금 나은 것은 미등록어의 생성시에 C-STAR IF가 영어 이름을 그대로 value로 사용하기 때문에 영어 생성에서는 미등록 value일 경우에도 특별한 처리 없이 생성할 수가 있기 때문이다.

SDU수	한국어 의미전달률	영어 의미전달률	일본어 의미전달률
753	85%	87%	84%

표1. 전사문장을 이용한 번역 시스템의 성능

2.2 음성인식 결과를 이용한 평가

음성인식 결과를 이용한 평가는 2.1과 같은 데이터에 대해 음성 인식기를 먼저 실행하여 얻은 결과를 이용한다. 그런데 음성인식기가 현재 가지고 있는 음성 데이타가 대화의 절반인 손님측만 있기 때문에 평가도 손님에 해당되는 부분으로 제한하였다. 그래서 사용된 SDU의 수가 322개이다. 음성 인식기의 결과는 오인식으로 인해 단어의 추가나 변경, 삭제 등이 일어날 수 있다. 평가 결과는 표2와 같다.

SDU수	한국어 의미전달률	영어 의미전달률	일본어 의미전달률
322	79%	79%	81%

표2. 음성인식결과를 이용한 번역 시스템의 성능

평가시 음성인식의 결과는 평가자에게 제공하지 않고, 2.1에 사용한 전사문장만을 보여 주었다. 의미전달이 실패한 경우에 그 원인이 음성인식의 오류 때문에 발생한 경우는 전체 실패의 33%에 해당된다. 이것은 음성인식의 오인식이 번역 시스템에 크게 영향을 미치지 못함을 의미하는 것으로 많은 음성인식의 오인식된 결과와 번역 시스템의 번역 실패의 결과가 상당부분 일치함을 의미한다.

3. 해석 시스템의 성능

전체 번역 시스템의 성능은 2장에서 알아보았다. 이 장과 다음 장에서는 번역 시스템의 두 부분인 해석 시스템과 생성 시스템 각각의 성능에 대해 기술한다.

한국어 해석 시스템의 성능을 알아보기 위해 해석문법 작성에 직접 사용한 160 대화와 사용하지 않은 20대화에 대해 각각 평가를 하였다. 평가의 기준은 SDU 단위로 올바른 IF가 만들어 졌는지를 판단하여 성공여부를 결정하였다.

평가 결과는 표3과 같다. 160대화는 1969개의 발화와 4909개의 SDU가 있고, 20대화에는 246개의 발화와 753개의 SDU가 들어 있다.

해석문법에 사용된 대화 수	사용대화 IF 해석성공률	해석문법에 미사용 대화 수	미사용 대화 IF 해석 성공률
160	93 %	20	80%

표3. 해석 시스템의 성능

표3에서 보는 바와 같이 해석문법에 사용된 160대화에 대한 시험 결과는 93%의 해석 성공이다. 여기서 발생한 에러의 대부분은 표현 자체가 가지는 모호성 때문이다. 예를 들면, "대한항공입니다" 라는 표현은 "대한항공과 아시아나항공이 있는데 어느 것으로 하시겠습니까" 라는 질문에 대한 답일 경우는 정보제공을 의미하는 give-information이란 화행에 해당되지만, 전화가 걸려왔을 때 전화를 받으면서 말한 것이라면 자기 소개를 의미하는 introduce-self라는 화행에 해당된다. 현재의 해석 시스템은 대화 관리를 하지 않고 문장 자체만 보고 IF를 결정하기 때문에 표현 자체가 갖는 모호성은 처리를 못하고 있다. 그 외에 복잡한 날짜 표현, 여행 안내 외의 다른 영역의 표현 등도 에러의 유형으로 나타나고 있다. 다음으로 해석문법 작성에 사용하지 않은 40대화에 대해 시험한 결과는 80%의 해석 성공률이었다. 같은 영역에 대한 데이타인데도 해석 성공률이 낮게 나오는 이유는 아직 해석문법이 완전하지 못하기 때문이다. 특히 미등록어 처리나, 간투사가 끼어 드는 발화에 대한 처리가 아직 미흡하다. 보다 나은 해석 시스템을 위해 한국어의 특징인 부분 자유어순을 처리할 수 있도록 해야 하며, 미등록 표현에 대한 대처 방안도 고려되어야 한다. 아울러 모호성 처리 부분도 추가되어야 한다.

4. 생성 시스템의 성능

생성 시스템의 평가는 해석문법 작성에 사용하지 않은 20대화를 이용한 평가와 1999년 7월 22일의 C-STAR 국제간 시연에 전 회원국이 사용한 IF를 대상으로 한 평가로 나누어 시행하였다.

4.1 해석문법 작성에 사용하지 않은 20대화를 이용한 평가

평가 기준은 첫째, 입력된 IF와 생성된 문장이 의미적으로 동일한지 아닌지가 된다. 동일한 경우에 성공, 동일하지 않으면 실패로 본다. 이때 IF에 들어있는 모든 정보(속성과 value)가 각 언어의 문장으로 정확하게 생성되어야 성공으로 본다. 둘째, 의미적으로는 동일하지만 생성문 자체가 문법에 맞지 않거나 부자연스러운 경우는 실패로 간주한다. 한국어, 영어, 일본어 모두 같은 기준을 사용한다. 평가 결과는 표4와 같다.

IF수	한국어 생성성공률	영어 생성성공률	일본어 생성성공률
753	98%	97%	95%

표4. 해석문법작성에 미사용한 20대화를 이용한 생성 시스템의 성능

평가 결과는 예상한 대로 95% 이상의 높은 성공률이다. 생성은 해석과는 달리 생성문법에 규칙이 있을 경우 100% 가까이 성공하기 때문이다. 실패한 경우는 해석시스템에서 잘못된 IF를 생성해 낸 경우가 있었고, 어떤 경우는 생성문법을 잘못 작성하여 실패한 경우도 있었다. 예를 들면, 한국어 생성의 경우 조사가 겹쳐 나왔거나, 잘못된 조사가 나온 경우이다. 이것은 생성문법을 수정하면 해결된다.

4.2 국제간 시연에 나타난 IF를 이용한 평가

C-STAR 회원국들의 국제간 시연에서 ETRI가 직접 시연에 참가한 경우도 있었고, 그렇지 않은 경우도 있었다. ETRI가 참가하지 않은 경우는 미국이 독일, 프랑스, 이탈리아, 일본과의 시연, 일본이 미국, 독일과의 시연, 프랑스가 미국, 독일과의 시연, 이탈리아가 미국, 독일과의 시연, 독일이 미국, 일본과의 시연 등이다. 평가 방법은 시연 전체에 사용된 IF를 대상으로 4.1에 제시한 평가 기준에 의해 평가하였다. 이론적으로 C-STAR IF는 DA(dialog act)[2]만을 대상으로 할 경우에도 100만개가 넘는다. 그러나 실제 각 회원국에서 처리할 수 있는 유일한 DA의 수는 500~2000 정도이다. ETRI 자체 테스트의 경우는 해석 시스템의 결과로 얻어진 IF에 대한 규칙이 이미 대부분의 생성문법에 반영되어 있기 때문에 4.1에서와 같은 높은 생성성공이 가능하다. 그러나, 다른 회원국들간의 데모는 비록 같은 여행계획 영역이라 할지라도 사용하는 DA는 많이 다를 수 있다. 그래서 이 방법은 ETRI 생성 시스템의 상대적 성능을 알 수 있는 좋은 시험 방법이 될 수 있다. 평가 결과는 표5와 같다. 3개 언어 모두 비슷한 결과가 나왔다. 여기서 IF수는 국제간 시연 당일에 사용된 모든 회원국들간의 시연에 나타난 IF의 총 수이다.

IF수	한국어 생성성공률	영어 생성성공률	일본어 생성성공률
1829	88.1%	87.3%	88.1%

표5. 국제간 시연에 나타난 IF를 이용한 생성 시스템의 성능

평가 결과가 90%에 육박한 것은 ETRI가 C-STAR 의 어느 회원국과 연결하여 번역 시스템 시연을 하더라도 많은 부분의 시스템 보완 없이도 시연을 보일 수 있음을 의미한다고 볼 수 있다. 생성 실패의 원인은 크게 두 가지로 볼 수 있는데 첫째는 ETRI 생성문법에 들어 있지 않은 미등록어(주로 고유명사)가 있는 경우이다. 둘째는, 생성문법에 해당 DA가 없거나, DA는 있지만 속성이 부족하여 자연스럽지 못한 문장을 생성하는 경우이다. 아래에 생성 에러의 유형별로 정리하여 기술하였다.

- 고유명사가 없어서 생성에 실패한 경우

각 언어에서 모두 나타나고 있는 생성 실패 원인중의 하나이다. 영어는 16개, 한국어는 41개, 일본어에서도 고유명사로 인한 생성 실패가 나타나고 있다. 영어에 비해 한국어에서 미등록어(주로 고유명사)로 인한 실패가 많이 나타난 이유는 IF가 기본적으로 영어를 바탕으로 이루어졌기 때문이다. 영어의 경우 유럽의 새 화폐 단위인 euro, 고유명사인 museum_of_modern_art 같은 명사들은 value로 등록되어 있지 않아도 그 value 자체를 출력해 주기 때문에 잘못된 생성인지 구분하기 어렵다. 그러나 한국어의 경우 euro, museum_of_modern_art 등의 단어들이 문법에 등록되어 있지 않으면 한글 문장 속에 영어로 나오기 때문에 생성에 실패하는 경우가 영어보다 더 많게 된다.

- 각 언어의 생성문법에 해당 DA가 없어서 생성 실패한 경우

영어에서는 31개, 한국어는 31 개, 일본어는 18개의 DA가 생성문법에 없었다. 이런 경우는 발생 가능한 DA지만, 그 동안 ETRI에서 다룬 표현에서는 나타나지 않은 것들이다. 생성문법에 해당 DA 자체가 없는 경우는 아무 것도 생성되지 않는다. 그래서 실제 데모 상황에서는 이런 경우에 상대방은 인식도 잘 되고, IF 생성도 잘 되어서 안심하고, 우리 측의 반응을 기다리고 있는데 우리측에서는 무슨 내용인지를 알

수 없기 때문에 다시 말해 달라고 부탁하는 수밖에 없다. 그러면 상대는 똑같이 말하고 또 기다리게 된다. 그래서 다른 회원국과의 시연에서는 이런 상황을 시연자에게 알리는 방법이 필요하다.

- 상대 기관에서 이전의 규칙을 사용한 경우

C-STAR IF 규칙이 99년 4월에 수정되었는데도 일부 회원국에서 수정된 것을 사용하지 않고, 기존의 규칙을 사용해서 에러가 발생한 경우이다. 이 경우는 여러 회원국들이 함께 작업을 하다 보니 상대적으로 진도가 뒤진 회원국이 있었기 때문이다.

- DA는 있지만 속성 부족이거나 문법적으로 잘못된 생성결과로 인해 생성 실패한 경우

DA가 각 언어의 생성문법 속에 있지만, 속성들을 포함하는 규칙이 없어서 부자연스러운 문장을 생성해 내는 경우이다. 이런 경우는 비록 의미는 어느 정도 파악 가능하더라도 자연스럽지 못하기 때문에 생성 실패로 판단했다. 그리고 영어에서는 단어의 불필요한 겹침(예: train train), 한국어에서는 어색한 조사의 사용(예: 쇼이 있습니다->쇼가 있습니다)등도 생성 실패의 주요 원인이었다. 이런 현상은 영어에서는 92개, 한국어는 74 개, 일본어는 61개의 IF에서 나타났다.

5. 마무리

이 논문은 C-STAR의 1999년 국제간 음성언어번역 시스템 시연에 사용된 ETRI의 언어번역 시스템을 시연 결과를 중심으로 평가한 것이다. 시연 영역이 여행 계획과 안내로 한정되어 있고, 다양한 표현을 아직은 처리하지 못하는 한계가 있지만, 6개 C-STAR 회원국들이 상호 연결하여 여러 언어의 번역 시연을 큰 어려움 없이 할 수 있음을 보인 것은 커다란 성과라고 볼 수 있다.

번역 시스템의 성능은 아직 실용화 시스템과는 거리가 많이 있다. 실제 상황에서는 현재 해석문법에 들어 있는 표현보다 훨씬 다양한 표현들이 사용된다. 그리고 음성인식의 에러도 가정해야 하기 때문에 보다 다양한 표현을 다룰 수 있도록 해석 시스템의 성능 개선이 요구된다. 생성 시스템의 경우는 적은 수의 DA를 이용하여 50만 개가 넘는 DA를 처리하기 위해 미등록 DA의 처리에 중점을 두고 연구가 지속되어야 한다.

감사의 글

이 연구는 정보통신부의 지원에 의해 이루어진 결과물입니다.

참고문헌

[1] 최운천, "CSTAR-IF를 이용한 다국어 대화체 번역시스템," 제10회 한글 및 한국어 정보처리 학술대회 논문집, pp. 159-163, 1998.

[2] 최운천, "다국어 대화체 음성언어 번역시스템을 위한 IF와 IF태깅" 제15회 음성통신 및 신호처리워크샵 논문집, pp.409-412, 1998.

[3] Jun park, Kyuwoong Hwang, Un-Cheon Choi, Junko Hosaka, Siong Hun Yi and Jae-Woo Yang, "Spoken language Translation System:Development and Demonstration," International Conference on Speech Processing(ICSP'99), Vol. 2, pp. 535-538, 1999.

[4] Nam-Yong Han, Un-Cheon Choi and Youngjik Lee, "An Implementation of a Partial Parser in the Spoken Language Translator", The 1998 International Conference on Acoustics, Speech and Signal Processing(ICASSP'98), Vol. I, pp. 205-208, 1998.

[5] 최운천, 한남용, 김재훈, "대화체 음성언어 번역시스템에서의 개념기반 번역 시스템" 한국정보처리학회 논문지, 제4권 제 8호, pp.2025-2037, 1997.