외국인 관광객을 위한 다국어 통번역 시스템

최승권, 김영길
한국전자통신연구원 언어처리연구실
e-mail : {choisk, kimyk}@etri.re.kr

Multilingual Speech and Machine Translation System for Foreign Tourists

Sung-Kwon Choi, Young-Gil Kim
Natural Language Processing Research Section, ETRI

요 약

본 논문은 현재 개발 중에 있는 외국인 관광객을 위한 다국어 통번역 시스템을 기술하는 것을 목표로 한다. 다국어 통번역 시스템에서 개발 중에 있는 언어는 한국어, 일본어, 중국어, 영어, 스페인어, 불어, 독일어, 러시아어이다. 이렇게 개발된 다국어 통번역 시스템은 2018 년 평창 동계 올림픽 때 다국어 통번역 서비스를 제공할 예정이다. 현재의 다국어 통번역 시스템의 성능은 번역만 보았을 때, 영한 87.63%, 한영 88.21%, 중한 85.38%, 한중 77.94%, 일한 89.00%, 한일 86.69%, 스한 76.90%, 한스 77.46%, 불한 76.28%, 한불 79.78%이다.

1. 서론

다국어 통번역 시스템은 2 개 이상의 언어를 대상으로 하는 자동 통역 및 자동 번역을 말한다 [1].
외국인 관광객 중에 중국과 일본인 관광객이 한국에서 쇼핑시 겪는 불편사항으로 언어소통이 가장 크다고 조사된 바 있다[2]. 또한 외국인 2 명의 '부여.전주 1 박 2 일 여행'을 동행해보니 영어로 의사소통을 거의 할 수 없어 외국인이 국내의 지방을 관광하는 것은 거의 불가능할 정도였다는 보도가 있었다[3]. 2018 년도의 평창 동계 올림픽에는 80 개국 이상이 참가할 예정인데 외국인들이 한국에서 겪을 불편사항은 여전히 언어소통일 것이라고 추정된다. 따라서 평창 동계 올림픽의 주요 국가들의 언어인 영어, 중국어, 일본어, 불어, 스페인어, 독일어, 러시아어 등의 통역이 필요할 것이다. 하지만 이런 수요에 비해 통역사의 숫자가 부족하고 또한 통번역에 소요되는 경제적 부담도 크다. 그리고 일반인의 경우에는 통역사의 도움을 받는 것 자체가 쉽지 않다. 최근 국내에서 자동 통역 및 자동 번역에 대한 일반인들의 관심이 커졌는데 그 이유는 자동 통번역에 대한 필요성이 커진 것과 더불어 구글(Google)의 자동 통번역 서비스나 한국전자통신연구원(ETRI)의 자동 통역기인 GenieTalk 의 자동 통번역 서비스로 기인한 바가 크다.
본 논문에서는 현재 개발 중에 있는 한/영/중/일/불/스 외국인 관광객을 위한 다국어 통번역 시스템을 기술하고자 한다. [1]

[1] 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신.방송 연구 개발사업의 일환으로 수행하였음. [R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발]

2. 다국어 통번역 시스템 구성도

다국어 자동 통번역은 여러 언어를 하나의 번역 엔진에서 처리해야 하기 때문에 양국어 자동 통번역의 방법론과는 다르다. 그림 1 은 다국어 통역 시스템의 인터페이스 및 간략한 구성도를 보인다.

(그림 1) 다국어 통역 시스템의 인터페이스 및 간략한 구성도

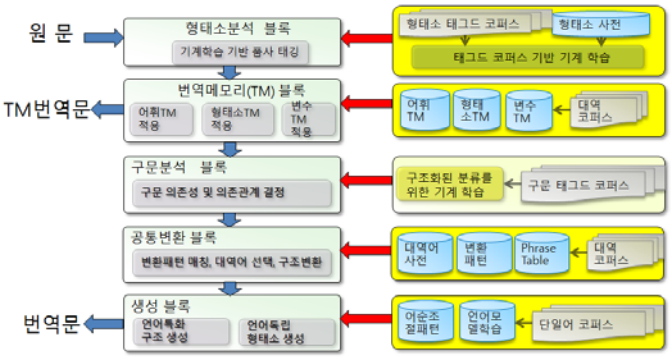
그림 1 의 다국어 통역 시스템에서는 다국어 음성 인식, 다국어 자동 번역, 개별적인 언어 합성에 의해 통역이 되는 것을 알 수 있다. 즉 다국어를 음성 인식하는 엔진이 하나이고, 다국어를 자동 번역하는 번역 엔진이 하나인 것을 알 수 있다. 다국어 자동 번역에 대해서는 다음 장에서 자세히 기술하도록 하겠다.

3. 다국어 자동 번역 방법 및 구성도

다국어 자동 번역은 다국어를 하나의 엔진에서 처리하는 것이 중요하다. 그래서 다국어를 자동 번역하기 위한 방법으로 다음과 같은 사항들을 고려하였다:

- 다국어에 동일한 공통 변환: 개별언어들에 상관없이 다국어 변환과정에서는 동일한 변환 포맷을 사용함으로써 다국어 확장이 용이하도록 한다.
- 하이브리드 번역 방법: 규칙기반 자동 번역 결과와 통계기반 자동 번역 결과 중에 더 좋은 번역결과를 선택하여 번역하도록 한다.
- 기계학습에 의한 번역지식 획득: 언어학자들이 만든 품사 또는 구문 부착 말뭉치로부터 분석지식과 다국어 번역용 변환지식을 자동으로 획득한다.

그림 2 는 다국어 자동 번역 시스템의 구성도이다.



(그림 2) 다국어 자동 번역 시스템 구성도

원문이 입력되면 형태소 분석 블록에서 형태소 분석이 이루어진다. 형태소 분석된 결과가 이미 저장되어 있는 번역메모리(TM:Translation Memory)와 일치하면 번역메모리에 의해 번역이 완료되어 TM 번역문이 출력된다. 만약 번역메모리에 의해 번역이 되지 않으면 번역메모리에 적용되지 않은 형태소 분석 결과는 구문분석 블록에서 의존 구조로 만들어 진 후에 공통변환 블록의 입력이 된다. 의존 분석 결과는 우선 특수 변환 규칙인 어휘 패턴과 일치 여부를 묻게 된다. 일치하는 공통 변환의 어휘패턴이 존재하면 언어 의존적인 변환이 이루어지며, 그렇지 않다면 어휘가 없는 다국어 공통 변환 규칙에 의해 변환이 이루어진다. 그 이후에 생성 블록에서 구조 및 형태소 생성에 의해 해당 언어의 번역문이 만들어지게 된다.

4. 실험

외국인 관광객을 위한 다국어 통번역 시스템의 평가는 다국어 번역만을 평가하였다. 원문의 의미가 정확히 번역되는 가를 측정하는 적합성 평가를 실시하였다. 평가문장 및 평가기준은 다음과 같았다.

- 평가문장: 여행 문장 중 임의로 추출한 언어별 500 문장.
- 평가기준: DARPA(Defense Advanced Research Projects Agency) 적합성 평가기준 [4]

<표 1> 적합성 평가 기준

점수	기 준
4	All meaning expressed in the source fragment appears in the translation fragment.
3	Most of the source fragment meaning is expressed in the translation fragment.
2	Much of the source fragment meaning is expressed in the translation fragment.
1	Little of the source fragment meaning is expressed in the translation fragment.
0	None of the meaning expressed in the source fragment is in the translation fragment.

적합성(adequacy) 평가 기준에 따른 번역률 식과 평가결과는 다음과 같았다.

$$\text{번역률}(\%) = 100 \times \frac{1}{|S|}\sum_{s\in S}\frac{1}{|T|}\sum_{t\in T}\frac{score_{s,t}}{4}$$

(S 는 번역문장수, T 는 번역자수, SCOREs,t 는 번역자 t 에 의해 평가된 문장 s 의 평가점수를 의미한다)

<표 2> 다국어 자동 번역 평가 결과

언어쌍	영한	한영	중한	한중	일한
번역률(%)	87.63	88.21	85.38	77.94	89.0
언어쌍	한일	스한	한스	불한	한불
번역률(%)	86.69	76.90	77.46	76.28	79.78

5. 결론

현재 개발 중에 있는 외국인 관광객용 다국어 통번역 시스템은 평창 동계 올림픽 때는 독일어와 러시아어도 포함하는 다국어 자동 통번역 서비스를 제공할 예정이다.

참고문헌

[1] 최승권, 홍문표, 박상규. “다국어 자동번역 기술”. 전자통신동향분석, 제 20 권 제 5 호, 16-27. 2005.

[2] 스포츠닷컴. “한국 방문 중일 쇼핑관광객 불만, 무엇인가?” http://www.sportsnews25.com/xe/index.php?mid=social&page=91&listStyle=webzine&document_srl=35903. 2014.5.10.

[3] 조선일보. “영어라곤 'No Smoking' 'Toilet'… 지방 여행은 오지 탐험 같더라” http://biz.chosun.com/site/data/html_dir/2016/09/02/2016090200248.html. 2016.9.2.

[4] Doyon, J., Taylor, K., & White, J.S. “The DARPA MT evaluation methodology: Past and present”. Proceedings of the Association for Machine Translation in the Americas, pp.1-4. 1988.