n-best 리랭킹을 이용한 한-영 통계적 음성 번역*

이동현* 이종훈 이근배
포항공과대학교 컴퓨터공학과 지능소프트웨어연구실
{semko, jh21983, gblee}@postech.ac.kr

Korean-English statistical speech translation Using n-best re-ranking

Donghyeon Lee*, Jonghoon Lee and Gary Geunbae Lee
iSoft Lab., Dept. of Computer Science and Engineering,
Pohang University of Science and Technology

요 약

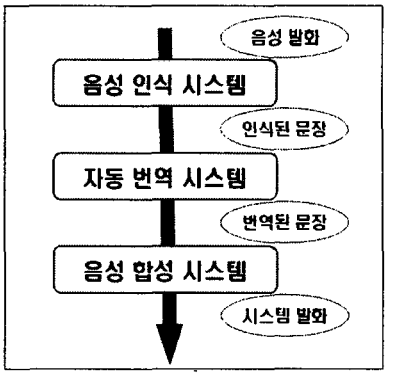
본 논문에서는 n-best 리랭킹을 이용한 한-영 통계적 음성 번역 시스템에 대해 논하고 있다. 보통의 음성 번역 시스템은 음성 인식 시스템, 자동 번역 시스템, 음성 합성 시스템이 순차적으로 결합되어 있다. 하지만 본 시스템은 음성 인식 오류에 보다 강인한 시스템을 만들기 위해 음성 인식 시스템으로부터 n-best 인식 문장을 추출하여 번역 결과와 함께 리랭킹의 과정을 거친다. 자동 번역 시스템으로 구절기반 통계적 자동 번역 모델을 사용하여, 음성 인식기의 발음 모델에서 기본 단어 단위와 맞추어 번역 모델과 언어 모델을 훈련시킴으로써 음성 번역 시스템에서 형태소 분석기를 제거할 수 있다. 또한, 음성 인식 시스템에서 상황 별로 언어 모델을 분리하여 처리함으로써 자동 번역 시스템에 비해 부족한 음성 인식 시스템의 처리 범위를 보완할 수 있었다.

1. 서론

세계화, 국제화 시대를 살고 있는 오늘날 국가간 교류가 점점 늘어남에 따라 다른 언어를 사용하면서 발생하는 의사소통 문제가 종종 발생한다. 전문 통역사가 이런 문제를 해결해줄 수 있지만, 늘어날 수요를 고려했을 때 인력이 부족한 편이고 비용 문제도 일반인에게는 부담스럽다. 이런 이유 때문에 음성 번역 시스템의 필요성은 점차 커지고 있다.

음성 번역 시스템은 군사적인 목적에 의해 처음 개발이 시도되었다. 전쟁 상황에서 현지인과의 의사소통을 위해서였다. 하지만, 지금은 다양한 목적으로도 음성 번역 시스템이 요구되고 있다. 대표적인 경우로 방송, 강의, 여행 등에서의 음성 번역 시스템은 사람들에게 큰 편리함을 가져다 줄 수 있다.

음성 번역 시스템은 <그림1>과 같이 크게 음성 인식 시스템, 자동 번역 시스템, 음성 합성 시스템으로 구성된다. 현재까지 대부분의 음성 번역 시스템은 순차적인 통합을 한다. 원시 언어로 된 사용자의 발화를 입력으로 받아 음성 인식 시스템이 문장으로 출력하고, 출력된 문장은 자동 번역 시스템을 거쳐 목적 언어로 번역된 문장이 된다. 이렇게 번역된 문장을 다시 음성 합성 시스템이 목적 언어로 된 시스템 발화로 출력해 낸다.



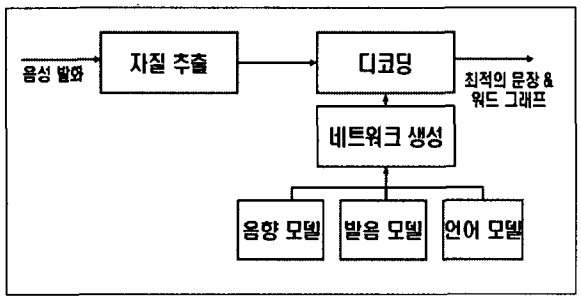
<그림 1> 순차적인 통합에 의한 음성 번역 시스템

* "본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음" (IITA-2005-C1090-0501-0018)

음성 언어를 다루는 시스템에서 음성 인식 시스템은 매우 중요한 역할을 한다. 하지만, 음성 인식 시스템은 종종 오류를 수반하므로 이러한 시스템을 오류에 강인하게 설계할 필요가 있다. 이 논문에서는 음성 인식 오류에 강인한 음성 번역 시스템을 제안한다. 여기서 제안한 음성 번역 시스템에서는 음성 인식 시스템의 1-best 출력 대신에 n-best 출력을 받은 뒤, 각각의 번역 결과를 고려한 리랭킹(Re-ranking)을 하였다.

2장과 3장에서는 음성 번역 시스템에서 사용한 한국어 연속 음성 인식 시스템과 구절기반 통계적 자동 번역 시스템에 대해 간략하게 소개한다. 4장에서는 음성 인식 시스템과 자동 번역 시스템이 어떻게 결합하였는지 전체적인 시스템 구조에 대해 소개하며, 5장에서는 음성 발화에 대한 실험을 수행하고 그 결과를 분석한다. 6장에서는 결론과 향후 과제에 대해 검토한다.

2. 한국어 연속 음성 인식 시스템



<그림 2> 음성 인식 시스템

본 연구에서 사용한 한국어 연속 음성 인식 시스템은 HTK(hidden Markov Model Toolkit)[1]를 기반으로 하였다. 이 시스템은 발성 화자에 관계없는 음성 인식을 하는 화자 독립 시스템이다. 인식 단위로는 음소 기반의 유사음소단위(PLU, Phoneme Like Unit)를 사용하며, 48개로 구성된 유사음소단위 세트를 이용하였다.

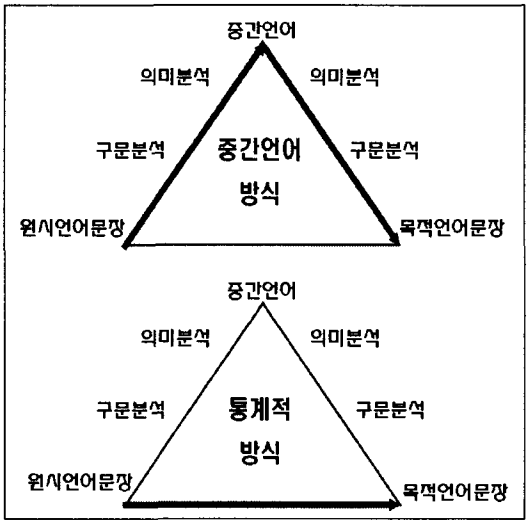
음향 모델은 상태 공유의 연속 히든 마코프 모델을 사용하였다. 모든 히든 마코프 모델은 각각 세 개의 상태를 가지고, 각 상태의 출력 확률 값은 다수의 가우시안 혼합분포로부터 계산된다. 발음 모델에서 기본 단어 단위는 의사 형태소 단위로 표현한다. 언어 모델로는 bigram을 사용하며, 워드 그래프가 생성된 이후에 trigram을 다시 적용한다.

주어진 모델에서 디코더는 토큰 전파 알고리즘에 기반한 비터비(Viterbi) 디코딩 방법을 사용하여 음성 발화에 대해 가장 확률이 높은 문장을 찾는다. 디코딩 과정에서 활성화된 히든 마코프 모델의 수를 줄이기 위해 빔 프루닝(Beam Pruning)을 사용한다. 최종 출력으로 최적의 문장과 워드 그래프를 생성하며, 워드 그래프로부터 N개의 최적 문장을 추출할 수 있다. 음성 인식 시스템의 구조는 <그림2>와 같다.

3. 구절기반 통계적 자동 번역 시스템

자동 번역 분야는 컴퓨터가 처음 만들어진 시점부터 오랜 기간 동안 많은 발전을 해왔다. 자동 번역의 가장 궁극적인 방법으로는 중간 언어(Interlingua) 방식을 이용한 번역 방법이 있다. 중간 언어 방식은 한 언어로부터 여러 언어로 번역하기에 적합한 방법으로 중간 언어 표현이 모든 언어에 무관하게 정의되어야 하며, 이는 매우 어려운 작업이다. <그림3>에 나타난 자동 번역 피라미드를 통해 알 수 있듯이 중간 언어 방식은 자연 언어 처리의 거의 모든 기술이 포함되어 있다. 따라서, 각 단계를 거치면서 오류가 계속 누적되는 단점이 있다.

이에 반해 통계적 방식은 가공 되지 않은 원시언어 문장에서 바로 목적언어 문장으로 번역을 해낸다. <그림3>에서 알 수 있듯이 다른 자연 언어 처리 기술이나 언어학적 지식은 크게 요구되지 않는다. 많은 병렬 말뭉치가 필요하지만, 자연 언어 처리 기술이 부족한 언어들에 대해서도 적용이 가능하다는 점에서 실용적이고 규칙 기반 방식에 비해서 적은 비용이 소모되고 빠른 프로토타이핑이 가능하다는 점에서 경제적이다. 이런 여러 가지 장점과 더불어 실제 번역 성능도 다른 방식에 비해 우수하다[2].



<그림 3> 자동 번역 피라미드

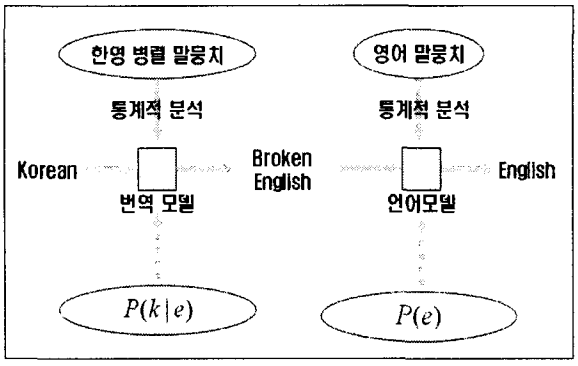
통계적 방식으로는 크게 예제기반(Example-based) 통계적 자동 번역, 단어기반(Word-based) 통계적 자동 번역, 구절기반(Phrase-based) 통계적 자동 번역,

구문기반(Syntactic-based) 통계적 자동 번역 등이 있다. 본 연구에서 사용한 시스템은 구절기반 통계적 자동 번역 시스템으로, 통계적 방식 중에서도 현재 가장 많이 쓰이고 뛰어난 번역 성능을 나타내고 있다.

3.1 구절기반 통계적 자동 번역

구절기반 통계적 자동 번역에서의 번역 단위로 구절을 사용한다. 여기서의 구절은 구문론에서 의미하는 구절이 아니라 단순히 연속된 단어들을 나타낸다. 대표적인 구절기반 통계적 자동 번역 시스템으로는 파라오(Pharaoh)[3]가 있다. 본 연구에서 사용한 구절기반 통계적 자동 번역 시스템도 파라오를 바탕으로 개발하였다.

통계적 번역 시스템은 언어 모델, 번역 모델, 디코딩 알고리즘으로 구성된다. 언어 모델은 문장의 확률을 제공하고, 번역 모델은 한 언어의 구절이 다른 언어의 구절로 번역될 확률을 제공한다. 디코딩은 주어진 문장과 번역 모델, 언어 모델에서 가장 높은 확률을 가진 문장을 찾는 과정이다. 전체적인 구조는 <그림4>와 같다.



<그림 4> 통계적 자동 번역의 구조

3.2 언어 모델, 번역 모델, 디코딩

언어 모델은 SRILM toolkit[4]을 사용하여 trigram 언어 모델을 생성했다. 번역 모델은 파라오 훈련 모듈과 GIZA++[5]를 통해 만들었다. GIZA는 통계적 자동 번역 시스템인 EGYPT의 부분으로서 단어 정렬 툴이다. GIZA++는 이런 GIZA를 확장한 것으로 IBM Model 5를 구현했다. GIZA++를 통해 단어 정렬(Word Alignment)을 언어 사이의 양방향으로 수행한 뒤 몇 가지 휴리스틱을 사용하여 양쪽 언어 사이의 구절쌍(Phrase Pair)을 추출한다.

이 시스템에서 번역은 그래프 탐색 문제로 볼 수 있다. 즉, 최적의 번역 문장을 찾는 것은 워드 그래프에서 최적의 패스를 찾는 것과 같다. 이런 과정을 수행하기 위해서 스택 디코딩 방식을 사용하며, 스코어는 언어 모델과 번역 모델로부터 계산 된다. 또한, 문장 길이에 따라 탐색 공간이 지수적으로 늘어나기 때문에 프루닝을 사용한다.

3.3 한영 번역에서의 성능 향상 방법

기본적인 구절기반 통계적 자동 번역 시스템에서 몇 가지 방법을 사용하여 한영 번역의 성능을 향상시킬 수 있다[6].

한국어에서는 영어와 달리 띄어쓰기 단위가 의미 단위와 일치하지 않는다. 번역 단위가 되는 구절을 이루는 최소 단위가 띄어쓰기 단위이므로 잘못된 단어 정렬이 될 수 있다. 훈련 단계 이전에 한국어 문장 말뭉치에 대하여 형태소 분석기를 돌리면, 형태소 단위로 분리할 수 있어 성능이 향상되었다.

구절은 연속된 단어에 의해서만 형성되므로 한국어와 영어 같이 문장 구조가 다른 경우에 제약된 한영 구절쌍을 추출하게 된다. 미리 정의한 룰을 이용한 파스 트리의 재구성(Restructuring)을 통해 한국어와 영어 사이의 구조를 비슷하게 맞추고 어느 정도의 파서 성능이 보장되면 번역 성능이 향상 된다. 하지만, 한국어 파서의 경우 아직 성능이 부족해 실험에서는 성능이 떨어졌다.

한국어 단어에서 격조사나 어미 같은 경우에는 의미적으로 영어 단어와 정렬이 될 수 없다. GIZA++에 의해 단어 정렬을 수행하면 이 같은 단어는 종종 엉뚱한 단어에 정렬이 된다. 따라서, 훈련 단위에서 이런 단어들을 제거하면 보다 나은 워드 정렬이 되어 전체적인 성능이 향상 되었다.

영어에서는 한국어와 달리 평서문과 의문문의 단어 순서에 차이가 있다. 따라서, 언어 모델을 문장의 형태에 따라 분리해서 적용하면 보다 적합한 문장 순서를 형성하여 성능이 향상 되었다.

실험 방법	%BLEU
B(기본 시스템)	26.97
B + T (형태소 분석기 사용)	31.54
B + T + R (문장 구조 재구성)	29.13
B + T + D (격조사/어미 삭제)	33.94
B + T + D + L(언어 모델 분리)	34.44
B + T + D + L + 사전 추가	35.19

<표 1> 한영 번역에서의 성능 향상 방법 및 성능

한영 사전은 한국어 단어와 같은 의미를 가진 영어 단어가 바로 대응되어 있다. 단어 정렬에 있어서 사

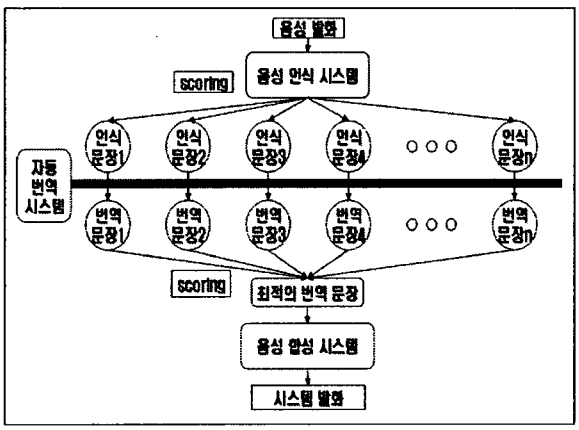
전에 대응된 단어들은 서로 정답이 될 수 있다. 따라서, 사전을 병렬말뭉치에 추가하여 훈련을 하면 보다 나은 단어 정렬이 되어 성능이 향상 되었다.

<표1>은 앞서 말한 방법들을 차례로 사용했을 때의 성능 향상 정도를 나타낸다.

4. n-best 리랭킹을 이용한 순차적 통합

4.1 n-best 리랭킹

음성 인식 시스템, 자동 번역 시스템, 음성 합성 시스템은 순차적인 접근(Cascading Approach) 방식을 이용하면 간단히 통합되어 음성 번역 시스템을 이루게 된다. 음성 번역 시스템을 구성하는 각각의 시스템은 입력에 대해 최적의 결과를 다음 시스템에 넘겨주기만 하면 된다. 하지만, 음성 인식 시스템에서부터 인식 오류가 발생했을 경우, 다음 시스템에도 영향을 끼치는 문제를 가지고 있다. n-best 리랭킹을 이용한 순차적인 시스템은 이런 문제점을 줄여줄 수 있다. <그림5>는 제안된 시스템의 구조를 나타낸다.



<그림 5> n-best 리랭킹을 이용한 음성 번역 시스템 구조

이전의 시스템은 음성 인식 시스템이 수행된 후 음성 인식 문장이 결정되고 음성 인식 문장이 자동 번역 시스템에 들어가 번역 문장이 결정된다. 반면, 이 시스템은 음성 인식 시스템의 수행 후에도 음성 인식 문장을 결정하지 않고 n-best 결과를 자동 번역 시스템에 넘겨 준 뒤 번역 결과까지 보고 n-best 리랭킹을 수행하여 최종적인 음성 인식 문장과 번역 문장을 결정한다.

이 시스템에서 사용한 음성 인식 시스템과 자동 번역 시스템은 모두 통계적 접근을 통한 확률 모델을 사용하여 스코어를 계산한다. n-best 리랭킹의 기준이 되는 스코어는 다음과 같이 계산한다.

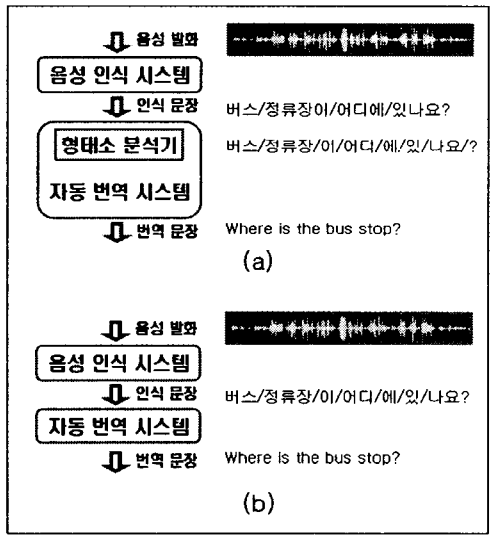
$$Score = ASRScore + (Weight \times SMTScore)$$

여기서 Weight는 음성 인식 스코어와 자동 번역 스코어 사이의 불균형을 조절하는 역할을 한다.

4.2 형태소 분석기 제거

자동 번역 시스템은 입력으로 완전한 문장을 받아 형태소 분석기를 돌린 후 디코딩을 수행한다. 따라서, 음성 인식 시스템의 출력을 완전한 문장 형태로 넘겨주기 위해서 <그림6>에서 (a)와 같이 발음 모델의 기본 단어 단위가 어절 단위가 되어야 한다. 하지만, 이 경우 단어의 개수가 늘어나 음성 인식 시스템에서의 탐색 공간이 매우 커진다.

이 문제를 해결하기 위해서 음성 인식 시스템에서 발음 모델의 기본 단어 단위는 의사 형태소 단위를 사용하면서, 자동 번역 시스템에서 한글 말뭉치를 의사 형태소 단위로 나타내어 훈련을 시킨다. 이러면, 자동 번역 시스템에서 음성 인식 시스템의 결과를 그대로 사용하여 디코딩을 수행할 수 있다. 따라서 <그림6>에서 (b)과 같이 형태소 분석기가 자동 번역 시스템에서 완전히 제거 될 수 있어 속도나 메모리 측면에서 효과를 볼 수 있을 뿐만 아니라 형태소 분석기 오류를 사전에 원천적으로 차단하게 된다.

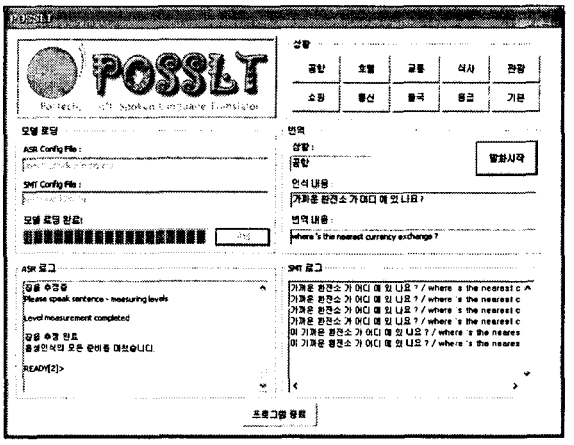


<그림 6> 형태소 분석기 제거

4.3 음성 인식 시스템에서의 언어 모델 분할

음성 인식 시스템은 상대적으로 자동 번역 시스템에 비해 탐색 공간이 매우 크다. 따라서, 음성 인식 시스템의 처리 범위가 자동 번역 시스템의 처리 범위에 비해 작을 수 밖에 없다. 이를 보완하기 위해서 음성 인식 시스템에서의 언어 모델을 분할 하였다.

사용자가 상황을 선택하면, 그 상황에서 정의된 언어 모델로 음성 인식을 수행한다. 사용자의 선택이 필요하지만, 음성 인식 처리 범위를 넓힐 수 있어 실용성을 높일 수 있다. <그림7>은 음성 번역 시스템의 실행 예제이다. 사용자가 상황을 선택한 뒤 발화를 하면 번역 결과에 대한 시스템 발화가 출력된다.



<그림 7> 음성 번역 시스템 실행 예제

5. 실험

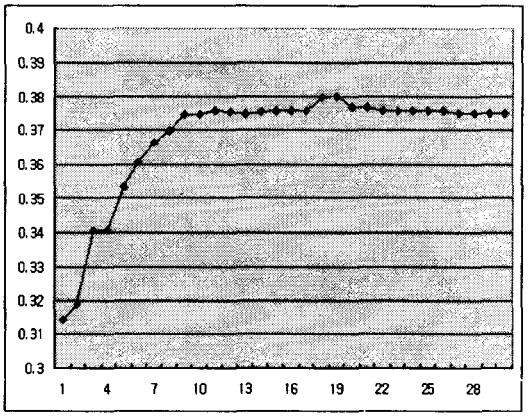
5.1 실험 환경

현재의 음성 번역 시스템은 여행 도메인에 맞추어져 있다. 4만 5천 문장의 한영 병렬말뭉치와 16만 단어의 한영 사전을 통해 자동 번역 시스템의 번역 모델과 언어 모델을 훈련 했다. 여기서 사용한 한글 말뭉치 중에서 1만 문장을 뽑아내어 10개의 상황에 나누어 음성 인식 시스템의 언어 모델을 훈련했다. 음성 인식 시스템의 단어 오인식률(WER)은 15.3%이며, 자동 번역 시스템의 BLEU 스코어는 0.352이다.

5.2 n-best 리랭킹 평가

실험을 위해 4만 5천 문장의 한영 병렬말뭉치 중에서 임의로 400개의 음성 발화를 사용하였다. 그 중 200개는 Weight를 학습하기 위해 사용했다. 선형회귀분석(linear regression)을 이용하여 최적의 Weight를 학습한 결과, 54.2가 나왔다. 그리고, 나머지 200개는 음성 번역 성능을 알아보기 위해 사용했다. 번역 성능을 측정하는 방법으로는 BLEU 스코어를 사용하였다 BLUE 스코어는 번역 성능을 자동으로 평가하기 위해 만들어진 것으로 현재 가장 널리 사용된다[7]. BLUE에서 스코어는 0에서 1사이로 나오며 좋은 번역일수록 수치가 높다.

<그림8>은 음성 인식 시스템에서 추출한 최적 문장의 개수에 따른 BLUE 스코어를 그래프로 나타는 것이다. 그래프에서 알 수 있듯이 전체적으로 최적 문장 개수를 많이 사용할수록 성능이 향상된다.



<그림 8> n-best 리랭킹에 따른 음성 번역 성능

사용자의 음성 발화를 음성 인식 시스템을 통해 텍스트로 인식하게 되는데, 음성 인식 시스템의 오류로 인해 인식된 텍스트는 본래 의도한 텍스트와 다소 차이가 있을 수 있다. 그 차이로 인해 음성 번역 시스템의 성능이 텍스트 번역 시스템보다 떨어지게 된다. <표2>에서도 알 수 있듯이 텍스트 입력을 번역 시스템에 사용한 경우 BLUE 스코어는 0.4902인데 반해, 음성 인식 시스템에서 1-best 결과만 사용하는 음성 번역 시스템의 경우 BLUE 스코어는 0.3143으로 크게 떨어진다.

여기서 제안한 음성 번역 시스템의 경우 음성 인식 시스템에서는 다수의 최적 문장을 뽑아서 번역 결과와 함께 리랭킹을 하여 떨어진 성능을 만회할 수 있게 된다. 즉, 음성 인식 오류에 보다 강건해 질 수 있다. <표2>에서 나타났듯이 n-best 리랭킹을 이용한 음성 번역 시스템을 통해 BLUE 스코어가 0.3143에서 0.3799로 향상되었다.

	BLEU 스코어
1-best	0.3143
n-best 리랭킹	0.3799
텍스트 입력	0.4902

<표 2> n-best 리랭킹의 효과

6. 결론

본 연구에서는 n-best 리랭킹을 이용한 순차적 통합에 기반하여 음성 번역 시스템을 개발하였다. n-best 리랭킹을 사용함으로써 음성 인식 오류에 보다 강인한 시스템을 구축할 수 있었다. 또한 형태소 분석기를 제거하여 속도를 향상 시키고, 음성 인식 시스템에서의 언어 모델을 분할하여 처리 범위를 향상 시켰다. 향후 계획은 성능을 보다 향상 시키기 위해 음성 인식 시스템과 자동 번역 시스템의 디코딩 방식이 유사한 점을 이용한 조인트 디코딩이 필요하다. 또한, 실용적으로 사용 될 수 있는 모바일 자동 통역 시스템을 위한 시스템 경량화 및 최적화도 요구된다.

참고 문헌

[1] S. Young, G. Everman, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK Book (for HTK Version 3.3)", *Cambridge University*, 2004.

[2] W. J. Hutchins, "Machine Translation: Past, Present, Future", *Chichester/New York: Ellis Horwood/Wiley*, 1986.

[3] P. Koehn, "Pharaoh: A Beam Search Decoder for Phrase based Statistical Machine Translation Models", *Proceedings of AMTA, Washington DC*, 2004.

[4] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", *Proceedings of ICSLP*, 2002.

[5] F. J. Och and H. Ney, "Improved statistical alignment models", *Proceedings of 38th Annual Meeting of the ACL*, pp. 440-447, *Hongkong, China, October* 2000.

[6] Jonghoon Lee, Donghyeon Lee, Gary Geunbae Lee, "Improving Phrase-based Korean-English Statistical Machine Translation", *Proceedings of Interspeech-ICSLP*, 2006.

[7] K. Papineni, S. Roukos, T. Ward, W. Zhu, "Bleu: a method for automatic evaluation of machine translation", *Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, York town Heights, NY, September*, 2001.