자동 평가 방법을 이용한 번역 지식 튜닝 시스템

박은진○, 김운○, 권오욱, 오영순, 김영길
한국전자통신연구원
{ejpark, wkim1019, ohwoog, suni, kimyk}@etri.re.kr

Translation Dictionary Tuning System By using of Auto-Evaluation Method

Eun-Jin Park○, Yun Jin○, Oh-Woog Kwon, Ying-Shun Wu, Young-Kil Kim
Natural Language Processing Research Team, ETRI

요 약

본 논문에서는 병렬 말뭉치에서 오류가 있을 것으로 추정되는 문장을 자동 추출하여, 다수의 번역 사전 구축 작업자가 자동 번역시스템을 직접 사용하면서 번역 사전을 튜닝하는 방법에 대하여 제안하고자 한다. 작업자는 병렬 말뭉치의 대역문을 이용하여 자동 번역 결과의 BLEU를 측정하고, 사전 수정 전과 후의 BLEU 차이를 정량적으로 제시해 줌으로써 양질의 번역 사전을 구축하도록 하였다. 대량의 번역 사전이 이미 구축된 자동 번역시스템에서 추가적인 성능향상을 위해 대량의 말뭉치에서 미등록어, 번역패턴 등을 추출하여, 대량으로 구축하는 기존 방법에 비해 사전 구축 부작용이 적으며, 자동번역 성능향상에 더 기여하는 것을 실험을 통해 증명하였다. 이를 위해 본 논문에서는 중한 자동 번역시스템을 대상으로, 중국어 문장 2,193문장에 대해, 사전 구축 작업자 2명이 2주간 튜닝한 결과와 15만 말뭉치에서 추출한 미등록어 후보 2만 엔트리를 3명의 사전 구축 작업자가 미등록어 선별, 품사 및 대역어 부착한 결과 7,200 엔트리를 대상으로 자동평가를 실시하였다. 실험결과 미등록어 추가에 의한 BLEU 성능향상은 +3인데 반해, 약 2,000문장 튜닝 후 BLEU를 +12 향상시켰다.

주제어: 자동평가, BLEU, 튜닝 시스템, 자동번역기

1. 서론

자동 번역 시스템에서 각 모듈의 알고리즘만큼 중요한 부분이 바로 번역 사전이다. 통계적 자동번역(SMT) 시스템에서는 말뭉치의 크기가 번역 성능에 정비례할 만큼 대량의 말뭉치를 필요로 하다. 규칙 기반의 자동 번역 시스템 역시 다양한 번역 사전(미등록어, 구문지식, 번역 패턴 등)이 있어야 하며, 번역 사전의 구축량, 정교함, 정확성 등이 자동번역 성능에 직접적으로 영향을 준다.

기존의 번역시스템에서는 대량의 말뭉치에서 신규 번역 사전을 자동 추출하여, 다수의 전문적인 작업자가 번역 사전을 구축하였다. 사전 구축 작업자는 기존에 구축된 고빈도 번역 사전 수정 작업하거나, 도메인 특화 작업하여, 번역 사전 오류를 제거하고 번역 사전을 특화함으로써 번역 성능을 향상하였다[1-3]. 이 방법의 장점은 짧은 시간에 대량의 번역 사전을 구축할 수 있다. 그리고 번역 성능도 크게 향상시킬 수 있어, 번역 시스템 구축 초기에는 이 방법이 매우 효율적이다. 반면, 이 방법은 대량의 번역 사전을 다수의 구축 작업자에 의해 구축하기 때문에 작업자의 전문성에 매우 의존적이다. 특히, 번역 사전에 어떻게 적용되는지 모르고 추출된 작업본만 보고 번역 사전을 구축하기 때문에 구축된 번역 사전에는 잠재적인 오류가 흔히 포함되어 있다. 또한, 번역 사전이 구축될수록 기존에 구축된 사전과 신규로 추가된 사전 간의 부작용도 종종 나타난다.

번역 사전을 구축하는 또 다른 방법으로서, 자동 번역 시스템 개발 엔지니어들이 산발적으로 추출한 문장을 번역 시스템을 사용하여 번역함으로서 번역 사전을 튜닝하는 방법이 있다. 이 방법의 장점은 번역시스템의 각 모듈 별 결과를 확인하면서 번역 사전을 튜닝하기 때문에 양질의 번역 사전을 구축할 수 있다는 장점이 있다. 하지만, 이 방법은 사전 튜닝 양이 매우 작다는 단점이 있다.

본 논문에서는 위와 같이 말뭉치에서 번역 사전 작업본을 추출하여 번역 사전을 구축는데 생기는 문제점을 해결하기 위하여, 자동 평가 방법을 이용한 다중 작업자용 자동 번역시스템 사전 구축 및 튜닝 방법을 제안하고자 한다. 본 논문에서 제안하고자 하는 방법은 크게 두 부분으로 나뉜다. 첫째, 번역 사전 튜닝 대상 문장을 추출 하는데 있어 잠재적인 오류가 내포된 문장만을 추출한다. 이를 위해, 각 원문에 대한 번역결과를 대량의 목적 말뭉치로 문장별 BLEU를 측정한다. 측정된 값이 작은 문장을 추출하여, 작업자에게 튜닝 문장으로 제시한다. 둘째, 번역 사전 구축 작업자는 자동 번역시스템을 이용하여 튜닝 문장을 번역하고, 번역 결과를 보고 오류가 있는 사전을 수정하거나 신규 구축할 수 있는 웹 워크벤치를 구축하였다. 구축된 웹 워크벤치는 작업자에게 자동 번역 성능을 BLUE(Bilingual Evaluation Understudy)[4-6]로 측정하여 제시해준다. 이때, 수정된 번역 사전은 자동번역을 통해 확인할 수 있을 뿐만 아니라 BLUE 측정을 통해 보다 정확히 정량적으로 알 수 있게 된다. BLUE 측정 도구 도입의 또 다른 목적은 대상 문장뿐만 아니라 대상 엔트리가 포함된 모든 문장에 대해 측정할 수 있도록 함으로써, 작업자의 사전 튜

닝을 극대화하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안한 방법과 관련이 있는 기존 연구에 대하여 살펴보고, 3장에서는 본 논문에서 제안한 방법에 대하여 기술한다. 4장에서는 본 논문에서 제안한 방법의 유효성을 증명하기 위한 실험 결과에 대해 간단히 분석하며, 마지막으로, 5장에서 본 논문에서 제안한 방법에 대한 연구 결과를 기술한다.

2. 관련 연구

규칙 기반 자동 번역 시스템의 번역 성능을 향상시키는 방법으로 대량의 말뭉치에서 번역 사전 후보를 추출하여 다수의 작업자가 번역 사전을 구축하여 번역 사전에 반영함으로써 성능을 향상시키는 방법이 있었다.[1-3]

일반 도메인의 영한 자동 번역기를 특허 도메인으로 특화하는 방법으로 대량의 특허 문서를 수집하고, 수집된 대량의 말뭉치에서 번역 사전(전문용어 대역어, 번역 패턴)을 추출하여 사전 구축 작업자들이 특허 분야에 맞게 번역 지식을 부착하여 특허 분야 번역율이 평균 81.03%로 향상되었다[1-2].

대역 말뭉치로부터 원문의 관용구 후보를 자동으로 추출한 후 어휘 번역 패턴을 작업자가 부착하였다[3].

통계 기반 자동 번역 시스템에서 번역 결과의 번역율을 측정하는데 널리 사용되는 BLEU 측정 방법[4-6]은 식 1과 같다.

$$BLEU = \exp\left\{\sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L_{ref}}{L_{sys}} - 1.0\right)\right\}$$

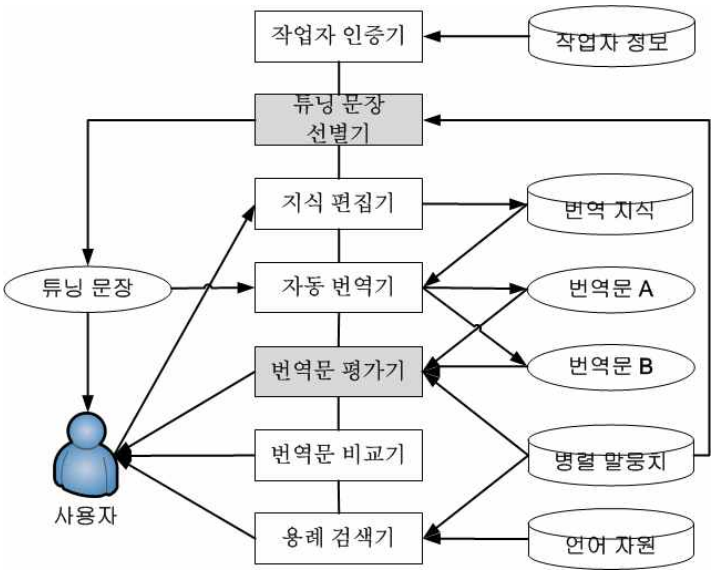
[식 1] BLEU

식 1에서 보면, $_n$은 ngram의 수를 의미하고, w_n은 길이에 대한 가중치로 $N-1$을 사용하고, p_n은 테스트 셋과 정답셋의 ngram 매칭 정도를 의미하고, L_{sys}는 테스트 셋의 길이를 의미하고, L_{ref}는 정답셋의 길이를 의미한다.

일반적으로 $N=4$를 사용하나 튜닝 시스템에서는 문장의 정답셋이 하나이고 문장의 길이가 짧은 것을 가만하여 $N=2\sim4$까지의 합을 사용한다. 결국 한 문장의 번역율은 0~3 사이의 값으로 표현된다.

3. 튜닝 시스템

튜닝 시스템 구성은 그림 1과 같다. 그림 1에서, 작업자 인증기는 작업자 정보를 바탕으로 등록된 작업자인지 판단하고 작업자 권한에 따라 튜닝 시스템 기능을 제한한다. 작업자 등급은 검수자, 튜닝 작업자, 개발자로 구분하여 접근 가능한 사전과 작업 이력 조회 등을 등급별로 제한한다. 튜닝 문장 선별기는 대량의 병렬 말뭉치에서 BLEU가 낮은 문장을 추출하는 모듈로 대량의 병렬 말뭉치를 자동 번역하여 병렬 말뭉치의 번역문을 이용 문장별 BLEU를 측정한다. 이렇게 측정된 문장 중에 BLEU가 낮은 병렬 말뭉치를 작업자에게 튜닝 문장으로 제시한다.



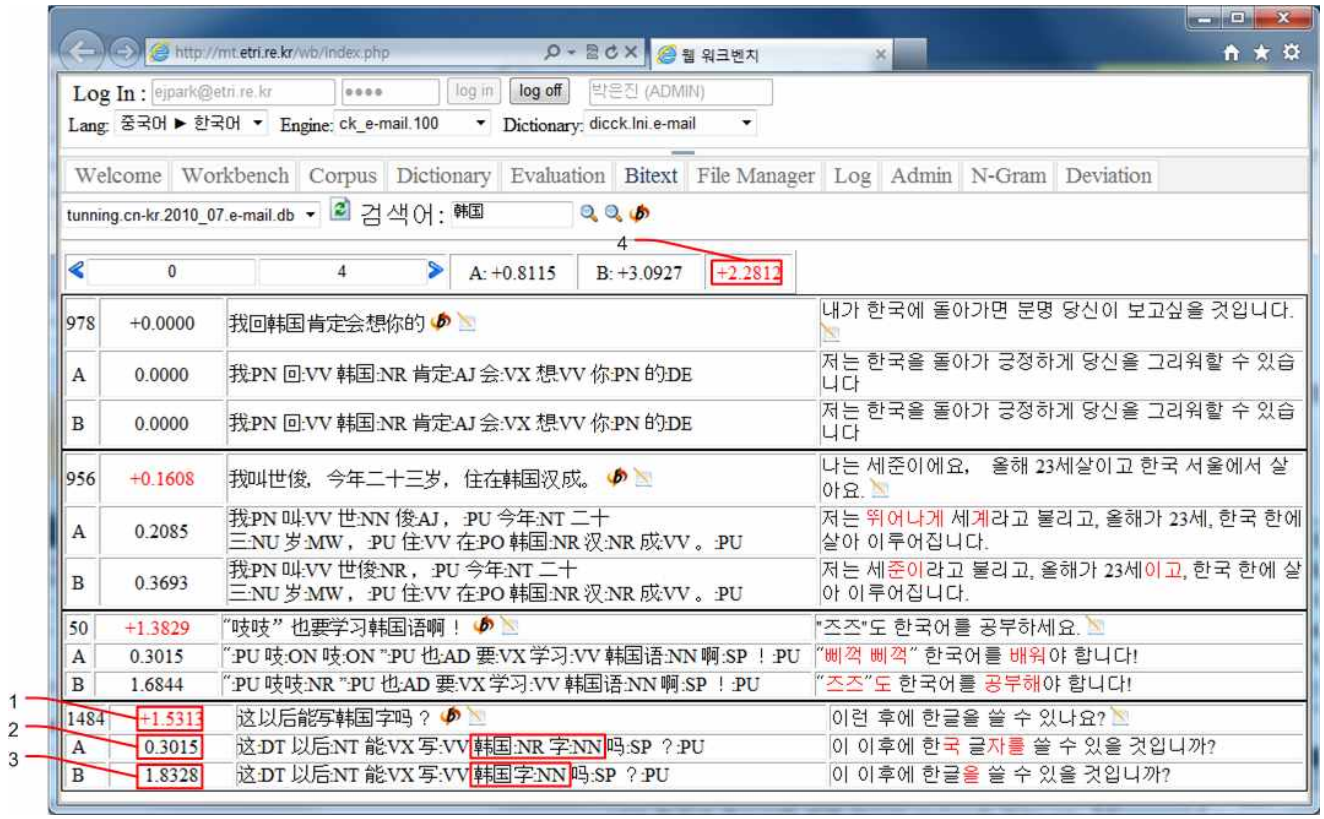
[그림 1] 시스템 구성

튜닝 작업자는 개별 문장이나 특정 어휘가 포함된 문장을 검색하여 사전이 수정되었을 때, BLEU 변화치를 보면서 작업할 수 있도록 인터페이스를 구성하였다. 그림 2는 특정 어휘 "韩国/한국"을 검색한 결과이다.

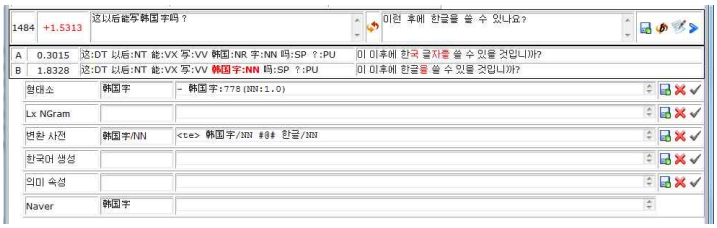
예를 들어 개별 문장 튜닝의 경우, 그림 2에서 보듯이, 중국어 문장 "这以后能写韩国字吗？(이후에 한글을 쓸 수 있나요?)"에 대하여, 자동 번역 결과는 "이 이후에 한국 글자를 쓸 수 있을 것입니까?"가 출력된다. 이렇게 번역되는 이유는 중국어 형태소 분석이 "这:DT 以后:NT 能:VX 写:VV 韩国:NR 字:NN 吗:SP ？:PU"로 중국어 원문의 "韩国字"가 "韩国"과 "字"로 분리되었기 때문이다. 이는 형태소 사전에 "韩国字/한글"이 등록되어 있지 않아서 생기는 오류이다. 튜닝 작업자는 "韩国字"를 형태소 사전에 등록하고, 대역 사전에 "韩国字"에 대한 대역어 "한글"을 등록하고 재번역하면, "이 이후에 한글을 쓸 수 있을 것입니까?"로 번역 되고, 사전 등록 전에 "한국 글자"로 번역 되는 것이 "한글"로 번역되는 것을 확인할 수 있다. 작업자는 사전을 수정하기 전 BLEU($N=2\sim4$의 합)가 0.3015(그림 2에서 2번)에서 사전을 수정하고 난 후 1.8328(그림2에서 3번)로 +1.5313(그림2에서 1번) 증가했다는 것을 정량적으로 알 수 있다.

또한 본 논문의 튜닝 시스템은 개별 문장뿐만 아니라 특정 어휘를 검색하고 검색 결과 내의 BLEU를 측정하여 그림2에서 4번 처럼 보여 준다. 그림 2에서 4번는 검색어 "韩国"에 대하여 검색된 문장들 전체의 BLEU가 +2.2812 증가하였음을 화면 상단에 표시하여 서전의 오적용을 방지하였다.

신속한 사전 편집을 위하여 형태소 분석 결과에서 형태소를 클릭하면 해당 형태소와 관련된 사전들을 화면에 표시하고 수정 가능하도록 그림 3와 같이 구성하였다.



[그림 2] 시스템 인터페이스



[그림 3] 사전 편집 화면

번역 사전의 일관성 유지를 위하여 본 논문의 튜닝 시스템은 두 개의 번역 사전을 유지 관리한다. 하나는 튜닝 작업자가 튜닝 문장을 튜닝하기 위한 사전이고, 다른 하나는 최종 사전이다. 튜닝 작업자들은 튜닝용 사전과 번역 시스템을 사용하여 튜닝 문장을 번역하고 자동 번역 결과의 오류를 사전 수정을 통해 바로 잡는다. 검수자 혹은 최종 관리자는 튜닝 작업자에 의해 튜닝된 사전과 최종 사전의 변경 이력을 확인하여 최종 사전에 반영 여부를 판단한다. 그림 4는 튜닝 사전과 최종 사전 간의 동기화하는 화면이다.

[그림 4] 사전 동기화

또한 동기화 기능에서는 개별 혹은 전체 동기화와 사전 편집자, 사전 변경 일자 등과 같은 기준으로 사전 변경 이력을 검색하여 최종 사전과 비교/검토한 뒤 사전을 동기화하는 기능을 제공한다.

4. 실험

본 논문에서 제안하는 병렬 말뭉치를 이용한 튜닝 방법의 유용성을 증명하기 위해, 기존의 방법인 대량의 말뭉치에서 작업본을 추출하여 번역 사전에 반영하는 방법과 병렬 말뭉치를 이용한 튜닝 방법을 중한 자동 번역 시스템을 사용하여 비교 평가해 보았다.

먼저 기존 방법으로, 대화체 분야 15만 말뭉치에서 추출한 미등록어 후보 2만 엔트리에 3명의 사전 구축 작업자가 3주간 번역 사전(미등록어, 품사, 대역어)을 부착하여 7,200개의 구축하였다. 이렇게 구축한 작업 결과를 대화체 자동 평가셋 7,000문장으로 자동 평가해본 결과 BLEU가 사전 반영 전 1609에서 작업 결과 반영 후 1612로 +3이 증가하였다.

그리고 튜닝 시스템을 이용하여, 2명의 사전 구축 작업자가 2주간 병렬 말뭉치 2,193문장을 튜닝 하였다. 그 결과 형태소 사전은 33개(단일 형태소 3,405개 중)가 수정되었고, 대역어 사전은 234개의 엔트리가 편집되었다. 이렇게 구축된 작업 결과를 대화체 자동 평가셋 7,000문장으로 자동 평가해본 결과 BLEU가 사전 반영 전 1609에서 사전 반영 후 1621로 +12 증가하였다.

5. 결론

본 논문의 자동 번역기 튜닝 시스템은 병렬 말뭉치에서 오류가 있을 것으로 추정되는 문장을 자동 추출하여, 다수의 사전 구축 작업자가 자동 번역시스템을 직접 사용하면서 번역 사전을 튜닝하는 방법에 대하여 제안하였다. 작업자는 병렬 말뭉치의 대역문을 이용하여 자동 번역 결과의 BLEU를 측정하고, 사전 수정 전과 후의 BLEU 차이를 정량적으로 제시해 줌으로써 양질의 번역 사전을 구축하도록 하였다.

대량의 번역 사전이 이미 구축된 자동 번역시스템에서 추가적인 성능향상을 위해 대량의 말뭉치에서 미등록어, 번역패턴 등을 추출하여, 대량으로 구축하는 기존 방법에 비해 사전 구축 부작용이 적으며, 자동번역 성능향상에 더 기여하는 것을 실험으로 증명하였다.

참고문헌

[1] 최승권, 권오욱, 이기영, 노윤형, 박상규. “도메인 특화 방법에 의한 영한 특허 자동 번역 시스템의 구축”, 정보과학회논문지 34권 2호, pp.95-187., 2007.

[2] 최승권, 권오욱, 이기영, 노윤형, 박상규, “웹 영한 번역기로부터 특허 영한 번역기로의 특화 방법”, 한글 및 한국어 정보처리 학술대회, pp.57-64., 2006.

[3] 최승권, 김영길, “번역 말뭉치로부터 추출한 어휘 번역 패턴의 의미 분류와 자동번역시스템에의 활용”, 번역학연구 제 11권 3호, pp.277-301., 2010.

[4] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation". pp. 311-318. 2001.

[5] Lin, C.-Y. and E. Hovy. “Manual and Automatic Evaluations of Summaries.” In Proceedings of the Workshop on Automatic Summarization, postconference workshop of ACL-2002, pp. 45-51, Philadelphia, PA, 2002.

[6] NIST. “Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics.”, 2002.