일영 통계기계번역에서 먼 거리 의존관계를 이용한 일본어 어순 조정

나휘동O, 이금희, 이종혁
POSTECH 컴퓨터공학과
leona@postech.ac.kr, ljj@postech.ac.kr, jhlee@postech.ac.kr

Reordering Heuristic using Long Distance Dependency for Japanese-to-English Statistical Machine Translation

Hwi-Dong NaO, Jin-Ji Lee, Jong-Hyeok Lee
POSTECH, Department of Computer Science and Engineering

1. 서론[1]

현재 통계기계번역에서 가장 좋은 성능을 보이는 방법은 구 기반 통계기계번역이고[1], 가장 많이 사용하는 구 기반 통계기계번역 도구는 Moses[2]이다. Moses는 위치정보를 기반으로 어순을 조정하는 방법(distortion model)을 사용하고 있다. Distortion 모델은 어순이 비슷한 언어에서는 비교적 좋은 성능을 보여준다. 하지만 SOV 어순을 가지는 일본어와 SVO 어순을 가지는 영어처럼 어순이 상이한 경우, 기존 구 기반 통계기계번역의 distortion 모델이 적합하지 않으며 언어학적 지식을 사용한 구 재배치 방법이 필요하다. 이러한 문제를 해결하고자 이 논문에서는 언어학적 지식을 사용하여 일본어를 영어 어순에 맞도록 조정하는 방법을 제안한다.

2. 본론

일본어 문장의 의존 관계를 분석했을 때, 서술어가 지배하는 주어와 서술어가 목적어 같은 구성요소를 사이에 두고 띄어지는 경우를 이 논문에서는 먼 거리 의존 관계라고 정의한다. 먼 거리 의존 관계를 가지는 문장을 번역할 때 기존의 distortion 모델은 올바른 서술어 위치를 결정하기 어렵다. 띄어진 사이에 구성요소가 얼마나 많을지 알 수 없기 때문이다. 이러한 문제점을 해결하고자 이 논문에서는 의존 관계 분석기를 사용해서 먼 거리 의존 관계를 파악하여 일본어를 영어 어순에 맞추는 방법을 제안한다.

파악한 의존 관계는 분절 단위로 의존 트리 구조로 나타낼 수 있다. 노드는 주어, 목적어, 서술어와 같은 분절을 나타내고, 분절이 지배하는 자식 노드를 여러 개 가질 수 있다. 자신보다 어순이 앞서는 분절은 왼쪽, 뒤지는 분절은 오른쪽으로 구분한다. 이 의존 트리 구조를 중위 순회(inorder traversal)하면 일본어 문장을 얻을 수 있다. 이 논문에서 제안한 방법은 일본어 의존 트리 구조를 재구성하여 중위순회로 얻은 문장이 영어 어순을 따르도록 조정한다. 위치가 아닌 의존 관계를 파악해서 어순을 조정하기 때문에 먼 거리 의존 관계를 갖는 문장도 올바르게 어순을 조정할 수 있다.

의존 트리 구조는 정해진 규칙에 따라 재구성된다. 재구성 단위는 노드와 노드에 직접 연결된 자식 노드로 정의한다. 예를 들어 서술어가 주어와 목적어를 지배하고 있다면 서술어는 왼쪽에 주어와 목적어를 자식 노드로 가진다. 이때 목적어의 위치를 서술어 다음으로 정의하는 규칙이 있다면, 목적격을 서술어 왼쪽에서 삭제하고 오른쪽으로 옮겨서 규칙을 반영한다. 이렇게 재구성한 결과를 중위순회하면 주어, 서술어, 목적어 순으로 영어 어순을 따르는 문장을 얻는다.

이 논문에서는 NTCIR-7 특허 번역 작업[2] 말뭉치를 사용하였다. 이는 일본어와 영어로 쓴 병렬 말뭉치로 특허

[1] 본 연구는 첨단정보기술 연구센터를 통한 과학재단 및 2008년도 두뇌한국21사업의 지원을 받았고 지식경제부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

내용을 담고 있다. 특허 문서이기 때문에 문장당 평균 스무 단어 넘게 나타날 정도로 문장이 길어 먼 거리 의존관계를 가지는 경향이 있어 제안한 방법을 사용하기에 적합하다. 이 논문에서 사용한 어순 조정 규칙은 학습 말뭉치에 등장하는 재구성 단위를 추출해서, 그 중에 어순 조정이 필요한 경우만을 사람이 판단해서 정의하였다.

실험에 쓰인 기본 모델은 두 가지이다. 첫째(distortion)는 Moses는 기본적으로 학습에 사용한 말뭉치에서 구를 추출할 때 그 위치정보를 파악해서 어순을 조정하는데 사용한다. 이 방법은 문장 구성 성분끼리 언어학적 관계를 따지지 않고 통계적인 위치만을 고려한다. 둘째(monotone)는 Moses 가 어순을 조정하지 않도록 설정한 다음 실험 말뭉치를 번역한다. 제안한 방법으로 학습용 말뭉치와 실험용 말뭉치 모두 어순을 조정한 경우(reordering both)와 실험용 말뭉치만 어순을 조정한 경우(reordering test)를 실험하였다. 두 경우 모두 일본어를 영어 순서에 맞게 바꾼 다음 Moses의 monotone 모델을 이용한다.

제안한 방법 reordering both로 번역한 결과의 BLEU 점수는 어순을 조정하지 않았을 때(monotone)보다 1.6%p 높고, 위치 기반으로 어순을 조정했을 때(distortion)보다 0.8%p 낮다. 하지만 제안한 방법에서 정답과 완벽하게 일치하는 번역문이 존재한다. 또한 점수가 0인 문장도 27%로 distortion 모델과 3%p 차이를 보인다. 제안한 방법 reordering test는 어순을 조정하지 않았을 때보다도 낮은 성능을 보여 결과 분석에서는 제외한다.

이 논문에서 제안한 방법은 BLEU점수로는 distortion 모델의 성능에 미치진 못하지만 사람이 보기에 정확한 번역 결과를 만들어 낼 수 있다. 이는 BLEU 점수가 구 단위로 배치만 되어 있다면 어순이 어떻게 되든 상관 없고 문장의 의미를 알 수 없는데도 높은 점수를 얻기 때문이다. 따라서 특히 먼 거리 의존관계를 가지는 경우에 제안한 방법이 효과적임을 실험결과 확인할 수 있다. 반면 제안한 방법이 distortion 모델보다 안 좋은 번역결과를 내는 경우는 다중문일 때이다.

3. 결론

기계 번역을 할 때에 어순을 결정하는 방법은 중요한 문제이고, 이를 통계적 기법으로만 해결할 수는 없다. 기존에 언어학 지식을 사용하여 어순을 조정하는 연구[4]와 유사하게, 이 논문에서는 정해진 규칙에 따라 어순을 조정하는 방법을 제안하였다. 실험 결과 제안한 방법이 통계적 방법보다 BLEU 점수가 낮았지만, 사람이 보기에는 더 정확한 번역을 얻을 수 있다. 앞으로는 다중문 같은 복잡한 문장 구조를 처리할 수 있도록 발전시키고, 규칙을 추가한다면 좀더 정교하게 어순을 조정할 수 있다.

참고문헌

[1] Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology - Volume 1, 2003

[2] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), 2007.

[3] Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, 2001

[4] Collins, M., Koehn, P., and Kučerová, I. Clause restructuring for statistical machine translation. In Proceedings of the 43rd Annual Meeting on Association For Computational Linguistics, 2005.

[2] http://if-lab.slis.tsukuba.ac.jp/fujii/ntc7patmt/index-en.html