문장 길이가 한영 통계기반 기계번역에 미치는 영향 분석

조희영○ 서형원 김재훈
한국해양대학교 컴퓨터공학과
{serensis, hws}@bada.hhu.ac.kr, jhoon@hhu.ac.kr

Empirical Impact Analysis of Sentence Length on Statistical Machine Translation

Hee-Young Cho○ Hyoung-Won Sou Jea-Hoon Kim
Department of Computer Engineering, Korea Maritime University

요 약

본 논문에서는 한영 통계기반 기계번역에서 한국어 문장 길이의 변화에 따른 번역 성능의 변화를 분석하고자 한다. 일반적으로 통계기반 기계번역은 정렬기법을 이용하는데 문장의 길이가 길수록 많은 변형(distortion)이 이루어진다. 특히 한국어와 영어처럼 어순이 매우 다를 경우, 문장 길이의 변화에 따라 그 변형이 더욱 심할 수 있다. 본 논문에서는 이러한 성질이 통계기반 기계번역에 어떠한 영향을 주는지를 실험적으로 살펴보고자 한다. 본 논문에서 비교적 잘 정렬된 203,310개의 문장을 학습데이터로 사용하였고, 세종 병렬 말뭉치로부터 89,309개의 문장을 추출하여 실험데이터로 사용하였다. 실험데이터는 한국어 문장의 길이에 따라 5구간(1~4, 5~8, 9~13, 14~19, 20~n 개)로 나뉘었다. 각 구간은 가능한 문장의 수가 비슷하도록 하였으며, 17,126, 18,507, 20,336, 17,884, 15,456개의 문장이 포함되었다. 데이터들은 모두 어절단위로 토큰을 나누었다. 본 논문에서는 한영 번역을 중심으로 평가되었다. 첫 번째 구간에서 가장 좋은 성능인 0.0621 BLEU를 보였으며, 마지막 구간에서 가장 좋지 않은 0.0251 BLEU를 보였다. 이는 문장의 길이가 길수록 변역 성능이 좋지 않음을 알 수 있었다. 문장이 길수록 구가 길어지고 구간의 수식이 복잡해지므로 번역의 성능은 점차 떨어진다. 이것을 볼 때, 구번역을 먼저 한 후, 다시 문장 번역을 한다면 좀 더 높은 기계번역의 성능을 기대할 수 있을 것이다.

1. 서 론

1960년대 이후 많은 연구자들이 자동번역시스템을 개발하기 위한 연구를 진행하였으나 아직 사용자들은 만족할 만한 자동번역시스템은 거의 없다고 해도 과언이 아니다[1]. 그러나 자동번역에 대한 응용 분야는 매우 다양하다. 예를 들면, 단순한 문서번역, 번역 업체의 초벌 번역, 웹 문서 번역, 기술 문서 번역, 전자우편 번역, 방송자막 번역, (휴대폰/PDA) 자동 통역, 다국어 정보검색 등이 있으며, 또한 많은 인터넷 사용자들은 검색엔진을 통해서 다양한 언어로 표현된 유용한 정보를 찾고 있다[1]. 이와 같은 다양한 응용 분야로 말미암아 여전히 기계번역에 대한 연구는 활발히 진행되고 있으며 최근에는 통계기반 기계번역에 대한 연구가 매우 활발히 진행되고 있다[1-3]. 통계기반 기계번역은 정렬기법을 이용하며, 여기에는 단어번역(translation), 위치변형(distortion), 단어대응(fertility) 등의 정보들이 이용된다[4]. 일반적으로 문장의 길이가 길수록 많은 변형(distortion)이 일어난다. 특히 한국어와 영어처럼 어순이 매우 다를 경우, 문장의 길이가 길수록 그 변형이 더욱 심해진다. 본 논문에서는 이러한 성질이 통계기반 기계번역에 어떠한 영향을 주는지를 실험을 통하여 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 통계기반 기계번역의 요소 기술들에 대해서 간략히 소개하고, 3장에서는 본 논문에서 사용된 기계번역 시스템을 간단히 기술한다. 4장에서는 문장 길이에 따른 통계기반 기계번역 시스템의 성능을 평가하고 그 결과를 분석한다. 마지막으로 5장에서 결론을 맺고 향후 연구 과제를 기술한다.

2. 관련 연구

2.1. 통계기반 기계번역

통계기반 기계번역은 기계번역이 처음 시작된 50년대부터 시도되었으나, 컴퓨터 기술과 성능의 한계로 크게 성공하지 못했다. 그러나 최근에 와서 컴퓨터의 기술이 발전되고 또 대량의 병렬 말뭉치들이 쉽게 구축할 수 있게 되면서 다시 활발한 연구가 시작되었다. 통계기반 기계번역의 기본적인 개념은 Shannon의 정보이론에 기반을 두고 있으며, 정렬기법을 이용하여 기계번역을 모델링하였다. (식 1)은 정렬기법을 이용한 통계기반 기계번역 모델이다.

$$\hat{e} = \text{argmax}_e P(e|f) = \text{argmax}_e P(e) \times P(f|e) \quad \text{(식 1)}$$

(식 1)에서 $P(e)$를 언어모델(language model)이라 하고, $P(f|e)$를 번역모델(translation model)이라고 한다. argmax_e는 통계기반 기계번역 시스템으로 일반적으로 번역기(decoder)라고 한다. 최근 통계기반 기계번역을 위한 많은 모델[1-2]들이 제안되었으나[1-2], 국내에서는 그다지 활발하게 연구되고 있지는 않다[3, 5-6].

2.2. 단어정렬

단어정렬은 병렬 말뭉치로부터 단어 단위의 대역을 찾는 것을 말한다. IBM Model 1-5[2, 4]를 시작으로 많은 연구들이 진행 되었으며 대체로 자율학습을 통한 단어 정렬 모델을 학습한다. 단어의 정렬은 1:1이 아닐 뿐만 아니라 단어 간 위치가 뒤집혀 정렬이 될 수 있어 모델 자체가 복잡하다. 그래서 특별한 언어들 사이에는 사전 지식을 이용하여 통계기반 기계번역 모델을 개선하였다. 단어 정렬은 이 자체로 많은 응용 분야를 가지고 있는데, 대역사전의 자동 생성[7], 의미모호성해소[8] 등이 있다. 단어 정렬의 도구로는 GIZA++[9], k-vec[10], PWA[11] 등이 있으며, 본 논문에서는 GIZA++를 사용하였다.

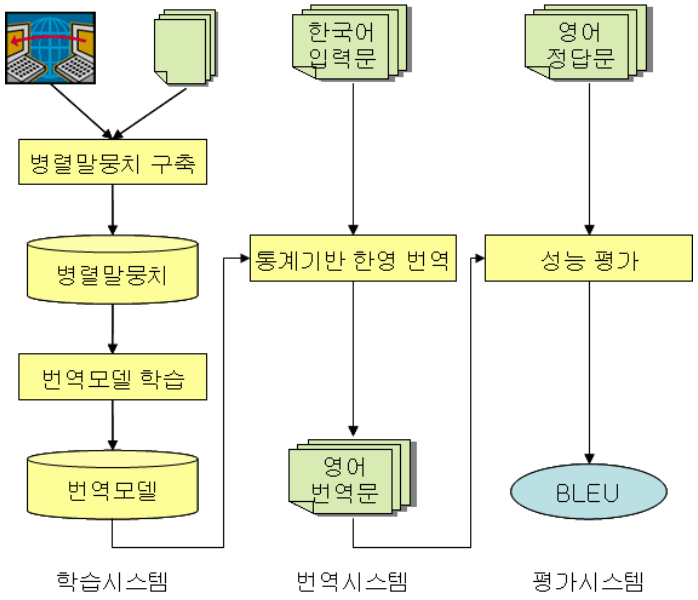
2.3. 번역기

번역기는 학습된 언어모델과 단어 정렬 말뭉치로부터 추정된 단어 정렬 모델을 이용해서 주어진 원시 문장을 목적 문장으로 번역한다. 통계기반 번역기는 탐색문제로서 주로 A* 알고리즘이나 탐욕알고리즘을 이용해 구현한다[12]. 또 다른 방법인 A* 알고리즘의 일종인 빔 탐색법은 계산된 확률 값이 허용된 범위 내에 있을 경우에만 탐색을 한다. 그리고 또 다른 방법으로는 FST를 이용하는 방법과 동적프로그래밍을 이용하는 방법들이 있으며, 본 논문에서는 PHARAOH[13]을 이용하였다.

2.4. 평가도구

BLUE(BiLingual Evaluation Understudy)는 IBM 연구소에서 제안한 번역 성능 평가 측도이며, n-gram의 공기빈도를 이용하여 번역의 질을 측정하는 방법이다[14]. 즉 기계로 번역된 문장이 사람이 번역한 문장에 가까울수록 높은 값을 가지는 측도로서 현재 통계기반 번역에서 널리 사용되고 있다.

3. 한영 통계기반 기계번역 시스템의 구성



(그림 1) 전체 시스템 구성도

본 논문에서 사용되는 통계기반 기계번역 시스템의 구성은 (그림 1)과 같으며 크게 학습시스템, 번역시스템, 평가시스템으로 구성된다. 학습시스템은 병렬말뭉치 구축과 번역모델 학습으로 이루어진다. 병렬말뭉치 구축은 [15]에서 제안된 방법으로 이용해서 다양한 병렬문서로터 한영 병렬말뭉치를 구축한다. 번역모델 학습은 널리 잘 알려진 도구인 GIZA++[9]을 이용해서 학습한다. (그림 1)에 자세히 표현하지 않았지만 번역모델은 GIZA++에 의해서 학습된 단어번역모델 이외에 언어모델도 포함되어 있는데 이는 [16]를 이용해서 학습되었다. 번역시스템은 한국어

문장을 입력하여 영어 문장을 생성하는 시스템으로 본 논문에서는 PARAOH[16]을 이용하였다. 평가시스템은 NIST에서 공개된 기계번역 평가시스템을 사용하였다[1].

4. 실험 및 평가

학습데이터는 [15]에 의해서 구축된 병렬말뭉치를 이용하였으며, 이 중에서 비교적 잘 정렬된 병렬문장 203,310개를 이용하였다. 이것은 기사 번역 내용을 추출한 것이 주가 된다. 실험데이터는 비교적 정확하게 구축된 세종 병렬말뭉치[2]를 이용하였으며, (표 1)과 같이 구성되었다. 길이에 따른 번역의 성능을 관찰하기 위해서 실험말뭉치는 어절의 수에 따라 5개의 구간으로 나누었고, (표 1)에서 보아 알 수 있듯이 각 구간의 문장 수는 가능한 한 비슷하도록 조절하였다. 학습데이터의 경우 203,310개의 문장이 각 구간별로 7,110, 22,724, 38,025, 64,023, 70,430개로 구간이 길어질수록 문장의 양이 많아지는 분포를 가진다. 구간은 어절을 단위로 나누고 실험과 학습을 할 때에는 번역의 성능을 높이기 위해 토근 단위로 나누었다. 이때 실험데이터와 학습데이터의 한글 데이터는 형태소 단위로 토큰을 나누고, 영어 데이터는 축약된 동사들을 분리하고 어절 단위로 토큰을 나누었다. 학습데이터의 토큰 수는 영어 120,833개, 한글 107,836개 이고 실험데이터의 경우 각 구간별로 영어는 9,443, 15,634, 21,167, 22,487, 23,644개, 한글은 9,198, 15,758, 21,141, 22,891, 25,187개이다.

(표 1) 문장의 길이 변화에 따른 누적 BLEU 변화

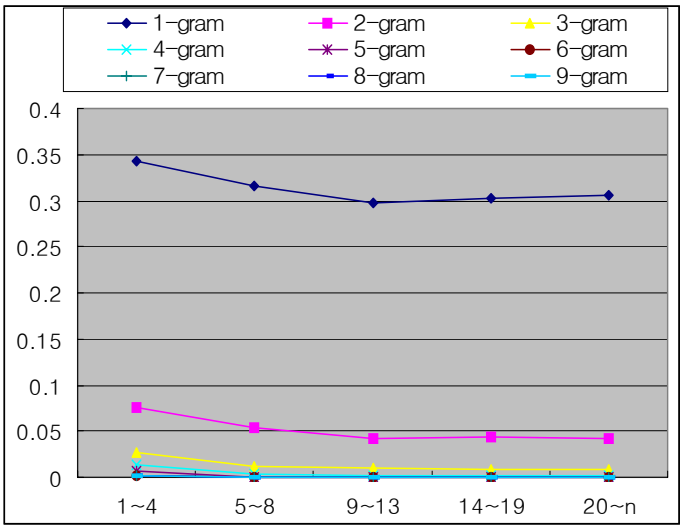
어절 수	구간 (단위: 어절)				
	1~4	5~8	9~13	14~19	20~n
실험 문장 수	17,126	18,507	20,336	17,884	15,456
BLEU	0.0621	0.0351	0.0275	0.0262	0.0251

이와 같은 방법으로 준비된 실험데이터에서 한국어 문장을 추출하여 번역시스템의 입력으로 사용하였으며 번역시스템의 결과는 평가시스템의 입력으로 사용되었다. 또한 실험데이터에서 영어문장을 추출하여 평가시스템의 입력인 영어 정답문장으로 사용하였다. 실험에서 사용된 번역 성능의 측도로는 누적 BLEU[3]를 사용하였다(표 1). (표 2)는 각 구간별 미등록어의 수를 나타낸다. 이것 역시 문장이 길어질수록 미등록어의 수도 늘어나는 것을 볼 수 있다.

(표 2) 구간별 미등록어의 수

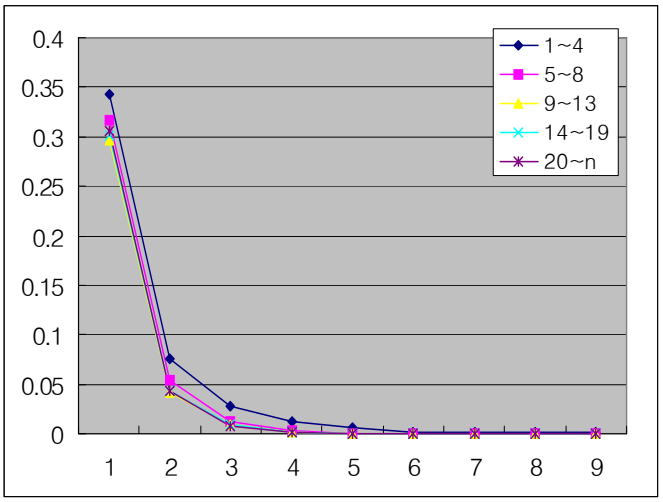
어절 수	구간 (단위: 어절)				
	1~4	5~8	9~13	14~19	20~n
실험 문장 수	17,126	18,507	20,336	17,884	15,456
미등록어 수	4,189	11,032	23,701	30,976	39,663

전체 학습데이터의 크기 그다지 많지 않고 실험데이터는 다양한 분야의 데이터이고 학습데이터는 기사 내용을 추출한 것이기 때문에 전체적인 성능이 높지는 않지만, 문장의 길이에 따른 성능의 변화는 관찰할 수 있었다. 즉, 문장의 길이가 길어질수록 전체 성능이 떨어졌고, 첫 번째 구간에서 두 번째 구간 사이의 값의 차이가 가장 크다. 여기서 가장 짧은 문장이 좀 더 번역의 성능이 높은 것을 알 수 있다.



(그림 2) 구간별 n-gram 정확률

1) http://www.nist.gov/speech/tests/mt/resources/scoring.htm
2) http://sejong.or.kr/sejong_kr/index.html
3) 누적 BLEU(cumulative BLEU)는 각 n-그램의 BLEU의 기하평균이다[14].



(그림 3) gram에 따른 구간별 정확률

(그림 2)와 (그림 3)은 누적 BLUE가 아닌 각 그램에 따른 개별 BLEU를 그래프로 표현한 것이다. (그림 2)에서 어절의 수가 길어질수록 전체적인 번역률이 떨어지는 것을 쉽게 관찰할 수 있었다. 전체적인 번역률이 떨어지는 것은 학습 코퍼스의 양이 적은 것과 원래 기계번역의 성능 자체가 높지 않은 데서 기인한다[4]. 또한 본 논문에서는 학습데이터는 대부분 신문기사에서 추출되었으나 실험 데이터는 성경 등 매우 다양한 분야에서 추출되어 두 데이터 간의 너무 상이한 것도 하나의 원인이다. (그림 3)에서는 그램이 증가함에 따라서 번역의 성능이 크게 감소됨을 알 수 있다. 이는 그램이 낮을수록 정확하게 일치되는 번역이 많이 존재함을 말하고 있다. 즉 1-그램의 경우 단어 자체의 번역으로 단어 자체를 정확하게 번역할 가능성이 약 37%정도임을 의미한다.

5. 결론 및 향후 과제

본 논문에서는 문장 길이 구간별 기계 번역의 성능을 평가하였다. 단어 정렬에는 GIZA++을 사용하였으며, 번역기로는 PHARAOH를 사용하였다. 약 8만개의 문장 대역쌍을 이용하여 학습하였으며, 약 9만개의 문장 대역쌍을 학습데이터를 문장 길이별로 5개의 구간으로 나누어 번역 성능을 평가하였다. 1~4개의 어절로 이루어진 첫 번째 구간이 0.0621 BLEU로 가장 큰 값을 가지고 문장의 길이가 길어질수록 그 값이 점차 낮아져서, 20개 이상의 어절로 이루어진 문장의 경우 그 번역 성능이 0.0251 BLEU로 가장 낮았다. 학습 데이터의 경우 신문기사의 영역에서 추출되었고, 실험 데이터는 전 영역에서 추출된 것이라, 서로 다른 영역이란 점에서 전체적인 성능이 낮은 편이다.

앞으로 더 많은 학습데이터로 구간별 성능 평가를 시도하고, 좀 더 세분화된 문장 길이별로 기계 번역의 성능을 평가하고, 언어 구조가 비슷한 언어 간에도 이와 같은 실험을 병행되어야 할 것이다. 본 논문 실험 결과를 적용하면 구번역이 선행된 문장 번역의 효용성이 평가 받을 수 있다고 생각된다.

4) http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html

참고 문헌

[1] Manning, C.D. & Schütze, H., *Foundations of statistical natural language processing*, Cambridge, MA: MIT Press, 1999.

[2] Brown P. F., Cocke, J., Della Pietra, S., Della Pietra, D., Jelinek, F., Mercer, R. and Roossin, P. "Statistical Approach to Machine Translation", *Computational Linguistics,* vol. 16, no 2, pp. 79-85. 1990.

[3] 신중호, 한국어/영어 병렬 말뭉치에 대한 단어단위 및 구단위 정렬 모델 , 한국과학기술원 전산학과 석사학위 논문, (1996).

[4] Brown, P. F., Della Pietra, S. A, Della Pietra, V. J., and Mercer, R. L., "The Mathematics of statistical Machine Translation: Parameter Estimation", *Computational Linguistics*, vol. 19, no.2, pp. 263-311, 1993.

[5] Huang J.-X., Choi, K.-S., "Chinese-Korea Word Alignment Based on Liguistic Comparison", *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 392-399, 2000.

[6] 리금희 , 김동일, 이종혁, "중-한 대조분석정보를 이용한 단어정렬", 제14회 한글 및 한국어 정보처리 학술발표 논문집, pp. 40-46, 2002.

[7] Smadja, F., McKeown, K. R., and Hatzivassiloglou, V., "Translating collocations for bilingual lexicons: A statistical approach", *Computational Linguistics*, vol. 22, no. 1, pp. 1-38, 1996.

[8] Diab, M. "An unsupervised method for

multilingual word sense tagging using parallel corpora: A preliminary investigation", *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, pp. 1-9, 2000.

[9] Och, F. and Ney, H. ,"Improved Statistical Alignment Models". *Proceedings of the ACL*, pp. 400-477, 2000.

[10] Fung, P. and Church, K. W., "K-vec: A New Approach for Aligning Parallel Texts", *Proceedings of the COLING,* pp. 1096-1102, 1994.

[11] Ahrenberg, L., Merkel M., Sågvall Hein, A., and Tiedemann, J., "Evaluation of Word Alignment Systems", *Proceedings of LREC,* pp. 1255-1261, 2000.

[12] Germann, U., "Greedy Decoding for Statistical Machine Translation in Almost Linear Time", *Proceedings of HLT-NAACL*, pp. 72-79, 2003.

[13] Koehn, P., "Pharaoh: a beam search decoder for phrase based statistical machine translation models", *Proceedings of the 6th Conf. of the Association for Machine Translation in the Americas*, pp. 115-124, 2004.

[14] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., "BLEU: a method for automatic evaluation of machine translation", Proceedings of ACL, pp. 311-318, 2002.

[15] 서형원, 김형철, 조희영, 김재훈, 양성일, "웹 문서로부터 한영 병렬말뭉치의 자동 구축", 제26회 한국정보처리학회 추계학술발표대회 논문집, 제13권, 제2호, pp. 161-164, 2006.

[16] Stolcke, A. "SRILM-An extensible language modeling toolkit", *Proceedings of Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, 2002.