병렬코퍼스자동구축체계에서 영－조병렬코퍼스려과의 한가지 방법

김준규, 신혁철

론문에서는 통계적인 방법들을 리용하여 신경기계번역체계구축에서 핵심으로 되는 병렬코퍼스려과에 대한 한가지 방법을 제안하였다.

1. 선행연구 및 문제설정

심층신경망과 신경기계번역기술의 발전에서 최근에 이룩된 성과들은 기계번역의 성능을 번역원의 수준에 거의 가깝게 비약시킬수 있게 하였으며 대용량의 병렬코퍼스에 대한 수요를 증대시키고있다.

질이 높은 병렬코퍼스를 수동적으로 구축하는것은 많은 인적노력과 자금을 요구한다. 이로부터 병렬코퍼스구축의 력사가 짧고 언어자원이 부족한 언어쌍인 경우 효과적이면서 효률적인 병렬코퍼스의 구축과 응용을 위한 연구가 활발히 진행되였다.

우선 병렬코퍼스의 구축에 필요한 토대자료자원들을 수집하고 그것을 자동처리함으로써 수동적인 노력이 없이 질이 높은 병렬코퍼스를 구축하는 방법에 대한 연구가 진행되였다.

다음으로 이미 개발한 기계번역기를 리용하여 1개 언어본문자료를 자동번역함으로써 합성병렬코퍼스를 구축하고 이것을 실제적인 병렬코퍼스와 통합하여 신경기계번역모형을 훈련시키는 방법에 대한 연구도 진행되였다.

이와 같이 병렬코퍼스의 자동구축체계가 확립되면서 그것에 대한 질을 높이는 문제가 제기되고있다.

WMT2018(세계기계번역경연대회2018)에서는 병렬코퍼스려과과제를 설정하고 이에 따르는 경연을 진행하였다.[1, 2] 경연의 목적은 인터네트상에서 수집한 방대한 규모의 병렬코퍼스자료들로부터 질이 높은 일정한 규모의 병렬코퍼스를 려과해낼수 있는 성능이 높은 방법들을 개발하고 평가하는것이였다.

경연방법은 조직자측에서 제공한 10억단어규모의 질이 낮은 Prarcrawl이라는 이름을 가진 도－영병렬코퍼스를 대상으로 려과를 진행하는것이다.

Paracrawl 도－영병렬코퍼스의 구성을 표 1에 보여주었다.

표 1. Paracrawl 도－영병렬코퍼스의 구성

문장정보	비률/%
기계번역의 학습에 리용할수 있는 문장쌍	23
정렬이 잘못된 문장쌍	41
영－영이거나 도－도인 문장쌍	20
도－영외에 다른 언어가 들어있는 문장쌍	3
번역되지 않은 문장쌍	4
너무 짧은 쪼각문장쌍(3－5단어)	5

이 경연에 17개 단체의 44건의 려과체계가 제출되였다.

표 1에서 보여준것처럼 자동구축된 병렬코퍼스들로부터 기계번역체계에 리용할수 있는 질이 높은 병렬코퍼스를 얻는 문제가 제기된다.

론문에서는 코퍼스자동구축체계로부터 얻어진 병렬코퍼스안에 들어있는 오유문장들을 제거하고 코퍼스의 질을 높이기 위하여 여러가지 규칙들을 적용하고 려과성능을 평가하였다.

2. 규칙을 리용한 영-조병렬코퍼스의 려과

병렬코퍼스자동구축체계의 구성을 그림에 보여주었다.

그림. 병렬코퍼스자동구축체계의 구성

론문에서 제안한 려과체계는 문장정렬부분에 속하게 된다. 그림에서 보여준것처럼 자동구축된 영-조병렬코퍼스는 선행연구들에 비하여 일련의 특징을 가진다. 우선 웨브문서들로부터 추출된 병렬자료보다도 PDF문서들이 절대적으로 많으며 PDF문서로부터 본문을 추출하는 과정에 서지구성과 수식으로부터 초래되는 오유들이 기본잡음으로 된다는것이다. 또한 기계번역토대의 자동정렬방법을 적용하였으므로 조선어문장과 영어문장의 쌍에 영어문장을 기계번역하여 얻은 문장들이 첨부되여있다는것이다.

제안한 려과체계는 이와 같은 자동구축된 병렬코퍼스를 입력받아 오유문장들을 제거하고 질이 높은 문장쌍들로 이루어진 병렬코퍼스를 만들어낸다.

구축된 영-조병렬코퍼스를 려과함에 있어 WMT2018의 병렬코퍼스려과공유과제에서 우수하게 평가된 방법[3, 4]들가운데서 질이 높은 병렬코퍼스로 훈련시킨 통계적모형을 리용할수 없다. 이로부터 구축된 병렬코퍼스의 조선어와 영어문장쌍들에서 경험적으로 추출한 통계적규칙들에 토대하는 려과방법만을 적용한다.

영-조병렬코퍼스려과에 리용된 규칙을 표 2에 보여주었다.

표 2. 영-조병렬코퍼스려과에 리용된 규칙

번호	규칙이름	설명
1	단어개수	영어 혹은 조선어문장이 3단어이하
2	평균단어길이	영어문장안의 단어들의 평균길이가 2-20에 포함
3	문장최대길이	영어문장의 단어개수가 60개이상
4	단어최대길이	영어단어중에서 최대길이가 50이상인것.
5	단어개수비률	조선어문장과 번역문장의 단어개수를 비교
6	특수단어비률	특수기호(문장기호포함)를 포함한 단어가 40%이상
7	특수기호포함	0x20이전의 문자포함
8	문장끝기호	영어, 조선어문장이 문장끝기호로 끝나는가.
9	언어식별	영어문장에 중어, 일본어, 조선어가 들어있는가.
10	괄호쌍관계	영어, 조선어문장안에서 열린 괄호와 닫긴 괄호들이 서로 대응되는가.
11	조선어토문장	조선어문장에서 토단어가 존재하는가.
12	단어의미판정	특정한 단어의미(will-의지)를 가지는 단어판정

표 2에서 보여준 규칙들을 리용하여 여러가지 형태의 전자문서들로부터 자동추출된 영－조병렬코퍼스들에 대한 려과를 진행하였다.

3. 실험 및 평가

론문에서 제안한 방법으로 자동추출된 영－조병렬코퍼스에 대한 려과를 진행한다.
통계적규칙을 리용한 영－조병렬코퍼스려과결과를 표 3에 보여주었다.

표 3. 통계적규칙을 리용한 영-조병렬코퍼스려과결과

규칙이름	제거된 문장/수
단어개수	159
평균단어길이	3
문장최대길이	280
단어최대길이	52
단어개수비률	432
특수단어비률	6 433
특수기호포함	17
문장끝기호	18 409
언어식별	29
괄호쌍관계	3 346
조선어토문장	16 571
단어의미판정	5

표 3에서 보여주는것처럼 자동추출된 226 414문장쌍의 영－조병렬코퍼스에서 45 736 문장이 제거되여 180 678문장을 학습에 리용할수 있게 되였다.

맺 는 말

통계적규칙들을 리용하여 자동추출된 영－조병렬코퍼스려과방법을 제안하고 영－조병렬코퍼스의 특성에 맞는 고유한 특징들을 추출하여 그것을 규칙으로 리용함으로써 병렬코퍼스의 질을 높일수 있게 하였다.

참 고 문 헌

[1] Philipp Koehn et al.; Proceedings of the Third Conference on Machine Translation(WMT), 2, 726, 2018.
[2] Nick Rossenbach et al.; Proceedings of the Third Conference on Machine Translation(WMT), 2, 946, 2018.
[3] Tom Ash et al.; Proceedings of the Third Conference on Machine Translation(WMT), 2, 853, 2018.
[4] Marcis Pinnis; Proceedings of the Third Conference on Machine Translation (WMT), 2, 939, 2018.

주체109(2020)년 11월 5일 원고접수

A Filtering Method of English-Korean Parallel Corpus in Parallel Corpus Automatic Production System

Kim Jun Gyu, *Sin Hyok Chol*

In this paper, we proposed a filtering method of parallel corpus, which was the core in building a neural machine translation system, using statistical methods, and then used this in improving the accuracy of training data.

In experiment, we showed that the translation accuracy improved by 1.2% than the original corpus.

Keywords: machine translation, parallel corpus, corpus filtering