신경기계번역에서 고유명사식별기술을 리용한 번역성능개선의 한가지 방법

김준규, 김광혁

경애하는 **김정은**동지께서는 다음과 같이 말씀하시였다.

《첨단돌파전을 힘있게 벌려야 나라의 과학기술전반을 빨리 발전시키고 지식경제의 토대를 구축해나갈수 있습니다.》

신경기계번역체계를 실현함에 있어서 학습자료에 존재하지 않는 고유명사(사람이름, 지역이름, 기관이름 등)들과 번역과정에 여러가지 원인으로 정확한 결과가 나타나지 않는 날자, 수자들을 처리하는것은 번역체계의 성능을 높이는데서 중요한 문제로 나선다. 이러한 문제들을 해결하기 위하여 자연언어처리분야에서는 고유명사식별(NER: Named Entity Recognition)을 진행하기 위한 연구가 심화되여왔다.

론문에서는 신경기계번역체계실현에서 난문제로 제기되는 미등록어처리를 진행하기 위하여 사전과 정규표현을 리용한 고유명사식별체계실현방법을 제안하였다.

대표적인 식별체계들로 NLTK(Natural Language Toolkit)[1]와 Stanford CoreNLP[2]를 들수 있다.

① NLTK는 자연언어처리를 위한 Python용서고이다. NLTK는 50개이상의 코퍼스자료들과 WordNet와 같은 어휘자원들을 제공하며 사용하기 편리한 대면부를 제공한다. 그리고 어종분류, 단어가르기, 말줄기가르기, 품사판정, 문장해석, 의미분석 등 여러가지 본문처리서고들을 제공한다.

NLTK를 리용하여 단어 및 말줄기가르기를 진행하고 품사판정을 진행하여 고유명사식별을 진행한 결과는 다음과 같다.

· 문장

At eight o'clock on Thursday morning, Arthur didn't feel very good.

· 단어 및 말줄기가르기

At, eight, o'clock, on, Thursday, morning, Arthur, did, n't, feel, very, good, .

· 품사판정

(At, IN), (eight, CD), (o'clock, JJ), (on, IN), (Thursday, NNP), (morning, NN), (Arthur, NNP), (did, VBD), (n't, RB), (feel, VB), (very, RB), (good, JJ), (., .)

· 고유명사식별

(At, IN), (eight, CD), (o'clock, JJ), (on, IN), (Thursday, NNP), (morning, NN), [PERSON, (Arthur, NNP)], (did, VBD), (n't, RB), (feel, VB), (very, RB), (good, JJ), (., .)

② Stanford CoreNLP는 자연언어처리를 진행하는 도구들을 제공해준다. 그리고 단어의 품사를 제공하며 그것들이 어떤 지역이나 기관, 사람의 이름인가, 날자, 시간인가 혹은 수값들의 표준형태인가를 식별한다. 이 체계는 확장가능하게 설계되였으며 선택적으로 기능들을 리용할수 있다. 제공되는 언어들은 아랍어, 중국어, 영어, 프랑스어, 도이취어, 에스빠냐어들이며 식별꼬리표들은 PERSON, ORGANIZATION, LOCATION, DATE, PRECENT 등이다.

론문에서는 선행방법들을 분석한데 기초하여 현재 구축된 자료기지로부터 추출한 고유명사사전과 정규표현을 리용한 고유명사식별체계를 개발하고 그 성능을 평가하였다.

1. 고유명사식별을 리용한 신경기계번역성능개선방법

신경기계번역체계에 결합된 고유명사식별기능의 동작과정은 다음과 같다.

걸음 1 고유명사사전을 구축한다.

걸음 2 문장이 입력되면 첫 단어로부터 시작하여 최장일치법으로 사전검색을 진행한다.

사전에서 단어가 검색되면 XNX(신경기계번역체계의 학습자료에도 존재하지 않으며 번역결과도 달라지지 않는 문자렬임.)로 원본단어를 치환하고 그뒤에 검색번호를 붙인다. 즉 XNX0, XNX1, … 형태로 영어단어가 치환되며 해당 단어에 대한 조선어대역은 사전에서 검색되여 목록화된다.

걸음 3 사전검색이 진행되여 변환된 문장에 대한 날자식별을 진행한다.

날자식별은 여러가지 형태로 장악된 날자형태들에 대한 정규표현을 리용하여 1차검색을 진행한다. 검색된 결과들은 XDX로 치환되며 그뒤에 검색번호를 붙인다.

걸음 4 XNX_i와 XDX_j로 치환된 문장을 신경번역체계에 보내여 번역문을 얻는다.

걸음 5 얻어진 번역문에 대하여 XNX_i들에 대한 사전에서 검색된 조선어대역치환을 진행한다.

조선어대역치환을 진행할 때 토처리(은|는, 와|과, 로|으로, …)를 진행한다.

걸음 6 얻어진 번역문에 대한 정규표현2차검색을 진행하여 정확한 날자번역결과를 얻고 최종번역문을 얻는다.

고유명사사전은 규칙토대의 번역기에 존재하던 110만개의 올림말들에서 추출된 78 821개의 고유명사들로 이루어져있다.

정규표현으로 서술된 날자형태들을 표 1에 보여주었다.

표 1. 정규표현으로 서술된 날자형태

No.	정규표현	번역결과
1	MONTH the NUM1{+ ORDNUM}, NUM2	NUM2년 MONTH월 NUM1일
2	MONTH NUM1{+ ORDNUM}, NUM2	NUM2년 MONTH월 NUM1일
3	MONTH NUM1, NUM2	NUM2년 MONTH월 NUM1일
4	NUM1{< 32} MONTH (,) NUM2	NUM2년 MONTH월 NUM1일
5	NUM1{+ ORDNUM} of MONTH in NUM2	NUM2년 MONTH월 NUM1일
6	NUM1{+ ORDNUM} of MONTH (,) NUM2	NUM2년 MONTH월 NUM1일
7	NUM1{+ ORDNUM} of MONTH	MONTH월 NUM1일
8	MONTH NUM	MONTH월 NUM일
9	MONTH(the) NUM{+ ORDNUM}	MONTH월 NUM일
10	(the) NUM{+ ORDNUM} MONTH	MONTH월 NUM일
11	NUM MONTH	MONTH월 NUM일
12	MONTH [in/of] NUM	NUM년 MONTH월
13	mid — MONTH	MONTH월 중순
14	beginning of MONTH	MONTH월초

2. 실험 및 평가

론문에서 제안한 방법으로 고유명사식별을 진행하기 위하여 2 200문장규모의 자료에 수동적으로 꼬리표를 붙여 고유명사(지역이름, 기관이름, 사람이름)와 날자들에 대한 정확한 시험자료를 만들고 앞에서 론의한 NLTK와 Stanford CoreNLP를 리용하여 고유명사식별을 진행하였다.

식별체계성능평가결과를 표 2에 보여주었다.

표 2. 식별체계성능평가결과

식별체계	문장수	정확한 개수		식별한 개수		성능/%
		고유명사	날자	고유명사	날자	
NLTK	2 200	267	98	198/22	0	48.2
Stanford CoreNLP	2 200	267	98	265/48	36/13	65.8
제안한 방법	2 200	267	98	272/39	92/11	86.0

표 2에서 보여준것처럼 선행방법의 식별체계들은 고유명사식별을 기본으로 하고 날자에 대한 처리는 대체로 진행하지 않는 결함을 가지고있었다. 론문에서는 사전과 정규표현들을 리용하여 날자식별을 진행하고 체계의 성능을 개선하였다.

제안한 방법으로 날자번역을 개선한 실례를 표 3에 보여주었다.

표 3. 제안한 방법으로 날자번역을 개선한 실례

영어	조선어	
	적용전	적용후
On average, such a potential ranges from 55% in April to 87% in October.	평균적으로 이러한 잠재력은 10월에 55%로부터 10월에 87%까지 범위에 있다.	평균적으로 그러한 포텐샬은 4월에 55%로부터 10월에 87%까지의 범위에 있다.
The space cooling period was from July 5, 2000 to August 29, 2000.	우주랭각주기는 2000년 7월부터 2000년 8월 29일까지 진행되였다.	공간랭각주기는 2000년 7월 5일부터 2000년 8월 29일까지였다.

맺는 말

신경기계번역체계실현에서 사전과 정규표현을 리용하여 고유명사식별을 진행하는 방법을 제안하였다. 이 방법은 여러가지 원인으로 하여 정확히 출력되지 않는 고유명사들과 날자들에 대한 처리를 진행함으로써 번역문의 질을 높일수 있게 한다.

참 고 문 헌

[1] Franck Dernoncourt et al.; Proceedings of the 2017 EMNLP System Demonstrations, 97, 2017.

[2] Marta R. Costa-jussa; Proceedings of the Sixth Named Entity Workshop, Joint with 54th ACL, 88, 2018.

주체110(2021)년 5월 5일 원고접수

A Method to Improve Translation Performance Using Named Entity Recognition in Neural Machine Translation System

Kim Jun Gyu, Kim Kwang Hyok

In this paper, we have proposed the method to process NER words in the neural machine translation system, using dictionary and regular expressions. And we have improved translation performance for proper nouns and dates that are not correctly outputted.

Keywords: machine translation, out-of-vocabulary, named entity recognition