중국어 학습자들을 위한 단어 분절 및 추천 파이프라인 시스템

PIAO SHENGMIN[O] 김경훈 박상현[1]
연세대학교 컴퓨터과학과
e-mail : {sungmin630, kkh115505, sanghyun}@yonsei.ac.kr

A Pipeline System about Chinese Word Segmentation and Recommendation for Chinese Learners

PIAO SHENGMIN[O] Kyeonghun Kim Sanghyun Park[1]
Dept.of Computer Science, Yonsei University

요 약

중국어 단어 분절(Chinese Word Segmentation)은 중국어 문장을 단어 단위로 나누는 작업이다. 분절된 단어를 중국어 학습자들의 실력에 맞는 단어로 대체하는 것을 이 논문에서는 추천(Recommendation)으로 표현한다. 본 논문에서는 Bi-LSTM 모델을 사용해서 중국어 단어 분절을 구현하고 난이도별 custom 사전을 기반으로 하여 rule-based 중국어 단어 추천 기법을 사용한다. 두 분산된 작업은 파이프라인(Pipeline)구조로 연결해서 구현한다. 중국어 단어 분절은 SIGHAN-2005 PKU 테스트 데이터 셋에서 F1점수가 95.5인 결과를 얻었고 중국어 단어 추천은 직관적인 좋은 결과를 얻었다.

1. 서 론

중국어 문장을 단어 단위로 나누는 작업을 지칭하는 중국어 단어 분절(Chinese Word Segmentation)은 중국어 자연어처리 영역에서 반드시 선행되어야 하는 작업으로써 주목받고 있다. 중국어는 영어, 한국어와 달리 단어 사이에 띄어쓰기와 같은 기호가 없이 문장이 구성된다. 그러므로 문장의 뜻에 맞는 단어를 잘 분별하는 것은 중국어 학습자에게는 쉽지 않다. 이와 같은 이유로 중국어 단어 분절은 중국어 자연어처리에서 중요한 작업이다. 문장의 의미에 맞는 단어를 정확하게 분절해야 추천, 번역 등 작업을 실행할 수 있다.

하지만 중국어 단어 분절은 주로 두 가지 큰 문제점이 있다. 첫 번째는 모호성(Ambiguity)다, 즉 단어 사이의 가능한 조합이 많아서 다양한 분절 결과가 나올 수 있다. 하지만 문법에 안 맞는 결과 혹은 문장 전체의 의미와는 다른 결과가 나올 수 있다. 두 번째는 OOV(Out Of Vocabulary)다. 이 문제의 원인은 사회 문화의 발전 그리고 지역성 언어문화 차이로 신조어, 지방 방언 등 많은 미등록된 단어가 생기기 때문이다. 분절 시스템이 효과적으로 이런 미등록 된 단어를 식별하는 것도 중국어 단어 분절이 극복해야 할 문제다.

중국어는 가장 오래 존재한 문자 중의 일부로서 긴 세월의 발전 중에서 아래와 같은 세 가지 특징이 있다. (1) 문자 수가 방대하다. (2) 거의 단어마다 네 가지 성조가 있다. (3) 단어마다 의미가 다양하다. 이런 특징은 중국어 학습자에게는 문장을 정확하게 분절해도 여전히 단어의 뜻을 이해하기에는 어려움을 가져온다. 현재 시장에는 중국어 초심자를 대상으로 한 교과서는 많다. 하지만 어느 정도의 기초가 있는 학습자들은 중국어를 공부하면서 어려운 단어를 접촉할 때 자신의 지식을 활용해서 이해하기에는 어렵고 이를 도와줄 수 있는 효율적인 교과서도 많지 않다. 본 논문에서는 한 언어를 배울 때 가장 좋은 방법을 자신의 모어가 아닌 그 언어로 배우는 것이 훨씬 더 효과가 있다는 비전으로 중국어 단어 추천을 한다. 이 작업은 학습 등급을 기준으로 사전을 만들고 이를 사용해서 분절된 어려운 단어를 학습자의 실력에 맞는 쉬운 단어로 대체하는 작업이다.

2. 시스템 구조

이 섹션에서는 파이프라인 구조로 실현한 시스템에 대해서 상세하게 설명해 드린다. 그림 1과 같이 이 구조는 중국어 단어 분절과 추천을 포함한다. 사용자는 분절과 추천을 할 문장과 자신의 중국어 실력 등급을 입력한다. 시스템은 먼저 훈련된 Bi-LSTM 모델로 문장을 분절하고. 추천 시스템은 사전과 사용자의 등급에 의하여 분절된 단어를 사용자의 등급에 맞는 단어로 대체한 후 결과를 출력한다.

2.1. 중국어 단어 분절

Bidirectional long short-term memory[1](Bi-LSTM) 모델은 중국어 단어 분절에서 좋은 성능을 보여줬다. 그러므로 본 연구에서도 그림 2와 같은 Bi-LSTM 모델을

---

[1] 교신저자(Corresponding author)

* 이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2017-0-00477, (SW 스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)과 국토교통부의 스마트시티 혁신인재육성사업의 지원을 받아 수행된 연구임

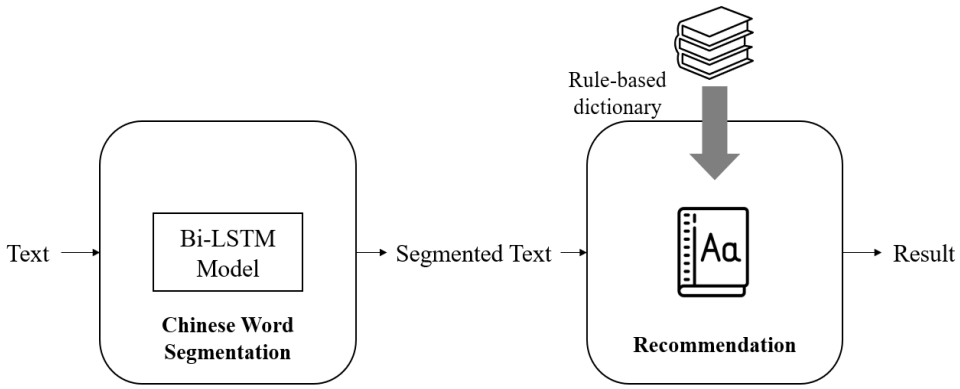

그림 1. 중국어 단어 분절 및 추천 파이프라인 시스템

사용한다. 단어의 표현력을 풍부하기 위해 단어의 임베딩은 유니그램(unigram), 바이그램(bigram), 트라이그램(trigram) 세 가지를 사용한다. 단어의 임베딩은 예비 훈련(pre-training) 단계에서 Word2Vec[2]를 사용해서 구한다. 그리고 임베딩을 연결(concatenate)한 후 Bi-LSTM 모델에 입력한다. 본 연구에서는 분절한 단어의 BIES 태그를 인코딩(encoding) 한다. BIES는 각각 시작(Begin), 중간(Inside), 끝(End), 단자(Single)를 표시한다. 단어의 BIES 태그를 구할 때 CRF 기법을 사용하지 않고 소프트맥스 레이어(softmax layer)를 사용한다.

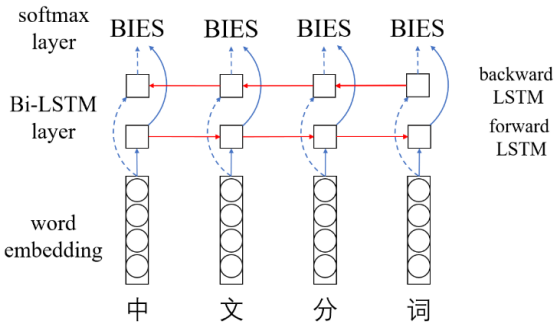

그림 2. Bi-LSTM 모델

2.2. 중국어 단어 추천

본 연구에서는 학습자들의 실력에 맞는 중국어 단어를 추천하기 위하여, 난이도별 custom 사전을 기반으로 하여 rule-based 추천 기법을 수행한다. 중국어 능력 시험(HSK)에서는 중국어 실력을 1-6의 등급으로 구분한다. 이에 따른 각 등급의 사전이 있지만 사전이 포함한 단어 개수는 표 1과 같이 적다는 단점이 있다. 또한, 각 등급 사이의 단어 개수의 배치도 불균형 하다. 이런 단점을 극복하기 위해 본 연구는 HSK의 사전을 사용하지 않고 “A Dictionary of Chinese Usage: 8000 Words Chinese Proficiency Test Vocabulary Guideline”[3](DCU)를 사용한다. 이 책의 단어 양은 표 2와 같이 HSK에 비해 상대적으로 많고 난이도 등급을 구분하는 방식도 HSK와 다르다.

효율적인 단어 추천 시스템을 위해 본 연구는 “HIT-SCIR Tongyici Cilin (Extended)”[4](TC-E)을 활용해 방대한 통합 단어 사전을 만든다. TC-E는 난이도와 상관없이 동의어를 한 인덱스로 표현한다. 그러므로 본 연구는 DCU, TC-E와 중국어 단어 분절에서 사용한 훈련 데이터 셋(Section 3.1.)을 활용하여 아래의 2단계 과정을 걸쳐 통합 단어 사전을 만든다.

첫 단계는 훈련 데이터 셋을 사용해서 TC-E의 각 인덱스의 단어를 빈도수가 높은 기준으로 재배열한다. 하지만 빈도수 기준 정렬은 단어의 난이도를 고려하지 않았기 때문에, 다음 단계에서 DCU를 사용하여 난이도를 매핑한다.

두 번째 단계는 DCU 사전을 통해 위에서 생성된 사전을 난이도별로 정렬한다. 이때, DCU의 단어 개수보다 통합 단어 사전의 개수가 많다. 하지만, DCU는 주로 사용되는 대부분의 중국어 단어를 포용할 수 있기 때문에, DCU에 포함되지 않는 단어는 빈도수가 낮아 첫 단계에서 이미 가장 후순위로 정렬된다. 이러한 DCU에 포함되지 않는 여분의 단어는 새로운 등급-무급(5급)로 분류한다.

최종적으로 통합 단어 사전에는 뜻이 유사한 단어끼리 같은 인덱스를 갖고 인덱스 내부의 단어들은 난이도를 기준으로 정렬되어 있다.

표 1. HSK 단어 개수

HSK등급	단어개수	HSK등급	단어개수
HSK 1급	150	HSK 4급	600
HSK 2급	150	HSK 5급	1300
HSK 3급	300	HSK 6급	2500

표 2. DCU 단어 개수

등급	단어개수	등급	단어개수
갑급(1급)	1033	병급(3급)	2202
을급(2급)	2018	정급(4급)	3569

표 3. 중국어 단어 분절 및 추천 예제

원본 문장	단어 분절 결과	추천 등급	단어 추천 예제
大伙(여러분)都**劳苦(노고)**一天了	大伙 // 都 // 劳苦 // 一 // 天 // 了	3	**大家(모두)**都**辛苦(고생)**一天了
我有一个小**本本(서적)**本来很干净	我 // 有 // 一个 // 小 // 本本// 本来// 很 // 干净	2	我有一个小**书 (책)**本来很干净
研究生命的**起源(기원)**	研究 // 生命 // 的 // 起源	1	研究生命的**根(뿌리)**

3. 실 험

3.1. 중국어 단어 분절 훈련을 위한 데이터셋

본 연구는 SIGHAN 2005 bake-off task[5]의 데이터 셋을 사용한다. 이 데이터 셋의 내용은 표 4과 같다.

중국어 단어 분절의 효율성을 울리기 위해 본 연구는 데이터 셋에 있는 4개 서브 데이터 셋을 통합한다. 그중 AS와 CITYU는 번체 문자이므로 간체로 바꿔 사용한다.

3.2. 중국어 단어 분절 및 추천 실험 결과

표 5는 중국어 단어 분절의 선행 연구와 본 연구의 단어 분절 모형의 실험 결과이다.

기존 연구[6][7][9]들은 CRF 레이어를 통해 높은 점수를 기록하였다. 그러나 본 연구에서 사용한 모델은 CRF 레이어를 제거한 가벼운 모형임에도 불구하고, 유니그램, 바이그램, 트라이그램 정보를 모두 활용하여 선행 연구들과 유사한 결과를 얻을 수 있었다.

표 3은 중국어 단어 분절과 난이도를 낮춰 단어를 추천하는 예제이다. 원본 문장은 단어 분절 모형을 통해 의미적으로 구분되고, 분절된 단어에서 사용자가 요청한 난이도(표3. 추천 등급)에 맞춰 통합 단어 사전을 통해 단어를 추천한다. 예를 들어, 大伙都劳苦一天了 문장의 경우, 大伙(여러분)와 劳苦(노고)는 주로 사용되는 단어가 아니므로, 난이도를 낮추어 大家(모두), 辛苦(고생)로 각각 추천된다.

표 4. 중국어 단어 데이터 셋

	Train	Develo-pment	Test
AS	4,903,564	546,017	122,610
CTIYU	1,309,208	146,422	40,936
MSR	2,132,480	235,911	106,873
PKU	994,822	115,125	104,372

표 5. 선행 연구 비교 (PKU 데이터셋)

Model	F1
Tian et al. (2020)[6]	**96.5**
Yang et al. (2017)[7]	96.3
Ma et al. (2018)[8]	96.1
Yang et al. (2018)[9]	95.9
Ours	**95.5**

4. 결론 및 향후 연구

본 연구는 Bi-LSTM 모델을 사용해서 중국어 단어 분절을 구현하고, 중국어 단어 추천은 난이도별 custom 사전을 기반으로 하여 rule-based 기법을 사용한다. 두 작업은 파이프라인 구조로 연결함으로써 시스템을 구축하였다. 추후 세분화된 등급을 가진 사전 구축도 의미 있는 연구가 될 것으로 사료된다.

참고문헌

[1] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory. " *Neural computation* 9.8 : 1735-1780. 1997.

[2] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *NIPS*. 2013.

[3] Liu, L. L. "A Dictionary of Chinese Usage: 8000 Words Chinese Proficiency Test Vocabulary Guideline." 2000.

[4] http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm.

[5] Emerson, Thomas. "The second international Chinese word segmentation bakeoff." *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*. 2005.

[6] Tian, Yuanhe, et al. "Improving Chinese Word Segmentation with Wordhood Memory Networks." *ACL*. 2020.

[7] Yang, Jie, Yue Zhang, and Fei Dong. "Neural word segmentation with rich pretraining." *arXiv preprint arXiv:1704.08960.* 2017.

[8] Ma, Ji, Kuzman Ganchev, and David Weiss. "State-of-the-art Chinese word segmentation with bi-lstms." *arXiv preprint arXiv:1808.06511*. 2018.

[9] Yang, Jie, Yue Zhang, and Shuailong Liang. "Subword encoding in lattice lstm for chinese word segmentation." *arXiv preprint arXiv:1810.12594*. 2018.