음절의 의미역 태그 분포를 이용한 Bidirectional LSTM CRFs 기반의 한국어 의미역 결정

윤정민O, 배경만, 고영중
동아대학교 컴퓨터공학과
{yjungmin2, kyoungman.bae, youngjoong.k}@gmail.com

Korean Semantic Role Labeling Based on Bidirectional LSTM CRFs Using the Semantic Label Distribution of Syllables

Jungmin YoonO, Kyoungman Bae, Youngjoong Ko
Donga University, Department of Computer Engineering

요 약

의미역 결정은 자연어 문장의 서술어와 그 서술어에 속하는 논항들 사이의 의미관계를 결정하는 것이다. 최근 의미역 결정 연구에는 의미역 말뭉치와 기계학습 알고리즘을 이용한 연구가 주를 이루고 있다. 본 논문에서는 순차적 레이블링 영역에서 좋은 성능을 보이고 있는 Bidirectional LSTM-CRFs 기반으로 음절의 의미역 태그 분포를 고려한 의미역 결정 모델을 제안한다. 제안한 음절의 의미역 태그 분포를 고려한 의미역 결정 모델은 분포가 고려되지 않은 모델에 비해 2.41%p 향상된 66.13%의 의미역 결정 성능을 보였다.

주제어: 의미역 결정, Korean Propbank, Bidirectional LSTM-CRFs

1. 서 론

의미역(Semantic role)은 문장 내에서 서술어에 의해 기술된 행동이나 상태에 대한 명사구의 의미역할을 말하며, 의미역이 부여된 각 명사구를 논항(argument)이라고 한다. 의미역 결정(Semantic role labeling, 이하 SRL)은 서술어의 의미와 그 서술어에 대한 논항들의 의미역을 결정하는 것을 말한다. 이는 각 명사구의 논항을 미리 정의된 의미역관계로 사상하는 문제라 볼 수 있다[1].

최근 의미역 결정 연구는 잘 정제된 코퍼스를 기반으로 기계학습을 이용한 연구가 주를 이루고 있고 형태소와 구문분석 결과를 기반으로 최적의 자질 조합을 생성하여 의미역 결정 성능을 향상 시킨 연구들이 이루어 졌다. 최적의 자질을 기반으로 Structural SVM을 이용한 연구가 있고[1], 추가로 의미 정보자질을 이용하여 성능을 향상시킨 연구가 있다[2]. 또한 단어에 대해 높은 수준으로 추상화 된 표상을 이용하는 Bidirectional Long Short Term Memory CRFs(bi-LSTM-CRFs)기반의 연구가 이루어 졌다[3]. 본 논문은 최근 SRL 영역에서 좋은 성능을 보이고 있는 bi-LSTM-CRFs 모델을 기반으로 음절의 의미역 태그 분포를 반영한 개선된 모델을 제안한다.

bi-LSTM-CRFs는 어절단위로 입력이 결정되며, 각 어절에 대한 어절 표상을 나타내는 벡터를 구성하여 입력으로 사용된다. 본 논문에서는 이를 위해 대용량 원시 말뭉치를 기반으로 어절 표상을 나타내는 64차원의 벡터를 생성하고, 여기에 어절에 포함된 형태소의 품사와 현재 어절과 의미적으로 연관된 서술어에 대한 정보가 반영된 벡터를 결합하여 입력으로 사용한다.

bi-LSTM-CRFs의 성능을 향상시키기 위해 본 논문에서는 각 어절에 포함된 음절이 전체 학습 말뭉치에서 출현한 의미역 태그의 분포를 반영하여 어절의 벡터를 구성하여 bi-LSTM-CRFs의 입력을 확장하였다. 각 음절은 학습 말뭉치에서 여러 서술어의 의미역에 포함이 될 수 있으며, 여러 의미역을 나타내는 다양한 품사를 가질 수 있다. 학습 말뭉치에서 음절이 포함된 의미역의 의미역 태그 빈도수를 계산한 후 softmax를 통해서 확률을 각 차원의 값으로 사용하였다. 각 음절별로 생성된 의미역 태그의 분포 벡터는 bi-LSTM 기반의 음절 의미역 태그 분포가 반영된 어절 벡터를 생성하는 모델의 입력으로 사용되며, 어절에 포함된 모든 음절에 대해 forward와

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2015R1D1A1A01056907)

backward 단계를 거쳐 생성된 벡터를 생성하는데 이용된다. 생성된 음절의 의미역 분포 정보가 반영된 어절에 대한 벡터는 어절단위로 SRL을 진행하는 bi-LSTM-CRFs 모델의 입력 벡터에 결합하여 확장함으로써 개선된 의미역 결정 모델을 제안한다.

음절의 의미역 태그 분포 벡터를 이용한 bi-LSTM-CRFs 모델을 의미역 결정 영역에 효과적으로 적용한 결과 의미역 태그 분포 벡터를 이용하지 않은 모델에 비해 2.41%p향상된 66.13%의 성능을 보였다. 이를 통해 제안한 음절의 의미역 태그 분포가 bi-LSTM-CRFs 기반의 의미역 결정 모델에 효과적이라는 것을 확인 할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 설명하고, 3장에서는 제안하는 음절의 의미역 태그 분포를 이용한 bi-LSTM-CRFs 기반의 의미역 결정 방법을 설명한다. 4장에서는 제안한 방법을 실험을 통해 평가하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

의미역 결정 연구는 격틀 사전에 기반을 둔 방법과 말뭉치에 기반을 둔 방법으로 나눌 수 있다. 격틀 사전에 기반을 둔 방법은 각 서술어에 필요한 논항들의 쓰임을 격틀사전으로 구축하여 사용하고, 격틀(frame)과 선택제약(selectional restriction)을 이용하여 서술어에 대한 논항을 결정한다. 이러한 방법은 의미역이 사전에 구축된 격틀에 의해 결정되므로 높은 정확률을 보이지만, 격틀 사전의 구축이 어렵고 이미 구축된 격틀이 없을 때 의미역을 적용하기가 힘들다[1,4].

말뭉치에 기반을 둔 방법은 의미역이 태깅 된 말뭉치를 구축하여 이 말뭉치로부터 기계학습을 위한 학습데이터를 생성하고 기계학습 모델을 통해 의미역을 결정하는 방법이다. 하지만 말뭉치를 구축하기 어렵고 각 기계학습모델에 대한 최적의 자질 조합을 찾는데 많은 비용이 든다. 최근에는 주로 의미역 말뭉치를 이용하여 학습데이터를 만들고 형태소, 구문분석 자질을 이용하여 자질조합을 설계한다. 순차 레이블링에 좋은 성능을 보이는 SSVM을 이용하여 잘 설계된 자질조합으로 성능을 향상시킨 연구가 있고, 이보다 추가적인 자질로 의미자질을 사용하여 성능을 향상 시킨 연구가 있다[1,2]. 또한 단어를 높은 수준으로 추상화 하고 보다 자질조합을 설계하기 쉬운 딥 러닝에 기반한 연구가 진행되어 왔으며 순차 레이블링에 높은 성능을 보이는 BLSTM-CRF를 이용한 연구가 있다[3,5,6].

3. 제안 모델

3.1. bi-LSTM-CRFs 모델

RNN(Recurrent Neural Networks)[7]은 순환신경망으로 입력 층, 은닉 층, 출력 층을 거치는 일반적인 신경망과 달리 현재 은닉 층의 결과가 다시 은닉 층의 입력으로 들어가는 신경망이다. 학습이 진행되는 동안 이전 학습의 정보를 잃지 않고 연속적인 정보의 흐름을 학습에 반영 가능하다. 하지만, 학습이 진행되는 동안 그래디언트 소멸 또는 발산 문제가 발생하여 장거리 의존성을 학습할 수 없게 된다.

LSTM(Long Short Term Memory)은 순차데이터 처리에 유리한 RNN의 문제점을 보완한 모델이다. LSTM에서는 오류역전파(Backpropagation) 과정에서 오류의 값이 소실, 발산 하지 않고 잘 유지되는데 LSTM의 게이트가 부착된 셀의 기능으로 정보를 가져올지, 상태를 다음 노드로 전이할지 혹은 유지할 지를 결정한다.

bi-LSTM-CRFs는 앞에서부터 순차적으로 진행된 정보를 기반으로 학습을 하는 forward 단계와 뒤에서부터 순차적으로 진행되는 정보를 기반으로 학습하는 backward 단계를 동시에 고려한 모델이다. 최근 순차적인 레이블링이 필요한 많은 영역에서 활용되고 있으며, SRL 영역에서도 좋은 성능을 보여 주고 있다. 그림 1은 bi-LSTM-CRFs를 이용한 의미역 결정 모델의 입력 및 출력 형태를 도식화한 그림이다.

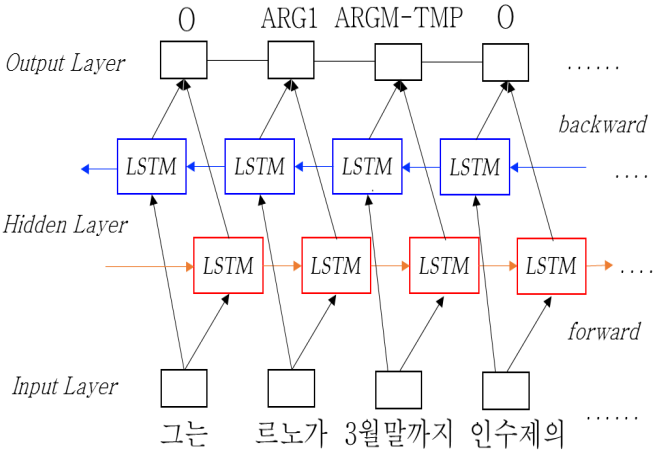

그림 1 bi-LSTM-CRFs

입력의 형태는 N차원의 벡터 형태이며, 현재 어절의 의미역 결정에 필요한 다양한 자질과 연관된 벡터들로 구성이 된다[3]. 아래 표 1은 제안하는 모델에서 사용한 자질을 나타낸다.

표 1. 구축한 bi-LSTM-CRFs 모델의 학습에 사용한 자질

자질	
자질 1	현재 어절, 서술어 어절, 현재 어절 바로 앞의 어절, 현재 어절 바로 뒤의 어절
자질 2	현재 어절과 바로 앞뒤 어절의 품사 정보
자질 3	서술어와 현재 어절 사이의 위치 관계 및 거리 정보

기본적으로 어절에 대한 정보가 반영된 벡터가 입력으로 사용된다. 본 논문에서는 64차원의 어절 임베딩 벡터를 기본입력으로 사용한다. 의미역 결정을 위해서는 관련이 있는 서술어의 정보가 중요하며, 서술어가 포함된 어절에 대한 임베딩 벡터를 현재 어절 벡터에 결합하여 확장된 벡터를 구성한다. 문장에서의 문맥 정보를 반영하기 위해 앞, 뒤 어절에 대한 벡터를 추가로 사용하여 입력 벡터를 구성한다. 그리고 각 어절에 포함된 형태소의 품사 정보를 반영하기 위해 어절에 포함된 형태소들의 품사 벡터를 표 2와 같이 구성하여 사용한다.

표 2. 어절별 품사 벡터(46차원)의 예

품사	NNG	...	NNP	JKS
빈도	1	...	0	1

가능한 형태소의 품사는 총 46개이고, 하나의 어절에 대해 출현한 형태소들의 출현 여부를 벡터의 값으로 사용한다. 그리고 현재어절과 술어에 관련한 위치 정보 및 거리를 벡터로 구성하여 최종적인 입력 벡터를 구성한다. 구성한 벡터를 bi-LSTM-CRFs에 입력 벡터로 사용하여 학습 말뭉치를 통해 학습을 진행하고, 새로운 문장에 대한 입력 벡터를 학습된 모델에 입력하면 그림 1과 같이 각 어절별로 의미역에 관한 태그가 결과로 출력이 된다.

3.2 음절의 의미역 태그 분포 이용한 bi-LSTM-CRFs 모델

본 논문에서는 bi-LSTM-CRFs를 이용한 의미역 결정 모델의 성능 향상을 위해 표 1에의 자질을 기반으로 구성된 입력 벡터를 확장한다. 이를 위해, 어절을 구성하는 음절들이 학습 말뭉치에서 출현한 의미역 태그의 분포를 벡터로 표현하여 입력 벡터를 확장함으로써 의미역 결정 성능을 개선한다. 이를 위해 의미역 결정을 위한 bi-LSTM-CRFs 모델과 별도의 bi-LSTM 모델을 구축한다. 음절의 의미역 태그 분포가 반영된 어절 벡터를 생성하기 위한 모델의 구성도는 그림 2와 같다.

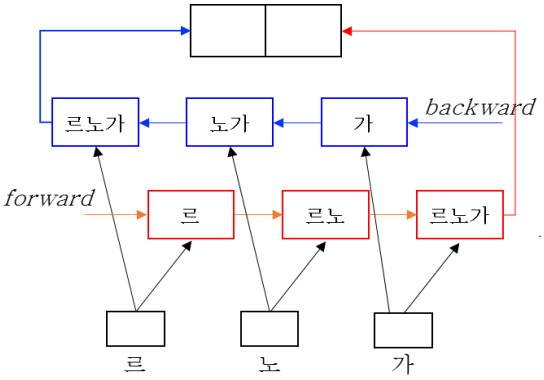

그림 2. 음절 기반 임베딩 벡터

어절은 여러 음절로 구성이 되며, 각 음절은 학습 말뭉치에서 서술어들에 대한 의미역 태그를 가지는 다양한 어절에 포함이 될 수 있다. 예를 들어, 문장 “300만마르크를 확보해 놓았다고 전했다.”는 서술어 “확보해”가 존재하며, “300만마르크” 1개의 논항을 가지고 있다. 또한, 문장 “르노가 3월말까지 인수제의 시한을 갖고 있다고 덧붙였다.”는 서술어 “갖고”와 “덧붙였다”를 가지고 있으며, “갖고”의 경우 “르노가”와 “시한을”인 2개의 논항을 가지고 있다. 이 경우 음절 “르”는 포함된 논항을 기준으로 I-ARG1, B-ARG0와 같은 태그를 가질 수 있다.

본 논문에서는 각 음절이 학습 말뭉치에서 출현한 의미역 태그의 분포를 반영하여 하나의 어절을 표현하는 벡터를 구성한다면 의미역 결정에 도움이 되는 정보가 반영되어 의미역 결정의 성능이 향상 될 것이라는 가정하에 음절의 의미역 태그 분포를 반영한 어절 벡터를 생성하는 모델을 제안한다. 제안하는 어절 벡터를 생성하는 모델은 bi-LSTM을 기반으로 하고 있으며, 입력은 음절에 대한 학습 말뭉치에서의 의미역 태그 분포가 반영된 벡터를 사용한다.

음절이 학습 말뭉치에서 출현한 의미역 태그의 분포를 벡터로 표현하기 위해서 음절의 의미역 태그 분포 벡터를 나타내는 벡터 차원은 모든 의미역 태그 18개에 B, I 태그를 반영한 36개의 차원에 의미역 태그가 아닌 경우를 나타내는 O 태그를 나타내는 1개의 차원을 더해 37차원의 벡터로 표현하였다. 이를 위해, 한 음절에 대해 학습 말뭉치에서 출현한 각 태그 별 빈도를 모두 구한 후 softmax를 통해 출현한 태그 별 빈도수를 확률 값으로 만들어 표 3과 같이 결정한다.

표 3. 음절 “가”에 대한 의미역 태그 분포 벡터의 예

	B-ARG0	B-ARG1	...	I-ARG0	O	...
softmax	0.02641	0.02664	...	0.03002	0.02734	...

표 3과 같이 생성된 어절에 포함된 음절의 의미역 태그 분포 벡터를 이용하여 그림 2에서와 같이 어절을 구성하는 음절이 출현한 순서로 bi-LSTM의 forward 와 backward 단계를 진행한다. 각 단계의 최종 상태를 나타내는 벡터를 결합하여 최종적으로 음절의 의미역 태그 분포가 반영된 어절의 벡터를 생성한다. 생성된 음절의 의미역 태그 분포가 반영된 어절의 벡터는 사전학습 된 어절 임베딩 벡터와 결합하여 그림 1의 의미역 결정 모델의 입력으로 사용이 된다. 제안하는 모델의 최종적인 구성도는 아래 그림과 같다.

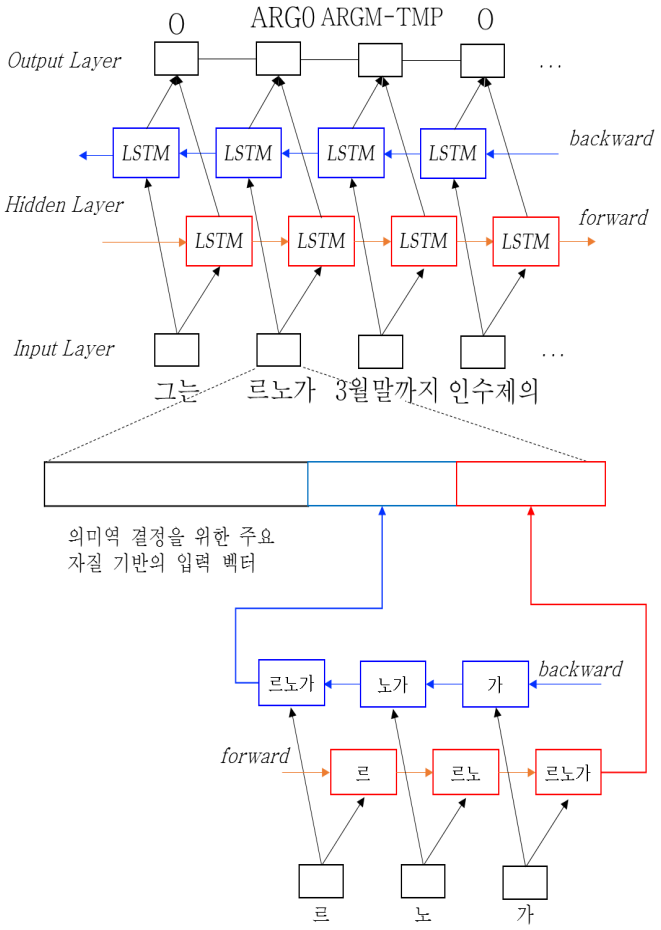

그림 3 음절의 의미역 태그 분포가 반영된 최종 의미역 결정 모델의 구성도

4. 실험

4.1 데이터 셋 및 평가 방법

본 논문에서는 실험을 위해 의미역 말뭉치인 Korean PropBank[8]를 사용하였다. Korean PropBank의 Newswire 경제 신문 도메인의 구구조를 의존구조로 변환한 4,714문장을 train에 3,771문장, test에 943문장을 사용하였다. 실험 결과의 성능은 정확률(precision)과 재현율(recall)의 조화평균인 F1을 사용하였다.

본 논문에서 제안하는 의미역 결정 모델은 한 문장에 2개 이상의 서술어가 존재하는 경우 각 서술어별로 의미역이 부여된 논항을 학습하였다. 의미역 분석에 대한 테스트 시에도 입력된 테스트 문장에 서술어가 2개 이상이라면 각 서술어별로 논항을 판단하여 평가를 진행하였다. 예를 들어, “그는 르노가 3월말까지 인수제의 시한을 갖고 있다고 덧붙였다.”라는 문장은 “갖고”와 “덧붙였다”라는 서술어가 포함된 어절이 존재하며, 각 서술어에 대한 의미역 결정 결과는 아래 표와 같다.

표 4. 서술어가 2개인 문장의 의미역 결정의 예

서술어	논항	정답	시스템 결과
갖고	르노가	ARG0	ARG0
	3월말까지	ARGM-TMP	ARGM-TMP
	시한을	ARG1	ARG1
덧붙였다.	그는	ARG0	ARG0
	있다고	ARG1	ARG1

“갖고”의 경우 “르노가”, “3월말까지”, “시한을” 3개 어절에 대해서 정답 레이블을 부여하여 학습을 진행한다. “덧붙였다”의 경우는 “그는”, “있다고”와 같은 2개 어절에 대해서만 정답 레이블을 부여하여 학습한다. 테스트는 문장에서 서술어를 인식한 후 각 서술어에 대해서 전체 문장의 어절들을 대상으로 의미역 결정을 진행하고, 위의 예제의 경우 “갖고”와 “덧붙였다” 서술어 각각에 대해서 의미역 결정을 진행한다. 표 6은 두 서술어에 대한 의미역 결정 결과가 모두 맞은 경우이다.

4.2 자질 별 의미역 결정 성능

본 논문에서 제안한 bi-LSTM-CRFs 기반의 의미역 결정 모델은 표 1에 언급한 것과 같이 여러 자질들을 기반으로 생성된 벡터를 입력으로 사용한다. 각 자질을 나타내는 것은 다음과 같다.

표 5. 사용한 자질

자질	의미
CE(current ejeol)	현재 어절
VE(verb ejeol)	서술어 어절
PE(previous ejeol)	이전 어절
NE(next ejeol)	다음 어절
CEP, PEP, NEP	어절 + 어절의 품사 집합
PEP1, NEP1	현재 어절과의 거리가 1인 이전 어절과 이전 어절의 품사집합
PEP12, NEP12	현재 어절과의 거리가 1, 2인 이전 어절들과 이전 어절들의 품사집합
PEP123, NEP123	현재 어절과의 거리가 1, 2, 3인 이전 어절들과 이전 어절들의 품사집합
LD(location & distance)	현재 어절과 서술어 어절의 문장 내 위치 및 두 어절의 거리

본 논문에서는 사용한 자질은 어절 자질(현재 어절, 서술어 어절), 품사 자질(어절의 품사 집합), 문맥 자질(현재 어절의 앞, 뒤 어절), 위치 자질(현재 어절과 서술어 어절의 문장 내 위치와 두 어절 사이의 거리)을 기반으로 생성된 벡터들을 모두 결합하여 bi-LSTM-CRFs의 입력으로 사용하였다. 각 자질에 대한 성능의 결과는 아래 표와 같다.

표 6. 자질별 의미역 결정 성능(%)

자질	Micro-F1
CE / VE	44.26
CEP / VE	46.95
CEP / VE / LD / PEP1 / NEP1	**55.52**
CEP / VE / LD / PEP12 / NEP12	54.35
CEP / VE / LD / PEP123 / NEP123	53.20

표 6에서 보는 봐와 같이 현재 어절과 서술어 어절 자질을 기본적인 자질로 성능 평가를 했을 때 44.26%의 성능을 보였다. 여기에 현재 어절의 품사 집합을 추가하였을 때 46.95%의 성능 향상을 보였으며, 앞과 뒤의 거리가 1인 어절과 어절의 품사 집합을 추가하였을 때 더 가장 좋은 성능을 보였다.

4.3 음절의 의미역 태그 분포를 이용한 의미역 결정 성능

본 논문에서는 구축한 bi-LSTM-CRFs 기반의 의미역 결정 성능을 개선하기 위해서 음절의 의미역 태그 분포를 이용한 의미역 결정 모델을 제안하였다. 일반적으로 논항은 문장에서 서술어의 앞에 존재하는 경우가 많기 때문에 본 논문에서는 문장에서 서술어의 앞에 존재하는 어절들만을 대상으로 의미역 결정을 진행하였다. 표 7에서 보는 것과 같이 서술어의 앞에 존재하는 어절에 대해서만 의미역 결정을 진행하였을 때 더 좋은 성능을 보였다.

표 7. 서술어의 앞에 존재하는 어절의 의미역 결정 결과 (%)

Model	Micro-F1
bi-LSTM-CRFs 모델	55.52
bi-LSTM-CRFs 모델+서술어 앞의 어절만 의미역 결정	**63.72(+8.2)**

표 7에서 보는 것과 같이 서술어의 앞 어절에 대해서만 의미역 결정을 진행하여 결과 기존의 모델에 비해 8.2%p 향상된 63.72%의 성능을 보였다. 이를 통해 서술어의 앞 어절에 대해서만 의미역 결정을 진행하는 것이 의미역 결정에서 효과적이라는 것을 확인할 수 있다. 서술어의 앞 어절에 대해서만 의미역 결정을 진행하는 방법을 기반으로 제안한 음절의 의미역태그 분포를 반영한 성능의 결과는 표8과 같다.

표 8. 음절 의미역 태그 분포를 이용한 실험결과(%)

Model	Micro-F1
bi-LSTM-CRFs 모델 + 서술어 앞의 어절만 의미역 결정	63.72
bi-LSTM-CRFs 모델 + 서술어 앞의 어절만 의미역 결정 + 음절의 의미역 태그 분포	**66.13(+2.41)**

표 8에서 보는 것과 같이 제안한 음절의 의미역 태그 분포를 반영한 결과 기존의 모델에 비해 2.41%p 향상된 66.13%의 성능을 보였다. 이를 통해, 본 논문에서 제안한 음절의 의미역 태그 분포가 의미역 결정 성능 향상에 효과적인 것을 확인 할 수 있다.

5. 결론

본 논문은 순차 레이블링 영역에서 뛰어난 성능을 보이고 있는 bi-LSTM-CRFs를 기반으로 음절의 의미역 태그 분포를 고려한 개선된 의미역 결정 모델을 제안하였다. 음절의 의미역 태그 분포를 반영한 의미역 결정 모델을 Korean Probank 말뭉치를 기반으로 학습 및 평가를 진행하였으며, 제안한 음절의 의미역 태그 분포를 고려하지 않은 모델에 비해 2.41%p 향상된 66.13%의 성능을 보였다. 이를 통해, 제안한 음절의 의미역 태그 분포를 반영한 어절의 벡터가 의미역 결정의 성능 향상에 효과적인 것을 확인 할 수 있었다.

향후 bi-LSTM-CRFs의 성능 향상을 위해 확장 가능한 추가적인 입력 벡터에 대한 연구를 지속적으로 진행할 것이다.

참고문헌

[1] 이창기, 임수종, 김현기. Structural SVM 기반의 한국어 의미역 결정. *정보과학회논문지* 제 42권 제2호, pp, 220-226, 2015

[2] 임수종, 김현기. 의미 정보를 이용한 한국어 의미역 인식 연구. *HCLT,* pp, 18-19, 2015

[3] 배장성, 이창기, Bidirectional LSTM CRF를 이용한 End-tod-end 한국어 의미역 결정. Master's Thesis, University of Kangwon National

[4] 김병수, 이용훈, 나승훈, 김준기, 이종혁. 부트스트랩핑 알고리즘을 이용한 한국어 격조사의 의미역 결정. *정보과학회논문지,* 제 33권, 제1호, pp, 4-6, 2006

[5] 배장성, 이창기, 임수종. Backward LSTM CRF를 이용한 한국어 의미역 결정. *HCLT,* pp, 194-195, 2015

[6] 배장성, 이창기, 임수종. 딥 러닝을 이용한 한국어 의미역결정. *한국컴퓨터종합학술대회, pp, 690-692,* 2015

[7] James Martens, Ilya Sutskever. "Learning Recurrent Neural Networks with Hessian-Free Optimization". *In Proceedings of International Conference on Machine Learing,* pp, 1033-1040, 2011

[8] Martha Palmer, S.Ryu,J. Choi, S.Yoon, Y.Jeon. "Korean Propbank", http://catalog.ldc.upenn.edu/LDC2006T03

[9] Zhou Jie, Wei Xu. "End-to-end Learning of Semantic Role Labeling Using Recurrent Neural networks".*Association Computational Linguistics,* pp, 1127-1137, 2015

[10] word2vector, [Online]. Available: https://code.google.com/archive/p/word2vec/