신경망 기반 텍스트 모델링에 있어 순차적 결합 방법의 한계점과 이를 극복하기 위한 담화 기반의 결합 방법

A Discourse-based Compositional Approach to Overcome Drawbacks of Sequence-based Composition in Text Modeling via Neural Networks

저자 (Authors)

이강욱, 한상규, 맹성현
Kangwook Lee, Sanggyu Han, Sung-Hyon Myaeng

출처 (Source)

정보과학회 컴퓨팅의 실제 논문지 23(12), 2017.12, 698-702 (5 pages)
KIISE Transactions on Computing Practices 23(12), 2017.12, 698-702 (5 pages)

발행처 (Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07319365

APA Style

이강욱, 한상규, 맹성현 (2017). 신경망 기반 텍스트 모델링에 있어 순차적 결합 방법의 한계점과 이를 극복하기 위한 담화 기반의 결합 방법. 정보과학회 컴퓨팅의 실제 논문지, 23(12), 698-702.

ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 23, No. 12, pp. 698-702, 2017. 12
https://doi.org/10.5626/KTCP.2017.23.12.698

신경망 기반 텍스트 모델링에 있어 순차적 결합 방법의 한계점과 이를 극복하기 위한 담화 기반의 결합 방법

(A Discourse-based Compositional Approach to Overcome Drawbacks of Sequence-based Composition in Text Modeling via Neural Networks)

이 강 욱 † 한 상 규 †† 맹 성 현 †††
(Kangwook Lee) (Sanggyu Han) (Sung-Hyon Myaeng)

요 약 자연 언어 처리(Natural Language Processing) 분야에 심층 신경망(Deep Neural Network)이 소개된 이후, 단어, 문장 등의 의미를 나타내기 위한 분산 표상인 임베딩(Embedding)을 학습하기 위한 연구가 활발히 진행되고 있다. 임베딩 학습을 위한 방법으로는 크게 문맥 기반의 텍스트 모델링 방법과, 기학습된 임베딩을 결합하여 더 긴 텍스트의 분산 표상을 계산하고자 하는 결합 기반의 텍스트 모델링 방법이 있다. 하지만, 기존 결합 기반의 텍스트 모델링 방법은 최적 결합 단위에 대한 고찰 없이 단어를 이용하여 연구되어 왔다. 본 연구에서는 비교 실험을 통해 문서 임베딩 생성에 적합한 결합 기법과 최적 결합 단위에 대해 알아본다. 또한, 새로운 결합 방법인 담화 분석 기반의 결합 방식을 제안하고 실험을 통해 기존의 순차적 결합 기반 신경망 모델 대비 우수성을 보인다.

키워드: 심층 신경망, 텍스트 모델링, 임베딩, 순차적 결합, 담화 기반 결합

Abstract Since the introduction of Deep Neural Networks to the Natural Language Processing field, two major approaches have been considered for modeling text. One method involved learning embeddings, i.e. the distributed representations containing abstract semantics of words or sentences, with the textual context. The other strategy consisted of composing the embeddings trained by the above to get embeddings of longer texts. However, most studies of the composition methods just adopt word embeddings without consideration of the optimal embedding unit and the optimal method of composition. In this paper, we conducted experiments to analyze the optimal embedding unit and the optimal composition method for modeling longer texts, such as documents. In addition, we suggest a new discourse-based composition to overcome the limitation of the sequential composition method on composing sentence embeddings.

Keywords: deep neural networks, text modeling, embedding, sequential composition, discourse-based composition

· 본 연구는 과학기술정보통신부 및 정보통신기술연구 진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [2013-0-00179, (액소브레인-3세부) 컨텍스트 인지형 Deep-Symbolic 하이브리드 지능 원천 기술 개발 및 언어 지식 자원 구축]
· 이 논문은 2017 한국컴퓨터종합학술대회에서 '텍스트 모델링에 있어 순차적 결합 기반 신경망 모델의 한계점 극복을 위한 담화 기반의 결합 방법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : KAIST 전산학부
chaximeer@kaist.ac.kr
†† 비 회 원 : KAIST 전산학부
hsg1991@kaist.ac.kr
††† 종신회원 : KAIST 전산학과 교수(KAIST)
myaeng@kaist.ac.kr
(Corresponding author임)

논문접수 : 2017년 9월 13일
(Received 13 September 2017)
심사완료 : 2017년 11월 1일
(Accepted 1 November 2017)

정보과학회 컴퓨팅의 실제 논문지 제23권 제12호(2017. 12)

1. 서 론

심층 신경망(Deep Neural Networks)을 자연 언어 처리 분야에 적용할 때, 텍스트의 의미를 나타내기 위해 다차원 벡터 형태를 갖는 분산 표상인 임베딩(embedding)이 사용된다. 대표적인 단어 단위 임베딩 학습 방법으로는 말뭉치 내 단어 간 공기 통계를 이용하는 Word2Vec [1]이 있다. 문장 또는 문서와 같이 여러 단어들로 구성된 장문의 텍스트 모델링을 위한 방법은 크게 문맥 기반의 텍스트 모델링 방법과 결합 기반의 텍스트 모델링 방법으로 구분된다. 문맥 기반 텍스트 모델링 방법으로는 Word2Vec을 문장 또는 문서 등의 가변 길이 텍스트 임베딩 학습이 가능하도록 확장한 Paragraph Vector[2]가 대표적이다. 결합 기반의 텍스트 모델링 기법에는 기 학습된 단위 텍스트의 임베딩을 장문 텍스트의 구조적 정보를 이용하여 결합하는 재귀 신경망(Recursive Neural Networks)을 이용한 방법 또는 단위 텍스트의 등장 순서에 따라 순차적으로 단위 텍스트 임베딩을 결합하는 순환 신경망(Recurrent Neural Networks) 기반의 방법이 대표적이다. 입력으로 사용되는 텍스트의 길이가 길 때 발생하는 장기 의존(Long-term dependency) 문제를 해결하기 위한 개량 순환 신경망 모델인 LSTM(Long-Short Term Memory)[4]도 제안되었다.

그러나 기존 결합 기반 텍스트 모델링 연구에서는 결합 방식에 따른 최적의 텍스트 단위가 무엇인지에 대한 고찰 없이 단어만을 기본 단위로 사용하여 연구가 진행되어 왔다. 본 논문에서는 장문의 텍스트를 결합 방식으로 모델링하고자 할 때 최적의 모델 설정을 찾을 수 있도록, 입력 텍스트 단위와 결합 기법의 차이에 따라 성능 변화 양상을 알아보는 실험적 연구를 수행한다. 실험을 위해 Word2Vec과 Paragraph Vector를 이용하여 단어/문장 임베딩을 학습했으며, 학습된 임베딩을 결합하기 위해서는 입력 임베딩을 순차적으로 결합하는 방법 중 하나인 LSTM을 이용했다. 또한, 문맥 기반 텍스트 모델링 기법과 결합 기반 텍스트 모델링 기법에 따른 성능 차이를 비교하고자 Paragraph Vector 방법만을 이용해 학습한 문서 임베딩도 비교군으로 설정했다. 여러 가지 방법으로 얻어진 문서 임베딩의 품질 비교를 위해서는 의미 이해가 필수적으로 선행되어야 하는 과업인 문서 단위 감정 분석(Sentiment analysis)과 비꼼 탐지(Sarcasm detection)를 이용했다.

실험 결과, LSTM을 이용해 문장 임베딩을 순차적으로 결합할 경우 성능이 저하되는 현상을 발견했으며, 그로부터 문장 임베딩의 결합에는 순차적 결합 방법이 적합하지 않다는 결론을 도출했다. 문장 배열의 이해를 위해서는 문장 간 담화(Discourse)에 대한 이해가 중요하다는 점에서 착안하여 새로운 문장 임베딩의 결합 방법인 담화 기반의 결합 방법을 제안했으며, 추가적인 실험을 통해 제안 방법의 유용성을 보였다.

2. 문맥 기반/결합 기반 텍스트 모델링

2.1 Paragraph Vector를 이용한 임베딩 학습

Paragraph Vector는 Word2Vec을 다양한 길이의 텍스트의 임베딩을 학습할 수 있도록 확장한 문맥 기반 텍스트 모델링 방법의 일종이다. Paragraph Vector는 그 학습 방식에 따라 Distributed Memory (DM) 모델과 Distributed Bag-of-Words (DBOW) 모델로 구분된다.

DM 모델은 학습하고자 하는 대상 텍스트의 임베딩과 문맥 윈도우 내 다른 단어를 이용하여 대상 임베딩을 유추하는 방식으로 동작한다. 이 때, 대상 텍스트의 임베딩과 유추한 임베딩 간의 오차를 줄이는 것을 목표로 학습이 이루어진다. DBOW 모델은 학습 하고자 하는 대상 텍스트의 임베딩을 이용하여 문맥 윈도우 내에 포함된 단어를 유추하는 방식으로 동작하며, 실제 문맥 내 존재하는 단어와 유추한 단어 간의 오차를 줄이는 것을 목표로 학습이 이루어진다.

본 연구에서는 문장/문서 모델링을 위해 DM 모델을 채택했다. 또한, DM 모델 학습 과정에서 주변 문맥 단어와의 공기 통계를 이용해 임베딩을 학습하는 Word2Vec 방법의 Continuous Bag-of-Words (CBOW) 모델과 비슷한 양상으로 단어 임베딩이 함께 학습되는데, 이를 실험을 위한 단어 임베딩으로 활용하였다.

2.2 LSTM을 이용한 임베딩 결합

문서 모델링을 위한 텍스트 결합 방법으로는 LSTM을 채택했다. LSTM은 순환 신경망 모델의 셀 연산 방법을 확장한 것으로서, 학습 과정 중 발생하는 장기 의존 문제를 해결하기 위해 정보의 흐름을 조정하기 위한 게이트 메커니즘을 도입하였다. LSTM 셀은 단위 셀에서 새로운 입력을 받아들일지 말지 결정하는 입력 게이트(input gate), 셀의 메모리 상태를 지속할지 말지 결정하는 망각 게이트(forget gate), 셀에서 계산된 값을 내보낼지 말지 결정하는 출력 게이트(output gate)로 이뤄져 있다.

본 연구에서는 임베딩 단위에 따른 결합 기반 텍스트 모델링 방법의 성능 변화 양상을 살펴보기 위해, 2.1에서 얻어진 단어/문장 임베딩을 LSTM 기반 순차 결합 방식에 적용하여 문장 임베딩을 산출한 뒤 실험에 이용하였다.

2.3 실험 결과 및 분석

임베딩의 품질을 직접적으로 평가하는 것은 불가능하기 때문에, 통상의 임베딩 학습 연구에서는 자연 언어 처리 과업에 산출한 임베딩을 적용하여 품질을 간접적

으로 평가하는 방법을 채택하고 있다. 본 연구에서는 텍스트 의미 이해가 필수적인 과업인 문서 단위 감정 분석과 비꼼에 적용하여 산출한 임베딩의 성능을 측정하였다. 감정 분석을 위한 데이터셋으로는 Cornell movie review dataset [5]과 Stanford movie review dataset [6]을 사용하였고 비꼼 탐지용 데이터셋으로는 Sarcasm dataset [7]을 사용했다. Cornell 데이터셋과 Sarcasm 데이터셋은 각각 8:1:1의 비율로 학습/개발/검증 데이터셋으로 나눠 사용했으며, 10-fold 교차 검증을 적용하였다. Stanford 데이터셋은 공개된 데이터에 이미 나눠져 있는 대로 학습/검증 데이터셋을 사용하였으며, 학습 데이터셋의 10%를 개발 데이터셋으로 이용했다. 각 데이터셋에 대한 통계는 표 1과 다음과 같다.

LSTM 기반 임베딩 결합 방법을 이용해 산출된 문서 임베딩의 분류를 위해서는 소프트맥스 분류기를 사용했으며, Paragraph Vector만을 이용해 산출한 문서 임베딩의 분류를 위해서는 로지스틱 회귀 분류기를 사용했다. 실험 성능은 정확도(accuracy)로 측정됐으며, 실험 결과는 표 2와 같다.

실험 결과, 모든 데이터셋에서 단어 단위 임베딩을 LSTM으로 결합한 모델이 가장 좋은 성능을 보였다. 문맥 기반 방법만을 이용하여 전체 문서의 임베딩을 학습하는 방법에 비해 단어 단위 임베딩을 LSTM으로 결합한 모델이 좋은 성능을 보인 것은 단어와 같이 세분화된 단위 텍스트의 임베딩을 결합하는 것이 문서의 내용 추상화에 보다 유리하다는 것으로 해석 가능하다.

허나, 문장 임베딩을 LSTM으로 결합한 방식은 극히 안좋은 성능을 보였는데, 이는 문장 간의 의미는 등장 순서에 따라 순차적으로 형성되는 것이 아닌 별도의 담화 구조를 이용해 형성되기 때문이라고 분석했다.

3장에서는 순차적으로 문장 임베딩을 결합할 때 나타나는 한계점을 극복하기 위한 방법으로 문서 내의 담화 구조를 활용하여 문장 임베딩을 결합하는 새로운 방법을 제안한다.

표 1 실험 데이터셋 통계
Table 1 Statistics of experimental datasets

	# documents	# words	# sentences
Cornell	2,000	1,402,412	162,482
Stanford	50,000	13,229,985	600,920
Sarcasm	1,254	341,034	16,383

표 2 검증 데이터셋에서 측정된 정확도
Table 2 Accuracies on test sets

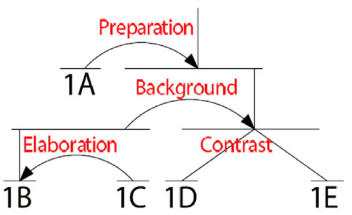
	Cornell	Stanford	Sarcasm
LSTM (Input: word embedding)	0.791	0.886	0.750
LSTM (Input: sentence embedding)	0.505	0.589	0.660
Document embedding via end-to-end Paragraph Vector	0.753	0.825	0.739

3. 담화 기반 문장 임베딩 결합 방법

3.1 Rhetorical Structure Theory 기반 담화 분석

문서 내의 문장들 간 담화 구조 분석을 위해 RST (Rhetorical Structure Theory)[8]를 이용했다. RST는 작은 담화 단위(Discourse Unit) 텍스트가 결합하여 더 큰 담화 단위 텍스트를 형성해 나간다고 가정하고 담화 관계(Discourse Relation)를 분석하는 이론으로써, 두 담화 단위 간의 담화 관계를 재귀적으로 유추하여 트리 형태의 담화 분석 결과를 산출한다. 그림 1은 RST를 이용하여 텍스트에 담화 분석을 수행한 예제이다. 아래 예제에서 담화 단위 1B와 1C는 부연설명(Elaboration) 관계를 형성하고 있으며, 이 때 화살표는 담화 관계 내에서의 수식 순서, 즉, 주(Nucleus)/부(Satellite) 역할을 나타낸다. 아래 예제에서는 1C가 1B를 수식하고 있으므로 1B가 주, 1C가 부 역할을 수행하고 있다고 말할 수 있다.

현재까지의 연구에서는 문장 단위의 담화 분석 결과만을 이용해왔으나[9,10], 본 연구에서는 문서에 대한 담화 분석 결과를 이용했다. 본 연구의 담화 분석을 위해서는 RST 기반 DPLP (Discourse Parsing from Linear Projection) 분석기[11]를 이용했으며, 편의상 최소 담화 단위(Elementary Discourse Unit)를 문장으로 설정하고 담화 단위 간 결합 관계인 담화 구조(Discourse structure)만을 이용했다.



[Lactose and Lactase.]1A
[Lactose is milk sugar;]1B
[the enzyme lactase breaks it down.]1C
[For want of lactase most adults cannot digest milk.]1D
[In populations that drink milk the adults have more lactase, perhaps through natural selection.]1E

그림 1 RST 기반 담화 분석 예제
Fig. 1 An example of RST-based discourse analysis

3.2 재귀 신경망을 통한 문서 모델링

3.1에서 얻어진 담화 구조를 활용하기 위해, 외부의 구조적 정보를 이용하여 신경망을 구성할 수 있는 재귀 신경망을 사용하였다. 재귀 신경망에서 자식 노드 임베딩들의 결합을 통해 얻어지는 부모 노드의 임베딩을 수식으로 나타내면 다음과 같다.

$$h_p = \tanh(W \cdot [h_{c1}, h_{c2}] + b) \quad (1)$$

이 때, $[h_{c1}, h_{c2}]$은 자식 노드 c_1과 c_2의 임베딩을 연결(Concatenation)한 벡터를 뜻하며, W와 b는 각각 가중치 행렬과 편향 벡터를 나타낸다.

입력에 해당하는 말단 노드(Leaf node)의 임베딩은 2.1장에서 학습한 문장 임베딩을 이용했으며, 문서 임베딩을 나타내는 최상위 노드(Root node)의 임베딩을 소프트 맥스 분류기의 입력으로 사용하여 예측한 문서의 레이블과 실제 레이블의 차이를 줄이는 방향으로 학습을 진행하였다.

3.3 실험 결과 및 분석

제안된 담화 구조 기반의 재귀 신경망을 통해 학습된 임베딩의 성능 평가를 위해 2.3장의 실험과 동일한 환경에서 실험을 진행했다. 실험 결과는 표 3과 같다.

결과에서 확인할 수 있듯이, 제안한 담화 기반 방법으로 문장 임베딩을 결합하여 문서 임베딩을 산출 했을 때 LSTM을 이용해 문장 임베딩을 결합했을 때에 비해 월등히 좋은 성능을 보였다.

특히 감성 분석을 위한 Cornell 데이터셋과 비꼼 탐지를 위한 Sarcasm 데이터셋에서는 비교 대상 가운데 최고의 성능을 보였다. 이는 문장 임베딩 결합 시 담화 기반의 결합 방법이 순차적 결합보다 효과적인 것으로 해석될 수 있다. Stanford 데이터셋에서는 제안한 방법이 LSTM을 이용한 단어 임베딩 결합 방법에 비해 낮은 성능을 보였다. 그러나, LSTM을 이용한 문장 임베딩 결합 방법에 비해서는 성능이 크게 향상된 것을 확인할 수 있다. 이러한 결과로부터 문장 임베딩의 경우에는 담화 기반 결합 방법이 보다 적합하다는 결론을 내릴 수 있다.

표 3 담화 기반 결합 방법을 포함한 검증 데이터셋에서 측정된 정확도(편의를 위해 제안 방법 외 실험 결과를 표 2에서 가져옴)

Table 3 Accuracies, including the proposed discourse-based composition, on test sets (Other results except the one from the proposed method are from Table 2 for the reader's convenience)

	Cornell	Stanford	Sarcasm
LSTM (Input: word embedding)	0.791	0.886	0.750
LSTM (Input: sentence embedding)	0.505	0.589	0.660
Document embedding via end-to-end Paragraph Vector	0.753	0.825	0.739
Proposed method (Input: sentence embedding)	0.828	0.856	0.754

4. 결 론

본 연구에서는 문서와 같은 장문의 텍스트 모델링에 적합한 모델링 기법과 임베딩 단위가 무엇인지 알아보기 위해 여러 가지 조합의 비교군을 구현한 뒤 텍스트의 함축적 의미를 파악하는 것이 중요한 문서 단위 감성 분석과 비꼼 탐지에 적용하여 성능 평가를 진행하였다. 실험 결과, 문맥 기반 학습 방법만을 이용하여 문서 전체 임베딩을 한꺼번에 학습하는 것에 비해 적합한 단위 텍스트 임베딩 및 결합 방법을 활용하는 것이 문서 임베딩 학습에 더 유리함을 알아볼 수 있었다.

또한, 문장 임베딩을 이용한 LSTM의 성능이 낮은 것으로부터 문장 임베딩의 결합에는 통상 LSTM의 순차적 결합 방식이 적합하지 않다는 결론을 도출했으며, 문장 임베딩의 결합을 위한 새로운 방법으로써 담화 기반의 결합 방식을 제안하여 순차적 결합 방법에 비해 성능이 크게 향상됨을 보였다.

추후에는 재귀 신경망에 내제되어 있는 장기 의존 문제를 해결하여 더욱 정교한 문서 임베딩을 산출할 수 있도록 트리 LSTM(tree LSTM) 등 새로운 트리 구조 신경망 모델을 적용하기 위한 방법에 대해 연구할 예정이다.

References

[1] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781. 2013.

[2] Q.V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *Proc. of the International Conference on Machine Learning*, pp. 1188-1196, 2014.

[3] C. Goller and A. Kuchler, "Learning task-dependent distributed representations by backpropagation through structure," *Proc. of the International Conference on Neural Networks*, pp. 347-352, 1996.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, Vol. 9, pp. 1735-1780, 1997.

[5] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pp. 271-278, 2004.

[6] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng and C. Potts, "Learning word vectors for sentiment analysis," *Proc. of the Annual Meeting of*

the Association for Computational Linguistics: Human Language Technologies, pp. 142-150, 2011.
[7] E. Filatova, "Irony and sarcasm: Corpus generation and analysis using crowdsourcing," *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC)*, pp. 392-398, 2012.
[8] W.C. Mann and S.A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary," *Journal for the Study of Discourse*, Vol. 8, pp. 243-281, 1988.
[9] P. Bhatia, Y. Ji and J. Eisenstein, "Better Document-level Sentiment Analysis from RST Discourse Parsing," *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 2212-2218, 2015.
[10] X. Fu, W. Liu, Y. Xu, C. Yu and T. Wang, "Long Short-term Memory Network over Rhetorical Structure Theory for Sentence-level Sentiment Analysis," *Proc. of Asian Conference on Machine Learning*, pp. 17-32, 2016.
[11] Y. Ji and J. Eisenstein, "Representation Learning for Text-level Discourse Parsing," *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pp. 13-24, 2014.

이 강 욱

2010년 한동대학교 전산전자공학부 졸업(학사). 2012년 KAIST 전산학부 졸업(석사). 2012년~현재 KAIST 전산학부 박사과정. 관심분야는 자연어처리, 정보검색, HCI

한 상 규

2015년 KAIST 전산학부 졸업(학사). 2017년 KAIST 전산학부 졸업(석사). 관심분야는 자연어처리, 머신러닝, 딥러닝

맹 성 현

1985년 미국 Southern Methodist University (SMU) 석사. 1987년 미국 Southern Methodist Univeristy (SMU) 박사 1987년~1988년 미국 Temple University 교수. 1988년~1994년 미국 Syracuse University 교수(tenured). 1994년~2003년 충남대학교 컴퓨터과학과 교수. 2003년~2009년 한국정보통신대학교 교수. 2009년~현재 KAIST 교수. 관심분야는 정보 검색, 텍스트 마이닝, 웹사이언스, 상황인지 컴퓨팅 등