

Ontofitting: 의미 표현을 위한 벡터 조정

오진영^o, 차정원
창원대학교

psycheojy@cwnu.ac.kr, jcha@cwnu.ac.kr

Ontofitting: Specialization of Word Vectors for Semantic Representation

Jinyoung Oh^o, Jeong-Won Cha
Changwon National University

요 약

우리는 단어 임베딩에 외부지식을 내재할 수 있는 Ontofitting 방법을 제안한다. 이 방법은 retrofitting의 한 방법으로 유의어, 반의어, 상위어, 하위어 정보를 단어 임베딩에 내재할 수 있다. 유의어와 반의어 정보를 내재하기 위해서 벡터의 각 유사도를 사용하였고 상하위어 정보를 내재하기 위해서 벡터의 길이 정보를 사용하였다. 유의어 사이에는 작은 각도를 가지고 반의어 사이에는 큰 각도를 가지게 된다. 하위어는 상위어보다 상대적으로 작은 길이를 가지게 된다. SimLex와 HyperLex로 실험하여 효과와 안정성을 검증하였다. 의미정보를 내재한 임베딩을 사용할 수 있다면 QA, 대화 등 응용에서 보다 좋은 성능을 보일 수 있을 것이다.

주제어: 단어 임베딩, Representation Learning, Retrofitting, 기계학습

1. 서론

단어의 표현을 학습하는 분야는 자연어 처리의 핵심 분야로 자리 매김되고 있다. 구문분석[1], 기계번역[2]과 다른 많은 응용[3,4]에서 단어 표현을 이용하고 있다. 단어 표현의 일반적인 방법은 분산표현(distributional representation)이라고 알려진 방법이다[5]. 이 방법은 대용량 문서에서 공기 정보를 이용하여 단어의 의미 표현을 학습하는 방법이다[6-9].

분산표현법의 단점은 크게 2 가지이다. 하나는 의미적 유사성(Semantic similarity)과 개념적 관련성(Conceptual association)이 같은 공간에 표현된다는 것이다. 예를 들어 '방위'를 나타내는 '동쪽'과 '서쪽'은 서로 의미적으로는 반의어이지만 개념적 관련성이 높아 유사 벡터를 가지게 된다. 다른 하나는 코퍼스로부터 학습되기 때문에 벡터의 표현이 도메인에 의존적이다. 이것을 해결하는 방법은 비지도학습으로 학습된 벡터를 외부 지식으로 수정하는 것이다. 영어권에서는 WordNet과 같은 지식을 이용하여 후처리(post-processing)하는 방법이라고 불린다[10].

이러한 방법은 단어들 사이의 특수한 관계를 표현하기 위해서 벡터를 수정하는 것이다. 이전 연구에서는 단어들의 유의어와 반의어 관계, 혹은 상하위어 관계만을 위해서 벡터를 수정했다. 본 논문에서는 두 가지 관계를 모두 수정하는 방법을 제안한다.

단어 단위의 유의어와 반의어 관계는 서로 대칭관계이다. 그렇지만 상하위어 관계를 비대칭관계이다[29, 30]. 비대칭관계의 핵심은 단어들 간의 의미구조를 결정하는 것이다[13]. 이와 같은 작업은 분류기준 자동 생성, 자연어 추론, 텍스트 생성, 은유 인식(metaphor detection) 등에 사용될 수 있다.

본 논문에서는 유의어-반의어 관계를 벡터내에서 표현하기 위해서 각 유사도를 사용하고 상하위 관계를 표현

하기 위해서 벡터의 길이를 사용한다. 그림1은 제안 개념을 설명한다. [26]이 제안했던 Attract-Repel 개념을 수정하여 두 관계를 모두 반영하는 방법을 제안한다.

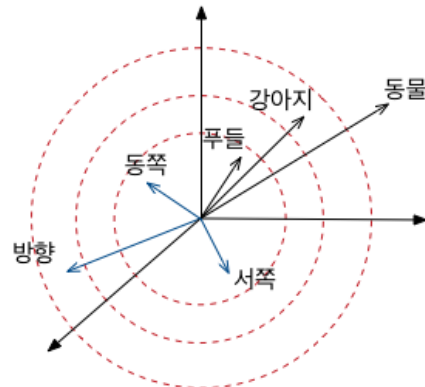


그림 1. Ontofitting의 개념. Ontofitting은 유의어 사이는 작은 각을 유지하고 반의어 사이는 큰 각을 유지한다. 그리고 상위어는 벡터의 크기가 크고 하위어는 벡터의 크기가 작다.

2. 제안 방법

각 단어의 벡터를 $V = \{v_1, v_2, \dots, v_N\}$ 라고 정의하자. 우리는 여기에 유의어-반의어, 상하위어의 의미 개념을 반영한다. 반영되어 수정된 벡터를 $V' = \{v'_1, v'_2, \dots, v'_N\}$ 라고 하자. 또한 유의어 집합을 S , 반의어 집합을 A 그리고 상하위어 집합을 H 라고 하자. 각 집합의 원소들은 단어들의 쌍으로 이루어져 있다. 예를 들어 유의어와 반의어의 각 (i, j) 쌍은 단어 집합에서 i 번째 단어와 j 번째

단어를 나타내고 이들은 각각 유의어와 반의어 관계에 해당한다. 또한 상하위어 집합의 각 (i, j) 는 i 번째 단어가 j 번째 단어의 하위어에 해당한다. 기 학습된 벡터를 수정하기 위한 목적함수는 세 가지 집합에 대해서 각각 다음과 같다.

1. 유의어 관계 수정: Ontofitting에서는 유의어 관계에 있는 단어 벡터들을 서로 가깝게 하기 위해 각도를 조절한다.

$$SA(V) = \sum_{(u,w) \in S} \tau(d(v'_u, v'_w) - \gamma) \quad (1)$$

여기서 d 는 각 벡터간의 라디안 각도를 의미하고 γ 는 이상적인 값인 0을 의미한다. 또한 τ 는 $\max(0, x)$ 를 의미한다.

2. 반의어 관계 수정: 반의어 관계에 있는 단어들의 각도를 멀리 하게 하기 위해서 각도를 조절한다.

$$AR(V) = \sum_{(u,w) \in A} \tau(\delta - d(v'_u, v'_w)) \quad (2)$$

여기서 δ 는 두 벡터가 정반대를 가리키는 값이다. 본 논문에서는 π 값을 가진다.

3. 상하위어 관계 수정: 상위어와 하위어 관계를 벡터에서 수정하기 위해서 두 벡터의 길이를 조절한다.

$$HL(V) = \sum_{(u,w) \in H} D(v'_u, v'_w) \quad (3)$$

여기서 D 는 두 벡터의 L2 norm의 차이를 구하는 것이다. 우리는 다음과 같이 2개의 다른 식을 사용하여 실험을 진행하였다.

$$D_1(v_x, v_y) = \frac{|v_x| - |v_y|}{|v_x| + |v_y|} \quad (4)$$

$$D_2(v_x, v_y) = \frac{|v_x| - |v_y|}{\max(|v_x|, |v_y|)} \quad (5)$$

수정된 벡터에서 유사도 평가는 식(6)을 사용하였다.

$$S(v_x, v_y) = \alpha * SA(v_x, v_y) + \beta * AR(v_x, v_y) \quad (6)$$

상하위 관계는 식(4), 식(5), 식(7)에 대해서 각각 실험하였다.

$$L(v_x, v_y) = d(v_x, v_y) + D(v_x, v_y) \quad (7)$$

식(7)은 대칭관계와 비대칭관계를 조합하여 사용한 식이다. 이 식에서 $d(v_x, v_y)$ 는 라디안 각을 계산하는 대칭관계식이며 $D(v_x, v_y)$ 는 두 벡터의 norm 차이를 계산하는 비대칭관계식이다. 실험에서 두 값을 함께 사용하는 것

이 결과에 도움을 주는 것을 보인다.

3. 실험

3.1. 실험 설정

벡터 수정의 안정성을 실험하기 위해서 잘 알려진 공개 영어 벡터를 사용하였다. 기본 벡터는 GLOVE Common Crawl[8]을 사용하였다.

벡터 수정을 위한 외부 지식으로는 다음과 같이 3가지 종류를 사용하였다. PPDB[28], WordNet 상하위어,¹⁾ 동의어, 반의어이다. 동의어와 반의어는 [11, 12]에서 만든 데이터를 사용하였다. 모두 31,850개의 동의어 쌍을 사용하였고 12,948개의 반의어 쌍을 사용하였다. 112,942개의 상하위어 쌍을 사용하였다.

실험을 위해서 사용한 파라미터 값은 각각 다음과 같다. $\alpha = 0.1$, $\beta = 0.1$, $\gamma = 0.0$, $\delta = \pi$ 를 사용하였다. 동의어, 반의어 벡터값 수정을 위해서 100번의 반복을 하였고 상하위 벡터값 수정을 위해서 50번의 반복을 하였다.

3.2. 실험 결과

유의어와 반의어의 성능을 보기 위해서 SimLex-999 데이터를 사용하였다[15]. 상하위어의 성능을 확인하기 위해서 HyperLex 데이터를 사용하였다[16]. 초기 단어 벡터로는 [8]에서 구축한 데이터를 사용하였다. 제안한 방법의 효용성을 검증하기 위해서 SimLex-999와 HyperLex에 대해서 Spearman's rank coefficient를 사용하였다.

표 1에 두 데이터 셋에 대한 성능을 비교 표시하였다. retrofitting, counter-fitting은 모두 초기 Glove보다 성능이 우수하지만 상하위 관계를 담고 있지 않다. [27]은 동의어, 반의어 그리고 상하위어를 담고 있지만 상관관계에 의해서 동의어, 하위어의 성능이 counter-fitting에 비해 하락한다고 보고하고 있다.

표 1. SimLex-999와 HyperLex에 대한 실험 결과. 상하위 관계 평가는 식(5)를 사용하였다.

	SimLex	HyperLex
Glove [8]	0.418	0.166
Glove + retrofitting	0.53	-
Glove + counter-fitting	0.58	-
LEAR[27]	-	0.686
Glove + Ontofitting	0.575	0.495

제안 시스템은 동의어와 반의어는 라디안 각으로 유사도를 계산하고 상하위어는 벡터는 크기로 계산하기 때문

1) WordNet 상하위어는 [31]에서 만든 데이터를 받아서 사용하였다.

에 이 둘이 의존관계가 없다. 따라서 상하위 관계를 계산하기 위해 벡터의 크기를 변경해도 동의어, 반의어에 영향을 주지 않는다.

그렇지만 결과를 분석해보면 제안 방법도 학습되지 않은 데이터에 대해서는 성능이 떨어지는 경향을 보였다. 이 부분은 보완을 해야 한다.

표 2. 다양한 평가식에 따른 SimLex-999와 HyperLex에 대한 실험 결과.

평가식			SimLex	HyperLex
SimLex	HyperLex	$D(v_x, v_y)$		
식(6)	식(4)		0.575	0.492
식(6)	식(5)		0.575	0.495
식(6)	식(7)	식(4)	0.575	0.462
식(6)	식(7)	식(5)	0.575	0.460

4. 관련 연구

외부 지식을 이용하여 벡터를 수정하는 방법은 기본적으로 유의어는 벡터를 가깝게 만들고 반의어는 벡터를 멀게 하는 방법을 이용한다. 초기의 연구는 학습 시에 외부 데이터를 이용하여 벡터를 수정하는 방법을 제안하였다[15-18]. 다른 방법은 학습 후에 벡터를 수정하는 방법을 사용한다. 이 방법은 retrofitting이라는 이름으로 잘 알려져 있다[19-24].

하지만 이 연구들은 동의어, 반의어 관계만을 이용하여 벡터를 수정하였다. [27]은 동의어, 반의어 관계와 상하위어를 동시에 사용하여 수정하는 방법을 사용하였다. 그런데 동의어와 반의어를 코사인 유사도를 사용하였기 때문에 상하위어 관계를 수정하면 동의어와 반의어도 수정이 되어 성능이 저하될 가능성이 있다.

본 연구는 동의어와 반의어를 이용하는 수정에 각도를 이용하여 상하위어를 변경하여도 영향을 받지 않는다.

5. 결론

본 논문에서는 기존에 존재하는 벡터에 유의어-반의어 관계와, 상하위어 관계 지식을 적용하여 벡터를 수정하는 방법을 제안하였다. 두 개의 상이한 관계를 표현하기 위해 각도와 길이를 이용하는 목적함수를 설계하였다. Ontofitting에 의해서 수정된 벡터를 평가하기 위하여 영어권에서 사용하는 SimLex와 HyperLex를 평가하여 경쟁력 있는 성능을 얻었다.

우리는 현재 개발된 방법을 다양한 언어에 적용할 예정이다. 다른 비대칭 관계(예를 들어 meronymy, holonymy 등)들에 대해서도 적용할 예정이다. 또한 언어뿐만 아니라 multi-modal을 동일한 의미 공간에 표현하는 연구도 진행할 예정이다.

Acknowledgement

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임

(No.2017R1D1A1B03033534).

참고문헌

- [1] Danqi Chen, Christopher D. Manning, "A fast and accurate dependency parser using neural networks", In Proceedings of EMNLP, pages 740-750, 2014
- [2] Will Y. Zou, Richard Socher, Daniel Cer, Christopher D. Manning, "Bilingual word embeddings for phrase-based machine translation", In Proceedings of EMNLP, pages 1393-1398, 2013
- [3] Joseph P. Turian, Lev-Arie Ratinov, Yoshua Bengio, "Word representations: A simple and general method for semi-supervised learning", In Proceedings of ACL, pages 384-394, 2010
- [4] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel P. Kuksa, "Natural language processing (almost) from scratch", Journal of Machine Learning Research, 12:2493-2537, 2011
- [5] Zellig S. Harris, "Distributional structure", Word, 10(23):146-162, 1954
- [6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, Jeffrey Dean, "Distributed representations of words and phrases and their compositionality", In Proceedings of NIPS, pages 3111-3119, 2013
- [7] Omer Levy, Yoav Goldberg, "Dependency-based word embeddings", In Proceedings of ACL, pages 302-308, 2014
- [8] Jeffrey Pennington, Richard Socher, Christopher Manning, "Glove: Global vectors for word representation", In Proceedings of EMNLP, pages 1532-1543, 2014
- [9] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, "Enriching word vectors with subword information", Transactions of the ACL, 5:135-146, 2017
- [10] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, Steve Young, "Counter-fitting word vectors to linguistic constraints", In Proceedings of NAACL-HLT, 2016
- [11] Jingwei Zhang, Jeremy Salwen, Michael Glass, Alfio Gliozzo, "Word semantic representations using bayesian probabilistic tensor factorization", In Proceedings of EMNLP, pages 1522-1531, 2014
- [12] Masataka Ono, Makoto Miwa, Yutaka Sasaki, "Word embedding-based antonym detection using thesauri and distributional information", In

- Proceedings of NAACL-HLT, pages 984-989, 2015
- [13] Christiane Fellbaum, WordNet, 1998
- [14] Barbara Ann Kipfer, “Roget’s 21st Century Thesaurus (3rd Edition)”, Philip Lief Group, 2009
- [15] Felix Hill, Roi Reichart, Anna Korhonen, “Simlex-999: Evaluating semantic models with (genuine) similarity estimation”, Computational Linguistics, 41(4):665-695, 2015
- [16] Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, Anna Korhonen, “Hyperlex: A large-scale evaluation of graded lexical entailment”, Computational Linguistics, 2017
- [17] Mo Yu, Mark Dredze, “Improving lexical embeddings with semantic knowledge”, In Proceedings of ACL, pages 545-550, 2014
- [18] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, Tie-Yan Liu, “RC-NET: A general framework for incorporating knowledge into word representations”, In Proceedings of CIKM, pages 1219-1228, 2014
- [19] Jiang Bian, Bin Gao, Tie-Yan Liu, “Knowledge-powered deep learning for word embedding”, In Proceedings of ECML-PKDD, pages 132-148, 2014
- [20] Douwe Kiela, Felix Hill, Stephen Clark, “Specializing word embeddings for similarity or relatedness”, In Proceedings of EMNLP, pages 2044-2048, 2015
- [21] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, Noah A. Smith, “Retrofitting word vectors to semantic lexicons”, In Proceedings of NAACL-HLT, pages 1606-1615, 2015
- [22] Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, “Ontologically grounded multi-sense representation learning for semantic vector space models”, In Proceedings of NAACL, pages 683-693, 2015
- [23] John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, “From paraphrase database to compositional paraphrase model and back”, Transactions of the ACL, 3:345-358, 2015
- [24] Kim Anh Nguyen, Sabine Schulte im Walde, Ngoc Thang Vu, “Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction”, In Proceedings of ACL, pages 454-459, 2016
- [25] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, Steve Young, “Counter-fitting word vectors to linguistic constraints”, In Proceedings of NAACL-HLT, 2016
- [26] Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young, “Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints”, Transactions of the ACL, 5:309-324, 2017
- [27] Ivan Vulić, Nikola Mrkšić, “Specialising Word Vectors for Lexical Entailment”, Proceedings of NAACL-HLT 2018, pages 1134 - 1145, 2018
- [28] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Benjamin Van Durme, Chris Callison-Burch, “PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification”, In Proceedings of ACL, 2015
- [29] Allan M. Collins, Ross M. Quillian, “Experiments on semantic memory and language comprehension. Cognition in Learning and Memory”, 1972
- [30] Richard Beckwith, Christiane Fellbaum, Derek Gross, George A. Miller, “WordNet: A lexical database organized on psycholinguistic principles”, Lexical acquisition: Exploiting on-line resources to build a lexicon, pages 211 - 231, 1991
- [31] George A. Miller, “WordNet: A Lexical Database for English”, Communications of the ACM, 1995