Week 2 [Subword]

1. **Enriching Word Vectors with Subword Information**

   Submitted on 15 Jul 2016 (v1), last revised 19 Jun 2017 (this version, v2)]

   NLP1607.04606v2.pdf

   ✍🏻 Review

Abstract

이 연구 논문은 서브워드(subword) 정보를 활용하여 단어 벡터를 강화하는 새로운 접근 방식을 제안합니다. 기존의 단어 표현 모델은 단어의 형태론적 구조를 무시해 왔으나, 이 방법은 각 단어를 문자 n-그램의 집합으로 표현하여 드문 단어에 대해서도 신뢰할 수 있는 벡터 표현을 구축할 수 있게 해줍니다. 빠른 모델 훈련이 가능하며, 데이터셋에 없는 단어에 대해서도 표현을 생성할 수 있는 장점을 지니고 있습니다. 본 논문을 통해 중요한 언어 처리 작업에서의 성능 향상을 확인할 수 있으며, 이는 형태론적으로 풍부한 언어에서 단어 표현을 개선하는 데 유익합니다.

✍️핵심 요약

1. 서브워드 정보를 활용한 **단어 벡터** 향상이 자연어 처리 성능을 높인다.
    - **서브워드** 단위로 단어를 표현함으로써 희귀 단어에 대한 신뢰성을 높이고, 훈련 데이터에 없는 단어에도 유연하게 대처할 수 있는 이점이 있다.
    - 이 접근법은 기존의 단어 벡터 모델보다 다양한 언어에서 뛰어난 성능을 발휘하는 것으로 확인되었다.
2. 형태소 기반 단어 표현 모델은 다양한 언어와 데이터셋에서 **우수한 성능**을 보인다.
    - 형태소 정보를 포함한 모델은 문법적 요소와 의미적 유사성을 효과적으로 학습하여 단어 유사성 및 유추 작업에서 인상적인 결과를 도출하였다.
    - 이 모델은 특정 언어에서 특히 강한 효과를 나타내며, 드물거나 복잡한 단어 처리에 효과적이다.
3. **훈련 데이터의 크기**가 모델 성능에 미치는 영향이 크다.
    - 더 많은 훈련 샘플을 사용할수록 드문 단어의 표현이 강화되며, 이는 모델의 전반적인 성능을 향상시키는 결과를 초래한다.
    - 모델은 훈련 세트에 없는 단어에도 효과적인 벡터를 생성할 수 있는 능력을 입증하였다.
4. 문자 **n-그램**의 크기는 단어 벡터의 성능에 중대한 영향을 미친다.
    - 3~6자로 구성된 n-그램이 영어와 독일어에서 성능을 극대화하는 것으로 나타났으며, 긴 n-그램은 복합 명사와 같은 구조적 특징을 잘 캡처한다.
    - n-그램 크기의 적절한 선택은 특히 의미적 유추 작업에서의 효과를 잘 보여준다.
5. OOV(Out-Of-Vocabulary) 단어에 대한 효과적인 처리 방법을 제안한다.
    - 훈련되지 않은 단어에 대해 n-그램의 벡터 평균을 사용하여 의미 있는 표현을 할당할 수 있다.
    - 이 접근법은 기존 방법보다 OOV 단어에서 훨씬 더 나은 성능을 나타내며, 다양한 언어에서 적용 가능성을 제시한다.

Contents

1. 🌟 서브워드 정보를 활용한 단어 벡터 향상

- Facebook AI Research 소속에서 연구가 진행되었다.
- 이 접근법은 단어의 **형태론적 구조**를 반영하여 벡터를 강화한다.
- 각 단어를 문자 n-그램의 집합으로 표현하여 드문 단어도 신뢰할 수 있는 **벡터**를 생성한다.
- 데이터셋에 없는 단어에도 적용 가능한 **표현 생성**의 장점을 제공한다.

2. 연속적인 단어 표현 학습 방법 제안

- 최근 대규모 비주석 말뭉치를 기반으로 한 연속 단어 표현은 자연어 처리에서 많은 작업에 유용하나, 기존의 모델은 단어의 형태소를 무시하고 각 단어에 대해 별도의 벡터를 할당하여 큰 어휘와 희귀 단어가 많은 언어에서는 제한적이다.
- 이 논문에서는 각 단어를 문자 n-그램의 집합으로 표현하는 새로운 스킵그램 모델 기반의 접근 방식을 제안하며, 각 문자 n-그램에 벡터 표현을 연관짓고 이를 통해 단어의 표현을 이들 벡터의 합으로 구성한다.
- 제안된 방법은 빠르게 대규모 말뭉치에서 모델을 훈련할 수 있도록 하며, 훈련 데이터에 나타나지 않은 단어의 표현도 계산할 수 있다.
- 아홉 개 언어에서 단어 유사성 및 유추 작업을 통해 우리 단어 표현의 성능을 평가하며, 최근 제안된 형태소 기반 단어 표현과 비교하여 뛰어난 성과를 달성했다.

3. 형태소 기반 단어 표현 연구

- 최근 몇 년 동안 형태소 정보를 통합한 단어 표현 방법이 여러 가지 제안되었다.
- Alexandrescu와 Kirchhoff(2006)는 희소 단어를 더 잘 모델링하기 위해 단어를 특징 집합으로 표현하는 **팩터화 신경 언어 모델**을 도입하였다.
- 이 방법은 터키어와 같은 형태소가 풍부한 언어에 성공적으로 적용되었으며, 여러 연구들이 형태소로부터 단어 표현을 도출하는 다양한 조합 함수를 제안하였다.
- 그 외에도, 자연어 처리에서 문자 수준 모델들이 단어 분할을 배제하고 문자로부터 직접 언어 표현을 학습하는 방식으로 여러 방향으로 연구되고 있다.
- 다양한 모델들이 음절 태깅, 텍스트 정규화, 감정 분석 등 다양한 작업에 적용되었으며, 최근 기계 번역 분야에서도 희소 단어 표현을 얻기 위해 **부분 단위**를 사용하는 방법이 제안되었다.

4. 모델

- 이 섹션에서는 형태소를 고려한 단어 표현 학습 모델을 제안합니다.
- 형태소는 서브워드 단위로 모델링하며, 단어는 문자 n-그램의 합으로 표현됩니다.
- 단어 벡터를 학습하기 위한 일반적인 프레임워크를 먼저 소개하고, 그 후 서브워드 모델을 설명합니다.
- 형태소 모델은 각 단어를 문자 n-그램의 집합으로 나타내어, 불규칙한 단어의 특성을 학습하도록 합니다.

5. 실험 설정 및 최적화 방법

- 대부분의 실험에서 우리는 **word2vec** 패키지의 스킵그램 및 CBOW 모델과 비교한다.
- 최적화 문제를 해결하기 위해 **확률적 경량 하강법**을 사용한다.
- 단어 벡터의 차원은 300이며, 각 긍정 예제에 대해 5개의 부정 예제를 샘플링한다.
- 데이터 세트는 아랍어, 체코어, 독일어, 영어, 스페인어, 프랑스어, 이탈리아어, 루마니아어, 러시아어로 구성된 Wikipedia에서 훈련한다.
- 모델은 C++로 구현되었으며, 공개적으로 제공된다.

6. 모델 평가 결과: 다섯 가지 실험

- 우리는 모델을 다섯 가지 실험을 통해 평가하며, 여기에 **단어 유사성**과 **단어 유추 평가**, 최신 방법과의 비교, 훈련 데이터의 크기 및 고려하는 문자 n-그램 크기의 효과 분석이 포함된다.
- 첫 번째로, 인간의 유사성 판단을 평가하며, **스피어만 순위 상관 계수**를 사용해 인간 판단과 벡터 표현의 코사인 유사성 간의 상관관계를 계산한다.
- 독일어에 대해서는 GUR65, GUR350, ZG222 세 가지 데이터셋을 비교하고, 영어는 WS353 데이터셋과 **희귀 단어 데이터셋**을 사용한다.
- 프랑스어 단어 벡터는 RG65 데이터셋에서 평가하며, 스페인어, 아랍어, 루마니아어 단어 벡터는 관련 데이터셋으로 평가된다.
- 러시아어 단어 벡터는 HJ 데이터셋을 사용하여 평가하며, 모든 데이터셋에 대한 결과를 테이블 1에 보고한다.

7. 단어 유사성 데이터셋에서 인간 판단과 유사성 점수의 상관관계

- 표 1을 보면, 제공된 모델(sisg)이 서브워드 정보를 사용하여 모든 데이터셋에서 기존 기준선보다 뛰어난 성능을 보이는 것을 확인할 수 있다.
- 특히, 비어 있는 단어에 대해 벡터를 계산하는 것이 항상 null 벡터를 사용하는 것보다 좋다는 점이 **서브워드 정보**의 장점을 입증한다.
- 또한, 아랍어, 독일어, 러시아어에서 문자 n-그램을 사용하는 효과가 영어, 프랑스어, 스페인어보다 더 중요함을 관찰할 수 있다.
- 독일어와 러시아어는 각각 4개의 경우와 6개의 경우를 가진 문법적 **격변화**를 가지며, 독일어의 복합어도 특징적이다.
- 마지막으로, 영어 희귀 단어 데이터셋(RW)에서는 서브워드 방식이 더 좋은 성능을 발휘하는 것을 확인할 수 있다.

8. 모델의 단어 유추 작업 정확도 비교

- 표 2에서는 체코어, 독일어, 영어 및 이탈리아어에 대한 단어 유추 작업에서 우리의 모델과 기준선의 정확도를 보고한다.
- 센터닉과 구문적 유추는 각각 별도로 보고하며, 영어 WS353 데이터셋에서는 성과가 좋지 않다.
- 영어 WS353 데이터셋의 단어들은 일반적인 단어로, 하위 단어 정보를 활용하지 않고도 좋은 벡터를 얻을 수 있기 때문이다.
- 덜 흔한 단어를 평가할 때, 단어 간 문자 수준의 유사성을 활용하는 것이 좋은 단어 벡터 학습에 도움이 된다는 것을 알 수 있다.
- 참고로, 형태학적 정보는 구문적 작업을 크게 개선하지만, 의미적 질문에는 도움이 되지 않으며, 독일어와 이탈리아어의 성능을 저하시킨다.

9. 단어 벡터와 형태소 표현 비교

- 우리는 단어 유사성 작업에 대한 형태소 정보를 포함한 이전 연구의 방법들과 우리의 접근 방식을 비교한다.
- 비교하는 방법으로는 Luong et al.(2013)의 재귀 신경망, Qiu et al.(2014)의 형태소 cbow, Soricut과 Och(2015)의 형태소 변환 방법을 사용했다.

- 비교의 공정성을 위해, 동일한 데이터 세트인 Shaoul과 Westbury(2010)가 발표한 영어 위키백과 데이터와 2013 WMT 공통 과제의 뉴스 크롤링 데이터를 사용하여 모델을 훈련시켰다.
- 우리는 또한 Botha와 Blunsom(2014)이 소개한 로그-바이린 언어 모델과 비교하며, Europarl과 뉴스 논평 코퍼스를 사용하여 훈련했다.
- 우리 모델을 사용하여, 어휘 외 단어의 표현을 문자 n-그램의 표현을 합산하여 얻었으며, 형태소 분할기에 기반한 기술들에 비해 간단한 접근법이 잘 작동한다는 것을 관찰했다.

10. 훈련 데이터 크기의 효과

- 우리는 **단어 간의 문자 수준의 유사성**을 이용함으로써 드문 단어를 더 잘 모델링할 수 있다.
- 따라서 훈련 데이터의 크기에 대해 더 강건해질 수 있으며, 이를 평가하기 위해 단어 벡터의 성능을 훈련 데이터 크기의 함수로 평가할 것을 제안한다.
- 위키피디아의 다양한 크기의 데이터를 이용하여 우리의 모델과 cbow 기준 모델을 훈련시켰고, 그 결과를 Fig. 1에 보고한다.
- 우리가 제안한 모델(sisg)은 이전에 보지 못한 단어에 대해서도 **의미 있는 벡터**를 할당한다.
- 예를 들어, 독일의 GUR350 데이터셋에서 우리의 모델은 5%의 데이터로 훈련했을 때, 전체 데이터로 훈련한 cbow 모델보다 더 나은 성능을 보였다.

11. n-그램 크기가 성능에 미치는 영향

- 제안된 모델은 단어를 벡터로 표현하기 위해 **문자 n-그램**을 사용한다.
- 3에서 6자로 구성된 n-그램을 사용하기로 결정하였으며, 이는 임의적인 선택이지만 다양한 정보를 포함할 수 있다.
- 실험을 통해 성능에 미치는 n-그램 크기의 영향을 검토한 결과, **영어와 독일어** 모두에서 3-6자의 선택이 적절한 결정임을 확인하였다.
- 긴 n-그램을 포함하는 것이 중요하며, 특히 독일어에서는 여러 단위로 이루어진 복합 명사를 캡처하는 데 유리하다.
- 유사성 작업에서는 더 큰 n-그램을 사용할 때 의미적 유추에 도움이 되며, 2-그램은 정보 제공에 한계가 있음을 보여주었다.

12. 언어 모델링 평가 및 결과

- 이 섹션에서는 **언어 모델링** 작업에서 저희 방법으로 얻어진 단어 벡터의 평가를 설명합니다.
- 우리는 Botha와 Blunsom(2014)이 도입한 데이터셋을 사용하여 다섯 개 언어(CS, DE, ES, FR, RU)에서 언어 모델을 평가했습니다.
- 모델은 LSTM 유닛 650개를 포함한 순환 신경망으로, 드롭아웃 확률 0.5와 가중치 감쇠(정규화 파라미터 10^-5)로 정규화됩니다.
- 훈련 집합의 단어 벡터는 문자 n-그램으로 학습하였으며, 이를 언어 모델의 조회 테이블 초기화에 사용했습니다.
- 테스트 결과, 서브워드 정보를 사용하여 학습된 단어 표현이 기존의 스킵그램 모델보다 성능이 월등함을 보여주며, 형태학적으로 풍부한 슬라브 언어에서 더 큰 향상이 나타났습니다.

13. 📊질적 분석 결과 요약

- 특정 단어에 대해 제안된 접근 방식을 사용한 벡터와 스킵그램 기준선에 대한 **코사인 유사도**에 따라 가장 가까운 이웃을 보고한다.
- 제안된 방법을 사용한 경우 복잡하고 기술적이며 사용빈도가 낮은 단어의 이웃은 기준선 모델에서 얻은 것보다 더 우수하다.
- 또한 단어에서 가장 중요한 n-그램이 형태소와 일치하는지 질적으로 평가하기 위해 각 단어를 n-그램의 합으로 나타내었다.
- 독일어, 영어, 프랑스어에서의 선정된 단어의 n-그램을 분석한 결과, 'Auto'와 'Fahrer'(자동차 운전사)와 같은 형태소로 나누어지는 것을 관찰했다.

14. 📊 OOV(OUT OF VOCABULALY)단어의 유사도 분석

- 우리 모델은 훈련 세트에 나타나지 않는 단어를 위한 **단어 벡터**를 생성할 수 있다.
- 이러한 단어에 대해서는 n-그램의 벡터 표현을 평균하여 사용하며, OOV 단어에 대한 n-그램의 매칭을 분석한다.
- 영어 RW 유사도 데이터셋에서 몇 개의 단어 쌍을 선택하여 훈련 어휘에 없는 단어가 n-그램으로만 표현되도록 한다.
- 이러한 단어 쌍에 대해 n-그램 간 코사인 유사도를 보여주며, OOV 단어 수를 늘리기 위해 **위키피디아 데이터의 1%**로 훈련한 모델을 사용한다.

- 분석 결과는 흥미로운 패턴을 보여주며, 예를 들어 'chip'이라는 단어는 'microcircuit'의 두 그룹 n-그램과 잘 매칭됨을 알 수 있다.

Conclusion

- 이번 논문에서는 **하위 단어** 정보를 고려하여 단어 표현을 학습하는 간단한 방법을 조사하였다.
- 문자는 n-gram을 스킵그램 모델에 통합하는 접근 방식으로, 이는 Schütze(1993)의 아이디어와 관련이 있다.
- 모델은 단순하여 빠르게 훈련되고 사전 처리나 감독을 필요로 하지 않으며, 하위 단어 정보를 고려하지 않을 때의 기준 모델들보다 성능이 우수하다.
- 향후 하위 단어 표현 학습에 대한 비교를 용이하게 하도록 모델 구현을 오픈 소스로 제공할 예정이다.