NimbleMiner:
An Open-Source Nursing-Sensitive Natural Language Processing System Based on **Word Embedding**

Topaz M, Murga L, Bar-Bachar O, McDonald M, Bowles K.
CIN: Computers, Informatics, Nursing. 2019 Nov

이동건
인하대학교 정보통신공학과
E-mail: time@inha.edu

2022. 04. 11.

- KEY WORDS:

Falls, Natural language processing, Nursing informatics, Open-access software, Word embedding

Contents

I. INTRODUCTION

About unstructured data

- 의료 분야의 데이터 중 80%는 비정형 데이터(unstructured data)
  - e.g. clinical note –
    discharge summaries, care coordination notes, radiology reports 등

- 의사와 간호사들이 insights를 뽑아내기 위해 노력하지만
  기하급수적으로 늘어나는 데이터를 다루긴 역부족

- NLP 시스템 개발이 이를 해결할 수 있음

About term similarity

- 본 연구의 focus
  - 임상 텍스트에서 유사한 용어를 식별하기 위한 새로운 접근법 개발

- 용어 유사성 (Term similarity)
  - 두 개 이상의 용어 사이의 유사성 (모양 또는 형태)
  - e.g. ""fall" and "patient collapsed"

- NLP 작업에서는 유사한 용어를 식별하는 것이 중요
  - e.g. 정규 표현식 검색, 텍스트 마이닝, 표준 용어 개발 및 유지보수 등

Tradition vs. NLP

- 유사 용어 식별 방법
  - 전통적 방식
    - 간호 표준 용어 사용
    - 전문가의 의견을 통해 식별
  - NLP
    - 지식 기반 접근 방식 (knowledge- based approaches)
    - 분배 기반 측정 방식 (distribution-based measures)

Knowledge-based approaches

- 지식 기반 접근 방식
  - 표준화된 용어와 같이, 인간이 만들고 큐레이션한 지식 소스 활용
  - e.g. Unified Medical Language System와 같은 표준화된 용어 사용
    - "fall"이라는 개념이 "fall" 또는 "falling" 등 여러 다른 개념과 연결됨을 확인

  - 그러나 실생활에서 발생하는 용어 약어나 오타를 포함하지 않으므로 실제 적용 가능성이 낮다.
  - 간호 분야의 경우 표준화된 용어가 상대적으로 적다.

Distribution-based measures

- 분포 기반 측정 방식
  - 일반적으로, 유사한 맥락에서 나타나는 용어들이 서로 관련이 있다는 가정에 기초
  - 이러한 접근 방식은 MEDLINE과 같은 데이터베이스에서 추출된 기사 또는 기사 요약 등 특정 분야의 말뭉치를 사용하여 구현

  - Biomedical 영역에서 유사한 개념을 식별하는 데 좋은 결과를 보임
  - Nursing 영역에서 이러한 방법을 적용하여 평가하는 연구는 없었음

Word Embedding Language Model

- 단어 임베딩 언어 모델은 각 단어가 corpus에서 인접 단어와 얼마나 자주 공존하는지 계산한 다음 이러한 카운트 통계를 각 단어에 대한 축약(condensed) 벡터로 매핑한다.

- 따라서 결과 언어 모델은 각 단어 또는 구문의 문맥에 민감하며 이웃을 기반으로 단어를 예측하는 데 사용될 수 있다.

Skip-gram[11] model of word embedding

- 현재 단어를 건너뛰면서(skip) 현재 단어의 주변 문맥(context)를 사용
- 이러한 방식으로 건너뛴 단어까지의 거리에 기초한 문맥 단어의 가중치를 사용하여 전체 구문(context words + current word)의 단어 순서를 보존
- 큰 텍스트 말뭉치를 표현하고 대상 단어에서 문맥 단어를 예측하는 데 좋은 성능
- Conceptually, 텍스트 말뭉치가 클수록, 강력한 skip-gram 모델이 생성됨

11. Mikolov T, Corrado G, Chen K, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of International Conference on Learning Representations (ICLR 2013). New York, NY: ACM publications; 2013: 1–12.

II. METHODS

Case Study:

Identifying Similar Terms for Fall History From Homecare Clinical Notes

- 데이터 출처
  - 89459명의 환자를 대상으로 대량의 기록(N = 149586)을 사용

- 평균 단어 길이
  - 150 words

Step-by-Step System Description - Preliminary Step:

Word Embedding Model Creation and Specifications

- 1. 사용자가 임상 기록이 담긴 CSV 파일을 업로드

- 2. word window width 지정
  - 클수록 모델이 특정 단어의 문맥에 대해 더 많이 배울 수 있음
  - 반면, 모델의 유사 개념 식별 능력을 잠재적으로 감소시킬 수 있음

Step-by-Step System Description - Preliminary Step:

Word Embedding Model Creation and Specifications

- 3. 전처리(pre-processing)
  - 텍스트 본래의 표현과 유사하게 보존하는 것이 중요하기 때문에 거의 필요하지 않음

  - (1) 구두점 제거
  - (2) 모든 알파벳 문자를 소문자로 변환
  - (3) 임상 노트에서 자주 동시에 발생하는 단어들을 최대 4개의 단어(4-gram)의 길이를 가진 문구로 변환
    - e.g. "pt_fall_yesterday” : 3-gram

Step-by-Step System Description - Preliminary Step:

Word Embedding Model Creation and Specifications

- “simclins” (SIMilar CLINincal terms)
  - 관심 개념 식별 시,
    긍정적 예측 가치가 높은(high positive predictive value) 단어 혹은 구문

- 잠재적 유사 용어는 코사인 거리(Cosine distance)라는 단어 임베딩 모델의 속성을 기반으로 자동으로 식별

Word Embedding Model Creation and Specifications

- 코사인 거리(Cosine distance)
  - 두 벡터 간의 각도 차이로 유사한 정도를 구하는 방법
  - $D = 1 - \frac{X \cdot Y}{\|X\| \times \|Y\|}$ (D : 코사인 거리, X, Y : 코사인 거리를 구하고자 하는 각각의 벡터들)
  - 범위는 0과 1 사이
  - 두 벡터가 직교이면 $D = 1$로 가장 유사성이 없다고 판단한다.
  - 두 벡터가 비슷한 방향일수록 D가 0에 가까워지므로 유사하다고 판단

- 단어 또는 구를 코사인 거리를 기준으로 정렬하면 사용자가 simclin을 식별하는 데 도움이 될 수 있음

참고 : https://dokpin.tistory.com/13

Step-by-Step System Description - Preliminary Step:

Word Embedding Model Creation and Specifications

- 본 연구에서는 잠재적 유사한 용어 목록의 개수를 25, 50, 75개로 설정하여 실험

- 유사한 용어를 많이 나타낼수록
  - 처리 시간이 길어짐
  - simclins를 더 확인해야 함

Step-by-Step System Description

FIGURE 2. NimbleMiner methodology steps.

Step 1: User inputs a concept of interest

Examples:
- Fall/fell
- Balance training
- Fall risk

Step 2: System suggests potential similar terms (based on word embeddings)

Fall synonyms:
1. slipped
2. hit her head
3. fainted
4. tripped

....

Step 3: User selects similar terms (and reviews sentences with examples)

Fall synonyms:
1. **slipped** ✓
2. hit her head ⊗
3. **fainted** ✓
4. **tripped** ✓

....

Sentence example: "... trying to get out of bed when she **slipped** to the floor...

Step 4: User stops the process

Once no more relevant synonyms are suggested by the system

System Evaluation Metrics

- ① Similar terms presented user (n = 25, 50, 75)
- ② Average discovery time (sec / simclin) (= ⑥ / ③)
- ③ True simclins identified by the user (n)
- ④ Simclin discovery precision ($\%$) (= ③ / ⑤ * 100)
- ⑤ System suggested similar terms (n)
- ⑥ Case study duration (min)
- ⑦ No. (%) of clinical notes with simclins (total n = 1 149 586)

III. RESULTS

Table 1. Differences Between Word Embedding Models

Table 1. Differences Between Word Embedding Models

	Word Window Width	3 Words			5 Words			7 Words			10 Words		
	Model	A	B	C	D	E	F	G	H	I	J	K	L
①	Similar terms presented user (n)	**25**	**50**	**75**	**25**	**50**	**75**	**25**	**50**	**75**	**25**	**50**	**75**
②	Average discovery time (s/simclin)	18.3	23.3	31.3	19.0	20.9	28.8	20.4	22.1	28.8	**14.8**	21.4	30.8
③	True simclins identified by the user (n)	131	139	161	164	215	217	221	214	233	170	233	**236**
④	Simclin discovery precision (%)	6.6	5.4	3.8	6.5	6.3	4.5	6.1	5.8	4.5	**7.7**	6	4.3
⑤	System suggested similar terms (n)	1970	2576	4256	2512	3436	4843	3614	3721	5199	2201	3900	**5500**
⑥	Case study duration (min)	40	54	84	52	75	104	75	79	112	42	83	**121**
⑦	No. (%) of clinical notes with simclins[a] (total n = 1 149 586)	102 629 (8.9)	103 098 (9)	**107 660 (9.4)**	60 358 (5.3)	103 492 (9)	72 514 (6.3)	103 918 (9)	105 419 (9.2)	104 453 (9.1)	65 898 (5.7)	105 923 (9.2)	106 155 (9.2)

Note. Digits in bold font indicate the largest values.

[a]There were 51 050 (4.4%) clinical notes that included words or expressions from the SNOMED-CT fall hierarchy (n = 240 unique terms).

Table 1. Differences Between Word Embedding Models

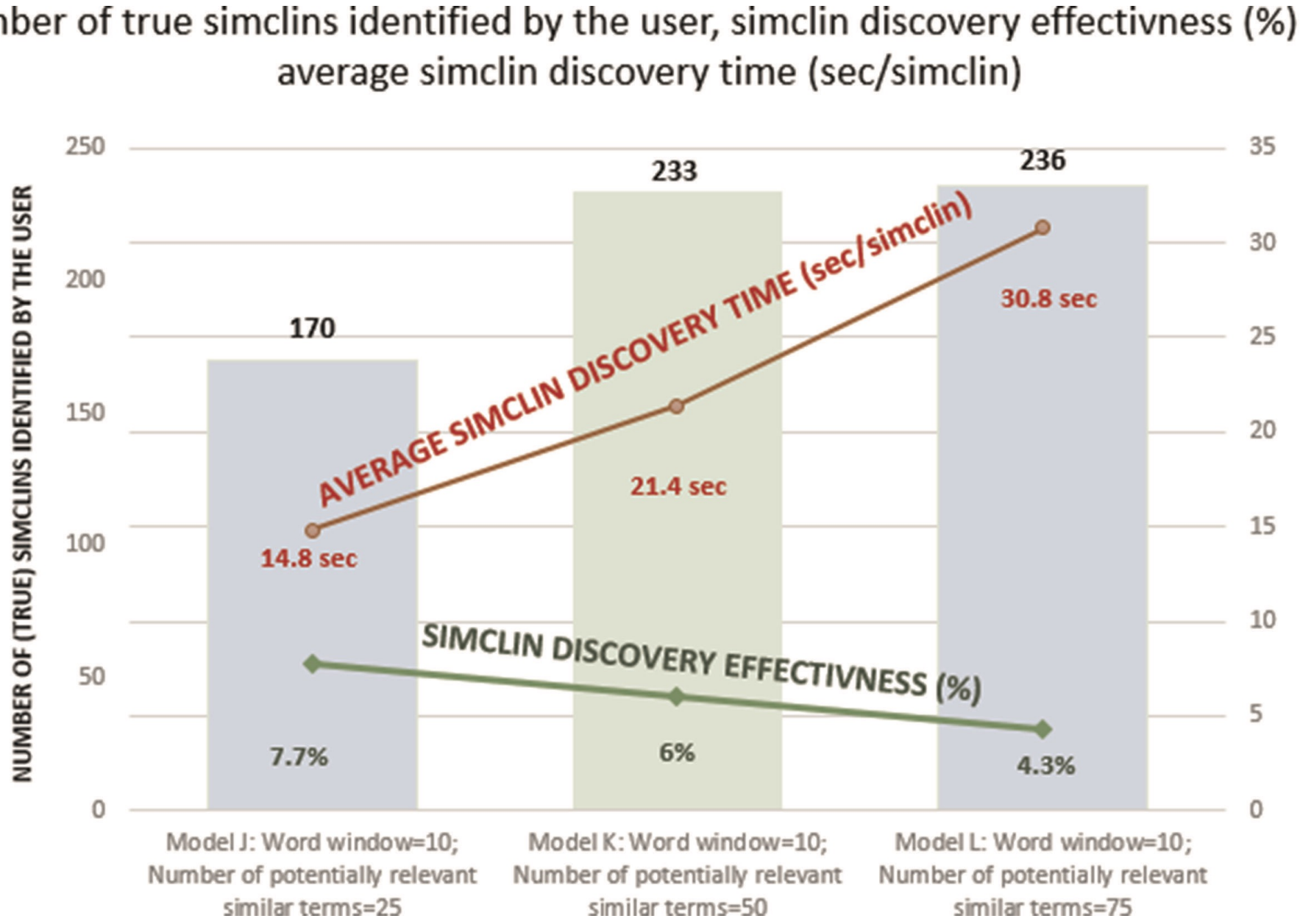
Table 1. Differences Between Word Embedding Models

	Word Window Width	3 Words			5 Words			7 Words			10 Words		
	Model	A	B	C	D	E	F	G	H	I	J	K	L
1	Similar terms presented user (n)	**25**	**50**	**75**	**25**	**50**	**75**	**25**	**50**	**75**	**25**	**50**	**75**
2	Average discovery time (s/simclin)	18.3	23.3	31.3	19.0	20.9	28.8	20.4	22.1	28.8	**14.8**	21.4	30.8
3	True simclins identified by the user (n)	131	139	161	164	215	217	221	214	233	170	233	**236**
4	Simclin discovery precision (%)	6.6	5.4	3.8	6.5	6.3	4.5	6.1	5.8	4.5	**7.7**	6	4.3
5	System suggested similar terms (n)	1970	2576	4256	2512	3436	4843	3614	3721	5199	2201	3900	**5500**
6	Case study duration (min)	40	54	84	52	75	104	75	79	112	42	83	**121**
7	No. (%) of clinical notes with simclins[a] (total n = 1 149 586)	102 629 (8.9)	103 098 (9)	**107 660 (9.4)**	60 358 (5.3)	103 492 (9)	72 514 (6.3)	103 918 (9)	105 419 (9.2)	104 453 (9.1)	65 898 (5.7)	105 923 (9.2)	106 155 (9.2)

Note. Digits in bold font indicate the largest values.

[a]There were 51 050 (4.4%) clinical notes that included words or expressions from the SNOMED-CT fall hierarchy (n = 240 unique terms).

FIGURE 5.


Number of true simclins identified by the user, simclin discovery effectivness (%) and average simclin discovery time (sec/simclin)
NUMBER OF (TRUE) SIMCLINS IDENTIFIED BY THE USER
250
200
150
100
50
0
35
30
25
20
15
10
5
0
170
233
236
AVERAGE SIMCLIN DISCOVERY TIME (sec/simclin)
14.8 sec
21.4 sec
30.8 sec
SIMCLIN DISCOVERY EFFECTIVNESS (%)
7.7%
6%
4.3%
Model J: Word window=10; Number of potentially relevant similar terms=25
Model K: Word window=10; Number of potentially relevant similar terms=50
Model L: Word window=10; Number of potentially relevant similar terms=75

Word Embedding Models Precision and Time Comparisons

- 서로 다른 모델에 의해 생성된 모든 simclins을 중복을 제외한 하나의 목록으로 컴파일했을 때, 371개의 고유한 simclins
  - 59% (n = 220) :
    - “fall”이나 “fell” 같은 단어를 포함하는 다른 표현들의 변형
    - 철자 오류
      - e.g. “felll,” “fals”
    - 잘못 쓰여진 표현
      - e.g. “felled”
    - 기타
      - e.g. “mechanical fall,” “ended up falling,” or “fell off ladder”
  - 41% (n = 151) :
    - 어휘의 사전적 변형 (lexical variation)
      - e.g. “tripping over,” “slipped off chair,” “slided down,” and “pt collapsed.”

IV. DISCUSSION, FURTHER WORK and LIMITATIONS

DISCUSSION about window width sizes

- 다른 연구와 유사하게[8, 12, 13],
  window width sizes가 더 큰 word embedding 모델이
  유사한 용어를 발견하는데 더 좋음

- 반면, 사용자가 작업을 완료하는 데 더 많은 시간 소요
  - e.g. 모델 L : 5500개 검토, 121분 / 모델 J : 2201개 검토, 42분

- 본 논문의 저자는 model K를 추천함
  - model K, 83 minutes with 3900 potential similar terms for user review, finding 9.2% of clinical notes to contain simclins

FURTHER WORK

- NimbleMiner가 대규모 어휘 작성 분야 중 과거에 거의 수행되지 않은 간호와 같은 영역에서 쉽게 적용될 것

- 또한, "사회적 지원의 부족, 약물 또는 알코올 남용, 열악한 생활 상태 및 기타 사회 행동 위험 요소" 등과 같이 이전에 구조화되지 않고 대부분 서술적 형태로 존재하는 개념에 대한 어휘를 쉽게 만드는 데 적용 가능

- 더 많은 word embedding model의 매개 변수에 대해 실험해야 함

FURTHER WORK

- 향후 유사 어휘 검색 프로젝트에서는 서로 다른 모델 간의 어휘 중복 검토 필요
  - 어휘 중복을 정량화하면 사용자의 유사 식별 시간을 줄일 수 있음

- 또한 여러 리뷰어들이 simclin 검색을 수행해야 함

- NimbleMiner는 언어에 구애받지 않음
  - 히브리어 검증, 러시아어 평가 중

LIMITATIONS

- 하나의 데이터 세트로 실험한 점
- 결과가 반복되거나 검증되지 않은 점
- 모델 훈련 시, 변경할 수 있는 제한된 매개 변수 세트만 사용
- SNOMED-CT 이외의 어휘에 대해 simclin 어휘를 평가하지 않음
- corpus에서 “fall history”를 설명하는 용어와 어구가 얼마나 많은지 파악 x
- Recall 또는 F-measure와 같은 시스템 성능 측정치를 추정할 수 없음