ETH zürich

Module III: Deep Learning, Applications to Biomedical Data

Felipe Llinares-López and Damian Roqueiro
Machine Learning & Computational Biology Lab
D-BSSE, ETH Zürich

Tutorial AM2: Machine learning methods in the analysis of genomic and clinical data. July 6, 2018

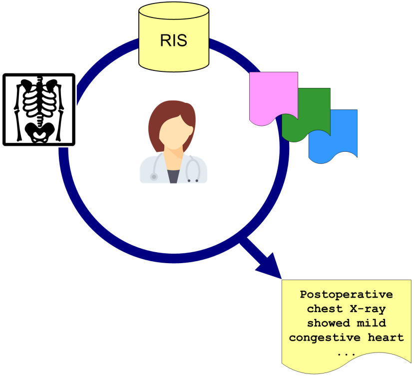
Part I. Introduction

What is an electronic health record?
Challenges in text mining of medical records

Electronic health records (EHRs)

A definition Reddy and Aggarwal [2015]

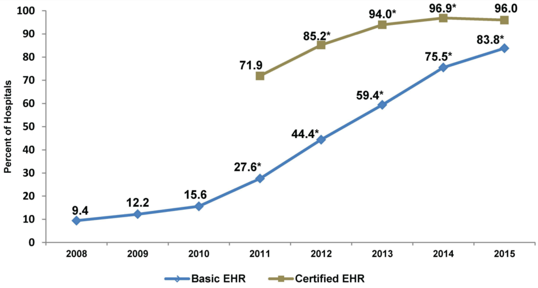
- Data related to a patient's care
  - Demographics
  - Medications
  - Vital signs
  - Medical history
  - Laboratory data
  - Reports (e.g. radiology)
  - Progress notes



Source: All icons in figures were downloaded & modified from: flaticon.com (designed by Freepik)

Electronic health records (EHRs)

Adoption in the United States



Source: ONC Data Brief 2016. Henry et al. [2016]

STAGE	HIMSS Analytics EMRAM EMR Adoption Model Cumulative Capabilities
7	Complete EMR; External HIE; Data Analytics, Governance, Disaster Recovery, Privacy and Security
6	Technology Enabled Medication, Blood Products, and Human Milk Administration; Risk Reporting; Full CDS
5	Physician documentation using structured templates; Intrusion/Device Protection
4	CPOE with CDS; Nursing and Allied Health Documentation; Basic Business Continuity
3	Nursing and Allied Health Documentation; eMAR; Role-Based Security
2	CDR; Internal Interoperability; Basic Security
1	Ancillaries - Laboratory, Pharmacy, and Radiology/Cardiology information systems; PACS; Digital non-DICOM image management
0	All three ancillaries not installed

Source: www.himssanalytics.org/emram

MIMIC-III database

Contents Johnson et al. [2016]

- 58,000 hospital admissions (2001–2012)
- 46,520 patients (38,645 adults; 7,875 neonates)
- Demographics
- Vital sign measurements
- Laboratory test results
- Medications
- Caregiver notes
- Imaging reports
- Mortality (both in and out of hospital)

Reports for adult patients

Type	Count
Radiology	507,326
Nursing/other	418,041
Nursing	223,546
ECG	208,413
Physician	141,617
Discharge summary	55,396
Echo	44,589
Respiratory	31,739
Nutrition	9,418
General	8,301
Rehab Services	5,431
Social Work	2,670
Case Management	966
Pharmacy	103
Consult	98

MIMIC-III database contd.

Sample radiology report Johnson et al. [2016]

```
[**Last Name (LF) 626** CT CHEST W/CONTRAST; CT ABD & PELVIS WITH CONTRAST
Clip # [**Clip Number (Radiology) 28851**]
Reason: f/u bilateral pneumothoraces + hemothorax. Please perform st
Admitting Diagnosis: S/P PEDESTRIAN STRUCK
Contrast: OMNIPAQUE Amt: 130CC
______________________________________________________________________________
[**Hospital 2**] MEDICAL CONDITION:
52M pedestrian struck w/ pelvic and acetabular fxs, scalp laceration w/
subgaleal hematoma, and L PTX s/p CT decompression.
REASON FOR THIS EXAMINATION:
f/u bilateral pneumothoraces + hemothorax. Please perform study with IV and
oral contrast, need to visualize the duodenum
No contraindications for IV contrast
______________________________________________________________________________
PFI REPORT
1.  Interval layering of hematoma with decreased component in the right
anterior pararenal space and tracking inferiorly into the right paracolic
gutter and pelvis.
2.  No definite evidence of solid organ injury.  No evidence of duodenal wall
hematoma.  No extraluminal oral contrast.
3.  Small left pneumothorax with mild interval increase in size compared to
prior.  Chest tube with tip terminating at the left lung base.
4.  Similar bibasilar opacities likely atelectasis and aspiration.  Subtle
increase in size of focal opacity in the left lower lung could be contusion.
5.  Known fracture of the right inferior pubic ramus and anterior of the right
acetabulum.
```

Mining medical records

Challenges

- Data heterogeneity
- Extensive use of quantitative data
- Misspelling detection and correction
- Negated and probable events
- Ambiguous acronyms and abbreviations

Sample admission report
BLOOD WBC-9.6 RBC-3.97* Hgb-11.8* Hct-34.4* MCV-87 MCH-29.7 MCHC-34.2 RDW-15.8* BLOOD Plt Ct-326 BLOOD Glucose-112* UreaN-23* Creat-0.7 Na-137 K-4.2 Cl-101 HCO3-23 AnGap-17 BLOOD TSH-2.0

Mining medical records

Challenges

- Data heterogeneity
- Extensive use of quantitative data
- Misspelling detection and correction
- Negated and probable events
- Ambiguous acronyms and abbreviations

Sample admission report
BLOOD WBC-9.6 RBC-3.97* Hgb-11.8* Hct-34.4* MCV-87 MCH-29.7 MCHC-34.2 RDW-15.8* BLOOD Plt Ct-326 BLOOD Glucose-112* UreaN-23* Creat-0.7 Na-137 K-4.2 Cl-101 HCO3-23 AnGap-17 BLOOD TSH-2.0

Misspelling detection and correction

Examples Johnson et al. [2016]

- Patient's **respirtory** status improved significantly over the first **hopsital** day and patient was transferred to the floor.
- A: Tolerating current feeding **regiman**. P: Continue to support **nutritioanl** needs.
- ...blood **culutes** was **postive** for GPC, he was started on 2 weeks of Vancomycin. He continued to have numerous cultures...

Related work Damerau [1964] , Lai et al. [2015]

- 80% of spelling errors: insertion, deletion, substitution, or two letters transposed or switched
- Detection: Unified Medical Language System (UMLS) lexicon + other sources (e.g. RxNorm)
- Correction: Scoring algorithms based on orthographic and phonetic edit distances

Misspelling detection and correction

Examples Johnson et al. [2016]

- Patient's **respirtory** status improved significantly over the first **hopsital** day and patient was transferred to the floor.
- A: Tolerating current feeding **regiman**. P: Continue to support **nutritioanl** needs.
- ...blood **culutes** was **postive** for GPC, he was started on 2 weeks of Vancomycin. He continued to have numerous cultures...

Related work Damerau [1964] , Lai et al. [2015]

- 80% of spelling errors: insertion, deletion, substitution, or two letters transposed or switched
- Detection: Unified Medical Language System (UMLS) lexicon + other sources (e.g. RxNorm)
- Correction: Scoring algorithms based on orthographic and phonetic edit distances

Negated and probable events

Examples Johnson et al. [2016]

- Reason: evaluate cough wheezing **?copd/pneumonia**
- ...**demonstrated no trace** pulmonic regurgitation and **no tricuspid** regurgitation...
- The Neurosurgery Team recommended a four-vessel angiogram to **rule out** any vessel damage.
- CARDIAC: RR, normal S1, S2. **No** murmurs but **apparent** pericardial friction rub with 2 knocks. **No** thrills, lifts. **No** S3 or S4.

Related work Harkema et al. [2009] , Kuhn and Eickhoff [2016]

- ConText/NegEx algorithm: sentences are labeled as negative, affirmative, certain or uncertain.
- There are algorithmic benefits to taking negated events into account

Negated and probable events

Examples Johnson et al. [2016]

- Reason: evaluate cough wheezing **?copd/pneumonia**
- ...**demonstrated no trace** pulmonic regurgitation and **no tricuspid** regurgitation...
- The Neurosurgery Team recommended a four-vessel angiogram to **rule out** any vessel damage.
- CARDIAC: RR, normal S1, S2. **No** murmurs but **apparent** pericardial friction rub with 2 knocks. **No** thrills, lifts. **No** S3 or S4.

Related work Harkema et al. [2009] , Kuhn and Eickhoff [2016]

- ConText/NegEx algorithm: sentences are labeled as negative, affirmative, certain or uncertain.
- There are algorithmic benefits to taking negated events into account

Ambiguous acronyms and abbreviations

Examples for the word **ACC** Johnson et al. [2016]

i) American College of Cardiology; ii) Agenesis of the Corpus Callosum; iii) Accident

- . . . dental work 2 weeks ago without antibiotic prophylaxis, per current **ACC**/AHA guidelines. . .
- Infant with prenatally diagnosed **ACC**, copolcephaly, microcephalic
- MOTORYCLE **ACC** OPEN FX'S, Admitting Diagnosis: TIB/FIB FRACTURE. . .

Word sense disambiguation (WSD). Related work Navigli [2009] Jurafsky and Martin [2017]

- It is considered an AI-complete problem
- Disambiguate a word given a lexicon with an inventory of senses for each entry

Ambiguous acronyms and abbreviations

Examples for the word **ACC** Johnson et al. [2016]

i) American College of Cardiology; ii) Agenesis of the Corpus Callosum; iii) Accident

- … dental work 2 weeks ago without antibiotic prophylaxis, per current **ACC**/AHA guidelines…
- Infant with prenatally diagnosed **ACC**, copolcephaly, microcephalic
- MOTORYCLE **ACC** OPEN FX'S, Admitting Diagnosis: TIB/FIB FRACTURE…

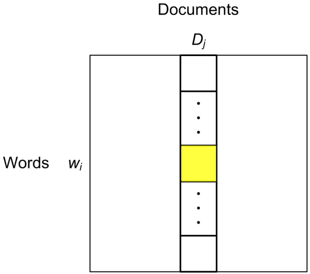
Word sense disambiguation (WSD). Related work Navigli [2009] Jurafsky and Martin [2017]

- It is considered an AI-complete problem
- Disambiguate a word given a lexicon with an inventory of senses for each entry

Part II. Word embeddings

Representation of documents and words
Words as vectors

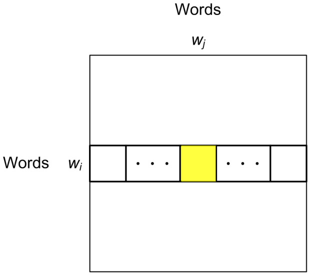
Documents as vectors



Linking terms to documents

- In the term-document matrix $\in \mathbb{R}^{|V| \times |D|}$
  - with a vocabulary of size $|V|$ and a corpus of size $|D|$
  - a row represents a word w_i
  - a column is a document D_j
  - the contents of each cell$_{ij}$ is a term weight, e.g. $\textit{tf-idf}(w_i, D_j)$
- General assumption: if two documents are similar, they tend to have similar words

Words as vectors



Linking terms to context

- In the term-context matrix $\in \mathbb{R}^{|V| \times |V|}$
  - with a vocabulary of size $|V|$
  - a row represents a word w_i (target)
  - a column is a word w_j (context)
  - each cell$_{ij}$ represents how frequently w_i co-occurs with w_j
- General assumption: if two words are similar, they tend to be present in similar documents. It is a representation of the meaning of a word based on the documents in which it occurs.

Words as vectors contd.

Different representations

- Pointwise mutual information (PMI) Church and Hanks [1990]
- Singular value decomposition (SVD) Eckart and Young [1936]
- Neural-network embeddings
  - Learned distributed feature vector Bengio et al. [2003]
  - Skip-gram with negative sampling (`word2vec`) Mikolov et al. [2013]

Pointwise mutual information (PMI)

Church and Hanks [1990]

Measure

- How often two events x and y occur when compared to the expectation of independence

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- Extending it to co-ocurrence of words

$$PMI(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

- Correcting for negatives values, obtain *positive* PMI

$$PPMI(w,c) = max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

Singular value decomposition (SVD) Eckart and Young [1936]

Procedure

- Factorization of a term-context matrix X as

$$X = L \cdot \Delta \cdot R^\top$$

- Alternative, one can truncate to the top d eigenvalues in Δ

$$X_d = L_d \cdot \Delta_d \cdot R_d^\top$$

- To finally obtain

$$W = L_d \cdot \Delta_d \qquad C = R_d^\top$$

where rows of W are word representations; rows of C are context representations

Words as vectors

General goal of learning neural word embeddings

- Define a predictive model, for a center word w_t and words in its context

$$P(w_c|w), w_c \in \text{context}(w)$$

- Specify a cost function to minimize (loss)

$$E = 1 - P(w_c|w)$$

- Perform this task for all words w in corpus
- Iteratively adjust the word vectors to minimize E

Stanford University, Lecture 2, CS224N/LING284, NLP with Deep Learning

Skip-gram with negative sampling (`word2vec`) Mikolov et al. [2013]

Motivation

- Traditional encoding of word (as vector) in a corpus

car = [0, 0, 0, ... , 0, 1, 0, ... , 0, 0, 0]

truck = [0, 0, 0, ... , 0, 0, 0, ... , 1, 0, 0]

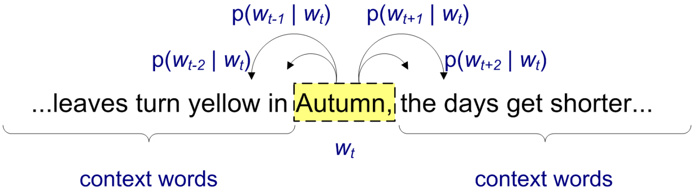
flower = [0, 1, 0, ... , 0, 0, 0, ... , 0, 0, 0]

- Number of elements is equal to size of vocabulary
- Referred to as one-hot vector
- Similarity between words with inner product is meaningless

Skip-gram with negative sampling (`word2vec`)

Goal

- Given a sequence of words $w_1, w_2, \ldots, w_T$
- Maximize the probability of any context word given the center word



Part II. `word2vec`

Objective function
Skip-gram model
Results

Mikolov et al. [2013]

word2vec

Objective function

- Given a sequence of words $w_1, w_2, \ldots, w_T$
- Maximize the probability of any context word given the center word

$$\prod_{t=1}^{T} \prod_{-c \leq j \leq c, j \neq 0} p(w_{t+j} | w_t)$$

- Equivalent to the negative log likelihood

$$-\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0}^{T} \log p(w_{t+j} | w_t)$$

word2vec contd.

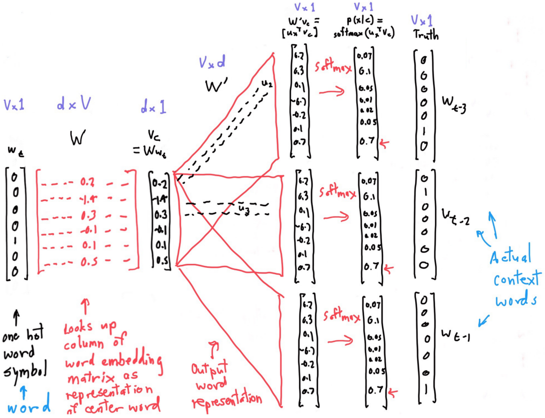
Objective function

- Given a sequence of words $w_1, w_2, \ldots, w_T$
- Maximize the probability of any context word given the center word
- Maximize the average log probability (Eq. 1) Mikolov et al. [2013]

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c \leq j \leq c, j \neq 0}^{T} \log p(w_{t+j} | w_t)$$

word2vec contd.

Skip-gram model *contd.*



Stanford University, Lecture 2, CS224N/LING284, NLP with Deep Learning (edited to fit slide)

`word2vec` contd.

Obtaining probabilities: softmax function

- Transforms elements of a vector (real numbers) to $[0, 1]$ range, adding up to 1

$$P(v) = \frac{e^v}{\sum_{i=1}^{M} e^{v_i}}$$

word2vec contd.

Putting the parts together

- We said before that the objective function is:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}^{T}\log p(w_{t+j}|w_t)$$

- The model computes the probability $p(w_{t+j}|w_t)$ as (Eq. 2) Mikolov et al. [2013]

$$p(w_O|w_I) = \frac{e^{{v'}_{w_O}^{\top} v_{w_I}}}{\sum_{w=1}^{W} e^{{v'}_{w}^{\top} v_{w_I}}}$$

word2vec contd.

Increasing predictive performance

- Apply sub-sampling of frequent words
- A word w_i is discarded with probability, (Eq. 5) Mikolov et al. [2013]

$$P(w_i) = 1 - \sqrt{\frac{t}{\text{freq}(w_i)}}$$

where

t is a parameter set by the user with default value 10^{-5}

$\text{freq}(w_i)$ is the frequency of w_i in the corpus

- Desirable side effect: Removing words effectively increases the size of context window

word2vec contd.

Analysis

- Trained model on internal Google dataset $\sim$1 billion words
- Discarded very infrequent words (<5 times)
- Final vocabulary size of 692,000 words

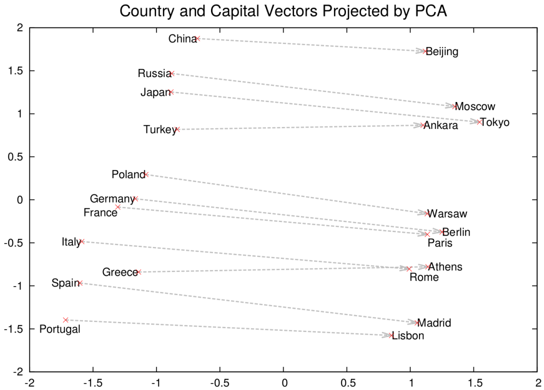
Tests

- Syntactic analogies: "quick" is to "quickly" as "slow" is to "slowly"
- Semantic analogies: "Germany" is to "Berlin" as "France" is to "Paris"
- Solved as:

  find the closest vector vec($\mathbf{x}$) to vec("Berlin") – vec("Germany") + vec("France")

word2vec contd.

Results



Mikolov et al. [2013], Figure 2

word2vec contd.

Results contd.

- More illustrative examples

Relation	Nearest token
sushi – Japan + German	bratwurst
Cu – copper + gold	Au
bigger – big + cold	colder

AI Summit Vienna 2017 – Tomas Mikolov, Neural Networks for Natural Language Processing

Part III. Mortality Prediction

Word embeddings from clinical reports
Prediction of mortality based on unstructured data
Attention mechanisms

Patricia Calvo Pérez

Mortality prediction

Endpoints

- In-hospital
- 30-day
- 1-year post-discharge

Importance Luo and Rumshisky [2016]

- Critical to understand and prevent future complications in order to discharge patients more safely and efficiently
- Majority of previous approaches are based on structured data

Added feature: attention mechanism

Example: Yelp reviews Lin et al. [2017]

- 1-star review

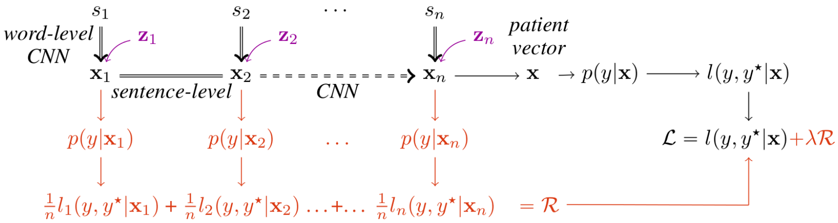
this place suck the food be gross and taste like grease I will never go here again ever sure the entrance look cool and the waiter can be very nice but the food simply be gross taste like cheap 99cent food do not go here the food shot out of me quick then it go in

- 5-star review

love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can taste it steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I have had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Cola

Example: mortality prediction contd.

Model Grnarova et al. [2016]



Preliminary results

Prediction

AUC	In-hospital	30-day	1-year
10-model ensemble	0.960	0.801	0.815

Interpretation

Top sentences, P(in-hosp survival=low)
not moving any extremities moving chin tongue with oral care no cough no gag no pupillary
pt continues to overbreathe though at times
frequent generalized myoclonus anoxic brain and nerve injury
found by fd unresponsive apneic in arrest

Top sentences, P(1-year survival=high)
low lung volumes accentuate normal pulmonary vasculature
no acute fractures are present
low lung volumes

Preliminary results

Prediction

AUC	In-hospital	30-day	1-year
10-model ensemble	0.960	0.801	0.815

Interpretation

Top sentences, P(in-hosp survival=low)
not moving any extremities moving chin tongue with oral care no cough no gag no pupillary
pt continues to overbreathe though at times
frequent generalized myoclonus anoxic brain and nerve injury
found by fd unresponsive apneic in arrest

Top sentences, P(1-year survival=high)
low lung volumes accentuate normal pulmonary vasculature
no acute fractures are present
low lung volumes

Conclusions

Deep learning and mining of electronic health records

- Text mining of EHRs faces many challenges due to unstructured nature of text
- Creation of word embeddings from EHRs
- Deep learning model gave good prediction performance in mortality prediction task
  - Not mentioned in the slides, but many baselines were tried (SVD, bag-of-words, and others)
- Attention mechanism(s) will assist physicians in the interpretation of the results

Acknowledgements

Machine Learning and Computational Biology Lab

Patricia Calvo Pérez

@MLCBResearch

@AGKBorgwardt

References

Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, Mar. 2003. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=944919.944966`.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16 (1):22–29, Mar. 1990. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=89086.89095`.

F. J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3): 171–176, Mar. 1964. ISSN 0001-0782. doi: 10.1145/363958.363994. URL `http://doi.acm.org/10.1145/363958.363994`.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sep 1936. ISSN 1860-0980. doi: 10.1007/BF02288367. URL `https://doi.org/10.1007/BF02288367`.

P. Grnarova, F. Schmidt, S. L. Hyland, and C. Eickhoff. Neural document embeddings for intensive care patient mortality prediction. *CoRR*, abs/1612.00467, 2016. URL `http://arxiv.org/abs/1612.00467`.

H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*, 42(5):839–851, Oct 2009.

J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. *ONC Data Brief*, (35), May 2016. URL `https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf`.

References II

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, may 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL http://www.nature.com/articles/sdata201635.

D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Draft, 3rd edition, 2017. URL https://web.stanford.edu/~jurafsky/slp3/.

L. Kuhn and C. Eickhoff. Implicit negative feedback in clinical information retrieval. *CoRR*, abs/1607.03296, 2016. URL http://arxiv.org/abs/1607.03296.

K. H. Lai, M. Topaz, F. R. Goss, and L. Zhou. Automated misspelling detection and correction in clinical free-text records. *J Biomed Inform*, 55:188–195, Jun 2015.

Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017. URL http://arxiv.org/abs/1703.03130.

Y.-F. Luo and A. Rumshisky. Interpretable Topic Features for Post-ICU Mortality Prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 827. American Medical Informatics Association, 2016.

References III

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf`.

R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, Feb. 2009. ISSN 0360-0300. doi: 10.1145/1459352.1459355. URL `http://doi.acm.org/10.1145/1459352.1459355`.

B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623732. URL `http://doi.acm.org/10.1145/2623330.2623732`.

C. K. Reddy and C. C. Aggarwal. *Healthcare Data Analytics*. Chapman & Hall/CRC, 2015. ISBN 1482232111, 9781482232110.