Semi-supervised Learning for Biomedical Named-Entity Recognition

Aleksandar Bojchevski
Supervisor: Prof. Burkhard Rost
Advisor: Juan Miguel Cejuela

Technische Universität München

Problem statement and motivation

- Named-entity recognition in the biomedical domain
  . . . mutation L173R of 3 beta-HSD type II was observed at . . .
- Recognition of **mutation** mentions expressed in **natural language**
  . . . where leucine at residue 173 was altered to an arginine in . . .
- Focus on natural language
  - Has not been considered before
  - We show it is significant

NL mention definition

subclass	example
ST: Standard	p.18G>K
SS: Semi-Standard	Gly 18 to Lys
NL: Natural Language	Glycine was substituted by Lysine at residue 18

- Algorithmic definition
- Heuristics
- **Exclusive** definer

Landscape of methods and corpora

- Corpora and methods

Corpus	Method capability	# Documents	Year
tmVar [1]	ST, some SS	500 abstracts	2013
MutationFinder [2]	ST	588 abstracts	2007
SETH [3]	mostly ST	630 abstracts	2014

- **IDP4** corpus [4]
  - **85** abstracts and **78** full texts
  - 3 annotators, **87.23%** inter-annotator agreement

Goals

1. Study significance of NL
2. Develop method
    - Create NL corpus
    - Better than state-of-the art for ST
    - Fairly well for NL and SS
3. Create useful tool for the community

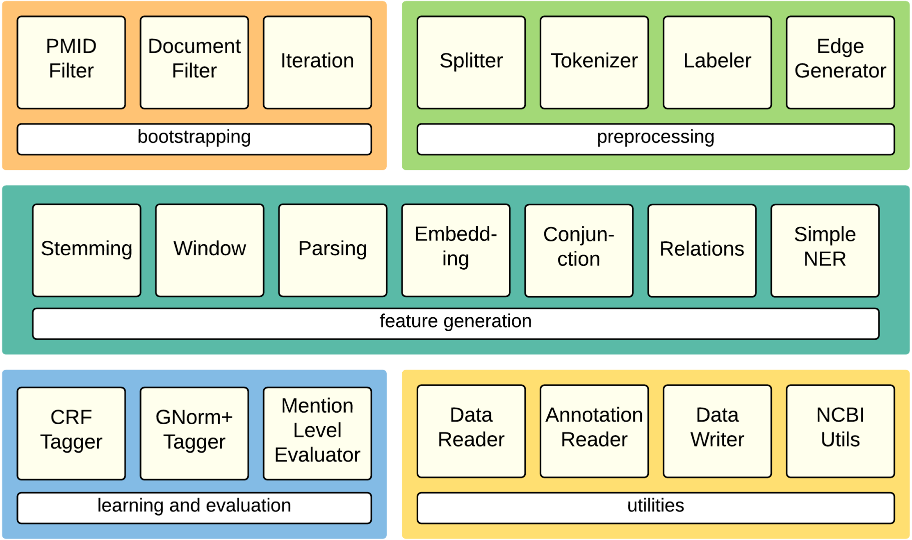
Goal 1: Significance of NL mentions

- Number of NL mentions
  - Only 3% for tmVar
  - Less than 5% for SETH
  - Between **16%** and **27%** for IDP4
- NL mentions that are not translated into an ST mention within the same text
  - Between **5%** and **12%** of all mentions
  - Between **32%** and **45%** of documents contain at least one
- Abstracts vs. Full texts
  - Around **8%** for full texts
  - Between **14%** and **19%** for abstracts

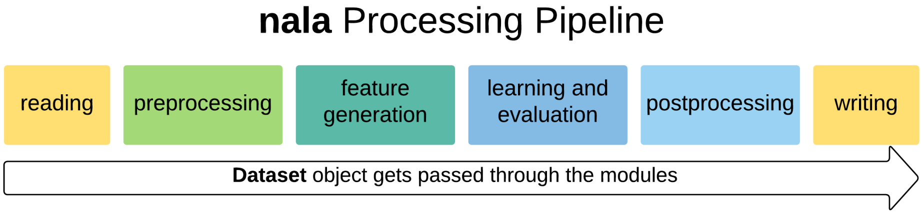
Goal 2: Develop state-of-the-art method

- Conditional random fields (CRFs)
- Semi-supervised learning
    - Unsupervised feature learning
    - Active learning
- Postprocessing

Goal 3: nalaf - (Na)tural (La)nguage (F)ramework



Goal 3: nala



- Built on top of nalaf
- Hand-crafted mutation specific features
- Natural language definers
- Postprocessing module

Unsupervised feature learning

- Latent representation
- Word representations
  - Clustering (Brown)
  - Distributional representations (LSA, LDA)
- **Neural word embeddings**
  - Train a neural network on unlabeled data
  - Continuous bag-of-words (CBOW) vs. skip-gram architecture

Continuous bag-of-words

- One-hot representation
- Predict the word at position t, based on the words in its surrounding context
- Each row in **W'** corresponding to an embedding for a particular word
- Infer $N = 100$ dimensional representation, with a context of 10 words
- Complete MEDLINE/Pubmed database

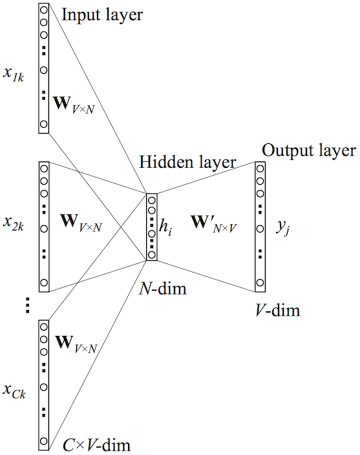
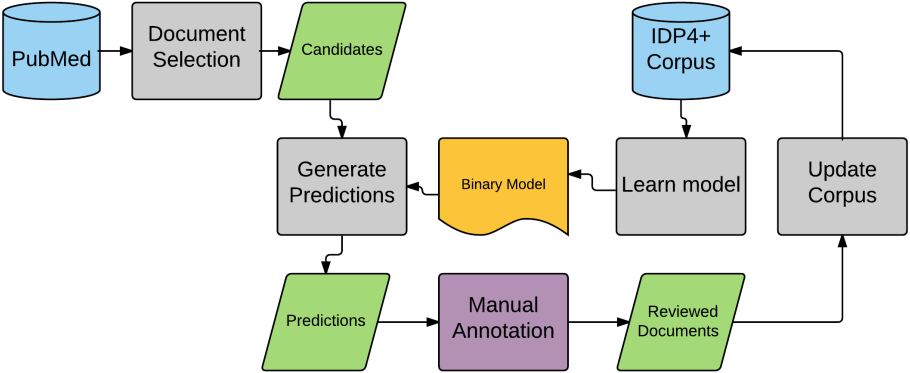


Figure: CBOW architecture [5]

Active learning



Postprocessing

- Fix wrong predictions
  - Split multiple mentions
    G187A/A283C $\Longrightarrow$ G187A / A283C
  - Fix boundary
    G75fsX1 07 $\Longrightarrow$ G75fsX107
    p.(Glu1500Val) $\Longrightarrow$ p.(Glu1500Val)
- Improve performance
  - **Positive** patterns: include false negatives
  - **Negative** patterns: remove false positive

nala performance

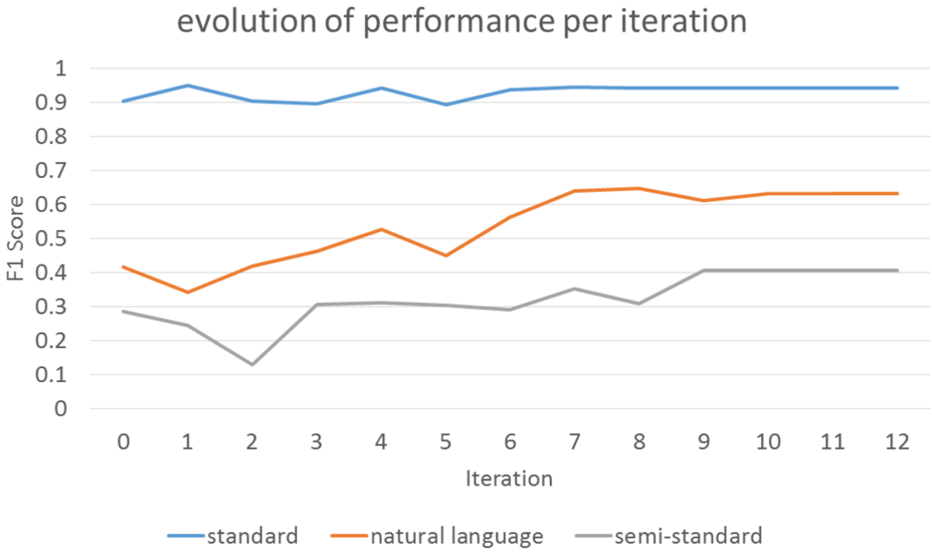
- tmVar corpus

method	precision	recall	F_1 score
tmVar	0.9138	0.9140	0.9139
nala	0.9487 ± 0.0011	0.9159 ± 0.0015	0.9320 ± 0.0012

- IDP4+ corpus

subclass	Precision	Recall	F_1 score
ST	0.9508 $\pm$0.0008	0.9374 $\pm$0.0018	0.9440 $\pm$0.0011
NL	0.7347 $\pm$0.0061	0.5538 $\pm$0.0059	0.6316 $\pm$0.0053
SS	0.5556 $\pm$0.0108	0.3191 $\pm$0.0092	0.4054 $\pm$0.0084
all	0.9566 $\pm$0.0007	0.9221 $\pm$0.0018	0.9390 $\pm$0.0011

Bootstrapping



Key findings

- Neural word embeddings
  - F_1 score increase of around 4.0% for NL
  - High modeling power compared to hand-crafted features
- Postprocessing
  - F_1 score increase of around 3.7% for tmVar corpus
  - F_1 score increase of around 5.3% for IDP4 corpus
- Use mutation specific tokenization
- Bootstrapping was helpful in increasing SS and NL performance
- Merging strategies: intersection and priority consistently outperforms other strategies

Conclusion

Achieved all of our goals:

1. NL mentions are indeed significant with more than **32%** of documents having at least one mention not translated to ST
2. We extended the IDP4 corpus to **IDP4+**, adding NL and SS mentions in a series of 11 iterations
2. nala outperform the state-of-the-art method for ST mentions with $F_1 = \mathbf{0.9444} \pm 0.0011$
2. nala performs fairly well for NL mentions with $F_1 = \mathbf{0.6316} \pm 0.0053$
3. The **nalaf** framework and the **nala** method, available at:
`https://github.com/Rostlab/nalaf`
`https://github.com/Rostlab/nala`

Thank you!

- Rostlab
- Prof. Burkhard Rost
- Juan Miguel Cejuela
- Carsten Uhlig
- DAAD
- My family

References

[1] Chih-Hsuan Wei, Bethany R Harris, Hung-Yu Kao, and Zhiyong Lu.
tmvar: a text mining approach for extracting sequence variants in biomedical literature.
Bioinformatics, page btt156, 2013.

[2] J Gregory Caporaso, William A Baumgartner, David A Randolph, K Bretonnel Cohen, and Lawrence Hunter.
Mutationfinder: a high-performance system for extracting point mutation mentions from text.
Bioinformatics, 23(14):1862–1865, 2007.

[3] Philippe Thomas, Tim Rocktäschel, Yvonne Mayer, and Ulf Leser.
SETH: SNP Extraction Tool for Human Variations.
http://rockt.github.io/SETH/, 2014.

[4] Juan Miguel Cejuela, Aleksandar Bojchevski, Rustem Bekmukhametov, Sanjeev Karn, and Shpend Mahmuti.
Idp4 corpus, 2015.

[5] Xin Rong.
word2vec parameter learning explained.
arXiv preprint arXiv:1411.2738, 2014.