National NLP Clinical Challenges n2c2 - Track 1

Medical University of Graz (Austria)
Institute for Medical Informatics, Statistics and Documentation

Michel Oleynik, **Amila Kugic**, Markus Kreuzthaler,
Zdenko Kasáč, Stefan Schulz

Overview

- Who are we?
- What was the task / the challenge?
- What was our approach to solving the challenge?
- What worked, what didn‘t?
- What can be improved?

Our Group

- BST (Biomedical Semantics Team) at the Medical University of Graz
  - led by Prof. Stefan Schulz, MD
  - in total about 13 members (including guest researchers)
  - five participants for the challenge (three computer scientists, two physicians)
- focus on
  - Applied Clinical NLP
  - Information Retrieval and Extraction
  - Information Models, Ontologies, Terminologies, Semantics, …

Our Approach

- three submissions
  - rule-based approach
  - machine learning
    - support vector machines (SVM)
    - neural network (NN)

Rule-Based Approach

- positive text markers
  - *"elevated creatinine"*
- negative text markers
  - *"Spanish", "with interpreter"*
- regular expressions for value extraction
  - *"HbA1C of 10.7"*
- basic negation and family history detection
  - *"no history of renal failure"*
  - *"rule out myocardial infarction"*

Mild anemia - repeat testing-hematocrit stable at 39 and hemogl
Mild elevated creatinine-repeat testing and check a urine and mi

SOCIAL HISTORY:
No tobacco, Occ Etoh, Spanish speaking, originally from Columbia

, including IDDM, with a HbA1C of 10.7 earlier this month

Renal/Genitourinary: no history of renal failure; denies hematuria,

1. Rule out myocardial infarction.

SVM

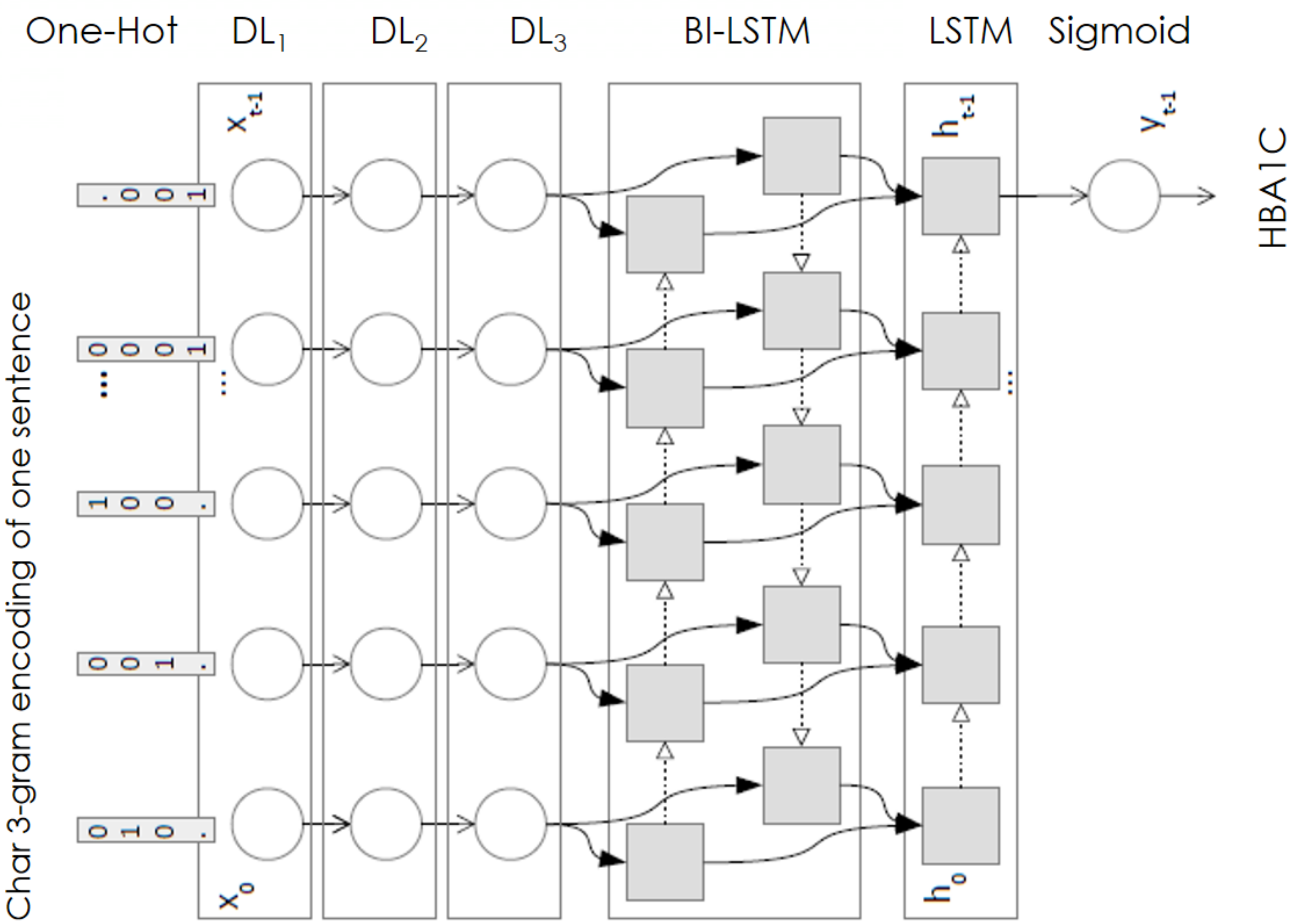
- data pre-processing
  - alphabetic tokens
  - lowercased
  - tf-idf weighting
  - 1000 most common tokens kept
- data modelling
  - bag-of-words representation
- classifier setting
  - linear kernel
  - optimizing cost parameter

Motivation for NNs

- capture sequences (time)[1]
- Google News vectors (word2vec)
  - low coverage rate
- corpus based recalculation
  - no robust vector representation
- instead used character 3-gram encoding [2]
  - of each sentence
- text cleansing and sentence detection

[1] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
[2] Arnold, Sebastian, et al. "Robust named entity recognition in idiosyncratic domains." arXiv preprint arXiv:1608.06757 (2016).

Neural Networks



Our Setup

- common framework
  - https://github.com/bst-mug/n2c2
  - Java Libraries: Weka, DL4J, libsvm
- continuous improvement
  - false-positive and false-negative analysis
  - feedback from rule-based approach influenced other strategies

Overall Results

Method	Overall (micro F1)	Overall (macro F1)
Rule-Based Classifier	0.9100	0.7525
Support Vector Machines	0.8035	0.5899
Neural Networks	0.6815	0.4118

Rule-Based Classifier

Criterion	Accuracy
ABDOMINAL	0.8837
ADVANCED_CAD	0.7906
ALCOHOL_ABUSE	0.9534
ASP_FOR_MI	0.8604
CREATININE	0.8372
DIETSUPP_2MOS	0.9186
DRUG_ABUSE	0.9651
ENGLISH	0.9418
HBA1C	0.9418
KETO_1YR	1.0
MAJOR_DIABETES	0.8372
MAKES_DECISIONS	0.9651
MI_6MOS	0.9651
OVERALL_MICRO	0.9123
OVERALL_MACRO	0.9123

** TRACK 1 ***

	met				not met			overall	
	Prec.	Rec.	Speci.	F(b=1)	Prec.	Rec.	F(b=1)	F(b=1)	AUC
Abdominal	0.8333	0.8333	0.9107	0.8333	0.9107	0.9107	0.9107	0.8720	0.8720
Advanced-cad	0.8000	0.8000	0.7805	0.8000	0.7805	0.7805	0.7805	0.7902	0.7902
Alcohol-abuse	0.0000	0.0000	0.9880	0.0000	0.9647	0.9880	0.9762	0.4881	0.4940
Asp-for-mi	0.8500	1.0000	0.3333	0.9189	1.0000	0.3333	0.5000	0.7095	0.6667
Creatinine	0.6786	0.7917	0.8548	0.7308	0.9138	0.8548	0.8833	0.8071	0.8233
Dietsupp-2mos	0.9111	0.9318	0.9048	0.9213	0.9268	0.9048	0.9157	0.9185	0.9183
Drug-abuse	0.5000	0.3333	0.9880	0.4000	0.9762	0.9880	0.9820	0.6910	0.6606
English	0.9359	1.0000	0.6154	0.9669	1.0000	0.6154	0.7619	0.8644	0.8077
Hba1c	1.0000	0.8571	1.0000	0.9231	0.9107	1.0000	0.9533	0.9382	0.9286
Keto-1yr	0.0000	0.0000	1.0000	0.0000	1.0000	1.0000	1.0000	0.5000	0.5000
Major-diabetes	0.8085	0.8837	0.7907	0.8444	0.8718	0.7907	0.8293	0.8369	0.8372
Makes-decisions	0.9651	1.0000	0.0000	0.9822	0.0000	0.0000	0.0000	0.4911	0.5000
Mi-6mos	1.0000	0.6250	1.0000	0.7692	0.9630	1.0000	0.9811	0.8752	0.8125
Overall (micro)	0.8784	0.9129	0.9120	0.8953	0.9376	0.9120	0.9246	0.9100	0.9124
Overall (macro)	0.7140	0.6966	0.7820	0.6993	0.8629	0.7820	0.8057	0.7525	0.7393

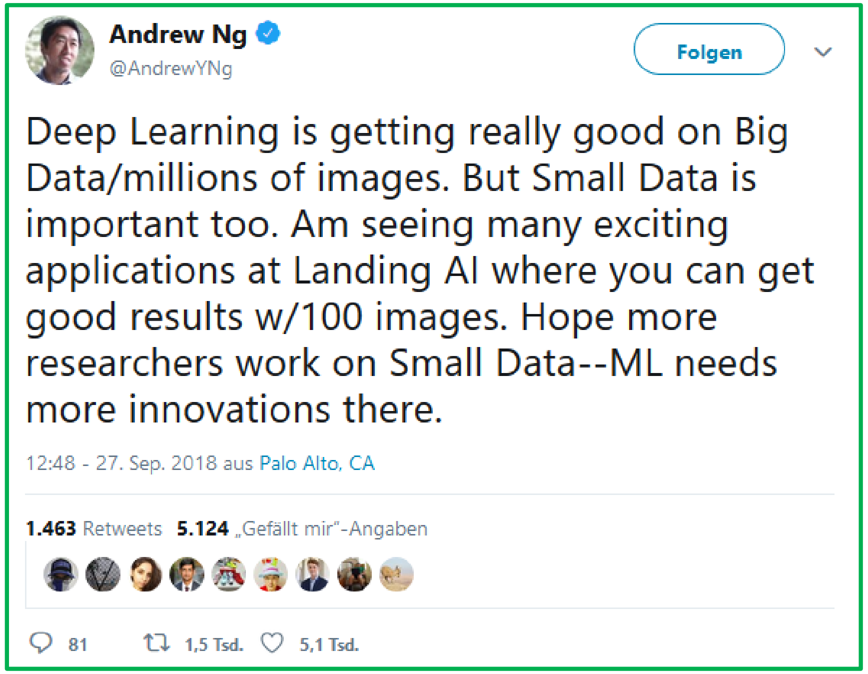
86 files found

Discussion

- What can be improved?
  - not enough data for Neural Networks (?)
  - better negation detection
- What was planned, but not implemented?
  - enhanced feature engineering
  - table processing

Conclusion

- better understanding of NNs used on small datasets
- rule based approach performed best



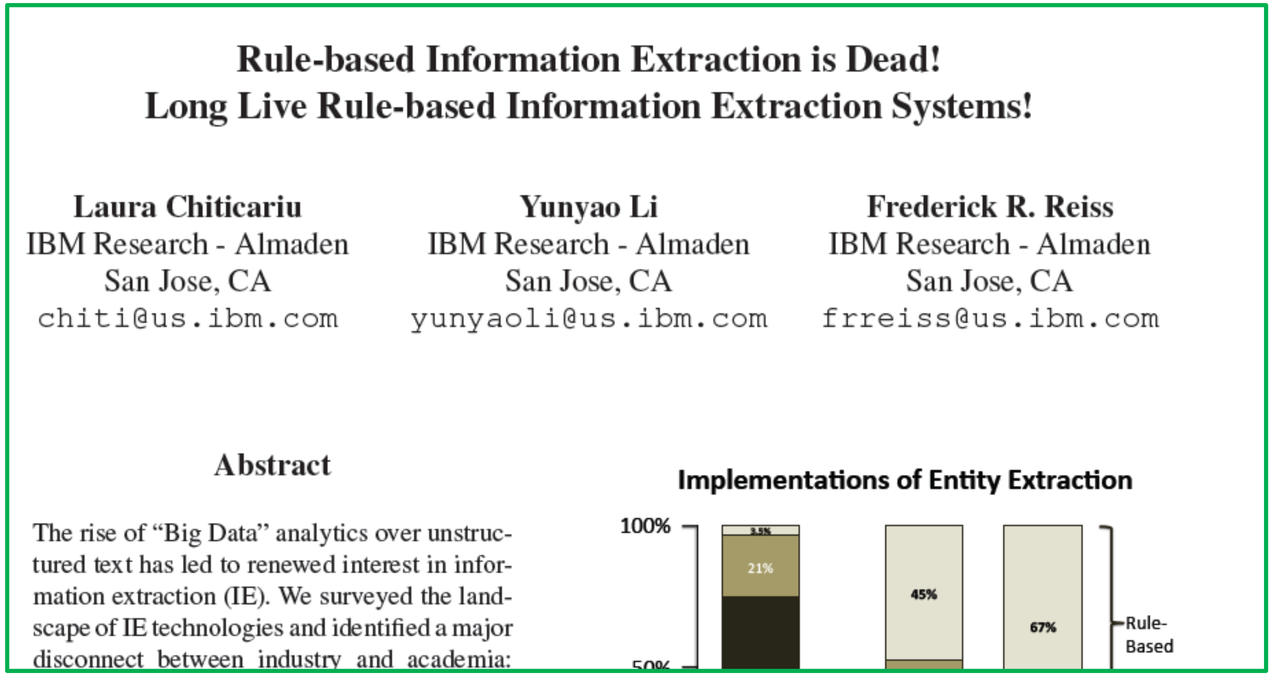
Andrew Ng
@AndrewYNg
Folgen

Deep Learning is getting really good on Big Data/millions of images. But Small Data is important too. Am seeing many exciting applications at Landing AI where you can get good results w/100 images. Hope more researchers work on Small Data--ML needs more innovations there.

12:48 - 27. Sep. 2018 aus Palo Alto, CA

1.463 Retweets 5.124 „Gefällt mir"-Angaben

81 1,5 Tsd. 5,1 Tsd.



Rule-based Information Extraction is Dead!
Long Live Rule-based Information Extraction Systems!

Laura Chiticariu, IBM Research - Almaden, San Jose, CA, chiti@us.ibm.com
Yunyao Li, IBM Research - Almaden, San Jose, CA, yunyaoli@us.ibm.com
Frederick R. Reiss, IBM Research - Almaden, San Jose, CA, frreiss@us.ibm.com

Abstract

The rise of "Big Data" analytics over unstructured text has led to renewed interest in information extraction (IE). We surveyed the landscape of IE technologies and identified a major disconnect between industry and academia:

Medical University of Graz (Austria)

Institute for Medical Informatics, Statistics and Documentation

Michel Oleynik, ***Amila Kugic***, Markus Kreuzthaler,

Zdenko Kasáč, **Stefan Schulz**

Amila Kugic

Computer Scientist

amila.kugic@gmail.com