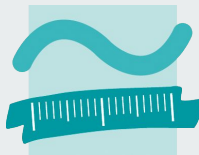


Neural Machine Reading for Domain-Specific Text Resources

Sebastian Arnold
sarnold@beuth-hochschule.de

Prof. Alexander Löser
Prof. Felix A. Gers
Prof. Philippe Cudré-Mauroux

28.10.2020



BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN

University of Applied Sciences

Data Science and Text-based
Information Systems (DATEXIS)
Beuth University of Applied Sciences
Berlin, Germany



UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Department of Informatics
Faculty of Science and Medicine
University of Fribourg
Fribourg, Switzerland

About Me

- 2006–2011 **Technische Universität Berlin**
Bachelor of Science in Computer Science
- 2011–2015 **Technische Universität Berlin**
Master of Science in Computer Science
Student Assistant @ DIMA, Prof. Volker Markl
- 2015–2020 **Beuth University of Applied Sciences Berlin**
Research Assistant @ DATEXIS, Prof. Alexander Löser
- 2016–2020 **Université de Fribourg (CH)**
Doctoral Program in Computer Science
PhD Student @ eXascale Infolab, Prof. Cudré-Mauroux
- since 07/2020 **Curalie GmbH, Berlin**
Machine Learning Expert



[@sebastianarnold](https://twitter.com/sebastianarnold)

PhD Timeline & Main Contributions

“Neural Machine Reading for Domain-Specific Text Resources” (since 06/2015)

Supervisors:

[Prof. Alexander Löser](#)

Beuth University of Applied Sciences Berlin (D)

[Prof. Philippe Cudré-Mauroux](#)

eXascale Infolab, University of Fribourg (CH)

Advisor:

[Prof. Felix A. Gers](#) DATEXIS group @ Beuth

External examiner:

[Prof. Laura Dietz](#) Univ. of New Hampshire (US)



Publications:

DOLAP Workshop @ CIKM 2015 Melbourne, **COLING Demos** 2016 Osaka, **WWW Demos** 2018 Lyon, **TACL** @ ACL2019 Florence, **WWW 2020** Taipei

Reviewer for:

WWW'18/19/20, CIKM'20, ACL'18/19, AAAI'19, EMNLP'19, ESSLLI'18

Talks:

German-Canadian Concourse @ Charité, Quo Vadis 2015, Zalando Meetup, Lange Nacht der Wissenschaften 2016, Holtzbrink AI Day 2017 @ SpringerNature, Dagstuhl Seminar 2019, Graduate School Data Science @ Einstein Center Berlin

Industry Projects:

Siemens, SAP, Zalando SE, SpringerNature, Mobile.de, Charité, BMWi, H2020

Open Source:

<https://github.com/sebastianarnold/>



Outline

1. Introduction
2. Background
3. **TASTY – A Robust Model for Efficient Entity Linking**
4. **SECTOR – Coherent Topic Segmentation and Classification**
5. **CDV – Contextual Document Representations for Answer Retrieval**
6. Systems
7. Conclusion and Future Work



*“Die Natur muß gefühlt werden, wer nur sieht und abstrahirt,
kann ein Menschenalter, im Lebensgedränge der glühenden
Tropenwelt, Pflanzen und Thiere zergliedern, er wird die Natur
zu beschreiben glauben, ihr selbst aber ewig fremd sein.”*

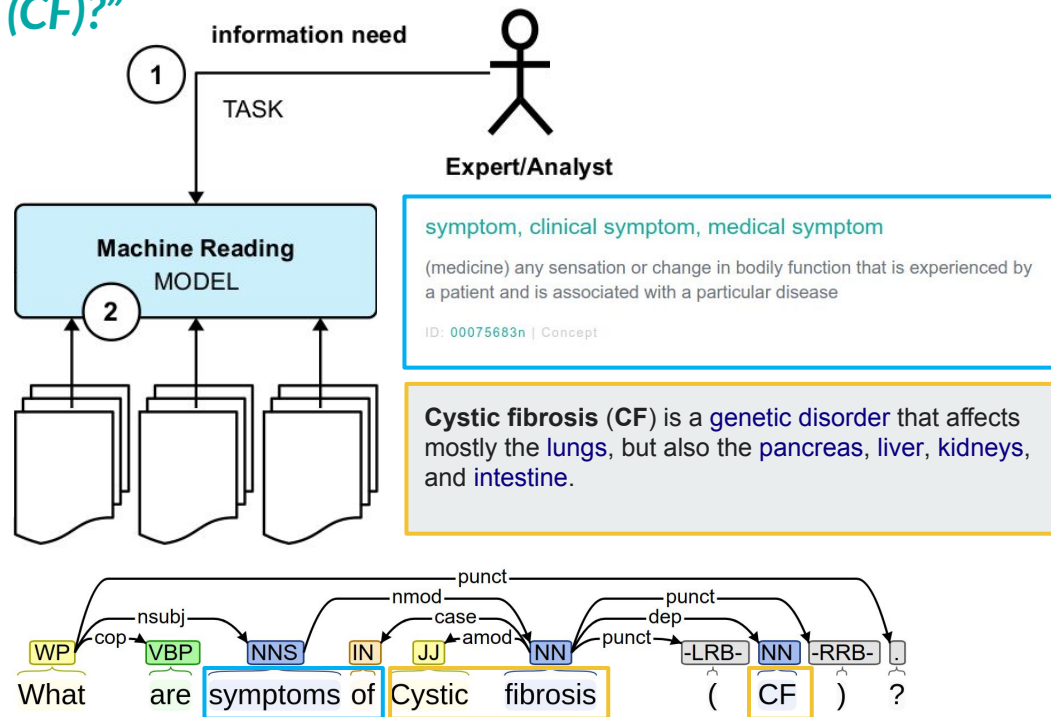
— Alexander von Humboldt, an Johann Wolfgang von Goethe, Paris 3. Januar 1810.

1. Introduction

Example: Information Search

“What are symptoms of Cystic fibrosis (CF)?”

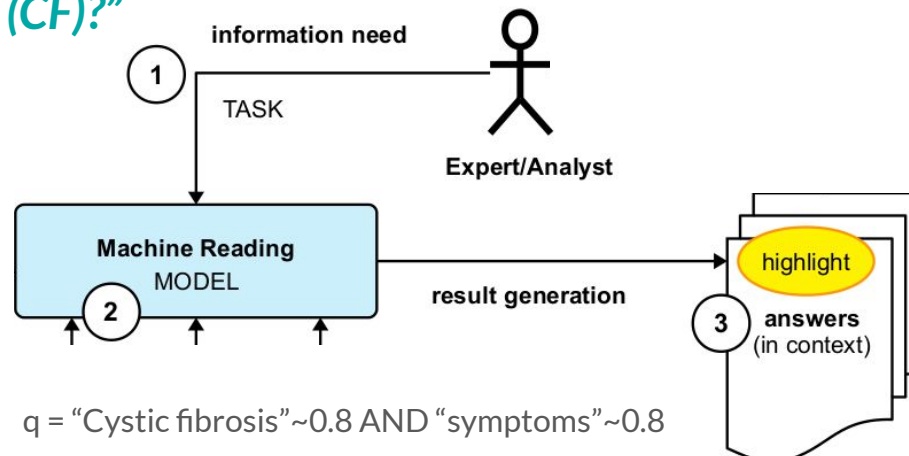
- Query intent classification
[Rose+Levinson 2004]
- Natural Language Processing
[Jurafsky 2008]
- Information Extraction (IE)
[Sarawagi 2008]
- Semantic representation
[Moro 2014]



Knowledge Processing and Retrieval

“What are symptoms of Cystic fibrosis (CF)?”

- Question Answering (QA)
- Information Retrieval (IR)
[Manning 2008]
- Feedback Loop
- Exploration and refinement
[Marchionini 2006]



Cystic fibrosis (CF) is a genetic disorder that causes mucus to build up and damage organs in the body, particularly the lungs and pancreas. Signs and symptoms may include salty-tasting skin; persistent coughing; frequent lung infections; wheezing or shortness of breath; poor growth; weight loss; greasy, bulky stools; difficulty with bowel movements; and in males, infertility. Over time, mucus buildup and infections can lead to permanent lung damage, including the formation of scar tissue (fibrosis) and cysts in the lungs. CF is caused by mutations in the *CFTR* gene and inheritance is autosomal recessive.^{[1][2][3]} Treatment aims to relieve symptoms and usually includes respiratory therapies, inhaled medicines, pancreatic enzyme supplement, and nutritional supplements. Newer medications such as *CFTR* modulators have been approved for use in the United States. Ongoing research is focused on finding a cure for the disease.^[2]

GoOLAP: Answering Analytical Queries with Tables [2012]

go_olap beta
Search. Explore. Decide.

current

EXPLORE **LIST** ← select your result

Organization **White House** position/employments of **White House**

We found 17,060 facts about the Organization **White House** in the World Wide Web.

Similar objects to White House

- [Whitehouse](#) (Company, 9 facts)
- [Whitehouse](#) (Person, 8 facts)
- [Whitehouse](#) (City, 3 facts)
- [White House](#) (Company, 1 fact)

Relations in this result

- generic relations (9371)
- ✓ position/employments (7256)
- ✓ communications (288)
- ✓ diplomatic relations (131)

position/employments of White House showing 1 to 10 of 3680 entries prev | next

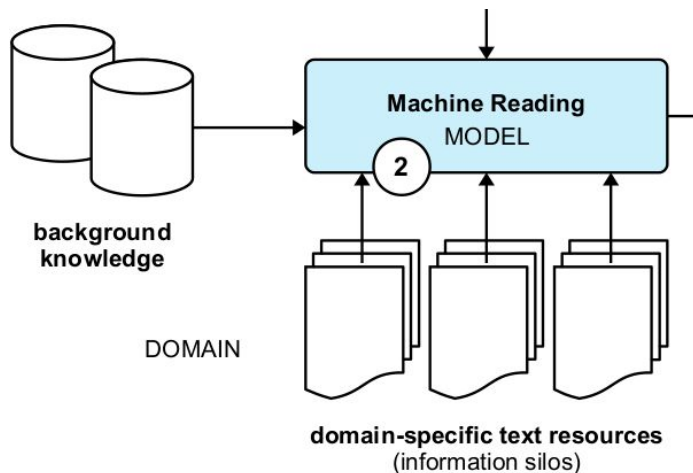
	country	city	person	organization	company	province/state	position
1			Robert Gibbs	White House			press secretar
2			Jay Carney	White House			press secretar
3			Rahm Emanuel	White House			chief of staff
4			Scott McClellan	White House			press secretar
5			Robert Gibbs	White House			spokesman
6			Jay Carney	White House			spokesman
7			Alberto Gonzales	White House			counsel
8			John Dean	White House			counsel
9			Scott McClellan	White House			spokesman
10			James Brady	White House			press secretar

communications of White House showing 1 to 10 of 219 entries prev | next

	date	person1	persondescription1	person	organizationorcompany	persondescription	facility
1	1993-01-27				White House	small delegation	
2				Eleanor Statues	White House		
3				Lloyd Blankfein	White House		
4					White House	CIA Director	
5				John S. McCain, Jr.	White House		
6				Ehud Olmert	White House		
7				Douglas	White House		

Vision: Machine Reading [Etzioni 2006]

MR is “the automatic, unsupervised understanding of text [achieved by] the formation of a coherent set of beliefs based on a **textual corpus** and a **background theory**.”



Our goal: Support the human information-seeking process with Machine Reading.

Challenges for Domain-specific Language Understanding

Traditional IE pipelines depend on task-specific training data or handcrafted rules.

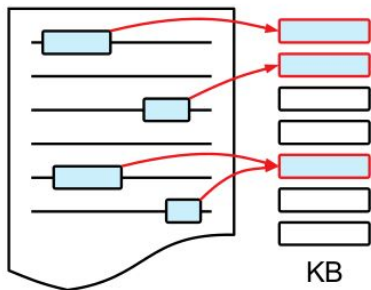
- Domain-specific language → adaptiveness and transferability
- Variations and noise → robustness, focus on recall
- Heterogenous document structure → coherence, local context sensitivity
- High task variance → generalization, multi-task, zero-shot
- Insufficient training data → efficiency, self-supervision, background knowledge
- Error propagation → end-to-end design, differentiable models, dense representations



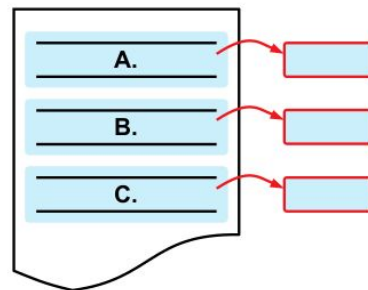
Hypothesis: Neural Machine Reading

1. **Deep Neural Networks** enable **self-supervised training** of language models based on distributional information.
2. **Machine Reading** increases error tolerance and reduces adaptation cost for reading **domain-specific text**.
3. **End-to-end models** combine general training objectives with background knowledge to fulfill **task-specific information needs**.

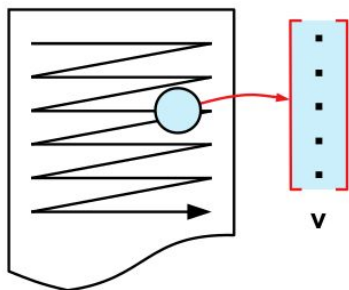
Research Questions



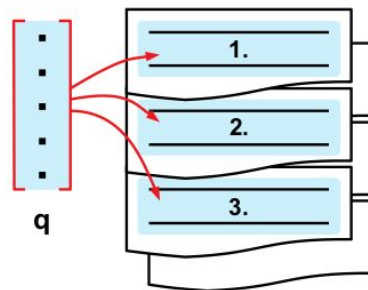
RQ1: Identify domain-specific named entities.



RQ2: Detect topics and structure in long documents.



RQ3: Embed discourse structure into document representations.



RQ4: Retrieve answer passages using vector representations.

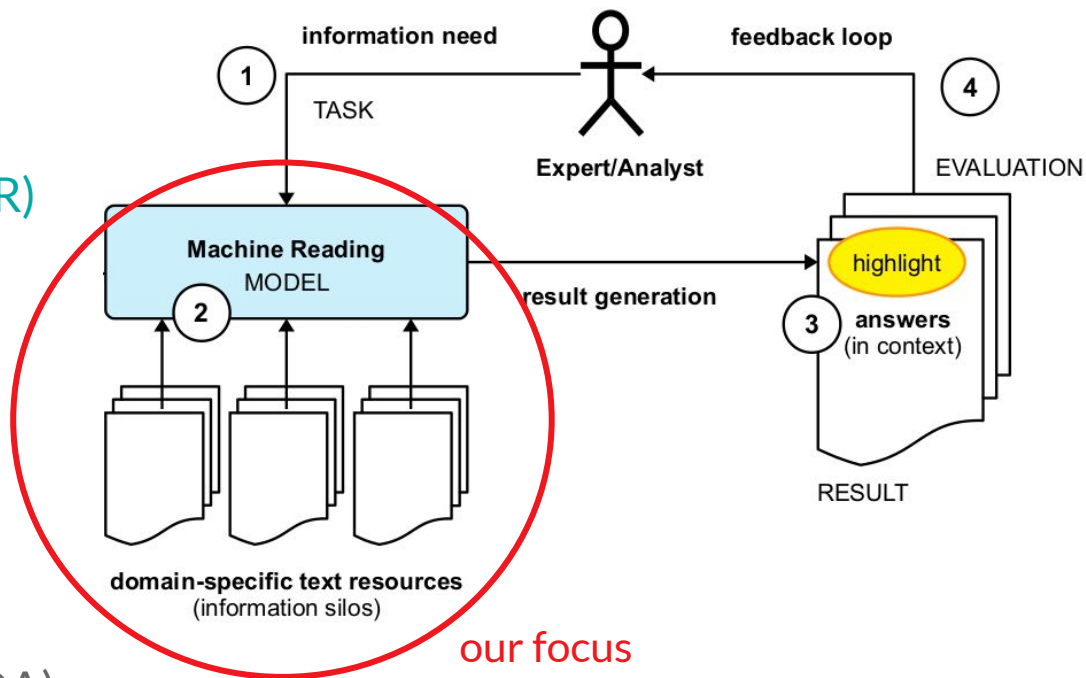
Focus of the Thesis

Tasks covered in this thesis:

- Named Entity Recognition (NER)
- Named Entity Linking (NEL)
- Topic Modeling
- Answer Passage Retrieval

Tasks NOT covered:

- Intent detection
- Relation Extraction (RE)
- Factoid Question Answering (QA)
- Machine Reading Comprehension (MRC)
- Interactive feedback loop



Main Contributions and Publications

TASTY – Named Entity Recognition and Linking

S. Arnold, F. A. Gers, T. Kiliyas and A. Löser (2016).

Robust Named Entity Recognition in Idiosyncratic Domains. **arXiv:1608.06757 [cs.CL]**

S. Arnold, R. Dziuba and A. Löser (2016).

TASTY: Interactive Entity Linking As-You-Type. **COLING 2016 (Demos)**

SECTOR – Topic Segmentation and Classification

R. Schneider, S. Arnold, T. Oberhauser, T. Klatt, T. Steffek and A. Löser (2018).

Smart-MD: Neural Paragraph Retrieval of Medical Topics. **WWW 2018 (Companion)**: 203–206

S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers and A. Löser (2019).

SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. **TACL Vol. 7**: 169–184 (presented at ACL)

CDV – Contextual Discourse Vectors for Answer Retrieval

S. Arnold, B. van Aken, P. Grundmann, F. A. Gers and A. Löser (2020).

Learning Contextualized Document Representations for Healthcare Answer Retrieval. **WWW 2020**: 1332–1343

J.-M. Papaioannou, S. Arnold, F. A. Gers, A. Löser, M. Mayrdorfer, and K. Budde (2021).

Aspect-Based Passage Retrieval with Contextualized Discourse Vectors. **Submitted to: ECIR 2021 System Demonstrations**

TeXoo – A Zoo of Text Extractors. Apache V2 License: <https://github.com/sebastianarnold/TeXoo>

2. Background

Distributional Hypothesis [Harris 1954]

	Paradigmatic relations Selections: “ x or y or...”			
Syntagmatic relations Combinations: “ x and y and...”	she	adores	green	paint
	he	likes	blue	dye
	they	love	red	colour

[Sahlgren 2008]



- **Neural Probabilistic Language Model:**

Learn to predict words based on their context [Bengio 2003]

$$p(w_1, \dots, w_T) = \prod_{t=1}^T p(w_t \mid w_1, \dots, w_{t-1})$$

- The **Skip-gram model** solves this problem more efficiently [Mikolov 2013]

$$\sum_{t=1}^T \sum_{c \in C_t} \log p(w_c \mid w_t)$$

Harris (1954): **Distributional Structure**. WORD Vol. 10, No. 2-3

Sahlgren (2008): **The Distributional Hypothesis**. Italian Journal of Linguistics 20, 33-54.

Bengio et al. (2003): **A Neural Probabilistic Language Model**. Journal of Machine Learning Research 3.

Mikolov et al. (2013): **Efficient Estimation of Word Representations in Vector Space**. arXiv:1301.3781

3. TASTY – A Robust Model for Efficient Entity Linking

S. Arnold, F. A. Gers, T. Kiliyas and A. Löser (2016).

Robust Named Entity Recognition in Idiosyncratic Domains. [arXiv:1608.06757 \[cs.CL\]](#)

S. Arnold, R. Dziuba and A. Löser (2016).

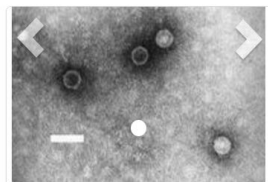
TASTY: Interactive Entity Linking As-You-Type. **COLING 2016 (Demos)**

Challenge: Recognize and Link Named Entities

RQ1: What are general solutions to identify named entities in domain-specific text?

- broad coverage
- domain and language-independent architecture
- focus on recall
- high robustness
- efficient training

Expression of the chemokine receptor BLR2/EBI1 is specifically Epstein-Barr virus nuclear antigen 2. In our attempt to identify ch are related to Burkitt's lymphoma receptor 1 (BLR1) and are ex lymphocytes we used RT-PCR resulting in the isolation of a cDN transmembrane receptor termed BLR2. The protein shows signil similarities to the family of G-protein coupled chemokine recepto identical to the recently described receptor EBI1. Northern blot a BLR2 mRNA could be highly stimulated in mitogen- and anti-Cr blood lymphocytes. BLR2-specific mRNA could be detected in a positive B cell lines. We show that transcription of the BLR2 gen induced in Epstein-Barr virus negative BL 41 cells via estrogen-r Epstein-Barr virus nuclear antigen 2, a key regulator of viral and immortalized B cells. Our data suggest an involvement of BLR2 migration in activated lymphocytes and in viral pathogenesis.

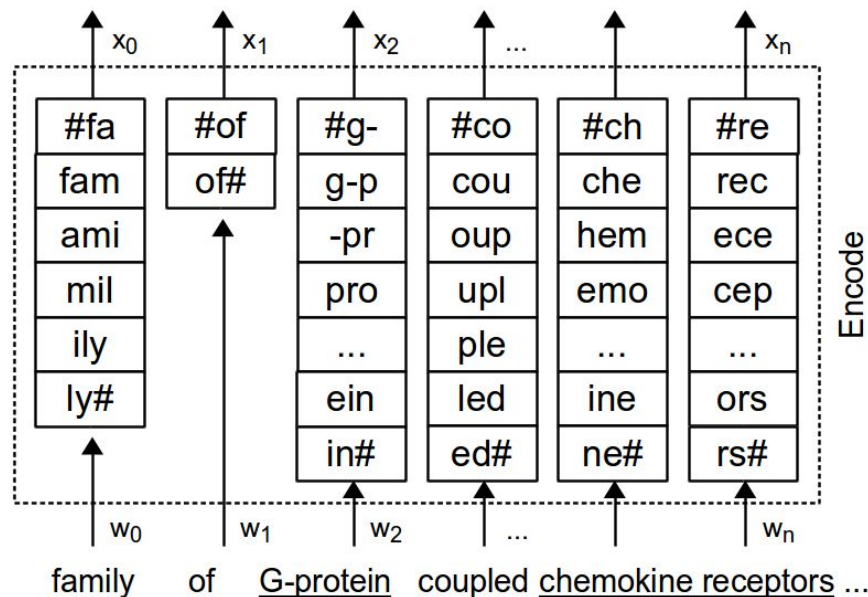


Epstein-Barr virus nuclear antigen 2

The Epstein-Barr virus nuclear antigen 2 (EBNA-2) is a viral protein associated with the Epstein-Barr virus. EBNA-2 is the main viral transactivator, switching

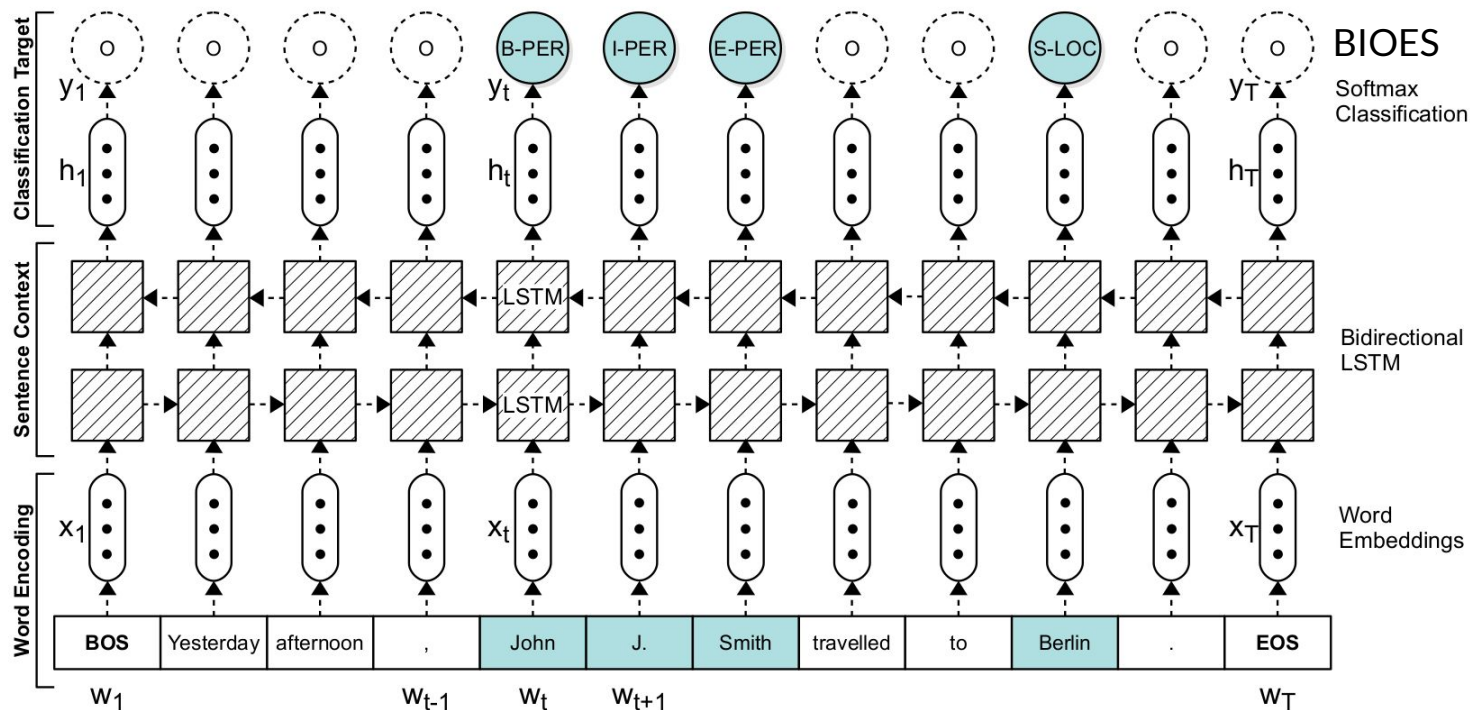
Extracting entities requires local, contextual and global features.

Encoding Local Sub-word Information with Letter-trigrams



Character-based word representations [Huang 2013] are a key component for robustness.

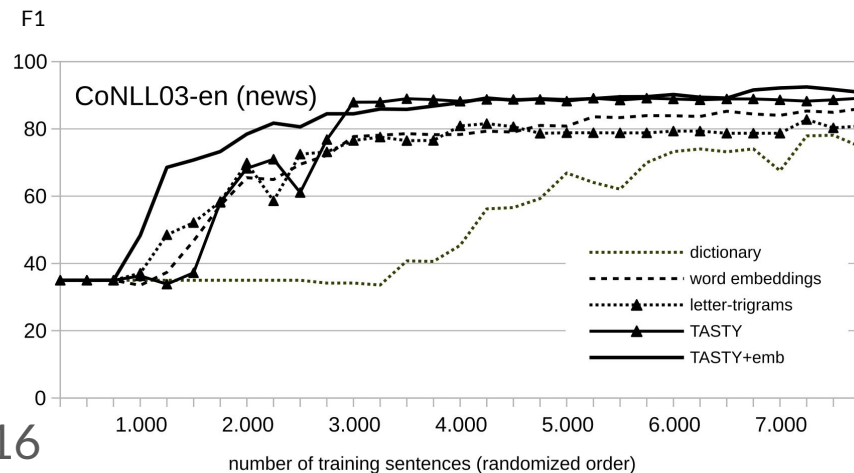
Capturing Sentence Context with Sequence Learning



Bidirectional LSTM [Hochr. 1997, Graves 2012] effectively captures long-range dependencies.

Key Results: TASTY

- TASTY is a **robust end-to-end model** for Entity Linking (NER+NEL)
- **Pushes down the costs** for a domain-specific model to labeling 4K sentences
- State-of-the-art performance for English and German NER: **91.1% F1 on CoNLL 2003** beats all models before 2016



Next steps on our journey towards Neural Machine Reading:

- We envision a general document understanding beyond isolated entities
- Model training should require as little supervision as possible

4. SECTOR – Coherent Topic Segmentation and Classification

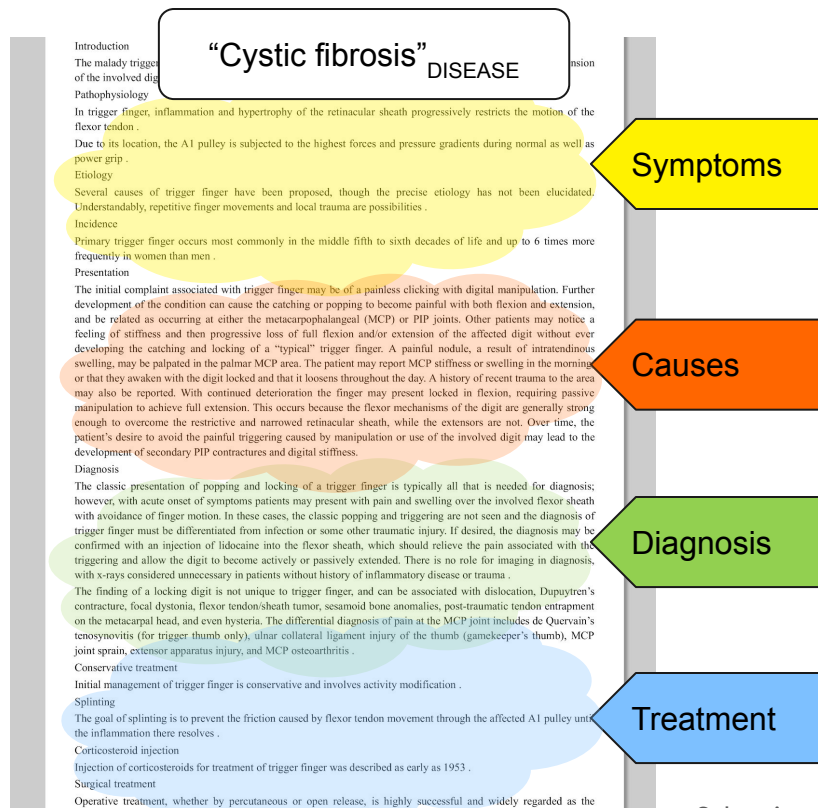
S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers and A. Löser (2019).
SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. **TACL Vol. 7**: 169–184 (presented at ACL)

Challenge: Understand Topics and Structure of a Document

RQ2: How can Machine Reading models **detect topics and structure** in long documents?

- topical information
- structural information
- coherent predictions
- sentence granularity

Distributional Hypothesis needs to be extended with **long-range structural information.**



Latent Topics can be Learned from Wikipedia Headings

Contents [hide]
1 Signs and symptoms
1.1 Diabetic emergencies
1.2 Complications
2 Causes
2.1 Type 1
2.2 Type 2
2.3 Gestational diabetes
2.4 Maturity onset diabetes of the young
2.5 Other types
3 Pathophysiology
4 Diagnosis
5 Prevention
6 Management
6.1 Lifestyle
6.2 Medications
6.3 Surgery
6.4 Support
7 Epidemiology
8 History
8.1 Etymology
9 Society and culture
9.1 Naming
10 Other animals
11 Research
12 References
13 Further reading
14 External links

preprocessing
and pruning

(from 8.5k
'noisy' labels,
bag-of-words)

en_disease (1.5k)

diagnosis
treatment
symptoms
signs
causes
epidemiology
management
cause
prognosis
history
pathophysiology
classification
prevention
genetics
research
types
culture
factors
mechanism
society
presentation
risk
differential
disease
surgery
therapy
clinical
pathogenesis

de_disease (1.0k)

therapie
symptome
behandlung
diagnose
ursachen
klinische
ursache
diagnostik
verbreitung
erscheinungen
einteilung
geschichte
epidemiologie
differentialdiag
prognose
klassifikation
formen
verlauf
haeufigkeit
vorbeugung
aetiologie
pathogenese
klinisches
pathologie
bild
klinik
entstehung
symptomatik

en_disease (27)

treatment
symptom
diagnosis
cause
classification
epidemiology
history
prognosis
management
pathophysiology
mechanism
prevention
research
genetics
tomography
culture
etymology
infection
fauna
risk
pathology
surgery
screening
medication
geography
complication
other

de_disease (25)

therapie
diagnose
symptom
ursache
kategorisierung
verlauf
epidemiologie
geschichte
prognose
praevalenz
vorbeugung
fauna
terminologie
pathologie
definition
klinik
komplikation
genetik
infektion
risiko
forschung
geographie
mensch
organe
sonstiges

clustering /
normalization
via BabelNet

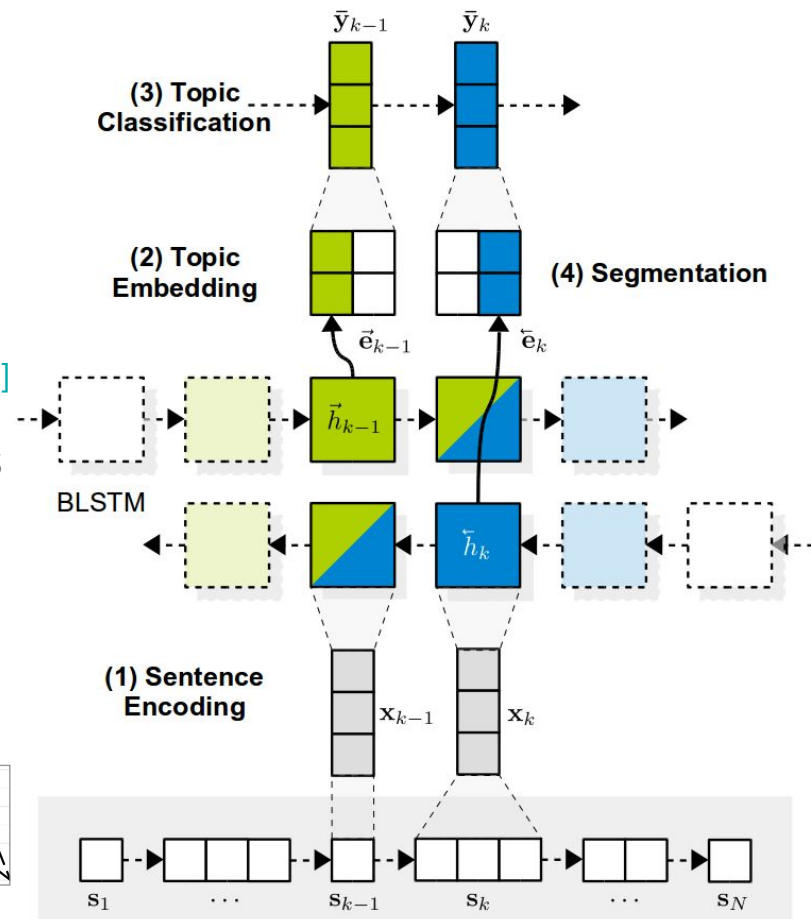
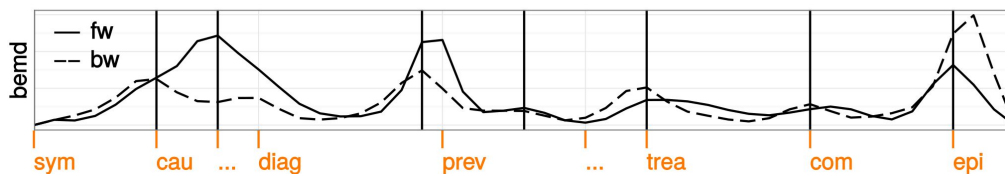
[Navigli 2012]

(85–95%
coverage)

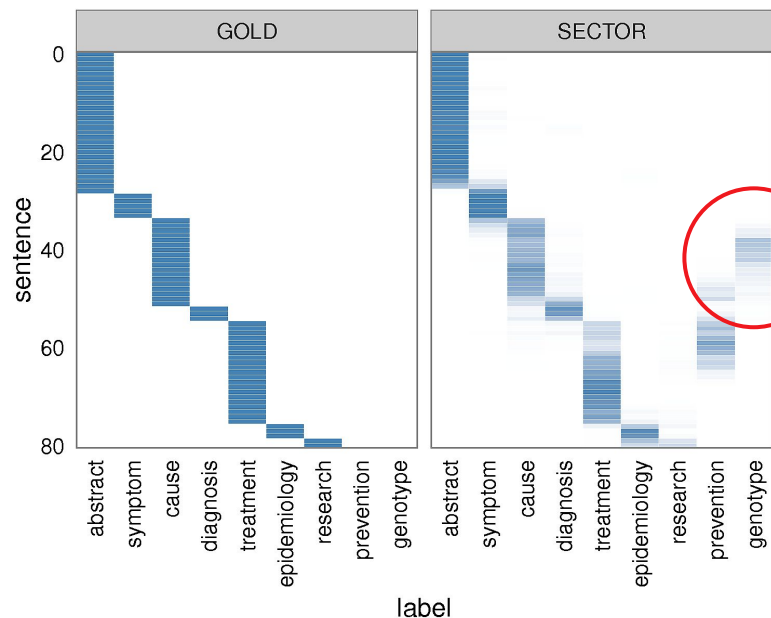
Capturing Entire Documents with Sentence Embeddings + BLSTM

Objective: predict topics on sentence level

- Bloom filter sentence compression [Serra 2017]
- Single-label classification into 25–30 topics
- Multi-label classification up to 603 words
- Segmentation based on deviation of the hidden layer topic embedding



SECTOR Prediction on Par with Wiki Authors for “Dermatitis”



Browse history ▼

Revision as of 02:24, 31 March 2017 (edit)
liibalesiii (talk | contribs)
(→Cause)
← Previous edit

Revision as of 02:27, 31 March 2017 (edit) (undo)
liibalesiii (talk | contribs)
(Causes was reorganized w/ better flow.)
(Tag: Visual edit)
Next edit →

Line 35:

== Causes ==

Many people with AD have a family history of [[atopy]]. Atopy is an immediate-onset allergic reaction (type 1 hypersensitivity reaction) as asthma, food allergies, AD or hay fever.<ref name = AFP/><ref name = MSR/> In 2006 it was discovered that [[mutation]]s in the gene for the production of [[filaggrin]] strongly increased the risk for developing atopic dermatitis. Most importantly two mutations were found that affect approximately 5% of people in Western Europe that may disrupt the production of filaggrin. Filaggrin is a protein that plays an important role in the retention of water in the [[stratum corneum]]. People who have these mutations often have [[xeroderma|dry

Line 35:

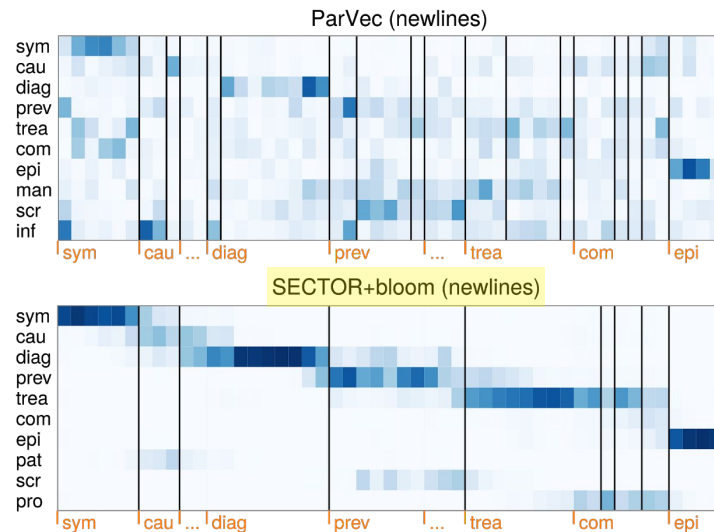
== Causes ==

=== Genetic ===

Source: https://en.wikipedia.org/w/index.php?title=Atopic_dermatitis&diff=786969806&oldid=772576326

Key Results: SECTOR

- Extends distributional models (top) with **topic and structure information** (bottom)
- Classifies **25–30 local topics** with up to 71.6% F1
- **Detects topic shifts** to segment long documents into coherent passages



Next steps on our journey towards Neural Machine Reading:

- We need to cover complementary information from named entities
- Hard passage segmentation should be replaced by continuous representations

5. CDV – Contextual Document Representations for Answer Retrieval

S. Arnold, B. van Aken, P. Grundmann, F. A. Gers and A. Löser (2020).
Learning Contextualized Document Representations for Healthcare Answer Retrieval. **WWW 2020**: 1332–1343

Challenge: Capture Document Discourse Structure

RQ3: How can we embed discourse structure into document representations?

Entities and aspects provide discourse information complementary to language models.

“Cystic fibrosis”

“Therapy”

of a locking digit is not unique to trigger finger, and can be associated with dislocation, Dupuytren's contracture, focal dystonia, flexor tendon/sheath tumor, sesamoid bone anomalies, post-traumatic tendon entrapment, carpal head, and even hysteria. The differential diagnosis of pain at the MCP joint includes de Quervain's tenosynovitis (for trigger thumb only), ulnar collateral ligament injury of the thumb (gamekeeper's thumb), MCP joint space, extensor apparatus injury, and MCP osteoarthritis .

treatment

Management of trigger finger is conservative and involves activity modification .

splinting is to prevent the friction caused by flexor tendon movement through the affected A1 pulley until inflammation there resolves .

injection

Corticosteroids for treatment of trigger finger was described as early as 1953 .

Surgical treatment

Operative treatment, whether by percutaneous or open release, is highly successful and widely regarded as the ultimate treatment for trigger finger. Indication for surgical treatment is generally failure of conservative treatment to resolve pain and symptoms. The timing of surgery is somewhat controversial with data suggesting surgical consideration after failure of both a single as well as multiple corticosteroid injections .

The percutaneous trigger finger release has been described and was first introduced by Lorthioir in 1958 .

Open release of trigger finger has been used as treatment for over a century .

Conclusion

Trigger finger is a long recognized condition characterized by a sometimes painful locking of the digit on flexion and extension. It is caused by the inflammation and subsequent narrowing of the A1 pulley through which the flexor

Hand therapy

In trigger finger, inflammation and hypertrophy of the retinacular sheath progressively irritates the tendon of the

conformed with an injection of lidocaine into the flexor sheath, which should relieve the pain associated with the triggering and allow the digit to become active or passively extended. There is no risk for triggering in diagnosis.

The timing of a trigger finger is not unique to trigger finger and can be associated with arthritis, rheumatoid arthritis, focal dystonia, flexor tendon/sheath tumor, sesamoid bone anomalies, post-traumatic tendon entrapment on the metacarpal head, and even hysteria. The differential diagnosis of pain at the MCP joint includes de Quervain's tenosynovitis (for trigger thumb only), ulnar collateral ligament injury of the thumb (gamekeeper's thumb), MCP joint space, extensor apparatus injury, and MCP osteoarthritis.

Conservative treatment

Initial management of trigger finger is conservative and involves activity modification.

Splinting

The goal of splinting is to prevent the friction caused by flexor tendon movement through the affected A1 pulley until inflammation there resolves.

Corticosteroid injection

Corticosteroids for treatment of trigger finger was described as early as 1953.

Surgical treatment

Operative treatment, whether by percutaneous or open release, is highly successful and widely regarded as the ultimate treatment for trigger finger. Indication for surgical treatment is generally failure of conservative treatment to resolve pain and symptoms. The timing of surgery is somewhat controversial with data suggesting surgical consideration after failure of both a single as well as multiple corticosteroid injections.

The percutaneous trigger finger release has been described and was first introduced by Lorthioir in 1958.

Open release of trigger finger has been used as treatment for over a century.

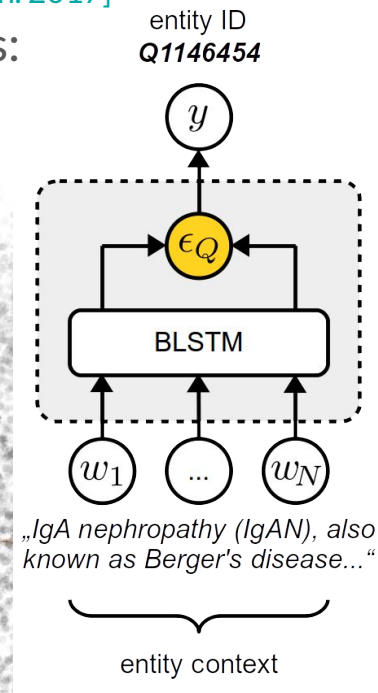
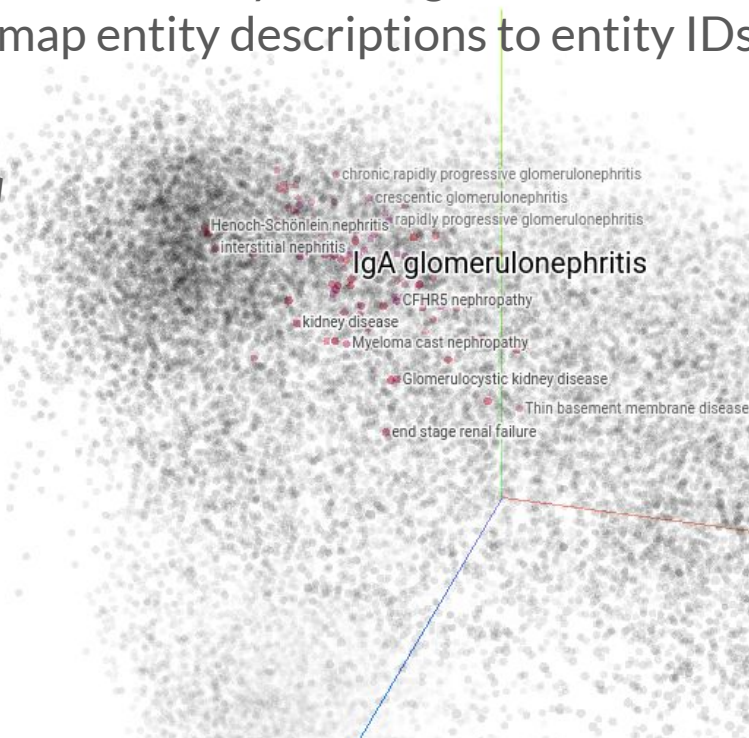
Conclusion

Trigger finger is a long recognized condition characterized by a sometimes painful locking of the digit on flexion and extension. It is caused by the inflammation and subsequent narrowing of the A1 pulley through which the flexor tendon passes. The timing of a trigger finger is not unique to trigger finger and can be associated with arthritis, rheumatoid arthritis, focal dystonia, flexor tendon/sheath tumor, sesamoid bone anomalies, post-traumatic tendon entrapment on the metacarpal head, and even hysteria. The differential diagnosis of pain at the MCP joint includes de Quervain's tenosynovitis (for trigger thumb only), ulnar collateral ligament injury of the thumb (gamekeeper's thumb), MCP joint space, extensor apparatus injury, and MCP osteoarthritis.

Covering Diseases and Health Problems in Vector Space

We create **entity embeddings** for diseases by training a Fasttext [Bojan. 2017] and BLSTM [Palangi 2016] model to map entity descriptions to entity IDs:

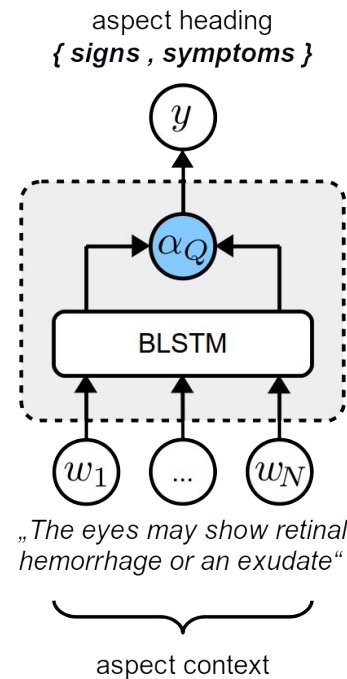
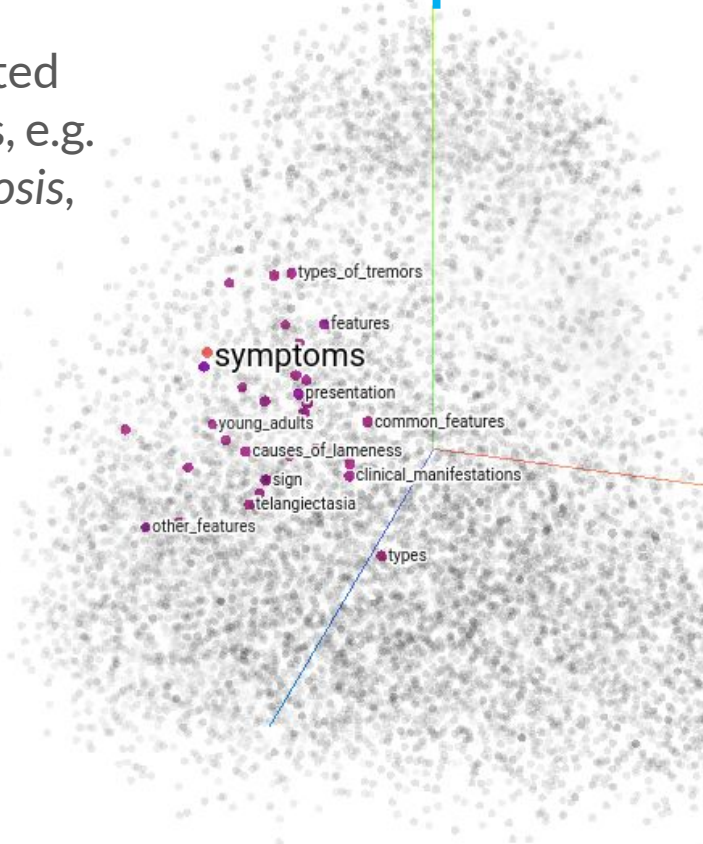
- Descriptions from *UMLS*, *GARD*, *DiseasesDB*, *Wikidata* and *Wikipedia*
- Covers 27,000+ diseases
- Provides fallback encoding for unseen entity names



Distributed Representation of Clinical Aspects

We extend SECTOR with distributed representations of clinical aspects, e.g. *symptoms, treatment, causes, diagnosis, prognosis, prevention, risk factors, ...*

- Covers 14,000+ aspects learned from headings of medical Wikipedia articles

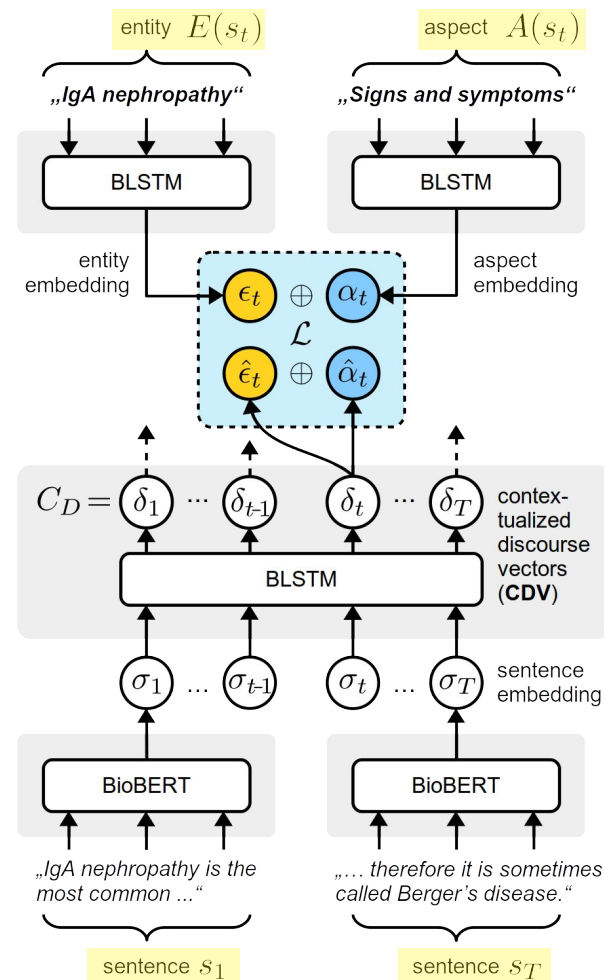


Training CDV with Multi-task Objective

- Goal: Learn discourse embedding to align **entities** $\epsilon_{1..T}$ and **aspects** $\alpha_{1..T}$ with all sentences $s_{1..T}$ in a long document:

$$\mathcal{L}_{\text{cdv}}(\Theta) = \frac{1}{T} \sum_{t=1}^T (\|\hat{\epsilon}_t - \epsilon_t\| + \|\hat{\alpha}_t - \alpha_t\|)$$

s	E(s)	A(s)
s0..1	Arteriosclerosis	description
s2	Arteriosclerosis	signs;symptoms
s3..5	Arteriosclerosis	pathophysiology
s6	Arteriosclerosis;Arteriolo sclerosis	pathophysiology
s7..8	Arteriosclerosis;Atherosclerosis	pathophysiology
s9	Arteriosclerosis;Monckeberg's_arter.	pathophysiology
s10	Arteriosclerosis	pathophysiology
s11	Arteriosclerosis;Hyaline_arterioscl.	pathophysiology
s12	Arteriosclerosis	diagnosis
s13..23	Arteriosclerosis	treatment
s24..26	Arteriosclerosis	epidemiology
s27..28	Arteriosclerosis	history
s29..31	Arteriosclerosis	society;culture



Contextualized Document Representation

- Sentence embeddings σ_t are preprocessed using BioBERT [Lee 2019] or Fasttext [Bojan. 2017]
- Distributional models cannot capture structural context from long documents [Arnold 2019]
- Therefore, we transform $\sigma_{1..T}$ into **contextual discourse vectors** $\delta_{1..T}$ by using BLSTM [Hochr. 1997]

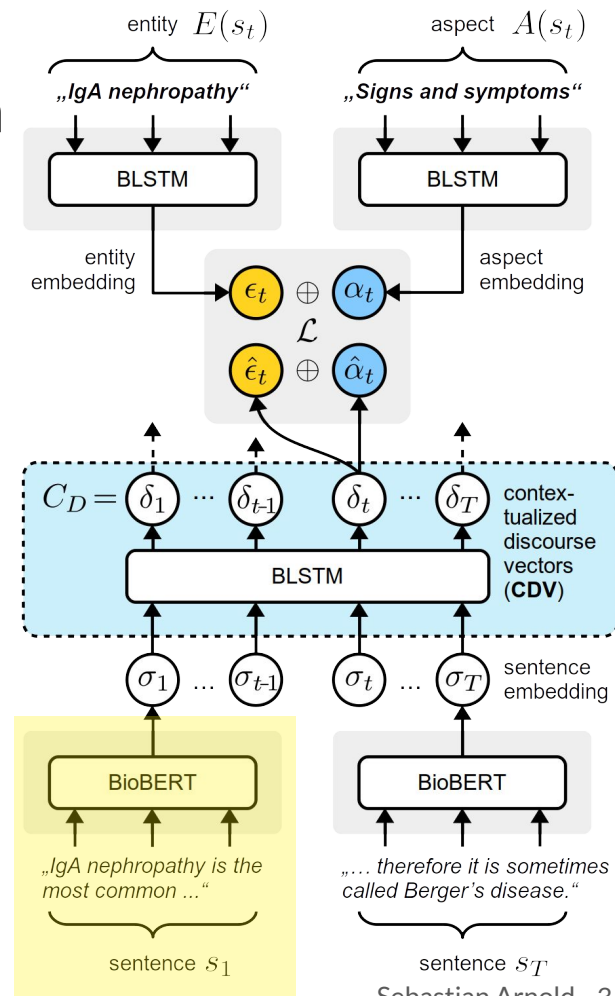
$$\vec{h}_t = \text{LSTM}_{\Theta}(\vec{h}_{t-1}, \sigma(s_t))$$

$$\tilde{h}_t = \text{LSTM}_{\Theta}(\tilde{h}_{t+1}, \sigma(s_t))$$

$$\delta_t = \tanh(W_{he}(\vec{h}_t \oplus \tilde{h}_t) + b_e)$$

$$\hat{\epsilon}_t = \tanh(W_{\delta\epsilon}\delta_t + b_{\epsilon})$$

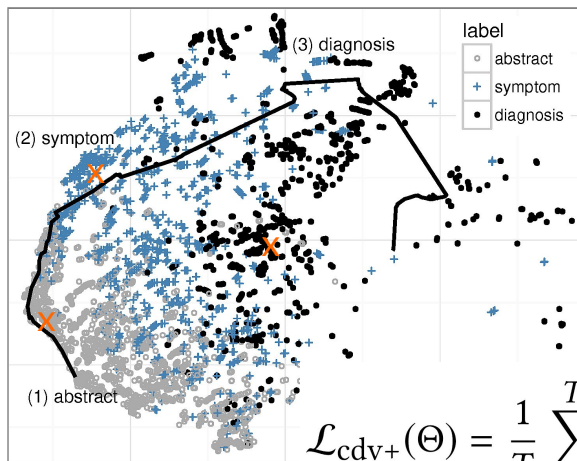
$$\hat{\alpha}_t = \tanh(W_{\delta\alpha}\delta_t + b_{\alpha})$$



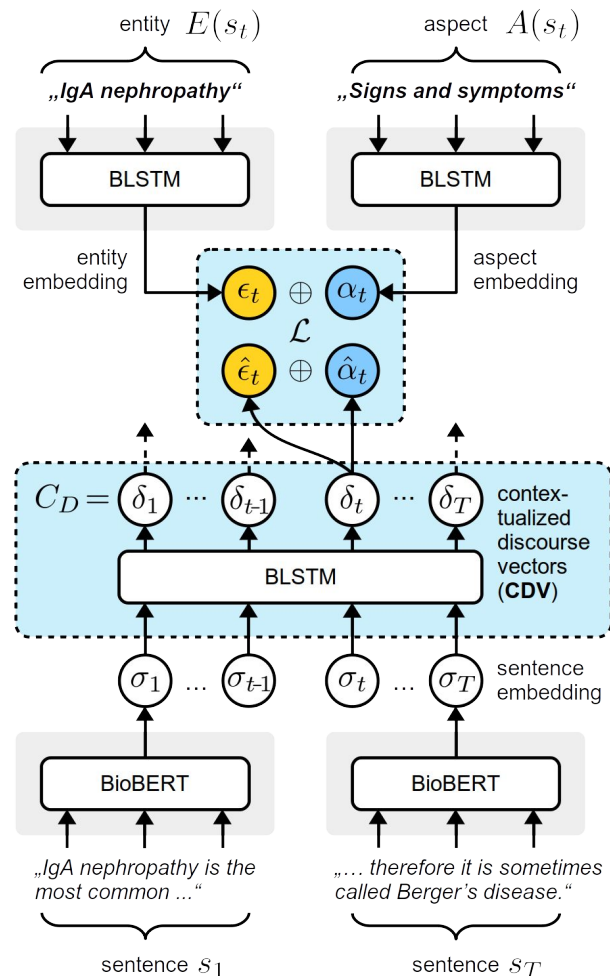
Loss Function

Our loss function needs to handle the large variance of documents.

- We optimize model parameters θ by minimizing Huber loss [Huber 1992]

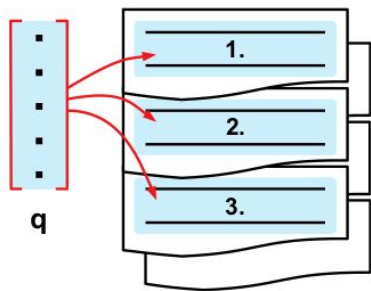


$$\mathcal{L}_{\text{cdv}+}(\Theta) = \frac{1}{T} \sum_{t=1}^T \left(\sqrt{1 + \left(\frac{\|\hat{\epsilon}_t - \epsilon_t\| + \|\hat{\alpha}_t - \alpha_t\|}{4} \right)^2} - 1 \right)$$



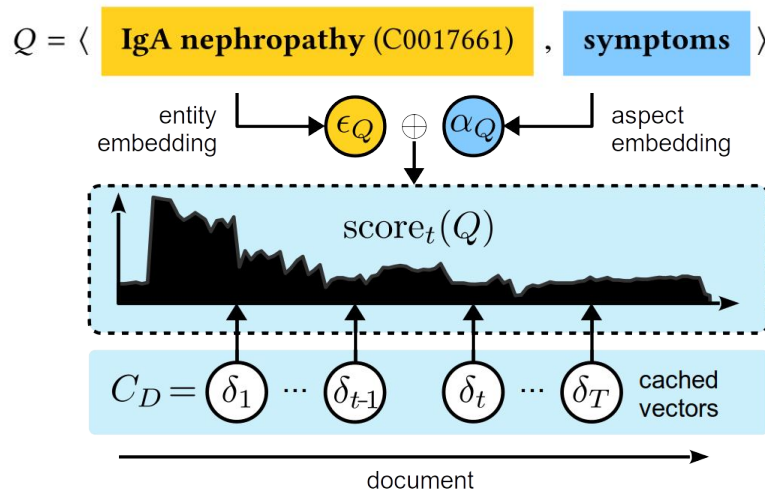
Challenge: Retrieve Answer Passages from Long Documents

RQ4: How effective are document representations for **retrieving answer passages**?



Cystic fibrosis (CF) is a genetic disorder that causes mucus to build up and damage organs in the body, particularly the lungs and pancreas. Signs and symptoms may include salty-tasting skin; persistent coughing; frequent lung infections; wheezing or shortness of breath; poor growth; weight loss; greasy, bulky stools; difficulty with bowel movements; and in males, infertility. Over time, mucus buildup and infections can lead to permanent lung damage, including the formation of scar tissue (fibrosis) and cysts in the lungs. CF is caused by mutations in the *CFTR* gene and inheritance is autosomal recessive.^{[1][2][3]} Treatment aims to relieve symptoms and usually includes respiratory therapies, inhaled medicines, pancreatic enzyme supplement, and nutritional supplements. Newer medications such as *CFTR* modulators have been approved for use in the United States. Ongoing research is focused on finding a cure for the disease.^[2]

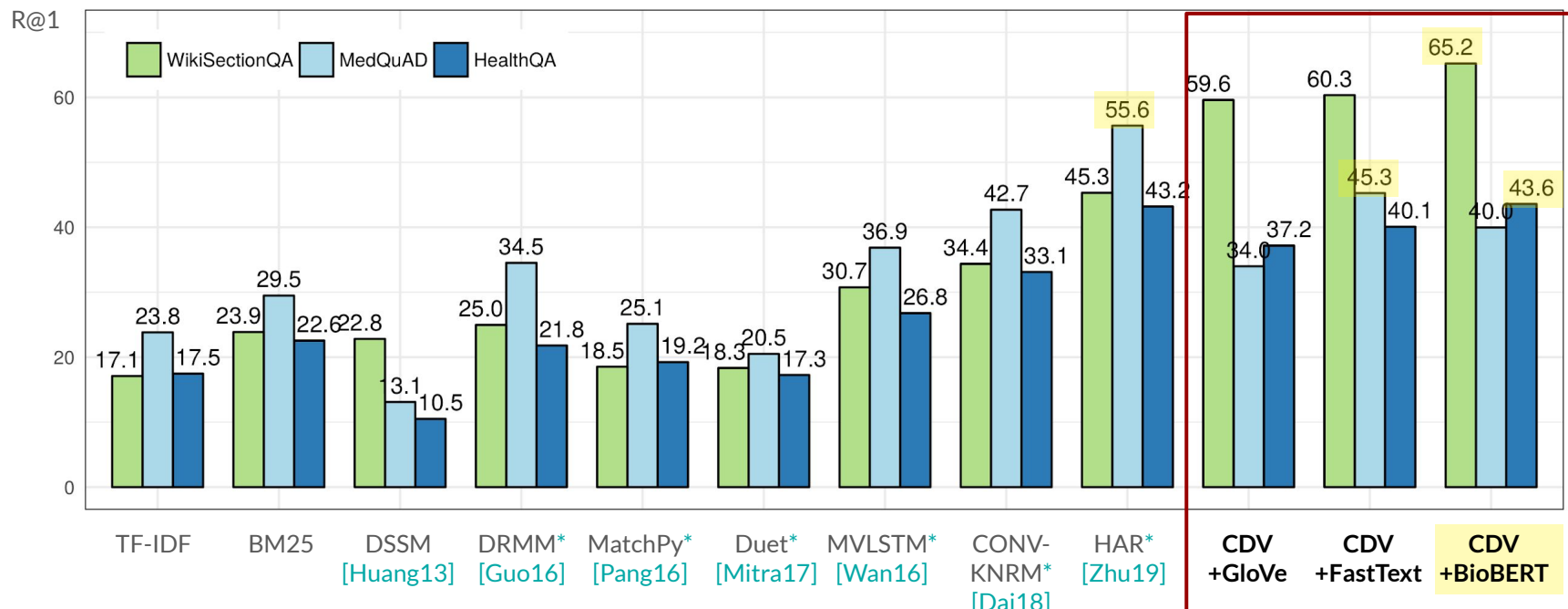
Answer Retrieval on Precomputed CDVs



$$\text{score}_t(Q(E, A), C_D) = \text{cosine}(\epsilon_Q \oplus \alpha_Q, \hat{\epsilon}_t \oplus \hat{\alpha}_t)$$

- No query/answer pairs are required to train the model
- CDV vectors C_D need to be precomputed only once

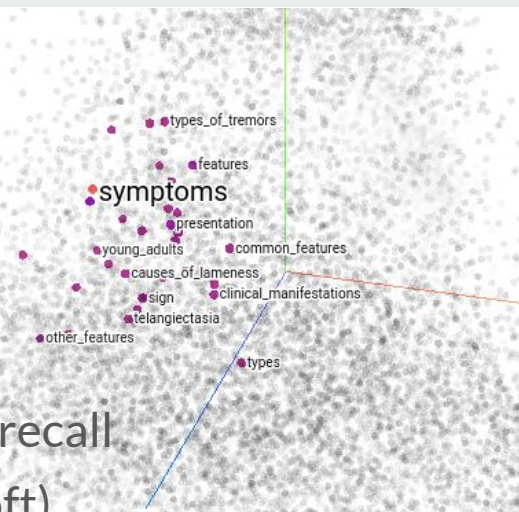
Results for Healthcare Answer Retrieval



We evaluate the retrieval of answer passages from over 2K healthcare articles from Wikipedia, NIH (including rare diseases) and Patient.info

Key Results: Contextual Discourse Vectors

- CDV **extends pre-trained Language Models** with long-range discourse information
- Trained with **self-supervision** on Wikipedia data
- Pre-computed vectors enable **efficient retrieval** with high recall
- Outperforms BM25 (Elastic Search), DSSM, Duet (Microsoft), MatchPyramid, Conv-KNRM, HAR and others with up to 65.2% R@1
- Potential errors caused by **hierarchical, related and overlapping information**

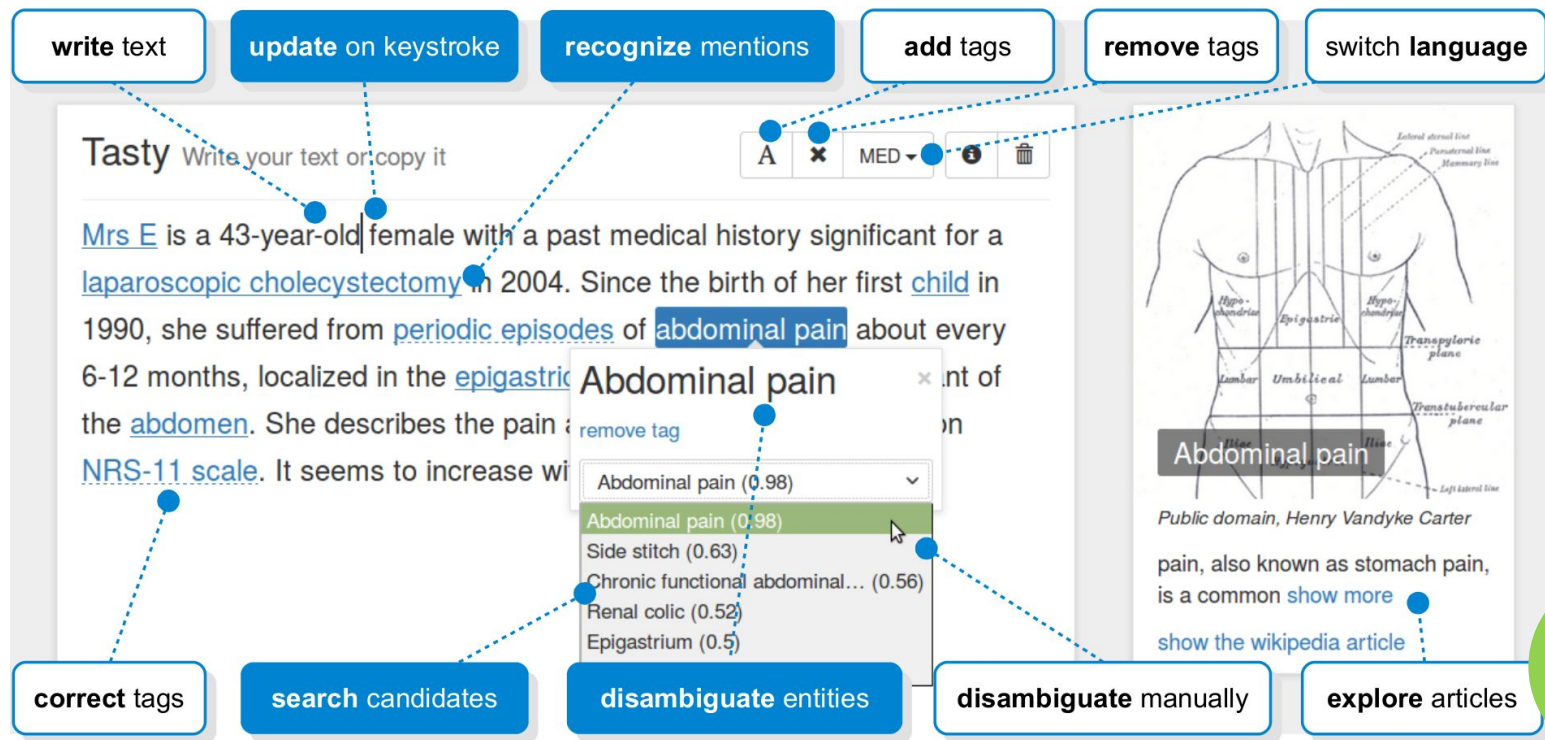


Our journey towards Neural Machine Reading:

- CDV fulfills all properties we have defined for automatic language understanding.
- This model satisfies our hypothesis of Neural Machine Reading.

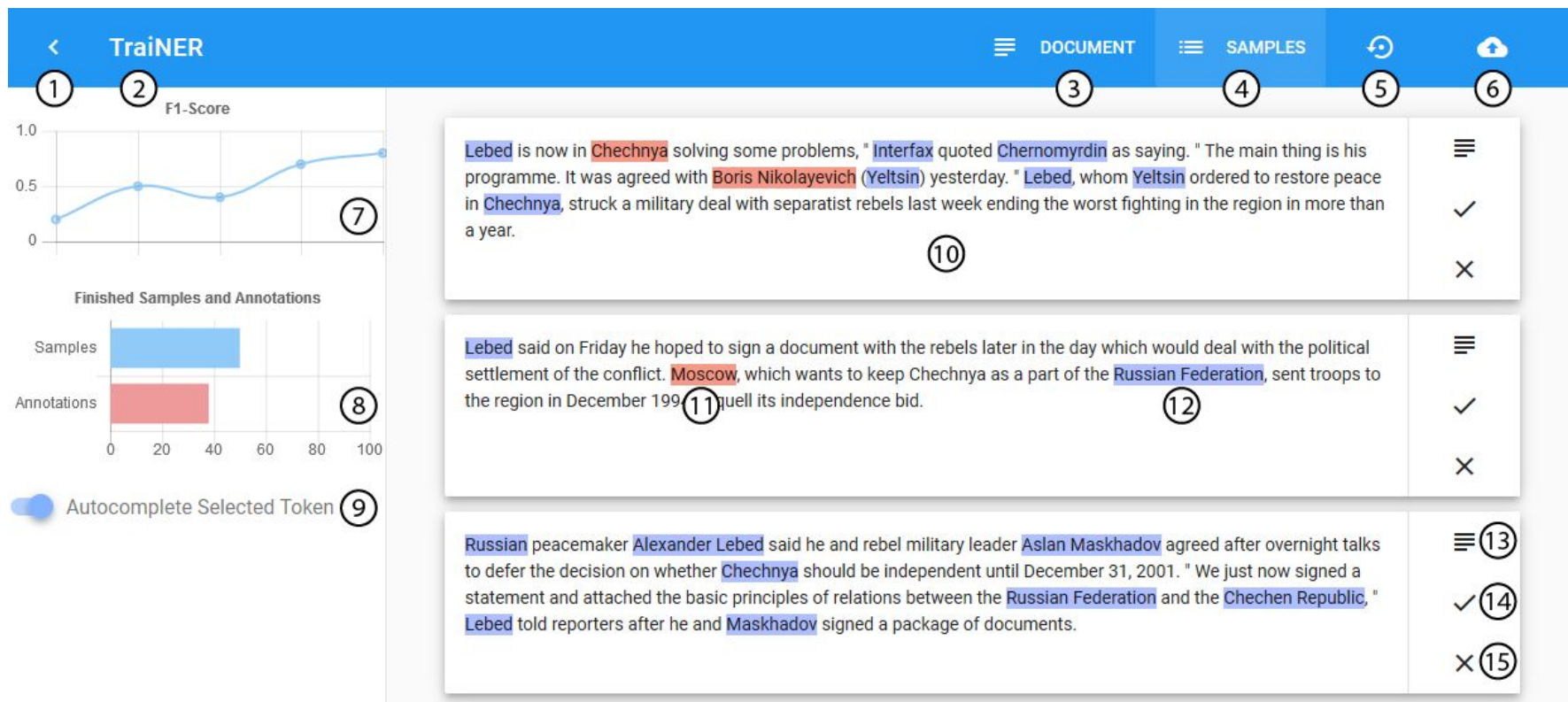
6. Systems

TASTY: Tag-as-you-type Entity Linking

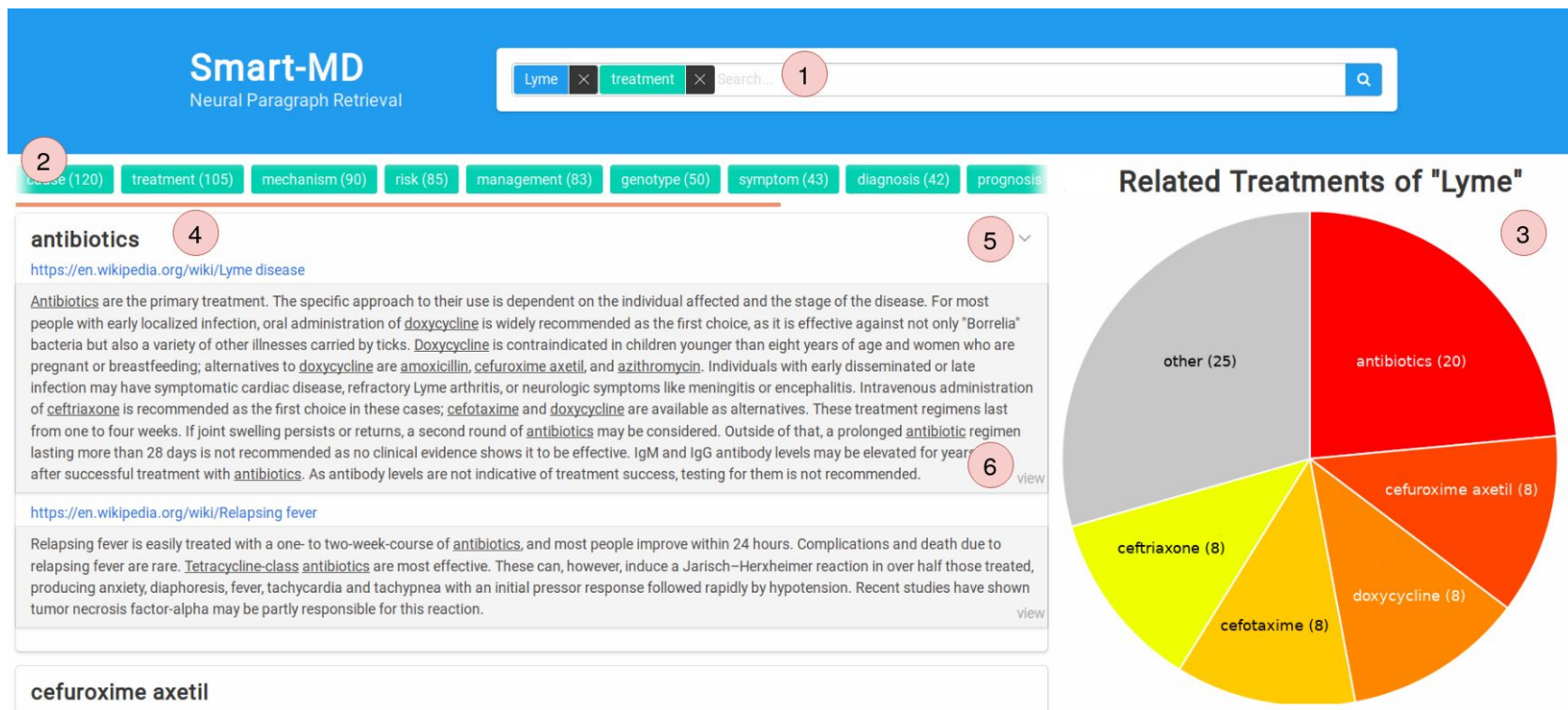


Online:
tasty.demo.dataxis.com

TraiNER: Bootstrapping Named Entity Recognition



Smart-MD: Clinical Decision Support System



CDV Healthcare Retrieval

CDV Search | Highlight

Find passages **1** **2** **3** **4** **5** **6** **7** **8** **9** **10** **11** **12** **13** **14** **15** **16** **17** **18** **19** **20** **21** **22** **23** **24** **25** **26** **27** **28** **29** **30** **31** **32** **33** **34** **35** **36** **37** **38** **39** **40** **41** **42** **43** **44** **45** **46** **47** **48** **49** **50** **51** **52** **53** **54** **55** **56** **57** **58** **59** **60** **61** **62** **63** **64** **65** **66** **67** **68** **69** **70** **71** **72** **73** **74** **75** **76** **77** **78** **79** **80** **81** **82** **83** **84** **85** **86** **87** **88** **89** **90** **91** **92** **93** **94** **95** **96** **97** **98** **99** **100** **101** **102** **103** **104** **105** **106** **107** **108** **109** **110** **111** **112** **113** **114** **115** **116** **117** **118** **119** **120** **121** **122** **123** **124** **125** **126** **127** **128** **129** **130** **131** **132** **133** **134** **135** **136** **137** **138** **139** **140** **141** **142** **143** **144** **145** **146** **147** **148** **149** **150** **151** **152** **153** **154** **155** **156** **157** **158** **159** **160** **161** **162** **163** **164** **165** **166** **167** **168** **169** **170** **171** **172** **173** **174** **175** **176** **177** **178** **179** **180** **181** **182** **183** **184** **185** **186** **187** **188** **189** **190** **191** **192** **193** **194** **195** **196** **197** **198** **199** **200** **201** **202** **203** **204** **205** **206** **207** **208** **209** **210** **211** **212** **213** **214** **215** **216** **217** **218** **219** **220** **221** **222** **223** **224** **225** **226** **227** **228** **229** **230** **231** **232** **233** **234** **235** **236** **237** **238** **239** **240** **241** **242** **243** **244** **245** **246** **247** **248** **249** **250** **251** **252** **253** **254** **255** **256** **257** **258** **259** **260** **261** **262** **263** **264** **265** **266** **267** **268** **269** **270** **271** **272** **273** **274** **275** **276** **277** **278** **279** **280** **281** **282** **283** **284** **285** **286** **287** **288** **289** **290** **291** **292** **293** **294** **295** **296** **297** **298** **299** **300** **301** **302** **303** **304** **305** **306** **307** **308** **309** **310** **311** **312** **313** **314** **315** **316** **317** **318** **319** **320** **321** **322** **323** **324** **325** **326** **327** **328** **329** **330** **331** **332** **333** **334** **335** **336** **337** **338** **339** **340** **341** **342** **343** **344** **345** **346** **347** **348** **349** **350** **351** **352** **353** **354** **355** **356** **357** **358** **359** **360** **361** **362** **363** **364** **365** **366** **367** **368** **369** **370** **371** **372** **373** **374** **375** **376** **377** **378** **379** **380** **381** **382** **383** **384** **385** **386** **387** **388** **389** **390** **391** **392** **393** **394** **395** **396** **397** **398** **399** **400** **401** **402** **403** **404** **405** **406** **407** **408** **409** **410** **411** **412** **413** **414** **415** **416** **417** **418** **419** **420** **421** **422** **423** **424** **425** **426** **427** **428** **429** **430** **431** **432** **433** **434** **435** **436** **437** **438** **439** **440** **441** **442** **443** **444** **445** **446** **447** **448** **449** **450** **451** **452** **453** **454** **455** **456** **457** **458** **459** **460** **461** **462** **463** **464** **465** **466** **467** **468** **469** **470** **471** **472** **473** **474** **475** **476** **477** **478** **479** **480** **481** **482** **483** **484** **485** **486** **487** **488** **489** **490** **491** **492** **493** **494** **495** **496** **497** **498** **499** **500** **501** **502** **503** **504** **505** **506** **507** **508** **509** **510** **511** **512** **513** **514** **515** **516** **517** **518** **519** **520** **521** **522** **523** **524** **525** **526** **527** **528** **529** **530** **531** **532** **533** **534** **535** **536** **537** **538** **539** **540** **541** **542** **543** **544** **545** **546** **547** **548** **549** **550** **551** **552** **553** **554** **555** **556** **557** **558** **559** **560** **561** **562** **563** **564** **565** **566** **567** **568** **569** **570** **571** **572** **573** **574** **575** **576** **577** **578** **579** **580** **581** **582** **583** **584** **585** **586** **587** **588** **589** **590** **591** **592** **593** **594** **595** **596** **597** **598** **599** **600** **601** **602** **603** **604** **605** **606** **607** **608** **609** **610** **611** **612** **613** **614** **615** **616** **617** **618** **619** **620** **621** **622** **623** **624** **625** **626** **627** **628** **629** **630** **631** **632** **633** **634** **635** **636** **637** **638** **639** **640** **641** **642** **643** **644** **645** **646** **647** **648** **649** **650** **651** **652** **653** **654** **655** **656** **657** **658** **659** **660** **661** **662** **663** **664** **665** **666** **667** **668** **669** **670** **671** **672** **673** **674** **675** **676** **677** **678** **679** **680** **681** **682** **683** **684** **685** **686** **687** **688** **689** **690** **691** **692** **693** **694** **695** **696** **697** **698** **699** **700** **701** **702** **703** **704** **705** **706** **707** **708** **709** **710** **711** **712** **713** **714** **715** **716** **717** **718** **719** **720** **721** **722** **723** **724** **725** **726** **727** **728** **729** **730** **731** **732** **733** **734** **735** **736** **737** **738** **739** **740** **741** **742** **743** **744** **745** **746** **747** **748** **749** **750** **751** **752** **753** **754** **755** **756** **757** **758** **759** **760** **761** **762** **763** **764** **765** **766** **767** **768** **769** **770** **771** **772** **773** **774** **775** **776** **777** **778** **779** **780** **781** **782** **783** **784** **785** **786** **787** **788** **789** **790** **791** **792** **793** **794** **795** **796** **797** **798** **799** **800** **801** **802** **803** **804** **805** **806** **807** **808** **809** **810** **811** **812** **813** **814** **815** **816** **817** **818** **819** **820** **821** **822** **823** **824** **825** **826** **827** **828** **829** **830** **831** **832** **833** **834** **835** **836** **837** **838** **839** **840** **841** **842** **843** **844** **845** **846** **847** **848** **849** **850** **851** **852** **853** **854** **855** **856** **857** **858** **859** **860** **861** **862** **863** **864** **865** **866** **867** **868** **869** **870** **871** **872** **873** **874** **875** **876** **877** **878** **879** **880** **881** **882** **883** **884** **885** **886** **887** **888** **889** **890** **891** **892** **893** **894** **895** **896** **897** **898** **899** **900** **901** **902** **903** **904** **905** **906** **907** **908** **909** **910** **911** **912** **913** **914** **915** **916** **917** **918** **919** **920** **921** **922** **923** **924** **925** **926** **927** **928** **929** **930** **931** **932** **933** **934** **935** **936** **937** **938** **939** **940** **941** **942** **943** **944** **945** **946** **947** **948** **949** **950** **951** **952** **953** **954** **955** **956** **957** **958** **959** **960** **961** **962** **963** **964** **965** **966** **967** **968** **969** **970** **971** **972** **973** **974** **975** **976** **977** **978** **979** **980** **981** **982** **983** **984** **985** **986** **987** **988** **989** **990** **991** **992** **993** **994** **995** **996** **997** **998** **999** **1000** **1001** **1002** **1003** **1004** **1005** **1006** **1007** **1008** **1009** **1010** **1011** **1012** **1013** **1014** **1015** **1016** **1017** **1018** **1019** **1020** **1021** **1022** **1023** **1024** **1025** **1026** **1027** **1028** **1029** **1030** **1031** **1032** **1033** **1034** **1035** **1036** **1037** **1038** **1039** **1040** **1041** **1042** **1043** **1044** **1045** **1046** **1047** **1048** **1049** **1050** **1051** **1052** **1053** **1054** **1055** **1056** **1057** **1058** **1059** **1060** **1061** **1062** **1063** **1064** **1065** **1066** **1067** **1068** **1069** **1070** **1071** **1072** **1073** **1074** **1075** **1076** **1077** **1078** **1079** **1080** **1081** **1082** **1083** **1084** **1085** **1086** **1087** **1088** **1089** **1090** **1091** **1092** **1093** **1094** **1095** **1096** **1097** **1098** **1099** **1100** **1101** **1102** **1103** **1104** **1105** **1106** **1107** **1108** **1109** **1110** **1111** **1112** **1113** **1114** **1115** **1116** **1117** **1118** **1119** **1120** **1121** **1122** **1123** **1124** **1125** **1126** **1127** **1128** **1129** **1130** **1131** **1132** **1133** **1134** **1135** **1136** **1137** **1138** **1139** **1140** **1141** **1142** **1143** **1144** **1145** **1146** **1147** **1148** **1149** **1150** **1151** **1152** **1153** **1154** **1155** **1156** **1157** **1158** **1159** **1160** **1161** **1162** **1163** **1164** **1165** **1166** **1167** **1168** **1169** **1170** **1171** **1172** **1173** **1174** **1175** **1176** **1177** **1178** **1179** **1180** **1181** **1182** **1183** **1184** **1185** **1186** **1187** **1188** **1189** **1190** **1191** **1192** **1193** **1194** **1195** **1196** **1197** **1198** **1199** **1200** **1201** **1202** **1203** **1204** **1205** **1206** **1207** **1208** **1209** **1210** **1211** **1212** **1213** **1214** **1215** **1216** **1217** **1218** **1219** **1220** **1221** **1222** **1223** **1224** **1225** **1226** **1227** **1228** **1229** **1230** **1231** **1232** **1233** **1234** **1235** **1236** **1237** **1238** **1239** **1240** **1241** **1242** **1243** **1244** **1245** **1246** **1247** **1248** **1249** **1250** **1251** **1252** **1253** **1254** **1255** **1256** **1257** **1258** **1259** **1260** **1261** **1262** **1263** **1264** **1265** **1266** **1267** **1268** **1269** **1270** **1271** **1272** **1273** **1274** **1275** **1276** **1277** **1278** **1279** **1280** **1281** **1282** **1283** **1284** **1285** **1286** **1287** **1288** **1289** **1290** **1291** **1292** **1293** **1294** **1295** **1296** **1297** **1298** **1299** **1300** **1301** **1302** **1303** **1304** **1305** **1306** **1307** **1308** **1309** **1310** **1311** **1312** **1313** **1314** **1315** **1316** **1317** **1318** **1319** **1320** **1321** **1322** **1323** **1324** **1325** **1326** **1327** **1328** **1329** **1330** **1331** **1332** **1333** **1334** **1335** **1336** **1337** **1338** **1339** **1340**

7. Conclusion and Future Work



Contributions

TASTY, SECTOR and CDV are Neural Machine Reading architectures for domain-specific language understanding:

- Variety of languages and domains
- Robust against noise and spelling variations
- Document and discourse structure
- Entity Linking, Topic Modeling, Answer Retrieval tasks
- Efficient to train
- In-depth error analysis

Implementation available under Apache V2 license
<https://github.com/sebastianarnold/TeXoo>



Limitations

Writing systems

- All experiments in Latin script
- Sentence and token normalization necessary

Self-supervision

- Preprocessing tailored to training data
- Holistic approach: Weak-supervision

Optimization objectives

- We focused on task accuracy (Prec/Rec/F1)
- Reality: result diversification, freshness, trust, popularity, etc.
- Dynamically changing objectives



Future Work

Hierarchical Knowledge Representations

- Capture **hypernym/hyponym relations**
- Model **semantic hierarchies** in vector space

Uncertainty Modeling

- **Data distribution** changes from training to inference time
- Report **level of confidence**

Continual Learning

- Continuously **learn from new data**
- Prevent **catastrophic forgetting**

Acknowledgements

Thanks to everyone who supported me during this time with inspiration, feedback and grounding.

- **Supervisors, Mentors and Committee**

Alexander Löser, Philippe Cudré-Mauroux,
Felix Gers, Peter Tröger, Amy Siu, Laura Dietz

- **Coauthors and Team Members**

Torsten Kiliyas, Robert Dziuba, Rudolf Schneider,
Christopher Kümmel, Robin Mehlitz, Tom Oberhauser,
Betty van Aken, Benjamin Winter, Paul Grundmann,
Michalis Papaioannou
DATEXIS Team @ BeuthHS
eXascale Infolab Team @ UNIFR

- **My Family and Friends**



Questions & Discussion



Our work is funded by the German Federal Ministry of Economic Affairs and Energy (BMWi) under grant agreements 01MD16011E (Medical Allround-Care Service Solutions) 01MD19003b (PLASS), 01MD19013D (Smart-MD) and H2020 ICT-2016-1 grant agreement 732328 (FashionBrain).

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Neural Machine Reading for Domain-Specific Text Resources

Speaker: Sebastian Arnold

sarnold@beuth-hochschule.de
[@sebastianarnold](https://twitter.com/sebastianarnold)

*Data Science and Text-based
Information Systems (DATEXIS)*
Beuth University of Applied Sciences
Berlin, Germany
www.datexis.de



BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Selected References (1)

- Bengio, Ducharme, Vincent and Janvin (2003): A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3.
- Bojanowski, Grave, Joulin and Mikolov, (2017): Enriching Word Vectors with Subword Information. *TACL* Vol. 5
- Devlin, Chang, Lee and Toutanova (2019): BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT* 2019
- Etzioni, Banko and Cafarella (2006): Machine Reading. *AAAI* 2006
- Firth (1957): A Synopsis of Linguistic Theory, 1930–1955. *Studies in Linguistic Analysis*
- Graves (2012): Supervised Sequence Labelling with Recurrent Neural Networks. Berlin Heidelberg: Springer
- Harris (1954): Distributional Structure. *WORD* Vol. 10, No. 2–3
- Hochreiter and Schmidhuber (1997): Long Short-term Memory. *Neural Computation*, Vol. 9, No. 8
- Huang, He, Gao, Deng, Acero and Heck (2013): Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. *CIKM* 2013
- Huber (1992): Robust Estimation of a Location Parameter. *Breakthroughs in Statistics*, Springer
- Jurafsky and Martin (2008): Speech and Language Processing. Vol. 2
- Lee et al. (2019): BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*
- Manning, Schütze and Raghavan (2008): Introduction to Information Retrieval. Cambridge University Press
- Marchionini (2006): Exploratory Search: From Finding to Understanding. *Communications of the ACM* Vol. 49, No. 4
- Mikolov, Chen, Corrado and Dean (2013): Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781
- Mitra, Diaz and Craswell (2017): Learning to Match using Local and Distributed Representations of Text for Web Search. *WWW* 2017
- Moro, Raganato and Navigli (2014): Entity Linking meets Word Sense Disambiguation: A Unified Approach. *TACL* Vol. 2

Selected References (2)

- Navigli and Ponzetto (2012): BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, Vol. 193
- Palangi et al. (2016): Deep Sentence Embedding using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *TASLP* Vol. 24, No. 4
- Peters, Neumann, Iyyer, Gardner, Clark, Lee and Zettlemoyer (2018): Deep Contextualized Word Representations. *NAACL-HLT* 2018
- Pink, Nothman and Curran (2014): Analysing Recall Loss in Named Entity Slot Filling. *EMNLP* 2014
- Prokofyev, emartini and Cudré-Mauroux (2014): Effective Named Entity Recognition for Idiosyncratic Web Collections. *WWW* 2014
- Rose and Levinson (2004): Understanding User Goals in Web Search. *WWW* 2004
- Sahlgren (2008): The Distributional Hypothesis. *Italian Journal of Linguistics* 20
- Sarawagi (2008): Information Extraction. *Foundations and Trends in Databases* 1, Nr. 3
- Serrà and Karatzoglou (2017): Getting Deep Recommenders Fit: Bloom Embeddings for Sparse binary Input/Output Networks. *RecSys* 2017
- Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin (2017): Attention Is All You Need. *NIPS* 2017
- Zhu, Ahuja, Wei and Reddy (2019): A Hierarchical Attention Retrieval Model for Healthcare Question Answering. *WWW* 2019