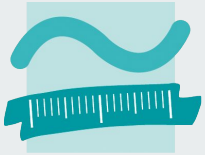
Neural Machine Reading for Domain-Specific Text Resources

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

UNI
FR
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

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.”

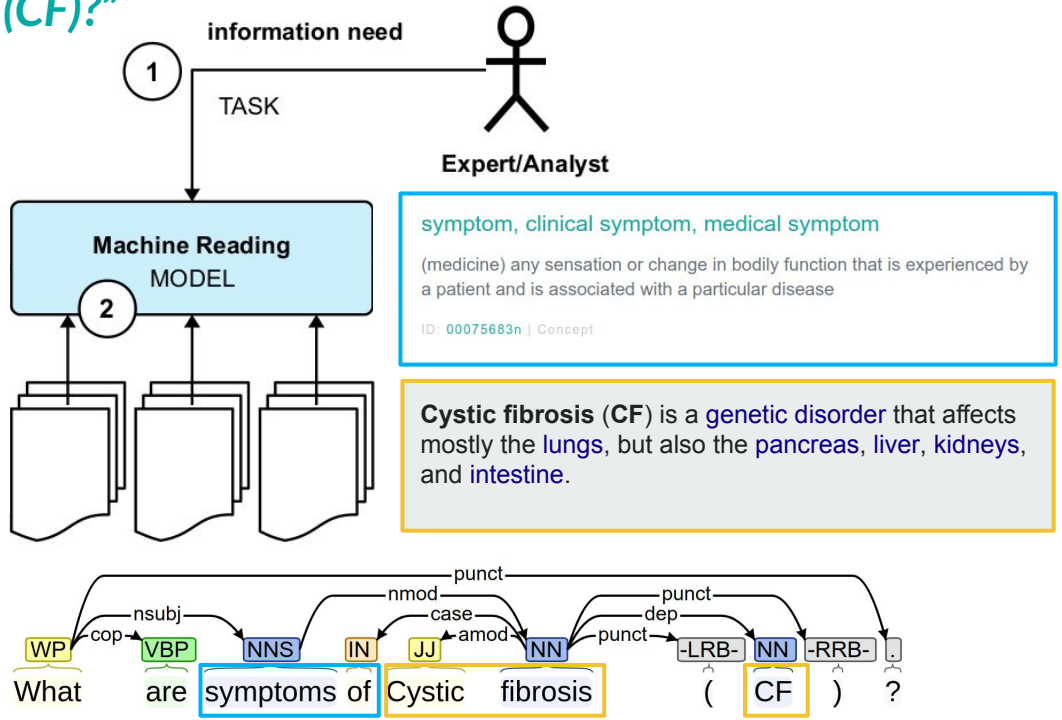
— Alexander von Humboldt, an Johann Wolfgang von Goethe, Paris 3. Januar 1810.

1. Introduction

Example: Information Search

"What are symptoms of Cystic fibrosis (CF)?"

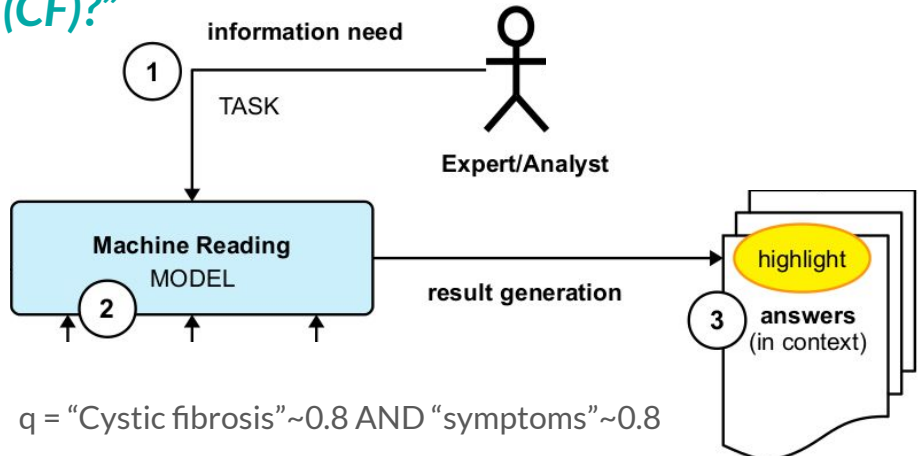
- Query intent classification
  [Rose+Levinson 2004]
- Natural Language Processing
  [Jurafsky 2008]
- Information Extraction (IE)
  [Sarawagi 2008]
- Semantic representation
  [Moro 2014]



Knowledge Processing and Retrieval

"What are symptoms of Cystic fibrosis (CF)?"

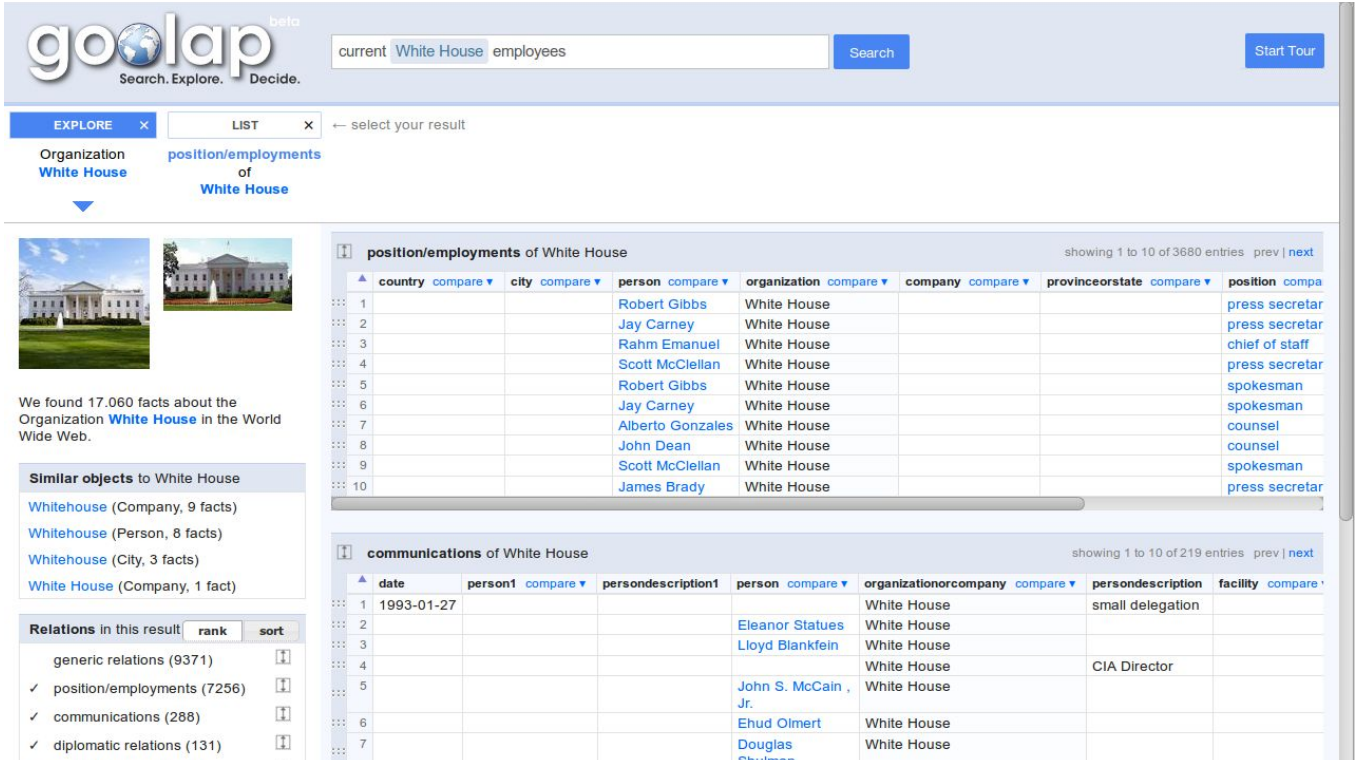
- Question Answering (QA)
- Information Retrieval (IR)
  [Manning 2008]
- Feedback Loop
- Exploration and refinement
  [Marchionini 2006]



q = "Cystic fibrosis"~0.8 AND "symptoms"~0.8

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]

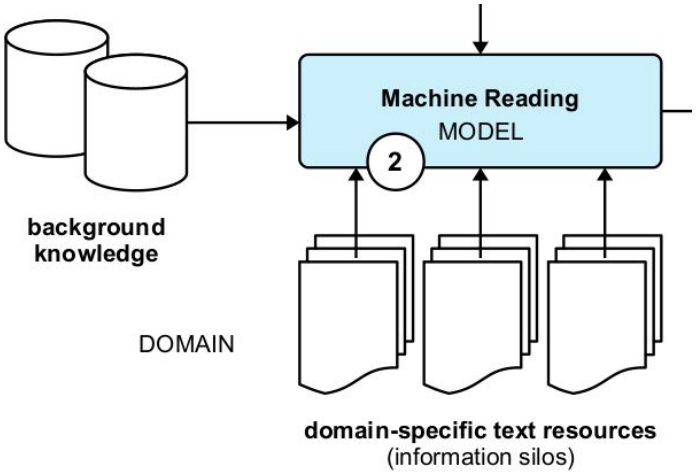
GoOLAP: Answering Analytical Queries with Tables [2012]



A. Löser, S. Arnold, T. Fiehn (2012): **The GoOLAP Fact Retrieval Framework.** Business Intelligence. Springer
S. Arnold, A. Löser, T. Kilias (2015):. **Resolving Common Analytical Tasks in Text Databases.** ACM DOLAP 2015

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.

Etzioni, Banko and Cafarella (2006): **Machine Reading.** AAAI 2006

Challenges for Domain-specific Language Understanding

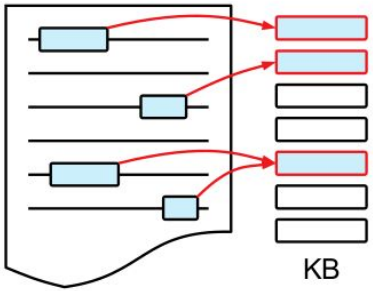
Traditional IE pipelines depend on task-specific training data or handcrafted rules.

- 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

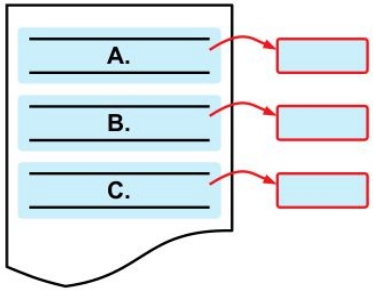
Hypothesis: Neural Machine Reading

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**.

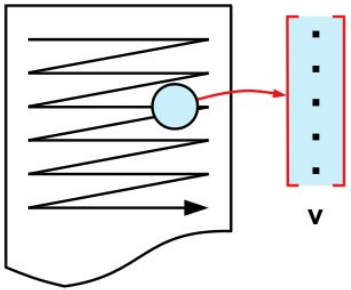
Research Questions



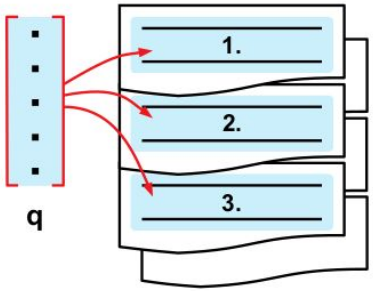
RQ1: Identify domain-specific **named entities**.



RQ2: Detect **topics and structure** in long documents.



RQ3: Embed **discourse structure** into document representations.



RQ4: **Retrieve answer passages** using vector representations.

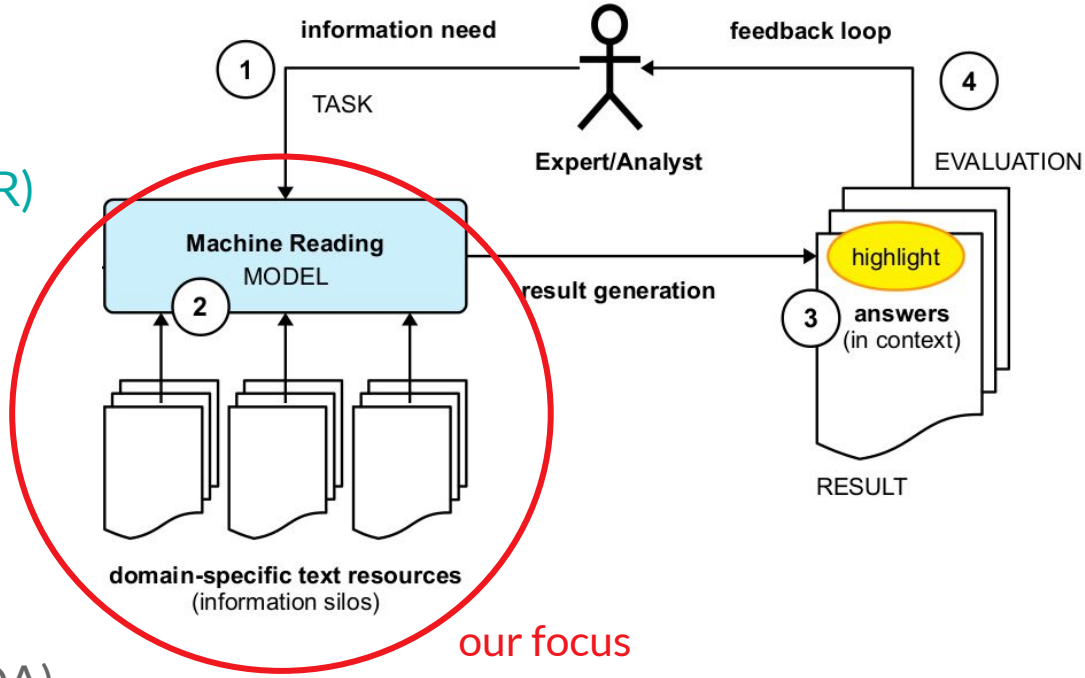
Focus of the Thesis

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. Kilias 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]

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

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

$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c | w_t)$$

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. Kilias and A. Löser (2016).
Robust Named Entity Recognition in Idiosyncratic Domains. **arXiv:1608.06757 [cs.CL]**

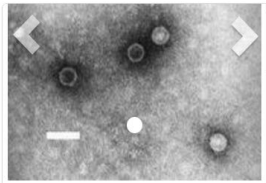
S. Arnold, R. Dziuba and A. Löser (2016).
TASTY: Interactive Entity Linking As-You-Type. **COLING 2016 (Demos)**

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 signif 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-C 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.

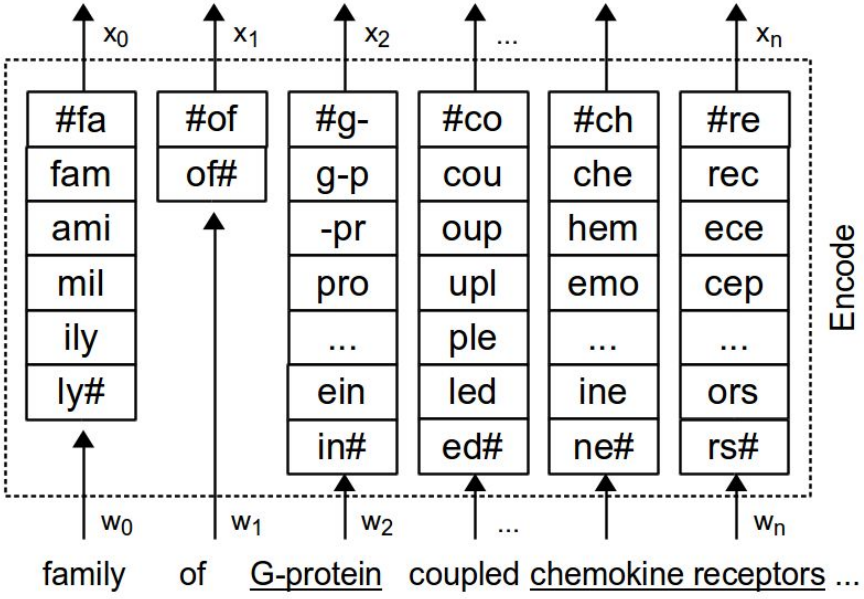


Epstein–Barr virus nuclear antigen 2

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

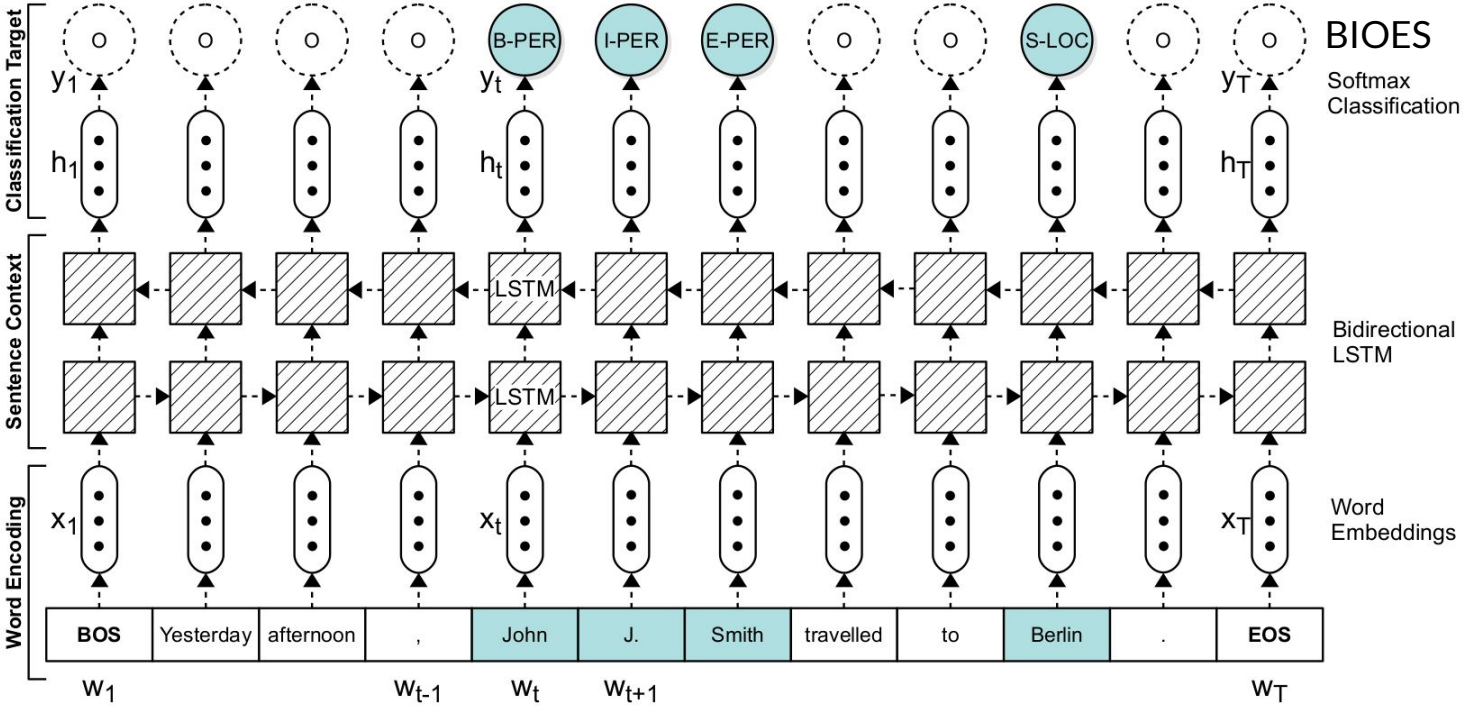
Extracting entities requires local, contextual and global features.

Encoding Local Sub-word Information with Letter-trigrams



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

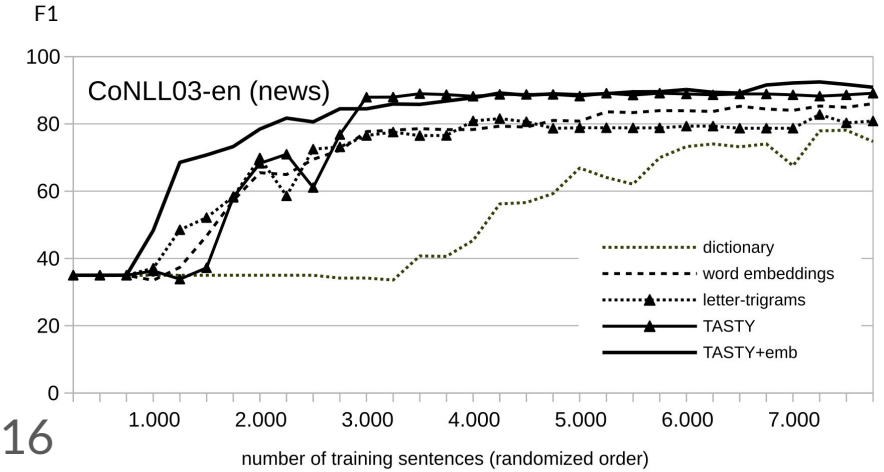
Capturing Sentence Context with Sequence Learning



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

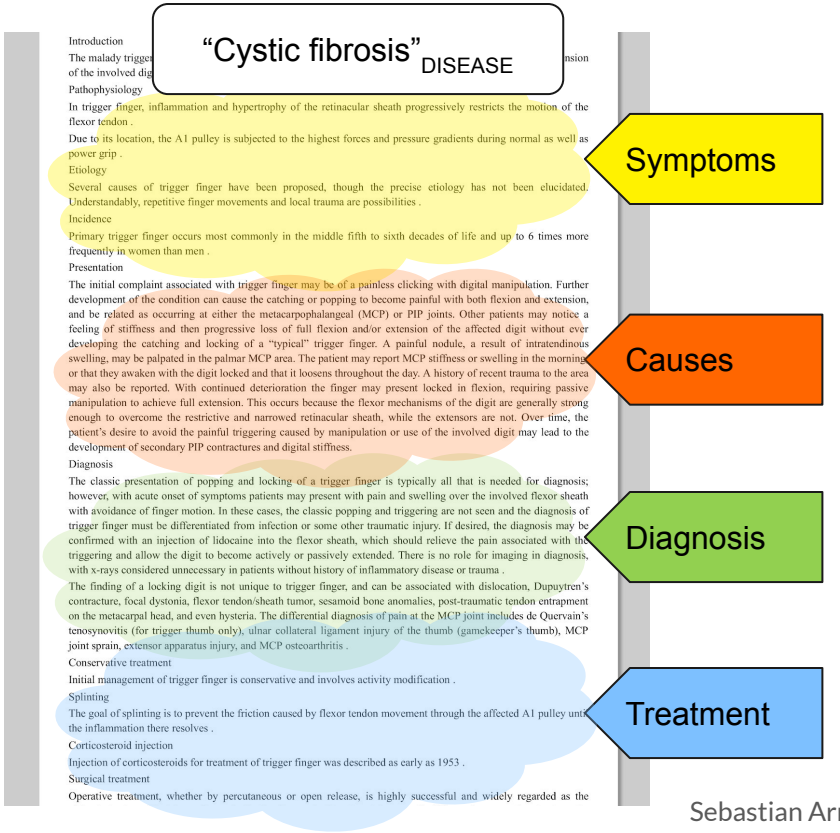
S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers and A. Löser (2019).
SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. **TACL Vol. 7**: 169–184 (presented at ACL)

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)	de_disease (1.0k)
diagnosis	therapie
treatment	symptome
symptoms	behandlung
signs	diagnose
causes	ursachen
epidemiology	klinische
management	ursache
cause	diagnostik
prognosis	verbreitung
history	erscheinungen
pathophysiology	einteilung
classification	geschichte
prevention	epidemiologie
genetics	differentialdiagr
research	prognose
types	klassifikation
culture	formen
factors	verlauf
mechanism	haeufigkeit
society	vorbeugung
presentation	aetiologie
risk	pathogenese
differential	klinisches
disease	pathologie
surgery	bild
therapy	klinik
clinical	entstehung
pathogenesis	symptomatik

clustering / normalization via BabelNet [Navigli 2012]

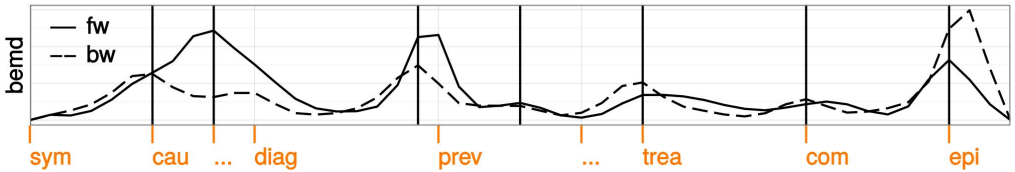
(85–95% coverage)

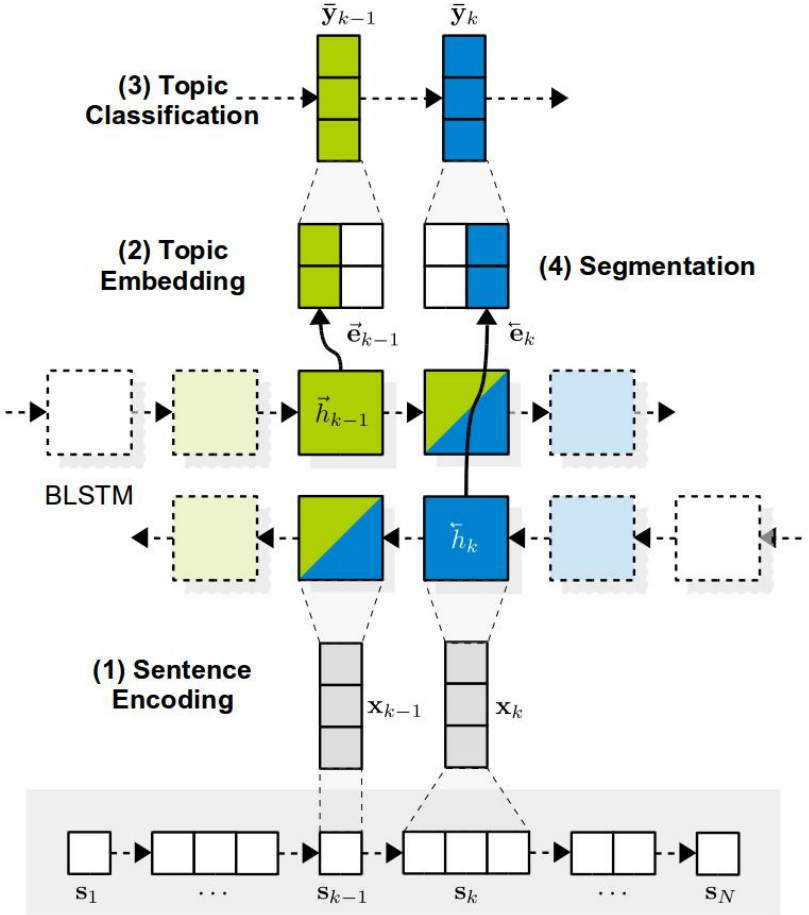
en_disease (27)	de_disease (25)
treatment	therapie
symptom	diagnose
diagnosis	symptom
cause	ursache
classification	kategorisierung
epidemiology	verlauf
history	epidemiologie
prognosis	geschichte
management	prognose
pathophysiology	praevalenz
mechanism	vorbeugung
prevention	fauna
research	terminologie
genetics	pathologie
tomography	definition
culture	klinik
etymology	komplikation
infection	genetik
fauna	infektion
risk	risiko
pathology	forschung
surgery	geographie
screening	mensch
medication	organe
geography	sonstiges
complication	
other	

Capturing Entire Documents with Sentence Embeddings + BLSTM

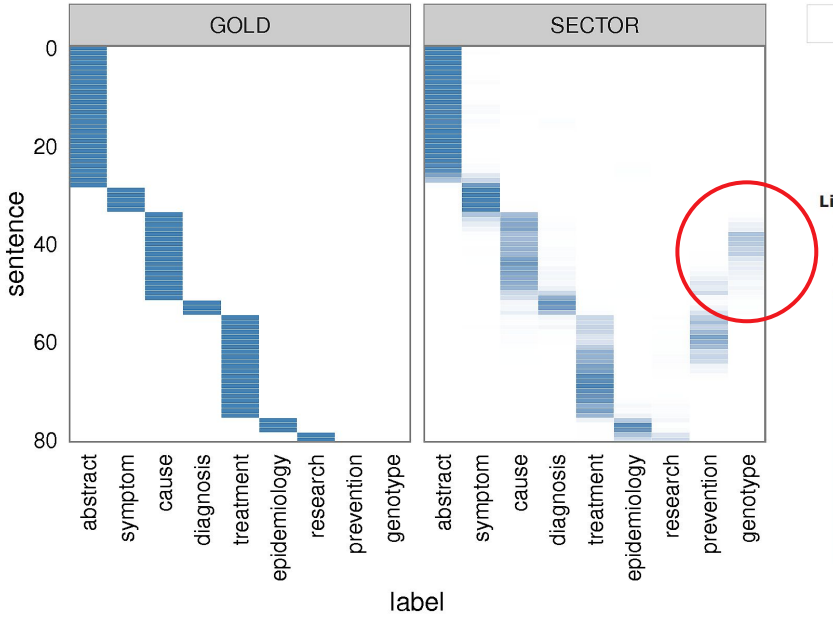
Objective: predict topics on sentence level

- 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)
Iiibalesiii (talk | contribs)
(→Cause)
← Previous 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
```

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

Line 35:

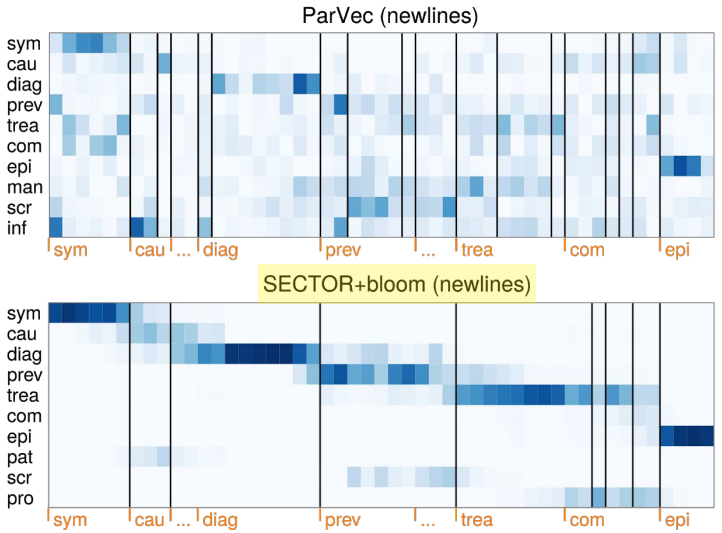
```
== Causes ==

=== 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

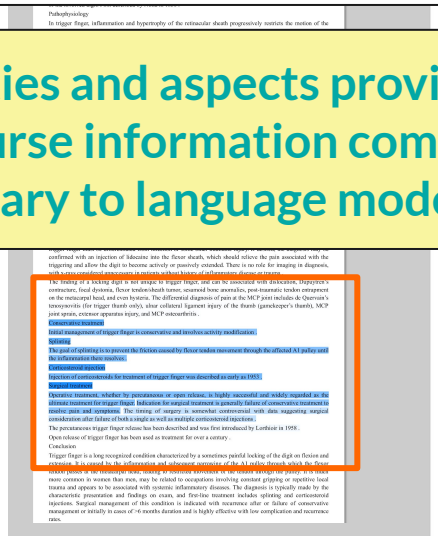
5. CDV – Contextual Document Representations for Answer Retrieval

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.



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

"Cystic fibrosis"

"Therapy"

treatment

ement of trigger finger is conservative and involves activity modification .

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

l injection

orticosteroids 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 .

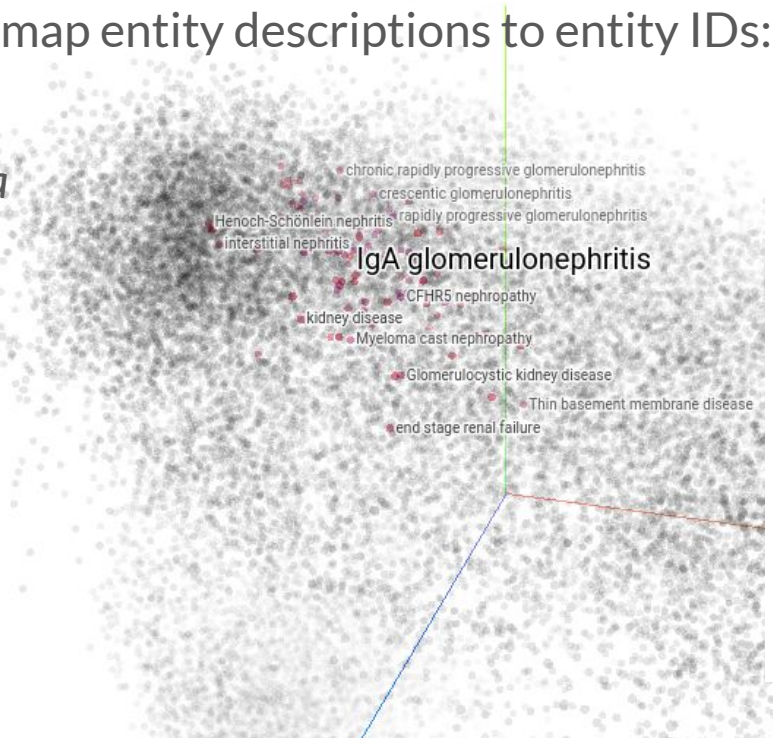
Conclusion

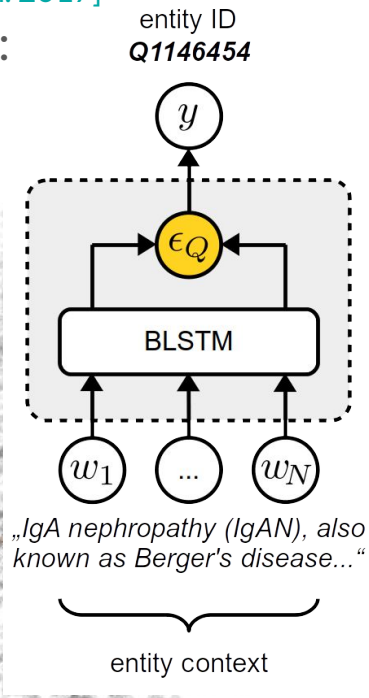
Trigger finger is a long recognized condition characterized by a sometimes painful locking of the digit on flexion and extension. It is caused by the inflammation and subsequent narrowing of the A1 pulley through which the flexor

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

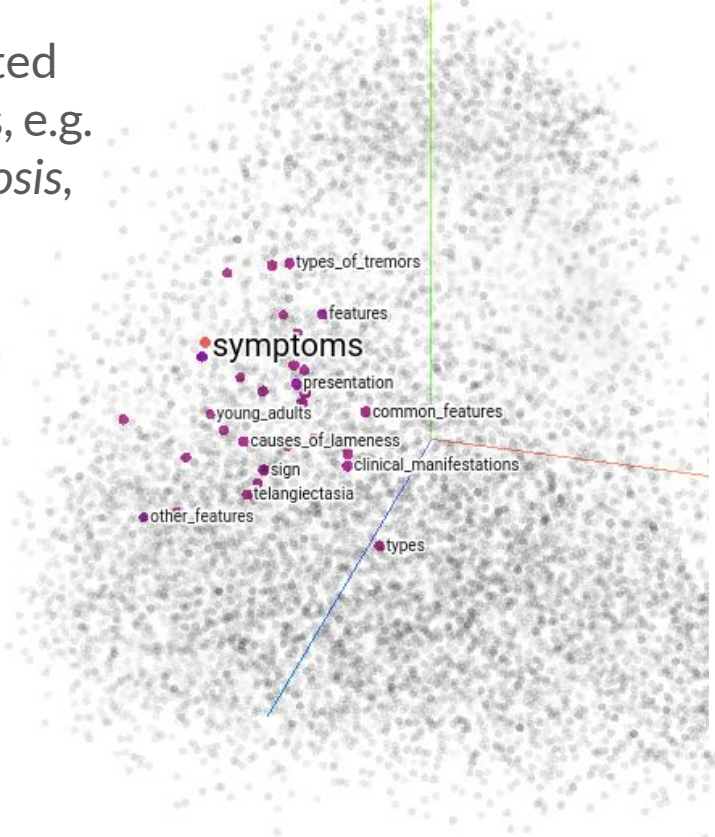


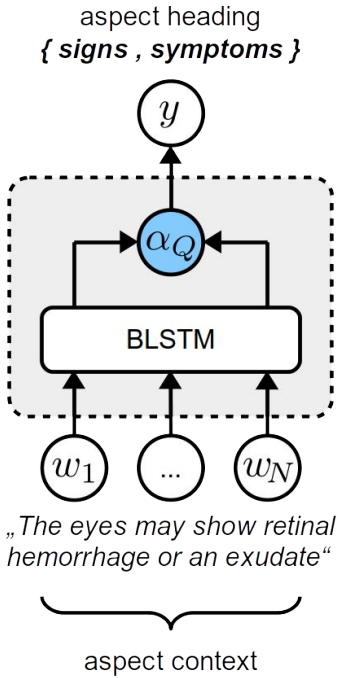


Distributed Representation of Clinical Aspects

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



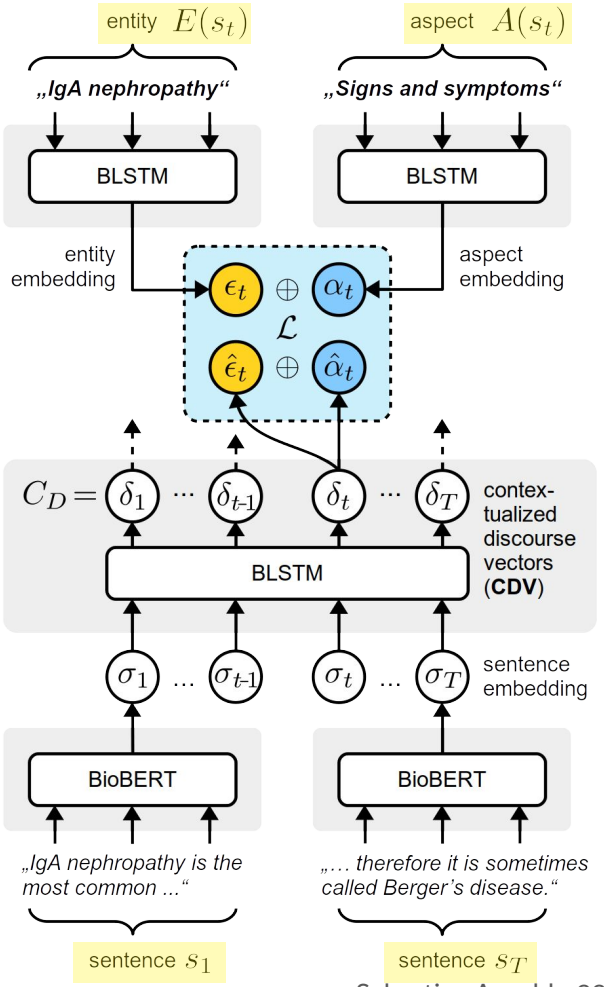


Training CDV with Multi-task Objective

- 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:

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

s	E(s)	A(s)
s0..1	Arteriosclerosis	description
s2	Arteriosclerosis	signs;symptoms
s3..5	Arteriosclerosis	pathophysiology
s6	Arteriosclerosis;Arteriolosclerosis	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]

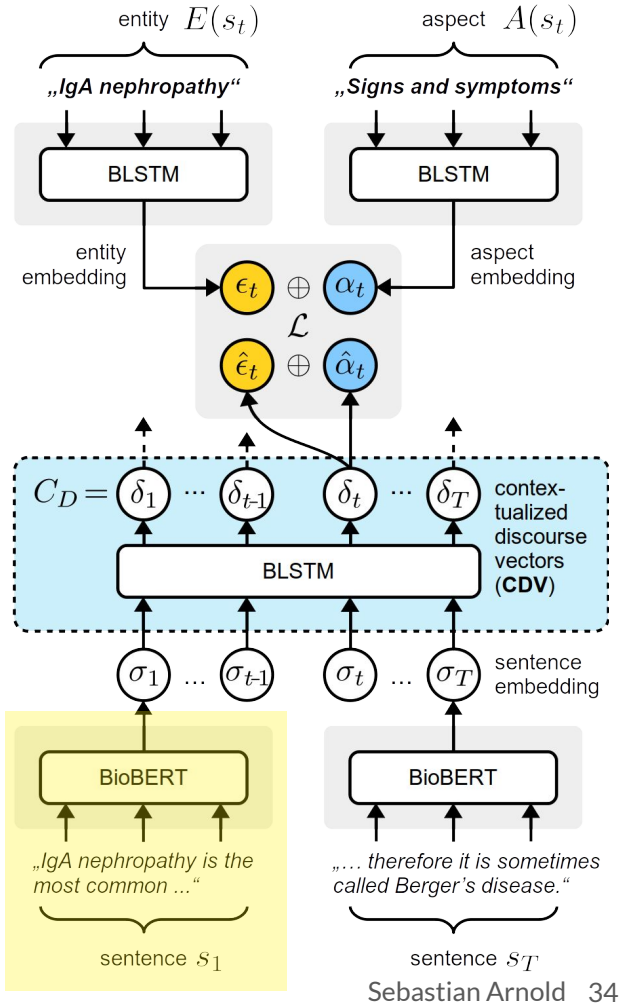
$$\vec{h}_t = \text{LSTM}_\Theta(\vec{h}_{t-1}, \sigma(s_t))$$

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

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

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

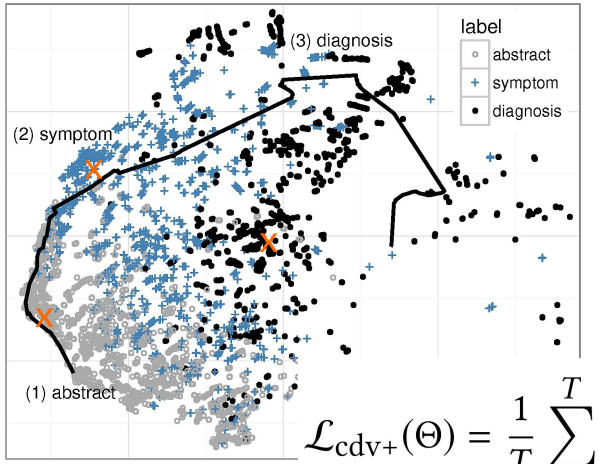
$$\hat{\alpha}_t = \tanh(W_{\delta\alpha}\delta_t + b_\alpha)$$



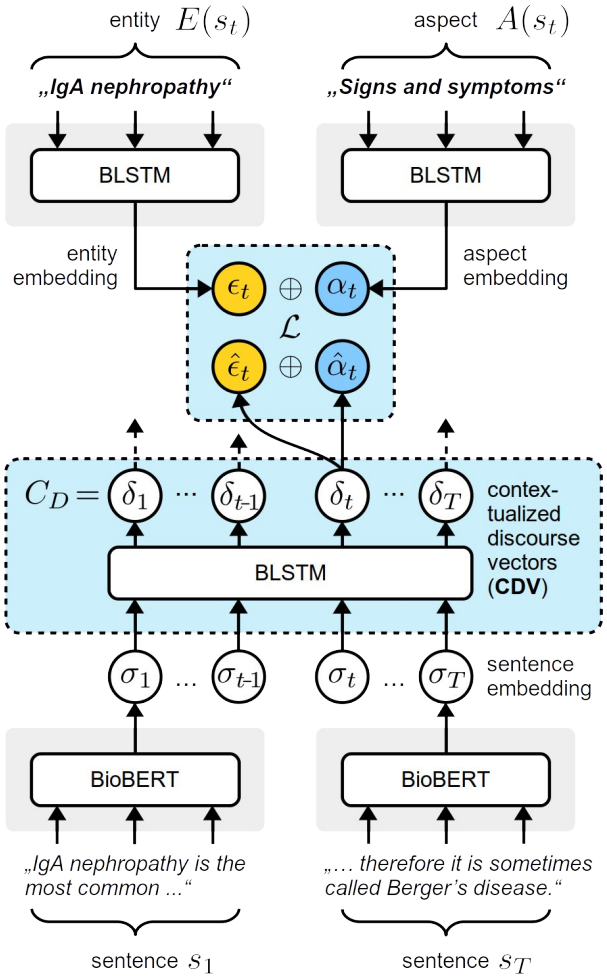
Loss Function

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

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

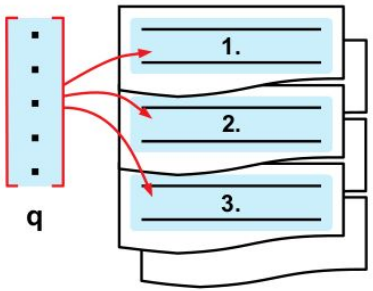


$$\mathcal{L}_{\text{cdv+}}(\Theta) = \frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{1+\left(\frac{\|\hat{\epsilon}_t - \epsilon_t\| + \|\hat{\alpha}_t - \alpha_t\|}{4}\right)^2} - 1\right)$$



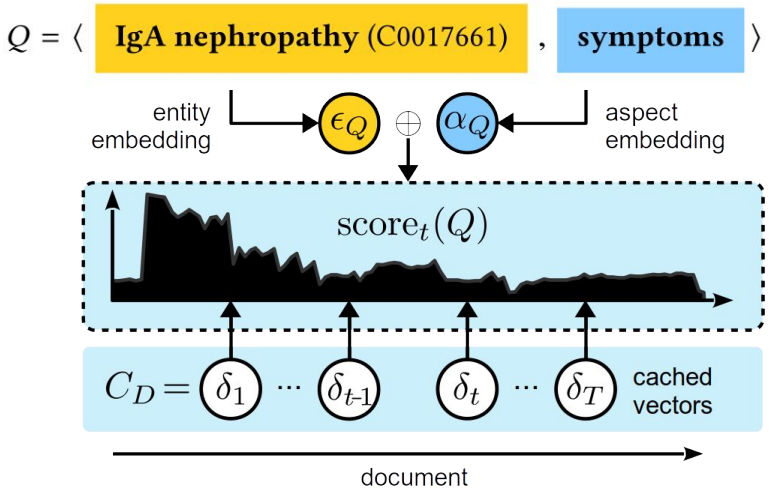
Challenge: Retrieve Answer Passages from Long Documents

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

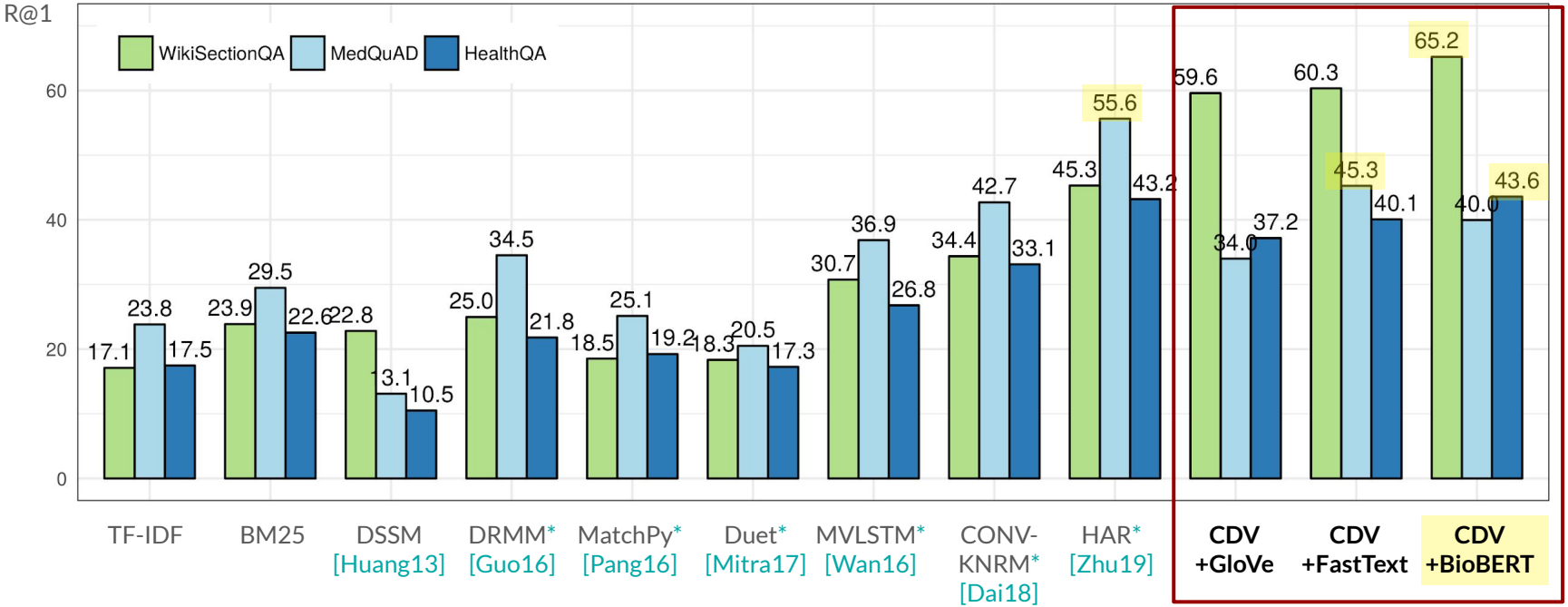


$$\text{score}_t\left(Q(E,A), C_D\right) = \text{cosine}\left(\epsilon_Q \oplus \alpha_Q, \hat{\epsilon}_t \oplus \hat{\alpha}_t\right)$$

➔ No query/answer pairs are required to train the model

➔ CDV vectors C_D need to be precomputed only once

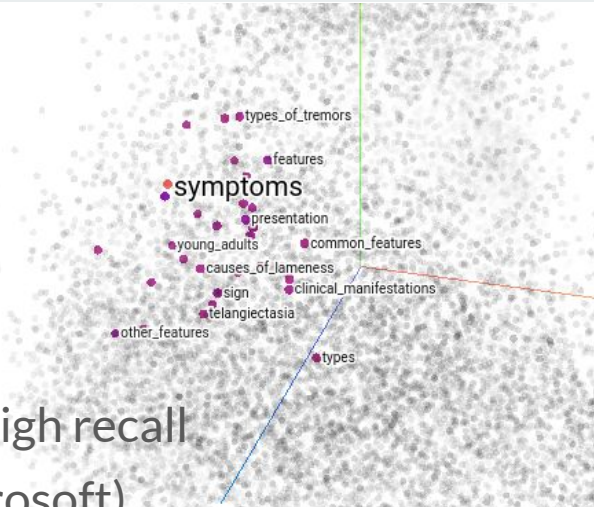
Results for Healthcare Answer Retrieval



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

* interaction-based models

Key Results: Contextual Discourse Vectors



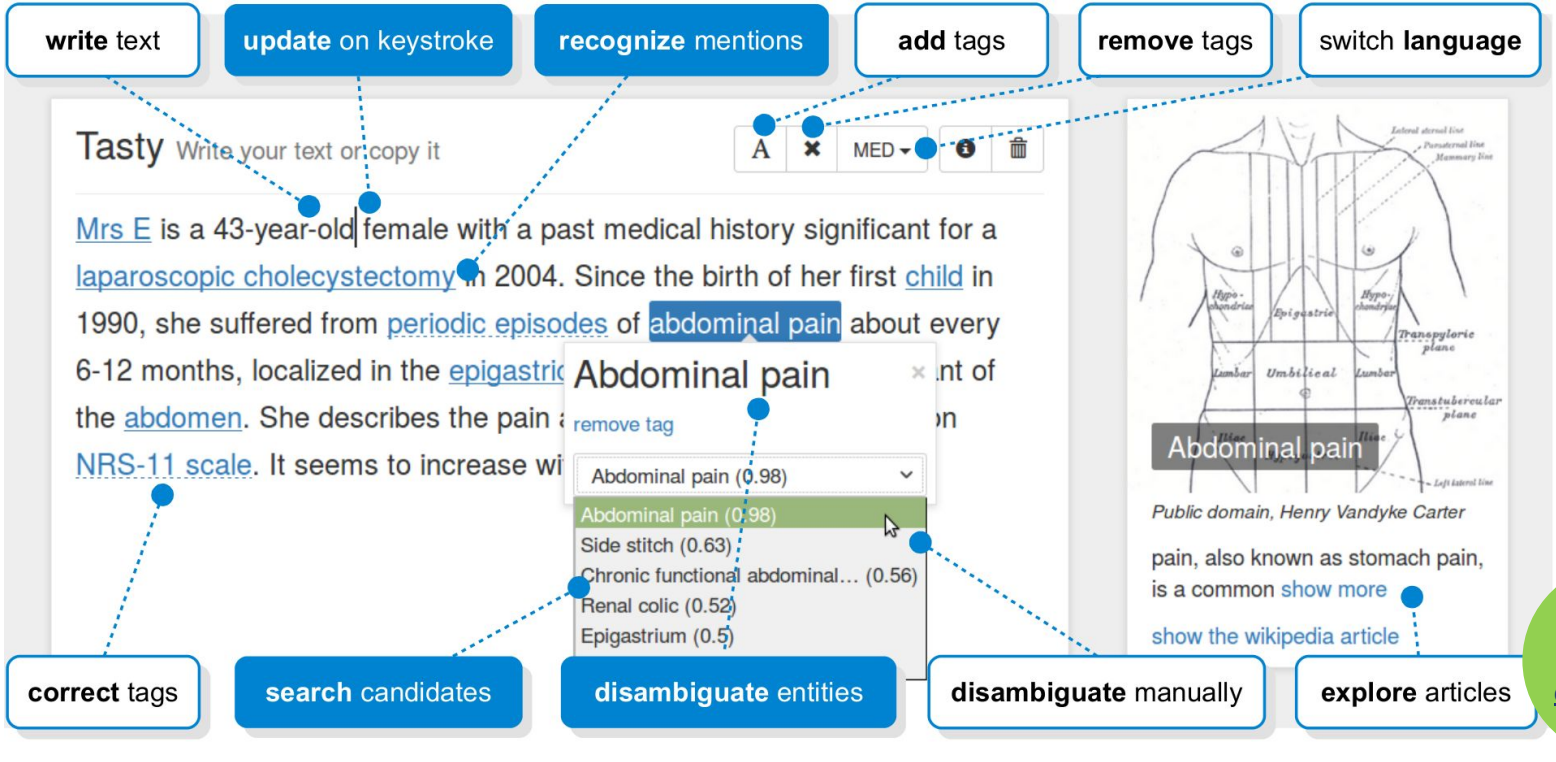
- CDV **extends pre-trained Language Models** with long-range discourse information
- Trained with **self-supervision** on Wikipedia data
- Pre-computed vectors enable **efficient retrieval** with high recall
- Outperforms BM25 (Elastic Search), DSSM, Duet (Microsoft), MatchPyramid, Conv-KNRM, HAR and 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.**

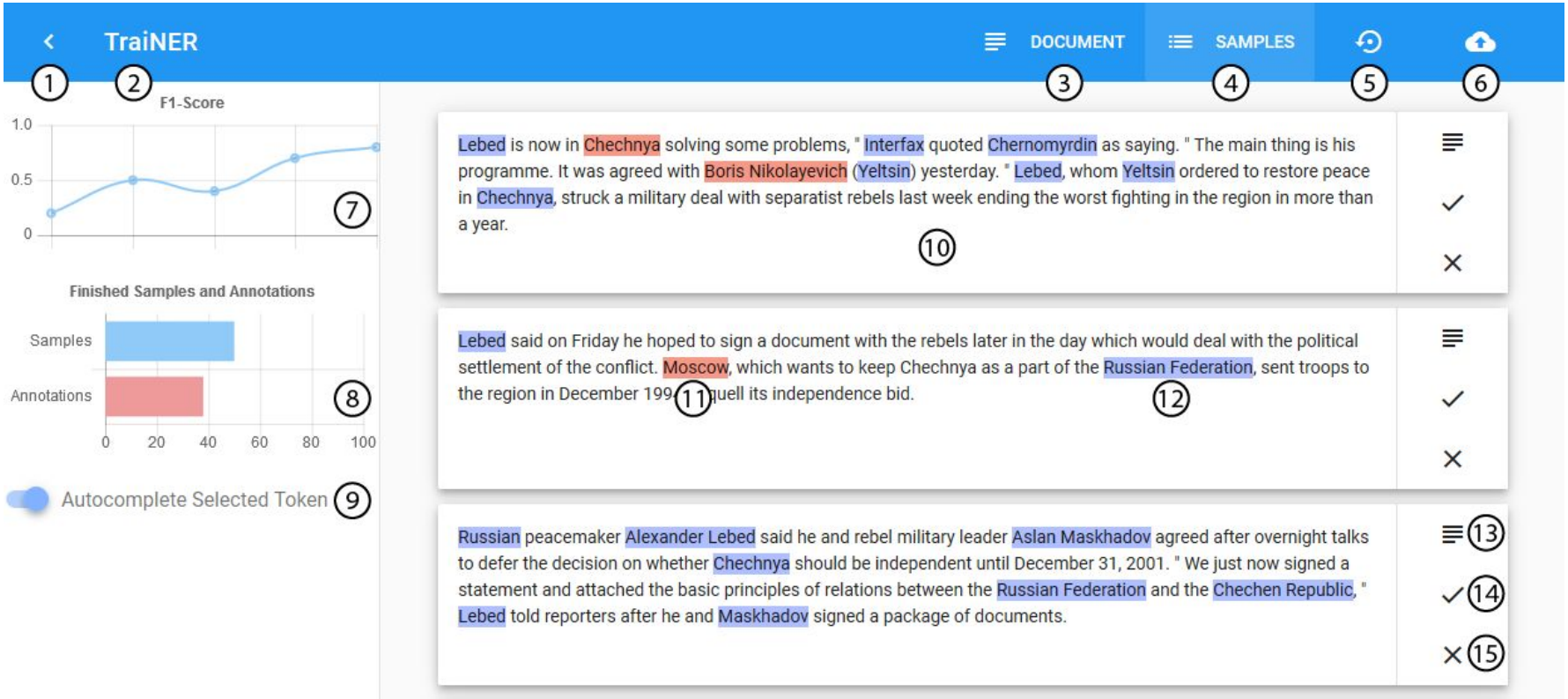
6. **Systems**

TASTY: Tag-as-you-type Entity Linking

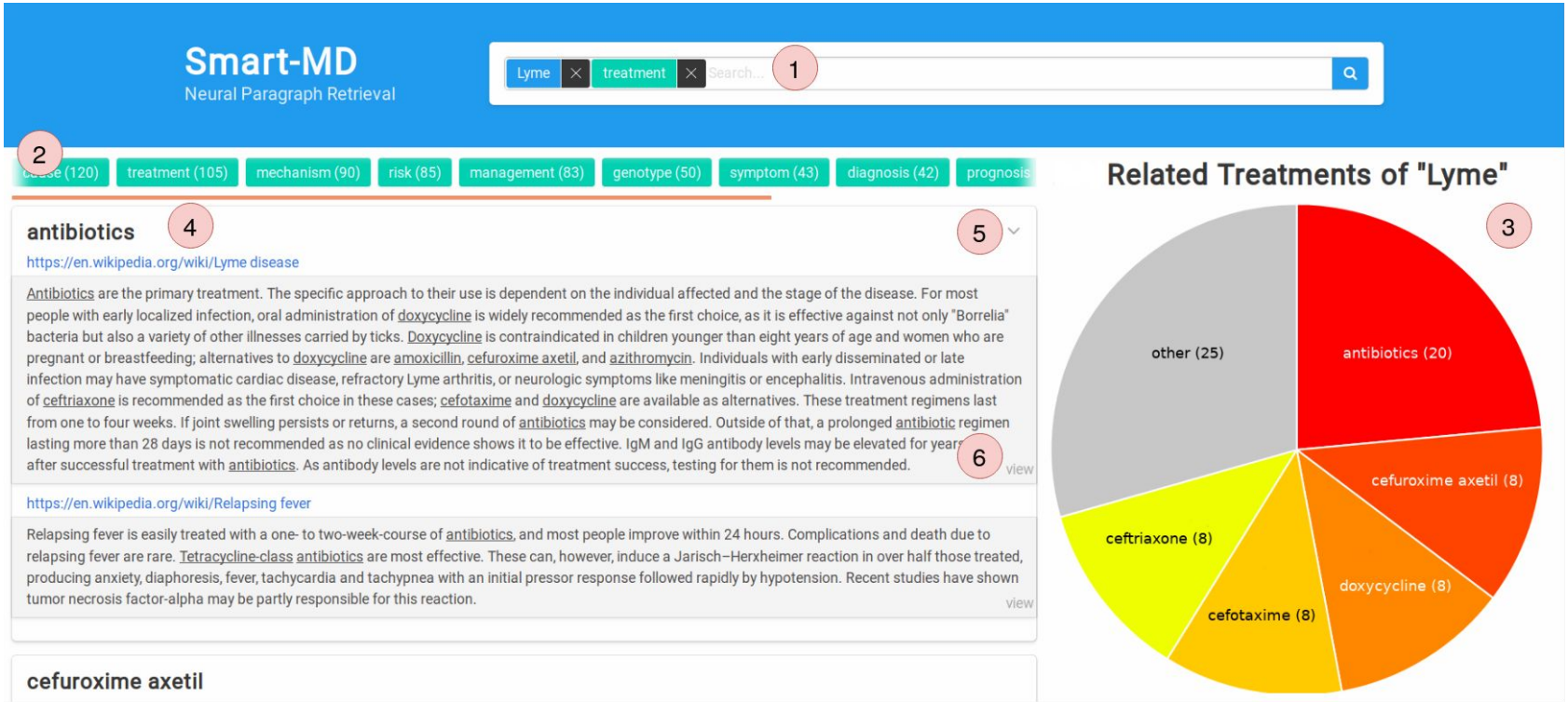


S. Arnold, F. A. Gers, T. Kilias and A. Löser (2016). **Robust Named Entity Recognition in Idiosyncratic Domains.** arXiv:1608.06757 [cs.CL]
S. Arnold, R. Dziuba and A. Löser (2016). **TASTY: Interactive Entity Linking As-You-Type.** COLING 2016 Demos

TraiNER: Bootstrapping Named Entity Recognition

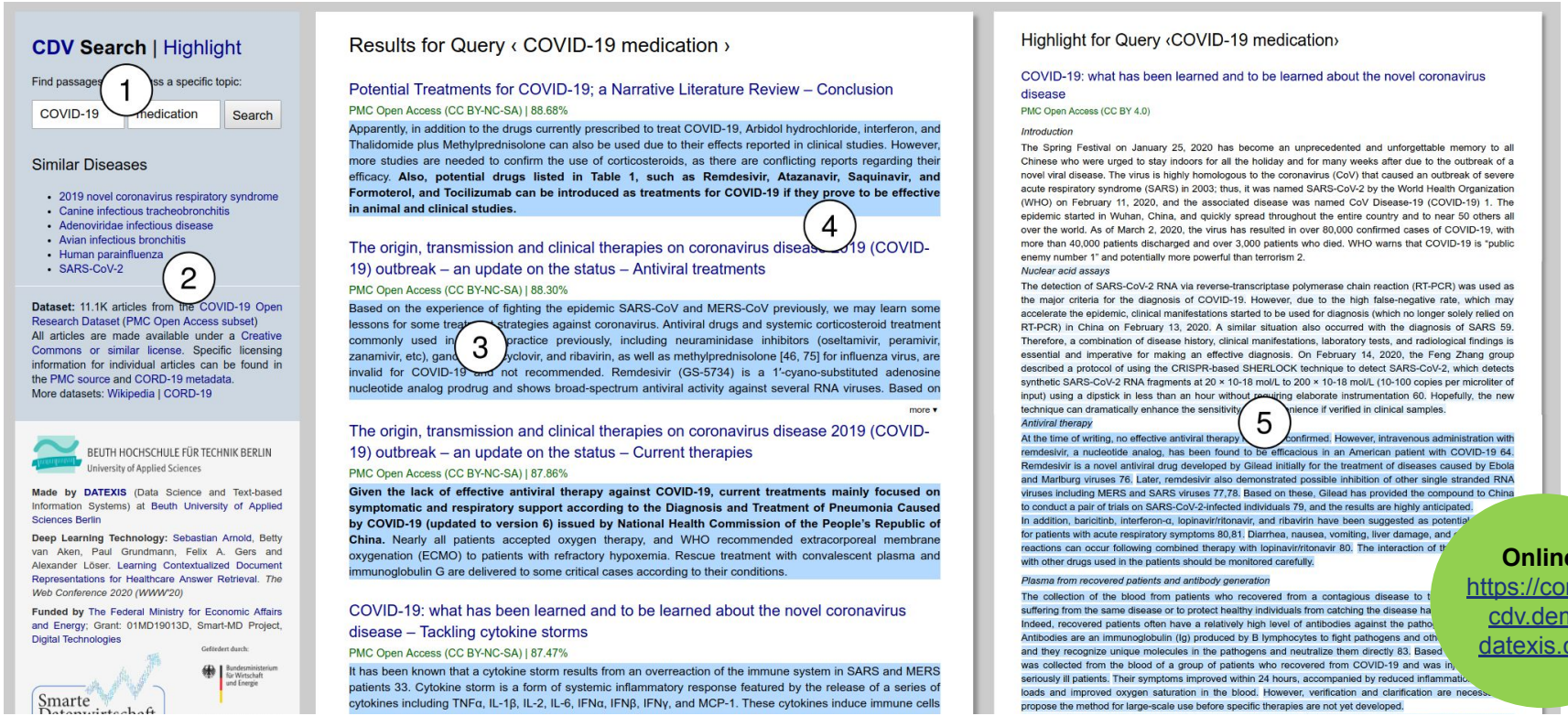


Smart-MD: Clinical Decision Support System



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

CDV Healthcare Retrieval



J.-M. Papaioannou, S. Arnold, F. A. Gers, A. Löser, M. Mayrdorfer, and K. Budde (2020). **Aspect-Based Passage Retrieval with Contextualized Discourse Vectors**. Submitted to: ECIR 2021 System Demonstrations

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 Kilias, 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

Neural Machine Reading for Domain-Specific Text Resources

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:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

Speaker: Sebastian Arnold

sarnold@beuth-hochschule.de
@sebastianarnold

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

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