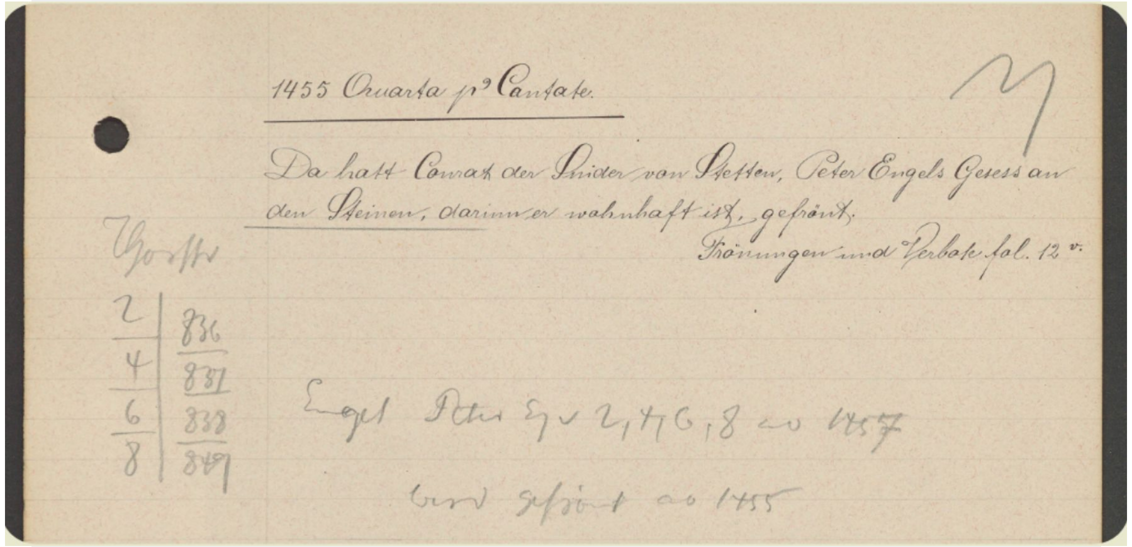
Economies of Space. Practices, Discourses and Actors on the Basel Real Estate Market (1400-1700)

Ismail Prada Ziegler

The Historical Land Register of Basel
1400-1700

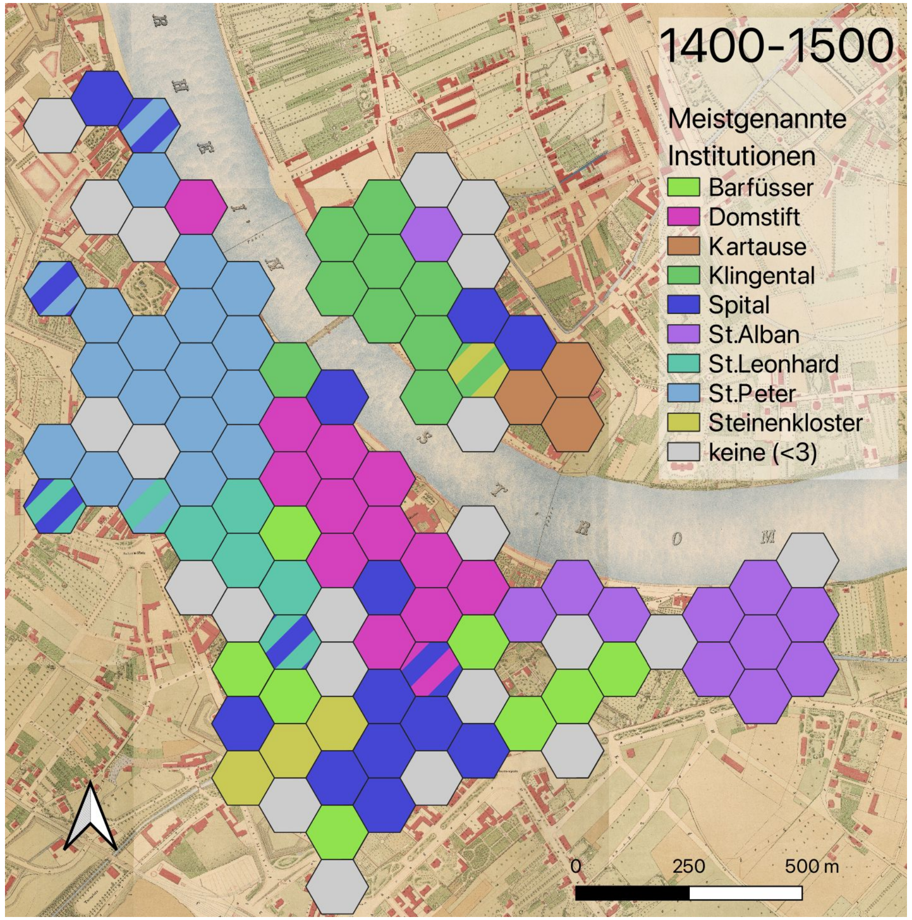
- ca. 80’000 in timeframe
- Usually one main event per text:
  - property purchase
  - seizures
  - rent purchase
  - testament / inheritance
  - court litigation
  - rent registries
  - etc.

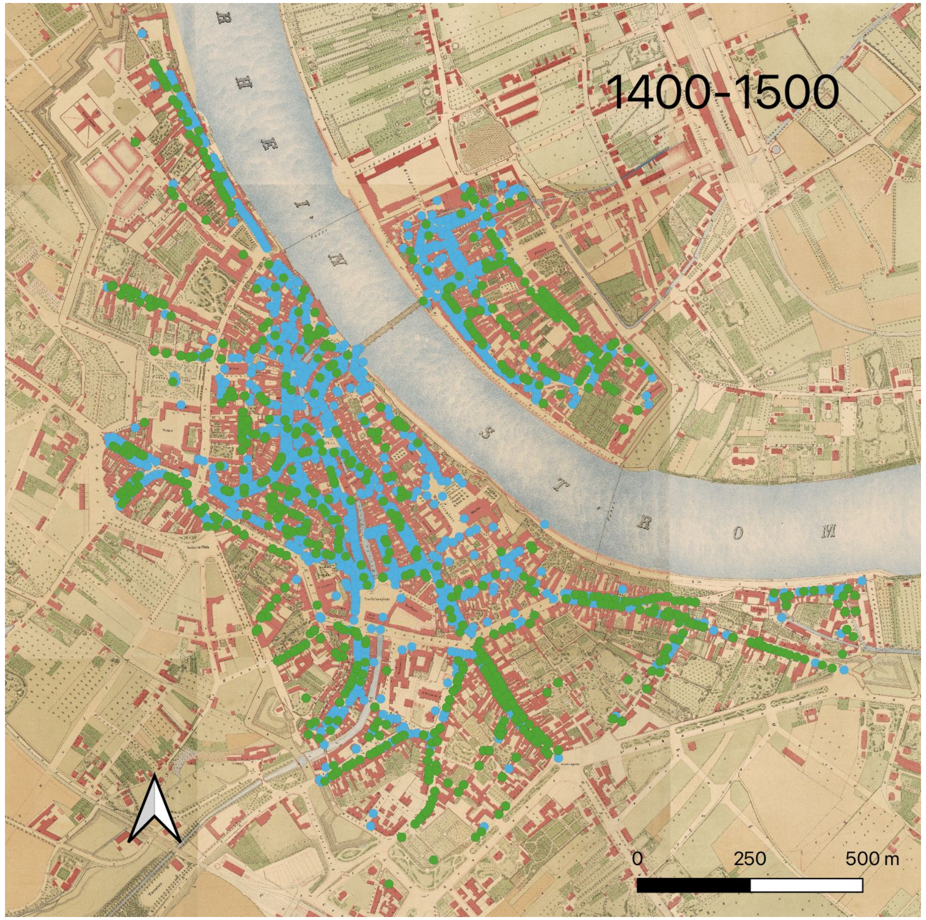


A typical document of the HLRB

Enneli Ebin and Lienhart Mornach the butcher as representative of Ulin Mornach, Ulrich Mornachs son and mentioned Ennelis son, sell Symon Sumer the baker and his wife Clara the house which is called Lemlis Hus located at Yengassen between the houses Zum Agstein on one side and the Roten Hus to the other, at the backside touching S. Martinshof, paying interest to the diocese 2 lb & 2 bags of grain to the church S. Martin, otherwise free, for 625 fl.

How do we get there?





Event Analysis

Enneli Ebin and Lienhart Mornach the butcher as representative of Ulin Mornach, Ulrich Mornachs son and mentioned Ennelis son, sell Symon Sumer the baker and his wife Clara the house which is called Lemlis Hus located at Yengassen between the houses Zum Agstein on one side and the Roten Hus to the other, at the backside touching S. Martinshof, paying interest to the diocese 2 lb & 2 bags of grain to the church S. Martin, otherwise free, for 625 fl.

Common NE-Annotation

Enneli Ebin and Lienhart Mornach the butcher as representative of Ulin Mornach, Ulrich Mornachs son and mentioned Ennelis son, sell Symon Sumer the baker and his wife Clara the house which is called Lemlis Hus located at Yengassen between the houses Zum Agstein on one side and the Roten Hus to the other, at the backside touching S. Martinshof, paying interest to the diocese 2 lb & 2 bags of grain to the church S. Martin, otherwise free, for 625 fl.

Long NE-Annotations + ~~Named~~ Entities

Enneli Ebin and Lienhart Mornach the butcher as representative of Ulin Mornach, Ulrich Mornachs son and mentioned Ennelis son, sell Symon Sumer the baker and his wife Clara the house which is called Lemlis Hus located at Yengassen between the houses Zum Agstein on one side and the Roten Hus to the other, at the backside touching S. Martinshof, paying interest to the diocese 2 lb & 2 bags of grain to the church S. Martin, otherwise free, for 625 fl.

BeNASCh - Quick Introduction

- **Text Layer**
  - Annotate Reference-Mentions and Values.

- **Description Layer**
  - +Head - required!
  - +Attributes - entity mention further describing parent
  - +Descriptors - non-mention further describing parent

- Nested: References & Attributes contain Description Layers, Descriptors contain Text Layers.

Layer Examples

Symon Sumer the baker

Description Layer:

Head: Symon Sumer

Attribute: the baker

Description Layer:

Head: baker

Layer Examples

the house located at Yengassen, paying interest to the diocese 2lb

Description Layer:

Head: house

Descriptor: located at Yengassen

Text Layer:

Reference: Yengassen

Description Layer:

Head: Yengassen

Descriptor: paying interest to the diocese 2lb

…Text Layer etc.

Symbiosis with Event/Relation Annotation

his wife Clara

Description Layer:

Head: Clara

Attribute: his wife ← Mention Subclass: Family / Wife

Description Layer:

Head: wife

Reference: his → Coreference to mention of husband

BeNASch is a "close-to-text" annotation

Symbiosis with Event/Relation Annotation

the house located at Yengassen, paying interest to the diocese 2lb

Description Layer:

Head: house

Descriptor: located at Yengassen ← Descriptor Class: located_at

Text Layer:

Reference: Yengassen

Description Layer:

Head: Yengassen

Descriptor: paying interest to the diocese 2lb ← Descriptor Class: dues

…Text Layer etc.

NE-Classes

- 4 classes:
  - Person
  - Place
  - Organization
  - Geo-political Entity
  - Considered: Miscellaneous
  - Custom Classes: e.g. Facility for HGB for Buildings

- Additional Mention Information
  - Mention Class, Mention Subclass, Specificity, Ordinality

Workflow

HTR

Transkribus

Unlock the past

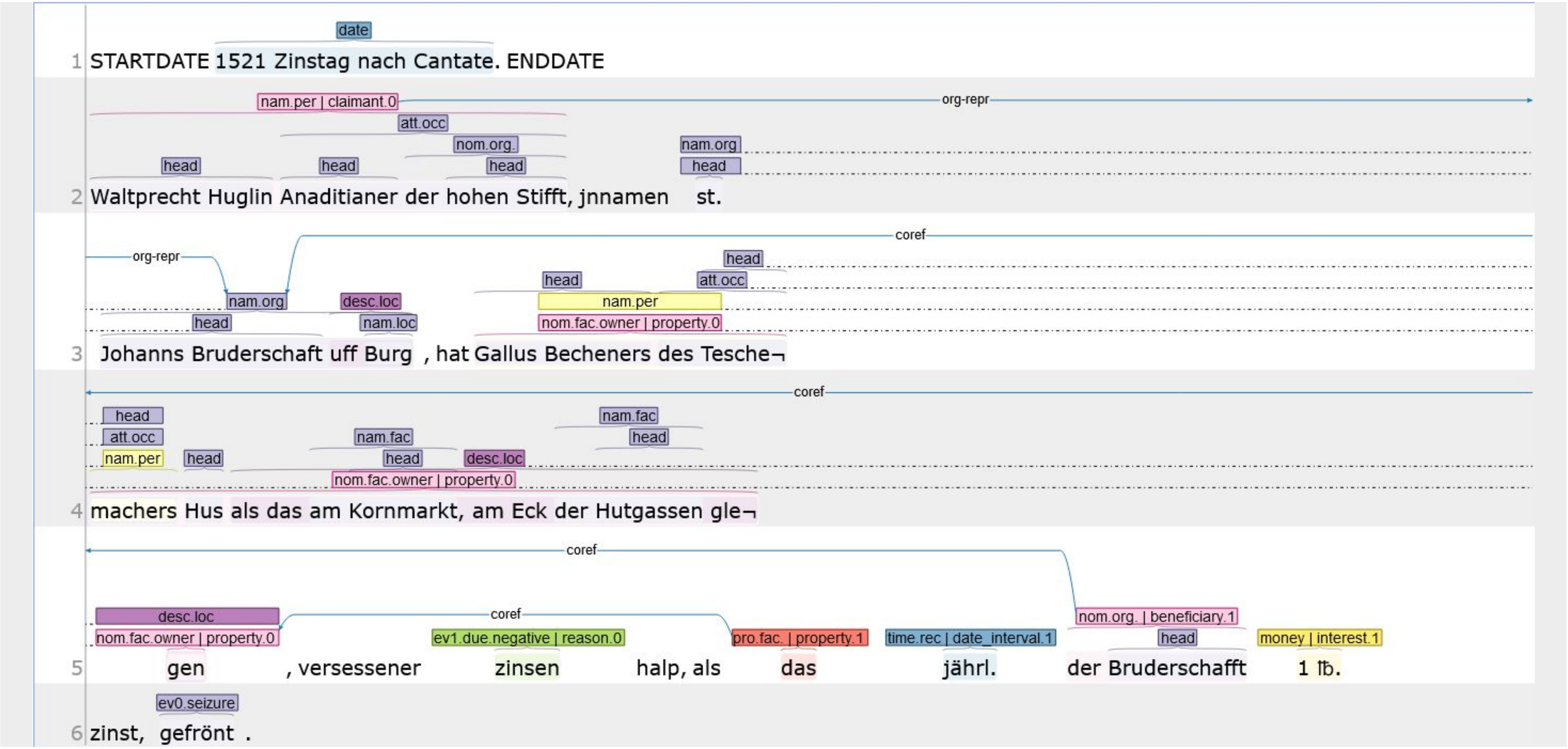
3.6% CER

Manual Annotation

INCEpTION

> 800 documents

Annotation in INCEpTION



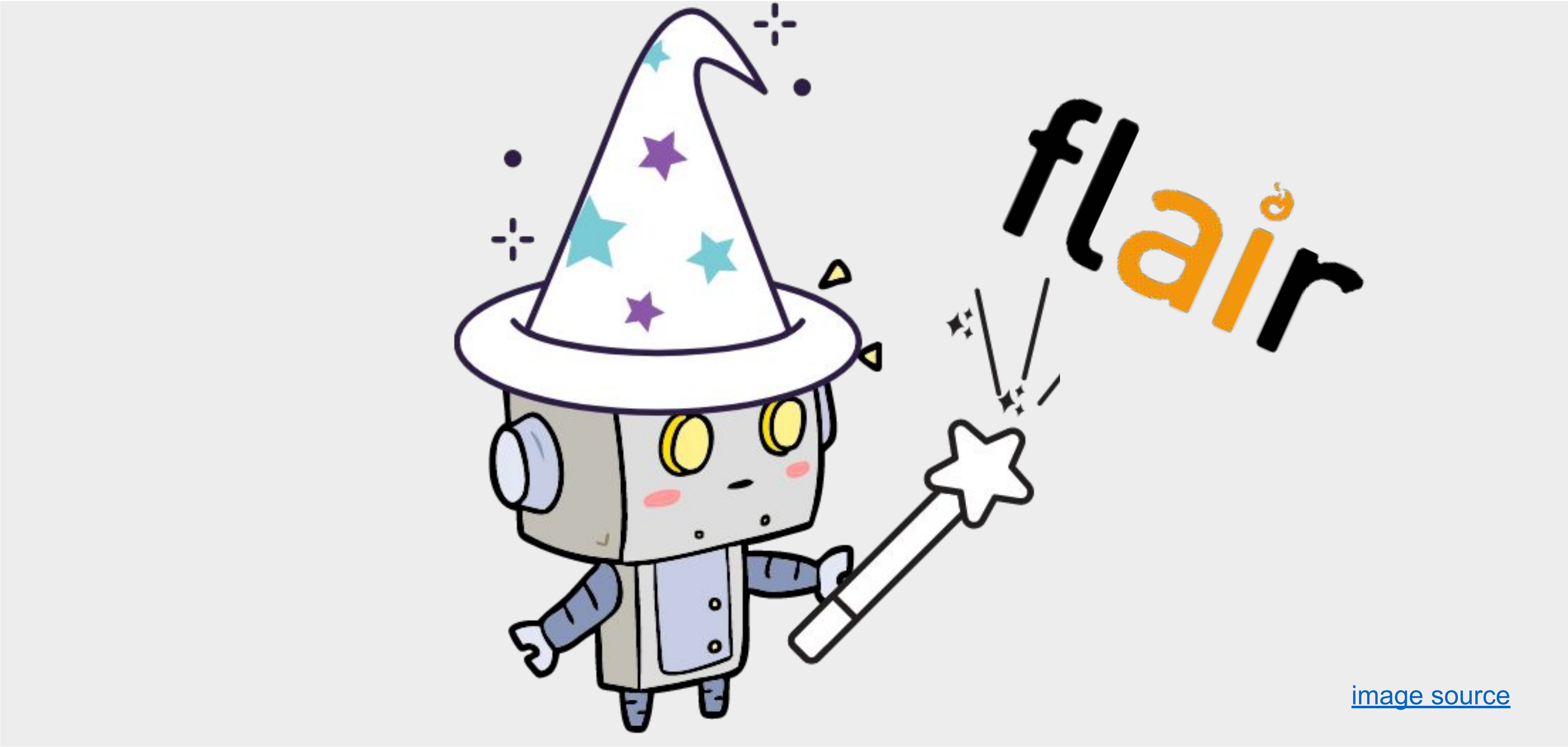
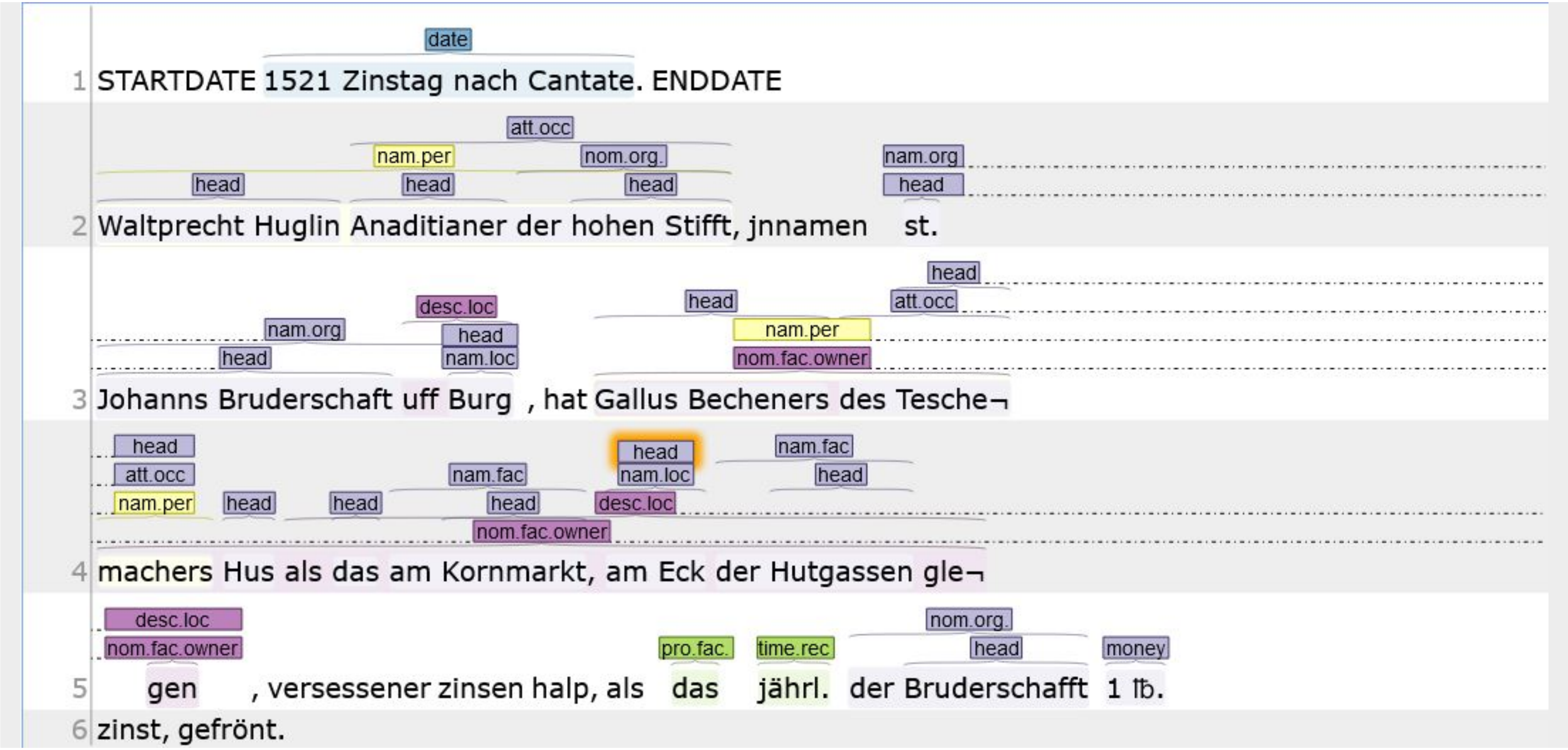


image source

Automated Annotation!



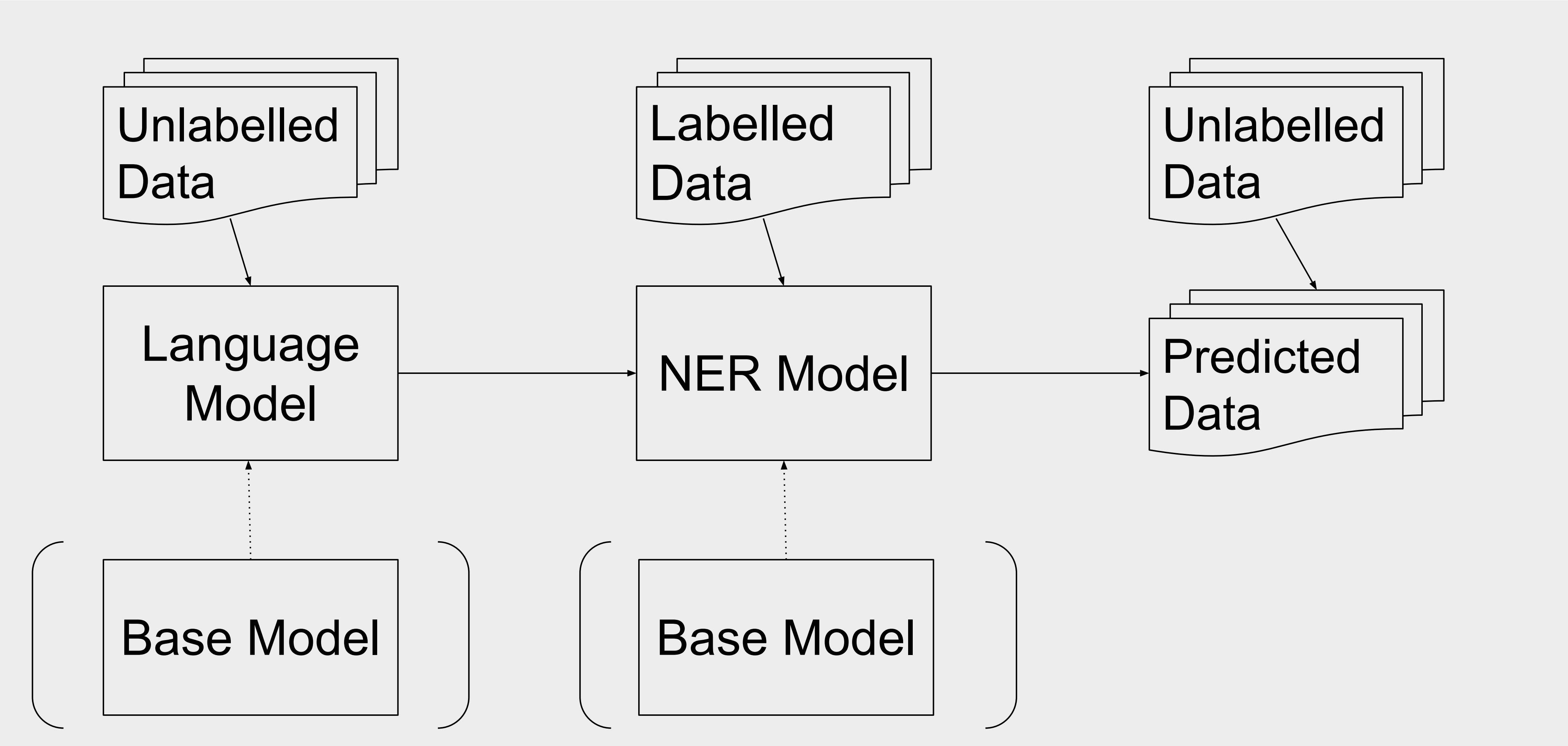
Next Steps

- Normalization and Entity Linking
- Relation Extraction outside of Nesting
  - esp. Coreference Resolution
- Event Extraction

How to Extract (Named) Entities in Historical Texts?

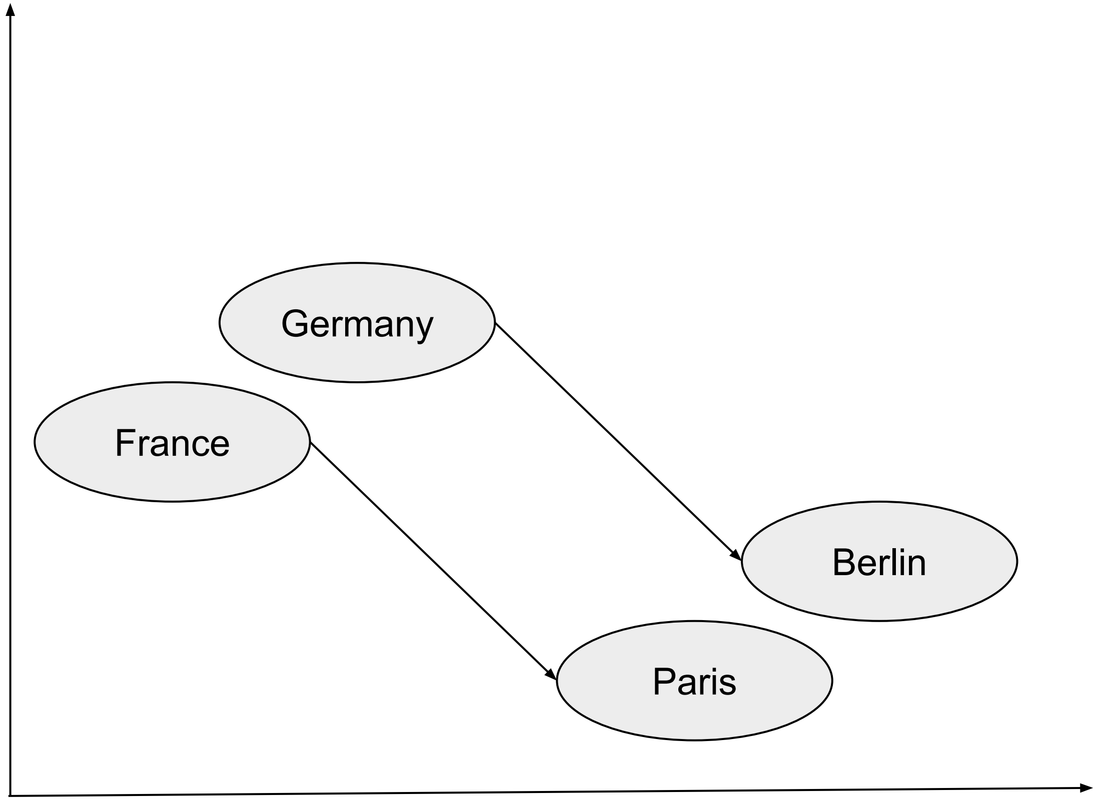
Ismail Prada Ziegler

In One Image: Sequence Tagging



Language Models

- Mathematical representation of a token / subtoken / character

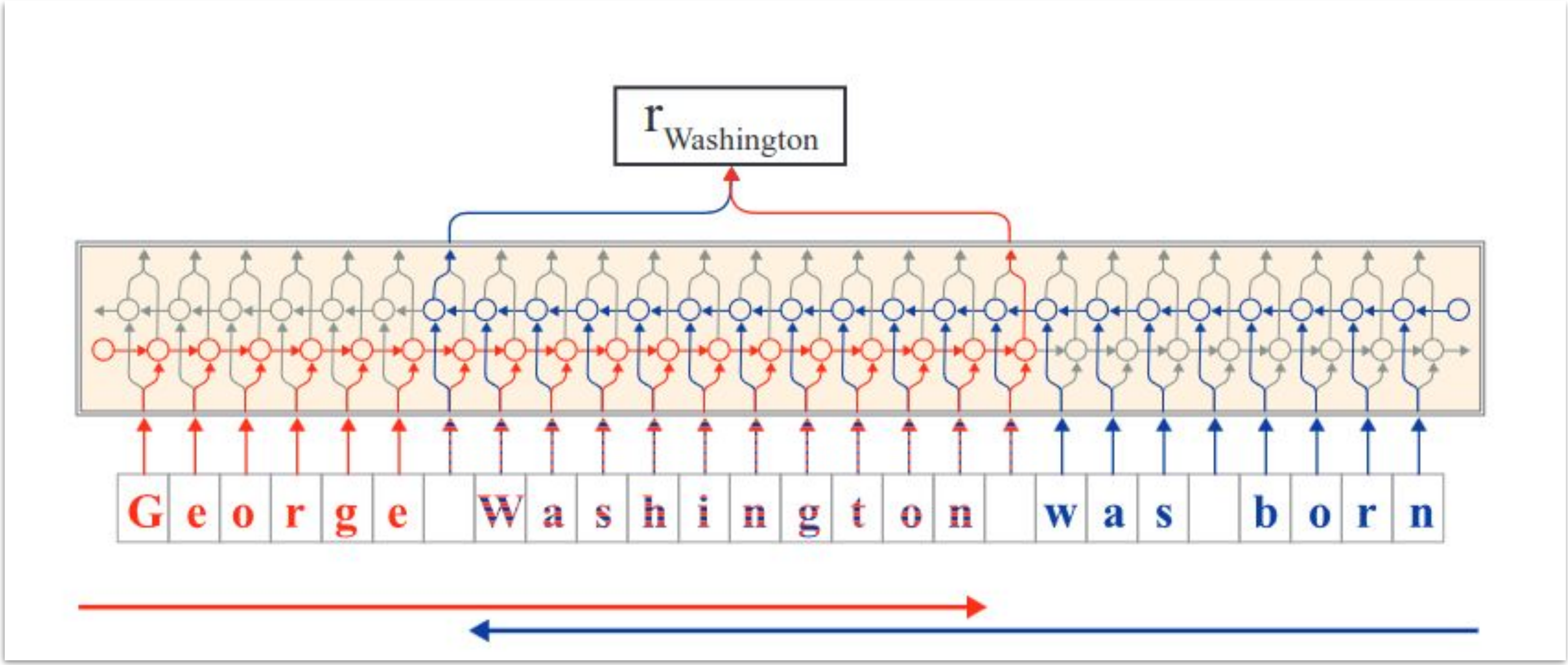


Language Models

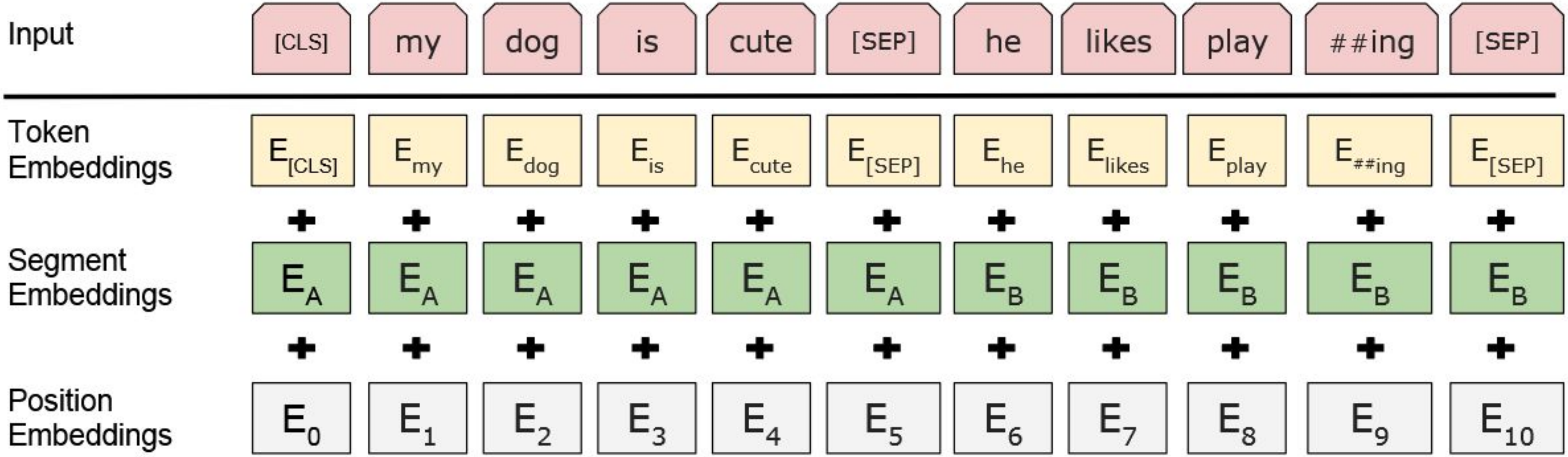
- Popular Embeddings:
  - word2vec (word-level)
  - fasttext (word-level + subword-information)
  - Contextualized Embeddings:
    - contextual character-emb (Flair) (character-level)
    - Transformers (BERT) (WordPiece-level)

- Trained by learning to predict likely words correctly from context (strongly simplified)

Contextual Character Embeddings



BERT (Transformer)

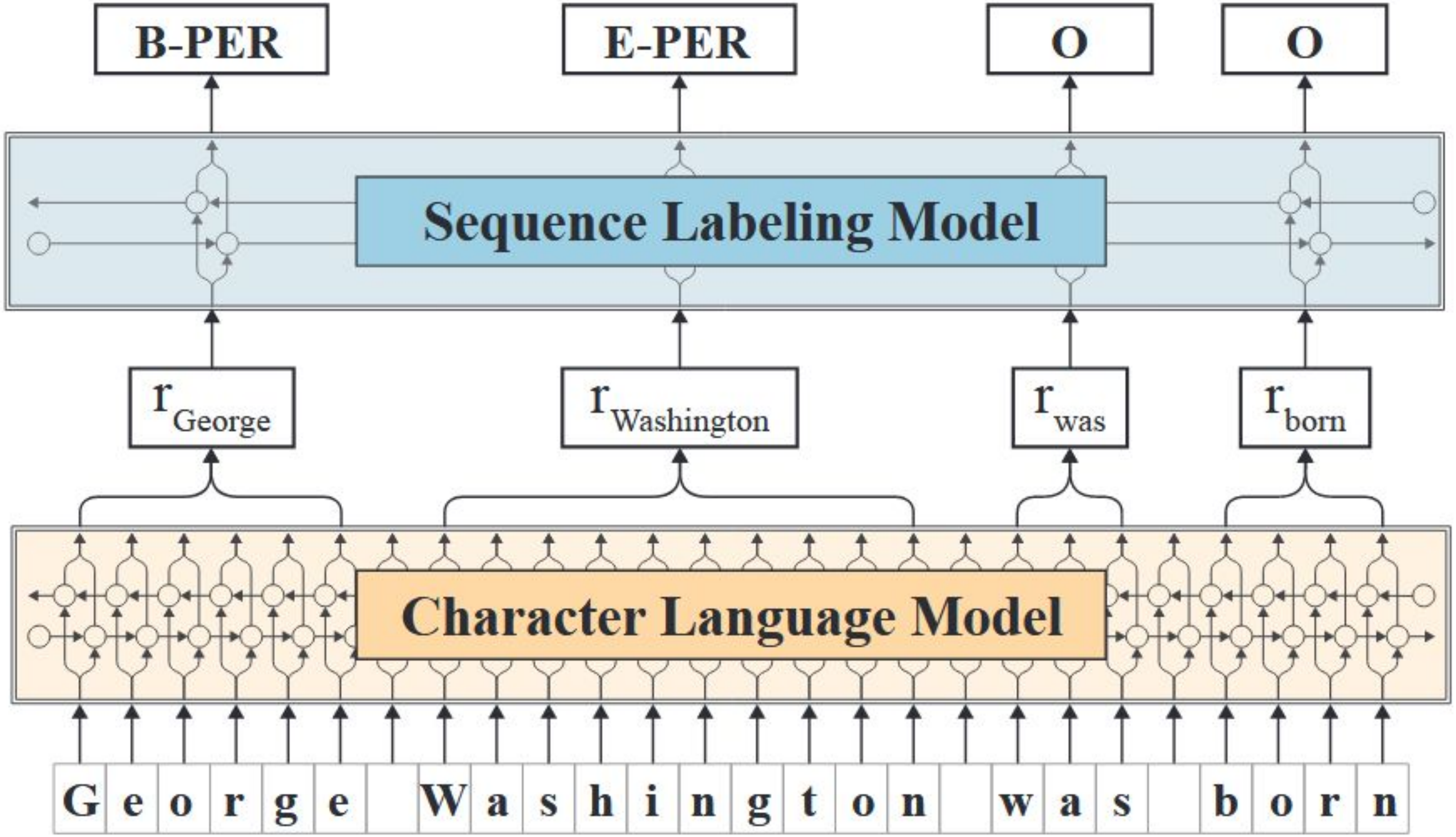


Labelled Data Format

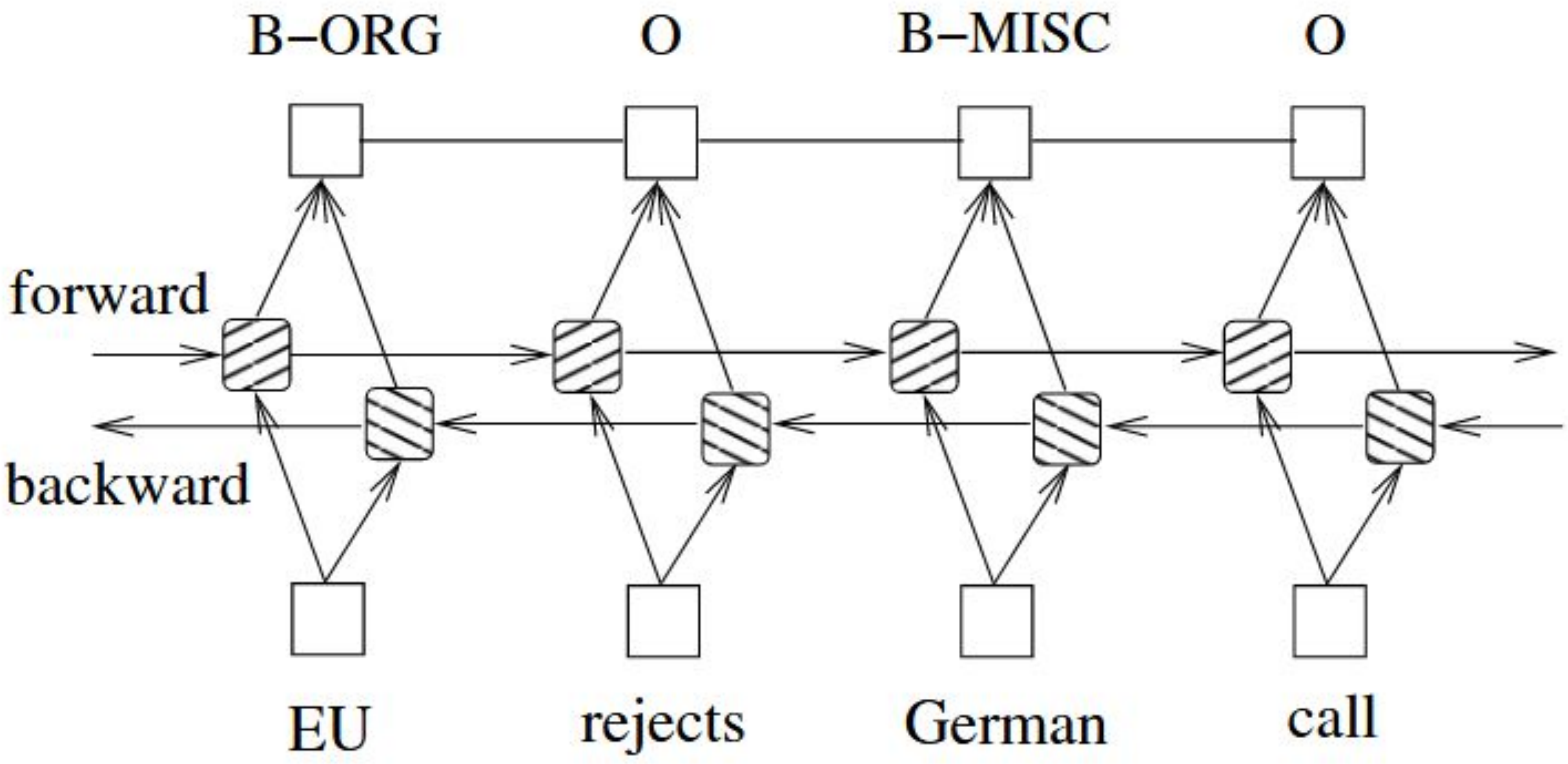
- Typical Format: BIO(ES)

Symon	B-PER
Sumer	I-PER
the	I-PER
baker	E-PER
lives	O
at	O
the	O
Yengassen	S-LOC

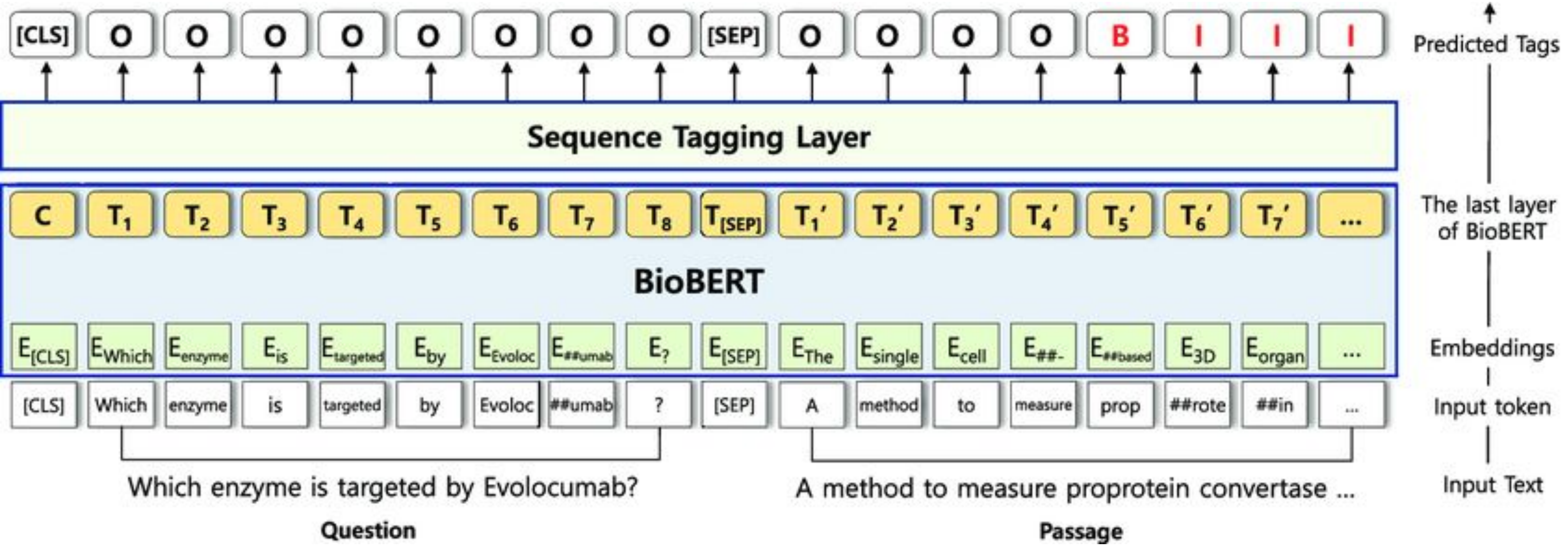
Flair Sequence Tagging Architecture



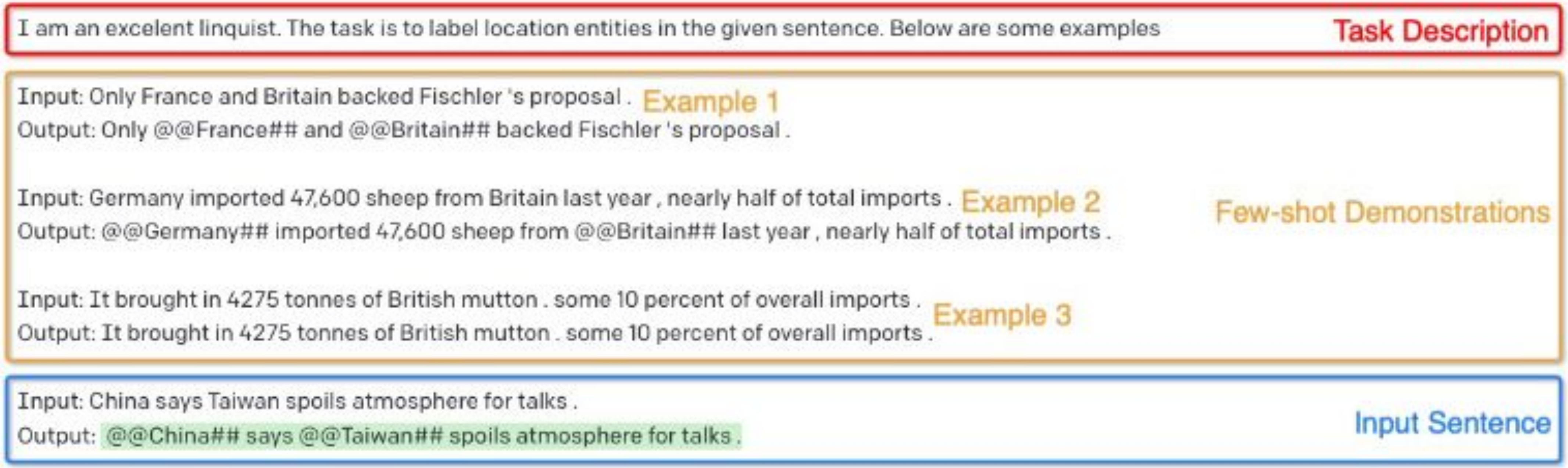
Bi-LSTM + CRF (Huang et al. 2015)



Transformer Sequence Tagging Arch.



Generative AI (LLMs)



Source: Wang et al. 2023

Main Training Strategies

- Training from scratch
  - Own training data needed

- Finetuning
  - Own training data needed
  - Harness the power of a pre-trained model

- Few-Shot Learning
  - Only show a few examples to “explain” the task to the model
  - Harness the power of a pre-trained model

Evaluation Metrics

Ground Truth	Prediction	Counts as
PER	PER	True Positive
O	O	True Negative
PER	O	False Negative
O	PER	False Positive

Recall → TP / (TP + FN)

Precision → TP / (TP + FP)

F-Score → 2 * ((Recall * Precision) / (Recall + Precision))

Challenges with Historical Data

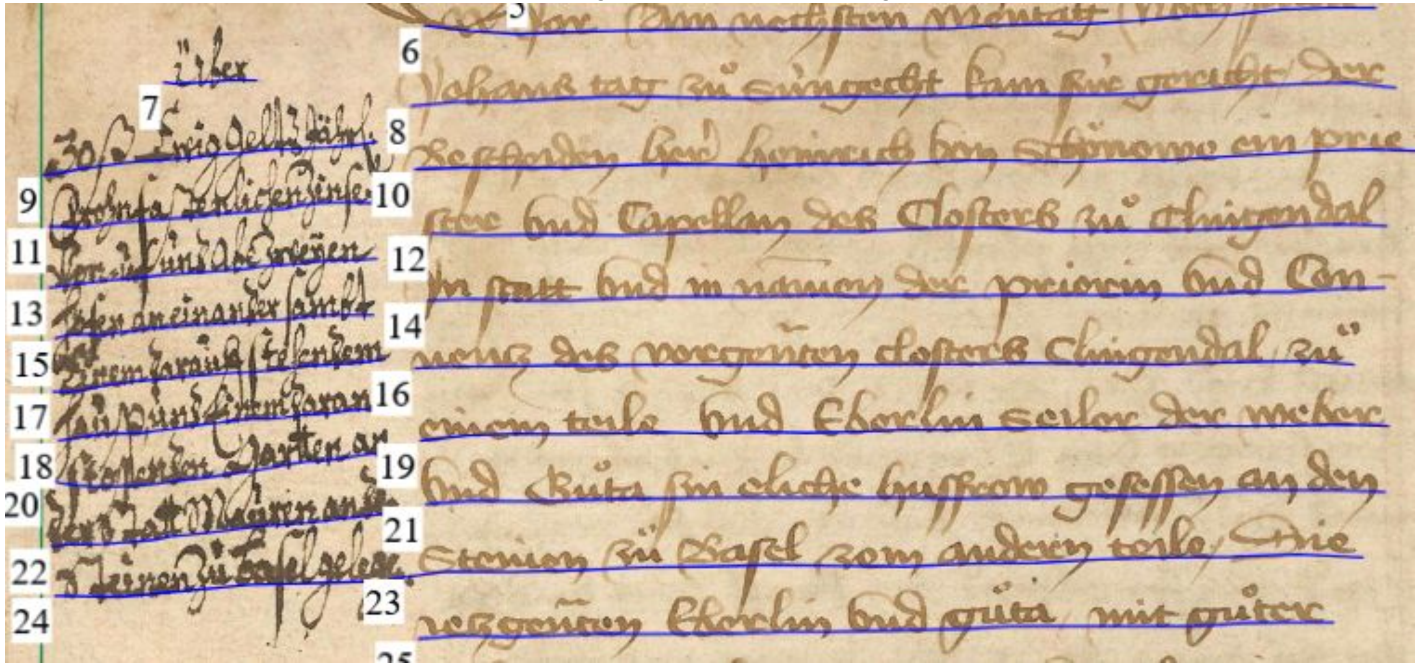
- Let’s collect some together!

- Great summary can be found in Ehrmann et al. 2023: “Named Entity Recognition and Classification in Historical Documents: A Survey”

Challenges: Noisy Input

- Errors in OCR/HTR
  - measured in *Character Error Rate* or *Word Error Rate*

- Errors in the Layout Analysis



Challenges: Dynamics of Language

- Historical Spelling Variation

Conratz, Cunratz, Cunraden, Cunraten, Conraten, Chunrad, Cunrad, Cunrads, Conrad, Conradt, Cunrat, Conrat, Cuntz, Contz

- Naming Conventions
- Entity and Context Drift

Challenges: Lack of Resources

Corpus	Doc. type	Time period	Tag set (# types)	Lang.	# NEs	Size	License
Quaero Old Press (2012) [175]	newspapers	19C	Quaero★ (8)	fr	147,682	XL	ELRA
Europeana (2016) [142]	newspapers	19C	PER,LOC,ORG (3)	fr, de, nl	40,801	L	CC0
De Gasperi (2016) [189]	various types	20C	PER,GPE (2)	it	35,491	L	CC BY-NC-SA
Latin NER (2016) [69]	literary texts	1C BCE–2C	PER,GEO,GRP (3)	la	7,175	S	GPL v3.0
HIMERA (2016) [198]	medical lit.	19C–21C	custom (7)	en	8,400	S	CC BY
Venetian references (2017) [42]	publications	19C–21C	custom (3 or 26)	Multi	12,879	M	CC BY
Finnish NER (2018) [178]	newspapers	19C–20C	PER,LOC,ORG (3)	fi	26,588	M	n/a
DROC (2018) [115]	novels	17C-20C	custom (?)	de	6,013	S	CC BY
Travel writings (2018) [187]	travelogues	19C–20C	LOC (1)	en	2,228	S	n/a
Coptic Scriptorium (2018)	literary texts	3C–5C	custom (10)	cop	88,068	L	CC BY
LitBank (2019) [17]	novels	19C–20C	ACE (w/o WEA) (6)	en	14,000	L	CC BY-SA
BIOfid (2019) [4]	publications	18C–20C	extended GermEval (5)	de	33,545	L	GPL v3.0
Cz. Hist. NE Corpus (2020) [101]	newspapers	19C	custom (5)	cz	4,017	S	CC BY-NC-SA
HIPE-2020 (2020) [62]	newspapers	18C–21C	*impresso*★ (5)	de, en, fr	19,848	M	CC BY-NC-SA
BDCamões (2020) [86]	literary texts	16C–21C	custom (6)	pt	144,600	XL	CC BY-NC-ND
GeoNER (2020) [113]	literary texts	16C–17C	GEO (1)	fr	264	S	LGPL-LR
NewsEye (2021) [93]	newspapers	19C–20C	*impresso*-comp. (4)	de, fr, fi, sv	30,580	L	CC BY
TopRes19th (2021) [12]	newspapers	19C	toponyms (6)	en	3,364	S	CC BY-NC-SA
Charters (2022) [202]	medieval charters	10C-15C	PER,LOC (2)	fr, la, sp	-	-	not stated
Est. Parish Court records (2022) [147]	court records	19C	custom (7)	et	27,540	M	not stated
AjMC (2022)	class. commentaries	19C	custom★ (6)	de, en, fr	7,482	S	CC BY
HIPE-2022 (2022) [63]	newspapers & classics	19C–20C	various★ (12)	de, en, fr, fi, sv	71,114	L	various

In the column *Tag set*, the star superscript indicates that the used typology is organised in a taxonomy. In such case, the number of types *(# types)* corresponds to the higher level.

Source: Ehrmann et al. 2023

Challenges: Lack of Resources

Corpus	Doc. type	Time period	Tag set (# types)	Lang.	# NEs	Size	License
Quaero Old Press (2012) [175]	newspapers	19C	Quaero* (8)	fr	147,682	XL	ELRA
Europeana (2016) [142]	newspapers	19C	PER,LOC,ORG (3)	fr, de, nl	40,801	L	cc0
De Gasperi (2016) [189]	various types	20C	PER,GPE (2)	it	35,491	L	CC BY-NC-SA
Latin NER (2016) [69]	literary texts	1C BCE–2C	PER,GEO,GRP (3)	la	7,175	S	GPL v3.0
HIMERA (2016) [198]	medical lit.	19C–21C	custom (7)	en	8,400	S	CC BY
Venetian references (2017) [42]	publications	19C–21C	custom (3 or 26)	Multi	12,879	M	CC BY
Finnish NER (2018) [178]	newspapers	19C–20C	PER,LOC,ORG (3)	fi	26,588	M	n/a
DROC (2018) [115]	novels	17C-20C	custom (?)	de	6,013	S	CC BY
Travel writings (2018) [187]	travelogues	19C–20C	LOC (1)	en	2,228	S	n/a
Coptic Scriptorium (2018)	literary texts	3C–5C	custom (10)	cop	88,068	L	CC BY
LitBank (2019) [17]	novels	19C–20C	ACE (W/O WEA) (6)	en	14,000	L	CC BY-SA
BIOfid (2019) [4]	publications	18C–20C	extended GermEval (5)	de	33,545	L	GPL v3.0
Cz. Hist. NE Corpus (2020) [101]	newspapers	19C	custom (5)	cz	4,017	S	CC BY-NC-SA
HIPE-2020 (2020) [62]	newspapers	18C–21C	*impresso** (5)	de, en, fr	19,848	M	CC BY-NC-SA
BDCamões (2020) [86]	literary texts	16C–21C	custom (6)	pt	144,600	XL	CC BY-NC-ND
GeoNER (2020) [113]	literary texts	16C–17C	GEO (1)	fr	264	S	LGPL-LR
NewsEye (2021) [93]	newspapers	19C–20C	*impresso*-comp. (4)	de, fr, fi, sv	30,580	L	CC BY
TopRes19th (2021) [12]	newspapers	19C	toponyms (6)	en	3,364	S	CC BY-NC-SA
Charters (2022) [202]	medieval charters	10C-15C	PER,LOC (2)	fr, la, sp	-	-	not stated
Est. Parish Court records (2022) [147]	court records	19C	custom (7)	et	27,540	M	not stated
AjMC (2022)	class. commentaries	19C	custom* (6)	de, en, fr	7,482	S	CC BY
HIPE-2022 (2022) [63]	newspapers & classics	19C–20C	various* (12)	de, en, fr, fi, sv	71,114	L	various

In the column *Tag set*, the star superscript indicates that the used typology is organised in a taxonomy. In such case, the number of types *(# types)* corresponds to the higher level.

Source: Ehrmann et al. 2023

Solutions: Noise

- **Character-level emb.** can handle **noisy input** and **spelling variation**
  - Spelling normalization and OCR Post-correction are usually not necessary anymore
  - WordPiece struggles here

- CharBERT introduces character-level features into BERT

Solutions: Low Resource

- **Finetuning / Few-shot-technique** don’t need as
  - GenAI has shown to outperform Transformer systems in low resource scenarios (Wang et al. 2023)

- Enough data for Transformer hard to come by
  - (enough for Latin though → Latin BERT)

- When using other (historical) datasets/models → Domain shift problem!

Example Experiments (16th-18th c. dutch)

Type	GT Layers	Embeddings	Prec.	Recall	F_1	support
PER	PER	Char, GysBERT	0.81	0.69	0.75	405.00
ATT	HOE	Char, FastText, GysBERT	0.57	0.56	0.56	573.00
COM	COM	Char	1.00	0.73	0.85	41.00
ORG	ORG	Char, FastText, GysBERT	0.82	0.71	0.76	283.00
LOC	LOC	FastText, GysBERT	0.79	0.76	0.77	570.00
DAT	DAT	Char	0.90	0.88	0.89	249.00
RES	All	FastText, GysBERT	0.82	0.70	0.75	57.00
OTH	All	Char, FastText, GysBERT	0.63	0.26	0.36	47.00

Source: Koolen et al. 2024

Now let’s get our own hands on it!

Image Sources

- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Yoon, Wonjin & Jackson, Richard & Lagerberg, Aron & Kang, Jaewoo. (2022). Sequence Tagging For Biomedical Extractive Question Answering. Bioinformatics. 38. 10.1093/bioinformatics/btac397.
- Wang, Shuhe & Sun, Xiaofei & Li, Xiaoya & Ouyang, Rongbin & Wu, Fei & Zhang, Tianwei & Li, Jiwei & Wang, Guoyin. (2023). GPT-NER: Named Entity Recognition via Large Language Models.
- Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, *56*(2), 27:1-27:47. https://doi.org/10.1145/3604931
- Koolen, M., Renkema, E., Groskamp, N., Smit, F., Reinders, J., Sluijter, R., ... & Oddens, J. Accessing the Republic. Entity extraction from the resolutions of the Dutch States-General.