Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Data Mining & NLP in e-rara

Semesterprojekt CAS Practical Machine Learning

Kathi Woitas, Universitätsbibliothek Bern

2021-04-08

Worum geht es?

e | rara

Titelansicht Inhaltsansicht Seitenansicht

Titelaufnahme

TITEL

Aktensammlung zur Geschichte der Berner-Reformation : 1521-1532 / hrsg. mit Unterstützung der bernischen Kirchensynode von R. Steck und G. Tobler

AUTOR, BETEILIGTE

Steck, Rudolf [1842-1924] G W ; Tobler, Gustav [1855-1921] G W

IMPRESSUM

Bern : Wyss, 1923

UMFANG

2 Bde. (IV, 1551 S.) ; 24 cm

Worum geht es?

www.e-rara.ch

- 84.6k Druckwerke des 15.-20. Jh. (60k Bücher, 19.4k mit OCR)
- **Bernensia des 18. bis frühen 20. Jh. -> 176 Bücher (de, OCR)**

Herausforderungen

- Keine Batch-Downloads für Metadaten, Images, Volltexte im GUI
- mangelnde OCR-Qualität!!!

Verfügbare Daten-Zugänge

- Metadaten per REST-API (OAI-PMH), allerdings undokumentiert

Motivation

- Nachnutzbarkeit der Digitalisierungsdaten untersuchen und verbessern
- einfache Hilfsmittel zu deren Nutzung entwickeln und bereitstellen

Zielsetzung

Kernfrage:

Können **Named-Entity-Recognition**-Ansätze auf Bernensia-Volltexte erfolgreich angewendet werden – trotz z.T. starker **OCR-Fehler**?

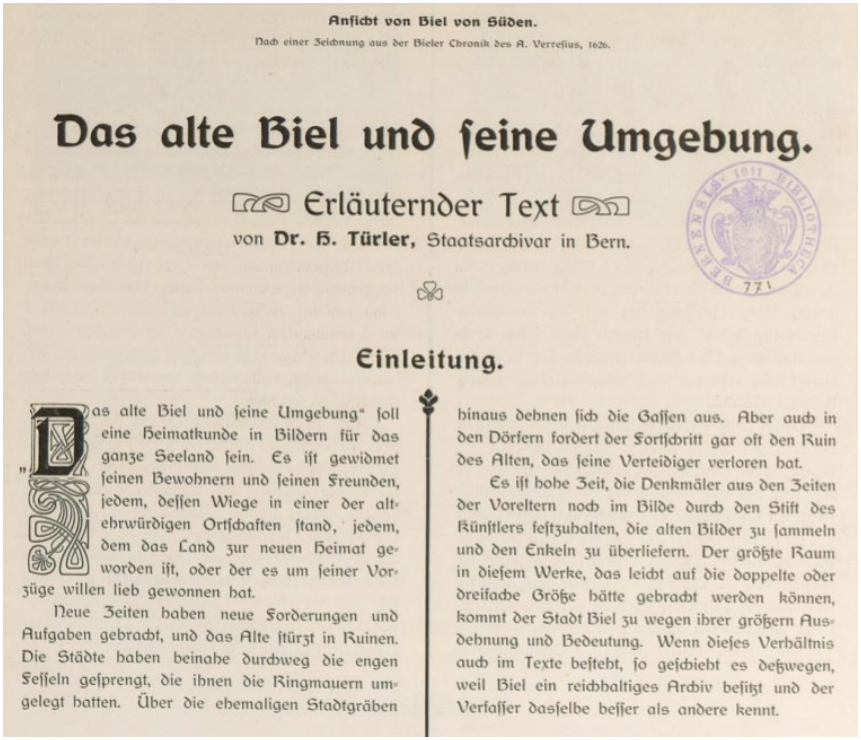
Ausgangslage: **realer Korpus** = problembeladen!

- deutsche, historische Sprache & z.T. schlechte OCR-Qualität
- inhomogen: Länge und Art der Dokumente (z.B. Adressenbücher, Ausstellungskataloge, «Ortsgeschichten»)
- keine Ground Truth vorhanden, d.h. eigenes Training nicht möglich

Ansatz: vergleichende Anwendung von NLP-Modellen für NER (LOC)

- spaCy («baseline»)
- Flair embeddings
- BytePair embeddings, trainiert auf historischen Text

Beispiel: e-rara 9119965



Ansicht von Biel von Süden.
Nach einer Zeichnung aus der Bieler Chronik des A. Verresius, 1626.

Das alte Biel und seine Umgebung.

Erläuternder Text
von **Dr. H. Türler,** Staatsarchivar in Bern.

Einleitung.

Das alte Biel und seine Umgebung" soll eine Heimatkunde in Bildern für das ganze Seeland sein. Es ist gewidmet seinen Bewohnern und seinen Freunden, jedem, dessen Wiege in einer der altehrwürdigen Ortschaften stand, jedem, dem das Land zur neuen Heimat geworden ist, oder der es um seiner Vorzüge willen lieb gewonnen hat.

Neue Zeiten haben neue Forderungen und Aufgaben gebracht, und das Alte stürzt in Ruinen. Die Städte haben beinahe durchweg die engen Fesseln gesprengt, die ihnen die Ringmauern umgelegt hatten. Über die ehemaligen Stadtgräben hinaus dehnen sich die Gassen aus. Aber auch in den Dörfern fordert der Fortschritt gar oft den Ruin des Alten, das seine Verteidiger verloren hat.

Es ist hohe Zeit, die Denkmäler aus den Zeiten der Voreltern noch im Bilde durch den Stift des Künstlers festzuhalten, die alten Bilder zu sammeln und den Enkeln zu überliefern. Der größte Raum in diesem Werke, das leicht auf die doppelte oder dreifache Größe hätte gebracht werden können, kommt der Stadt Biel zu wegen ihrer größern Ausdehnung und Bedeutung. Wenn dieses Verhältnis auch im Texte besteht, so geschieht es deßwegen, weil Biel ein reichhaltiges Archiv besitzt und der Verfasser dasselbe besser als andere kennt.

Heinrich Tuerler: Das alte Biel und seine Umgebung: [Tafeln u. Abb. im Text]. Biel 1902. https://doi.org/10.3931/e-rara-28664. S. 1

Auszug Volltext-File

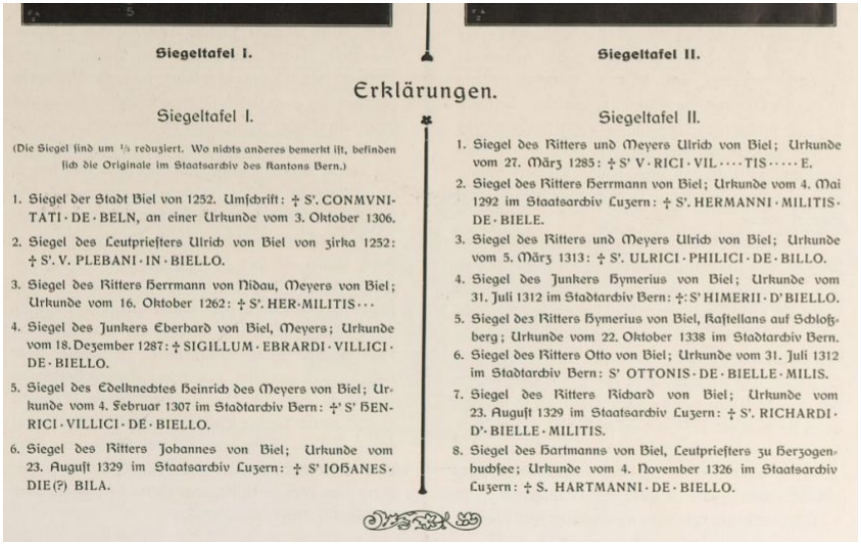
```
MM

Das alle kiel und seine Umgebung
rze^zesrzesnesnesrzwesrjesrzesrzanqrzis rzesrzesntsnisnisrzantsn
es Lexl

.jHV ■. ■ 'v f, ; r / ? r.fi S , J.  i < A > r* * * ■ " - '-' 4'
J 1 ’ V : vü ; : ,  i ■;• .’ " •/? \ ^V;i ‘.•■ ; .' £ v r si .;
''I A'v AA.LMM ''' •'/ •• " • '• ;>r -Ä.A jSV*"'! ■" 1;  Ä '•'' -
. A ».•AAv'-äivflÄ ; - !;V A 'j ■ y, ‘,iAf'SVA ‘ ° *' ‘ , ~>'%r
5; '.7V a.. . ' 1 .. i fifTU ,...-  '

Ansieht von viel von Büöen. Flach einer 3eicbnung aus der Kieler
Lkronik des Verresius, 162b. Das alle viel und seine Umgebung.
Erläuternder Text lMLi von Dr. ß. Türler, Staatsarcbivar in Bern.
Einleitung. ^)as alte Biel und seine Umgebung“ soll eine
Beimatkunöe in Bildern für dos gange Seelanö sein. Es ist
gewidmet seinen Bewohnern und seinen freunden, jedem, dessen
Wiege in einer der alt- ehrwürdigen Ortschaften stand, jedem, dem
das Land zur neuen Beimat geworden ist, oder der es um seiner
Vorzüge willen lieh gewonnen hat. Heue Zeiten haben neue
forderungen und Rufgaben gebracht, und das Rite stürzt in Ruinen.
Die Stääte haben beinahe durchweg die engen segeln gesprengt, die
ihnen die Ringmauern umgelegt hatten. Über die ehemaligen
Stadtgräben * hinaus dehnen sich die (Zassen aus. Rber auch in
den Dörfern fordert der fortschritt gar oft den Ruin des Riten,
das seine Verteidiger verloren hat. Es ist hohe Zeit, die
```

Beispiel: e-rara 9119965



Siegeltafel I. Siegeltafel II.

Erklärungen.

Siegeltafel I.

(Die Siegel sind um ⅓ reduziert. Wo nichts anderes bemerkt ist, befinden sich die Originale im Staatsarchiv des Kantons Bern.)

1. Siegel der Stadt Biel von 1252. Umschrift: † S'. CONMVNI-TATI · DE · BELN, an einer Urkunde vom 3. Oktober 1306.
2. Siegel des Leutpriesters Ulrich von Biel von zirka 1252: † S'. V. PLEBANI · IN · BIELLO.
3. Siegel des Ritters Herrmann von Nidau, Meyers von Biel; Urkunde vom 16. Oktober 1262: † S'. HER·MILITIS···
4. Siegel des Junkers Eberhard von Biel, Meyers; Urkunde vom 18. Dezember 1287: † SIGILLUM · EBRARDI · VILLICI · DE · BIELLO.
5. Siegel des Edelknechtes Heinrich des Meyers von Biel; Urkunde vom 4. Februar 1307 im Stadtarchiv Bern: † S' BEN-RICI · VILLICI · DE · BIELLO.
6. Siegel des Ritters Johannes von Biel; Urkunde vom 23. August 1329 im Staatsarchiv Luzern: † S' IOßANES· DIE (?) BILA.

Siegeltafel II.

1. Siegel des Ritters und Meyers Ulrich von Biel; Urkunde vom 27. März 1285: † S' V · RICI · VIL···· TIS····· E.
2. Siegel des Ritters Herrmann von Biel; Urkunde vom 4. Mai 1292 im Staatsarchiv Luzern: † S'. HERMANNI · MILITIS· DE · BIELE.
3. Siegel des Ritters und Meyers Ulrich von Biel; Urkunde vom 5. März 1313: † S'. ULRICI · PHILICI · DE · BILLO.
4. Siegel des Junkers Bymerius von Biel; Urkunde vom 31. Juli 1312 im Stadtarchiv Bern: †: S' HIMERII · D' BIELLO.
5. Siegel des Ritters Bymerius von Biel, Raftellans auf Schloßberg; Urkunde vom 22. Oktober 1338 im Stadtarchiv Bern.
6. Siegel des Ritters Otto von Biel; Urkunde vom 31. Juli 1312 im Stadtarchiv Bern: S' OTTONIS · DE · BIELLE · MILIS.
7. Siegel des Ritters Richard von Biel; Urkunde vom 23. August 1329 im Staatsarchiv Luzern: † S'. RICHARDI · D'· BIELLE · MILITIS.
8. Siegel des Bartmanns von Biel, Leutpriesters zu Herzogenbuchsee; Urkunde vom 4. November 1326 im Staatsarchiv Luzern: † S. HARTMANNI · DE · BIELLO.

'7-77 1 ' -^11 ^ I '77^ --77 I 77 7 ^ ,^n^, Die ältesten Siegel der Stadt und der IZerren von Siel. Sisgsllafsl I. a Siegeltafel II. sfs \ii\ifj <1 < M'S -i '• -u> Erklärungen. Siegeltafel I. # (Die Siegel sind um i 2 3 4 5 6 /n reduziert. Wo nichts anderes bemerkt ist, befinden sich die Originale im Staatsarchiv des Kantons Been.) 1. Siegel der Stadt Siel von 1252. Umschrift: f S'.CONMVNI- TATI • DE • BELN, an einer Urkunde vorn 3. Oktober 1306. 2. Siegel des Ceutpriefters Ulrich von Siel von zirka 1252: f 8'. V. PLEBANI ■ IN - BIELLO. 3. Siegel des Ritters Berrmann von Nidau, CDeyers von Siel; Urkunde vorn 16. Oktober 1262: f S\ HER-MILITIS• • • 4. Siegel des Junkers Eberhard von Biet, CDeyers; Urkunde vorn 18.Dezember 1287: f SIGILLUM • EBRARDI - VILLICI ■ DE-BIELLO. 5. Siegel des Edelknechtes ßeinrieb des CDeyers von Diel; Urkunde vorn 4. Sebruar 1307 im Stadtarchiv Bern: f' S' BEN- RICI - VILLICI - DE - BIELLO. 6. Siegel des Ritters Johannes von Diel; Urkunde vorn 23. August 1329 im Staatsarchiv Luzern: f S'IOBANES- DIE (?) BILA. Siegeltafel II. 1. Siegel des Ritters und CDeyers Ulrich von Siel; Urkunde vom 27. März 1285: J- 8' V - RICI • VIL • • • • TIS.E. 2. Siegel des Ritters Berrmann von Siel; Urkunde vom 4. Mai 1292 im Staatsarchiv Luzern: f S'. HERMANNI • MILITIS- DE - BIELE. 3. Siegel des Ritters und CDeyers Ulrich von Diel; Urkunde vorn 5. März 1313: f 8'. ULRICI ■ PHILICI • DE ■ BILLO. 4. Siegel des Junkers Bymerius von Siel; Urkunde vorn 31. Juli 1312 im Stadtarchiv Bern: f:S' HIMERII-D'BIELLO. 5. Siegel des Ritters Bymerius von Diel, Raftellans auf Schlotz- berg; Urkunde vorn 22. Oktober 1338 im Stadtarchiv Bern. 6. Siegel des Ritters Otto von Siel; Urkunde vorn 31. Juli 1312 im Stadtarchiv Bern: 8' OTTONI8 • DE • BIELLE • MILIS. 7. Siegel des Ritters Richard von Siel; Urkunde vorn 23. August 1329 im Staatsarchiv Luzern: J- 8'. RICHARDS D'-BIELLE-MILITIS. 8. Siegel des Bartmanns von Siel, Ceutpriefters zu Berzogen- buchfee; Urkunde vorn 4. November 1326 im Staatsarchiv Luzern: f S. HARTMANNI • DE • BIELLO. DZB 3

Heinrich Tuerler: Das alte Biel und seine Umgebung: [Tafeln u. Abb. im Text]. Biel 1902. https://doi.org/10.3931/e-rara-28664. S. 3

Auszug Volltext-File

Vorgehen

Datenzugang:

- Metadaten-Bezug, z.B. Publ.-Jahr: REST-API + Parsing der xml-files
- Identifizierung der Bernensia-Items: Webscraping
- Volltext-Bezug: Webscraping der txt-Files

Data Preprocessing:

- Beseitigung gröbster OCR-Fehler (Normalisierung)
- Satz-Splitting (NLTK) für spaCy
- Verzicht auf weitere manuelle Vorprozessierung

Anwendung von NER-Modellen

Auswertung/Analyse

Anwendung von NER-Modellen: spaCy

spaCy v3.0

- umfassendes Standard NLP Framework

Modell bzw. processing pipeline: **de_core_news_lg**

- inkl. 4-Klassen-NER (loc, per, org, misc)
- trainiert auf dt. Zeitungstexte + Wikipedia-Artikeln (= akt. Sprache)
- 500k unique vectors (300 dimensions)
- optimiert auf CPU-Nutzung
- F1-score: 85.12
- Anwendung in Google Colab

Anwendung von NER-Modellen: Flair embeddings

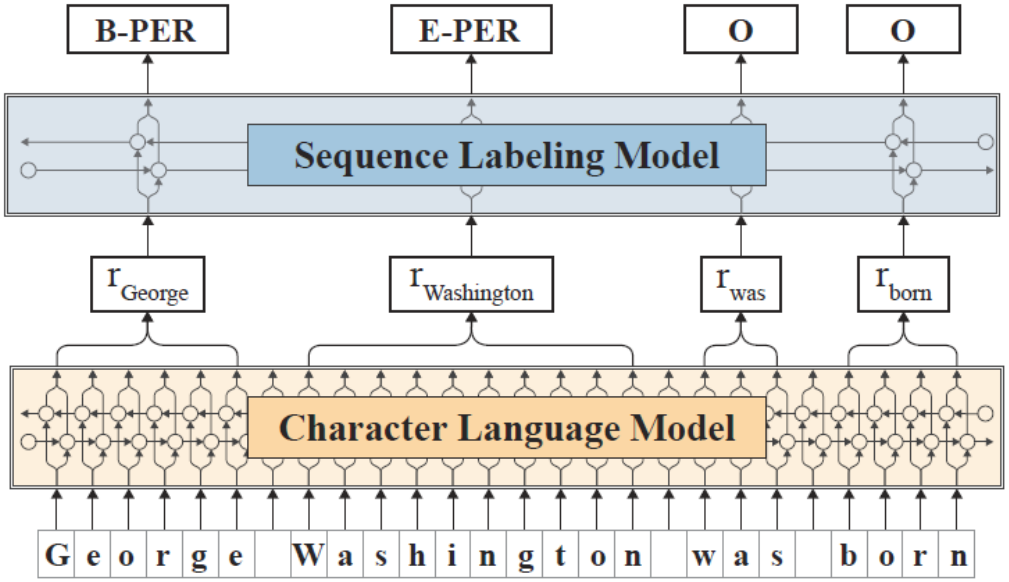
Flair

- state-of-the-art NLP Framework, entwickelt u.a. von Zalando, HU Berlin
- bietet Wrapper um verschiedene Embedding-Methoden
- einfache Kombination dieser („stacked embeddings“)
- einfacher Zugriff auf viele vortrainierte Modelle (via HuggingFace) inkl. BERTs und ELMos

Flair embeddings

- = „contextual string embeddings“ + LSTM-CRF (Paper)
- contextualized character-level word embeddings
- Modell: ner-multi-fast
- 4-Klassen-NER, mulitlingual anwendbar, auf CPU-Nutzung optimiert
- trainiert auf CoNLL-03 in en, de, nl, es (= akt. Sprache)
- F1-score: 85.72 (CoNLL-03 revised)
- Anwendung in Google Colab + HPC Uni Bern

Contextual string embeddings (Flair e.)



Alan Akbik / Duncan Blythe / Roland Vollgraf: . Santa Fe, New Mexico, USA 08.2018. https://www.aclweb.org/anthology/C18-1139. S. 1639

1. Sätze werden als character sequence durch ein vortrainiertes bidirektionales character language model verarbeitet -> contextal embeddings für alle Wörter ->
2. BiLSTM-CRF sequence labeler zum NER tagger

Anwendung von NER-Modellen: BP embeddings

BytePair embeddings

- subword-level word embeddings
- verfügbar in Flair
- Modell: historic-ner-onb
- erstellt von Bayerischer Staatsbibliothek (BSB)
- trainiert auf **historischen österr. Zeitungstexten** (1710-1873)
  - nur ca. 35k Token, OCR-Fehler-behaftet, Austrizismen
- optimiert auf GPU-Nutzung
- F1-score: 85.69
- Anwendung in HPC Uni Bern (24 nodes à 50GB statt GPU)

BytePair embeddings

Merge ops	Byte-pair encoded text
5000	豊 田 駅 (と よ だ え き) は 、 東京都 日 野 市 豊 田 四 丁目 にある
10000	豊 田 駅 (と よ だ えき) は 、 東京都 日 野市 豊 田 四 丁目にある
25000	豊 田駅 (とよ だ えき) は 、 東京都 日 野市 豊田 四 丁目にある
50000	豊 田駅 (とよ だ えき) は 、 東京都 日 野市 豊田 四丁目にある
Tokenized	豊田 駅 （ と よ だ え き ） は 、 東京 都 日野 市 豊田 四 丁目 に ある
10000	豐 田 站 是 東 日本 旅 客 鐵 道 (JR 東 日本) 中央 本 線 的 鐵路 車站
25000	豐田 站是 東日本旅客鐵道 (JR 東日本) 中央 本 線的鐵路車站
50000	豐田 站是 東日本旅客鐵道 (JR 東日本) 中央 本線的鐵路車站
Tokenized	豐田站 是 東日本 旅客 鐵道 （ JR 東日本 ） 中央本線 的 鐵路車站
1000	to y od a _station is _a _r ail way _station _on _the _ch ū ō _main _l ine
3000	to y od a _station _is _a _railway _station _on _the _ch ū ō _main _line
10000	toy oda _station _is _a _railway _station _on _the _ch ū ō _main _line
50000	toy oda _station _is _a _railway _station _on _the _chū ō _main _line
100000	toy oda _station _is _a _railway _station _on _the _chūō _main _line
Tokenized	toyoda station is a railway station on the chūō main line

Benjamin Heinzerling / Michael Strube: . Miyazaki, Japan 05.2018. https://www.aclweb.org/anthology/L18-1473. S. 2989

Byte-Pair-Encoding:

- Text wird als character/symbol sequence verarbeitet -> iterativ werden die häufigsten symbol pairs gemergt
- Anzahl der Merge-Operationen ist der Hyperparameter

Auswertung/Analyse

Performance/Handling

- spaCy
  - sehr schnell, ressourcenarm, komplette morpho-syntakische Analyse, bietet auch NER-Lemmata
- Flair multi-ner-fast
  - guter Trade-off zw. nötiger Rechenleistung und qualit. Ergebnis
  - läuft leider nicht ohne Probleme durch, wenige Texte fehlen
- BP historic-ner-onb
  - braucht mit Abstand am meisten Rechenleistung
  - aber: läuft ohne Probleme durch

Auswertung/Analyse – aber wie?

Constraints

- Beschränkung auf 50 Dokumente, keine bewusste Auswahl
- keine Ground Truth – keine Berechnung von Precision/Recall/Fx-score
- (stichprobenweise GT-Erstellung und Scoring kritisch bei inhomogenen Dokumenten -> zunächst gute Typisierung nötig)

Beurteilung durch Anzahl und Schnittmengen der LOC-Entities?

Vorgehen

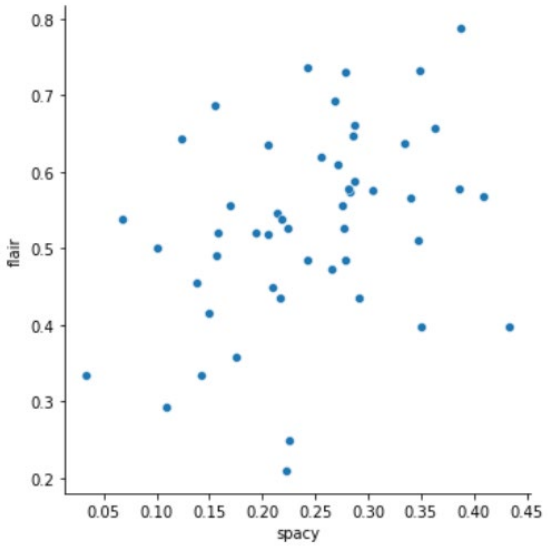
- für LOC-Entities wurden IOB- bzw. IOBES-tags rekombiniert
- Erstellung von LOC Sets
- Berechnung der Schnittmengen, symmetrischen Differenzen
- manuelle Begutachtung der Entities, insbesondere «Fragwürdige»

Überblick

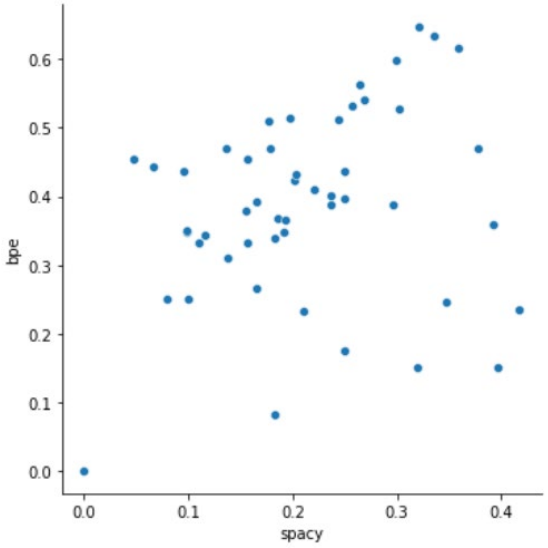
Korpus (n=50)	
Textlänge (char)	8k – 1.8M (mean: 219k, median: 95k)
Publikationsjahr	1815 – 1928 (mean: 1873, median: 1876)

	spaCy (de news lg)	Flair multi-ner-fast	BP historic-ner-onb
LOC count	[13, 23'095]	[5, 17'362]	[4, 19'694]
LOC Set (unique counts)	**[10, 10'417]**	**[2, 7099]**	**[2, 6369]**
Schnittmenge zu spaCy	-	[1, 3623]	[0, 2464]
Symm. Differenz zu spaCy	-	[10, 10'270]	[12, 11'858]
Schnittmenge zu Flair	-	-	**[1, 2415]**
Symm. Differenz zu Flair	-	-	**[4, 8638]**

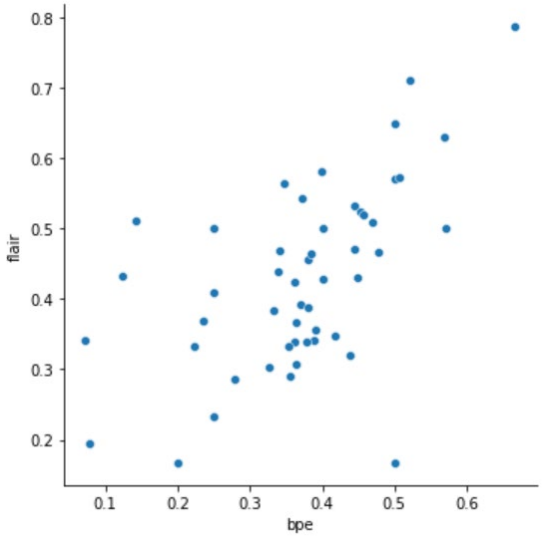
Grösse Schnittmenge / Grösse LOC Set



spaCy vs. Flair



spaCy vs. BPE



BPE vs. Flair

Begutachtung: Flair & BP embeddings

generell:

- Flair: keine Tendenz hinsichtlich Single vs. Compound Entities
- BPE: Tendenz zu Single-Entities?

Stärken:

- wenig Rauschen in den erkannten E. (vermutl. hohe precision)
- robust gegen OCR-Fehler mit geringer Levenstein-Distance

```
'Bartblomebof', 'S-LOC', 'Bircbe', 'S-LOC', 'Bircbturm', 'S-LOC',  'Büöen', 'S-LOC',
                                                          'Gurteugassc', 'S-LOC',
```

Systematisch auftretende Fehler:

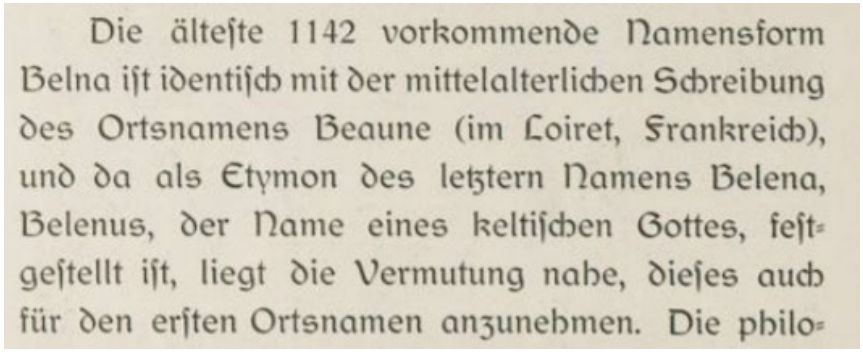
- selten: Differenzierung zu Personen-E.

```
'Christ', 'S-LOC', 'Jesuskind', 'S-LOC',  'Werner', 'S-LOC', 'Verresius', 'S-LOC',
```

- BPE: E. mit Trennstrich werden erkannt, aber nicht ganz korrekt:

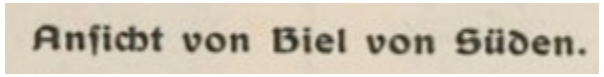
```
'Deuen', 'S-LOC', '-', 'B-LOC', 'burg', 'E-LOC',
'Em', 'S-LOC', '-', 'B-LOC', 'menthal', 'E-LOC',
```

Nochmal «Das alte Biel und seine Umgebung»



Die älteſte 1142 vorkommende Namensform Belna iſt identiſch mit der mittelalterlichen Schreibung des Ortsnamens Beaune (im Loiret, Frankreich), und da als Etymon des letztern Namens Belena, Belenus, der Name eines keltiſchen Gottes, feſtgeſtellt iſt, liegt die Vermutung nahe, dieſes auch für den erſten Ortsnamen anzunehmen. Die philo-

...vorkommende Bamensform...



Anſicht von Biel von Süden.

...Ansieht von viel von Büöen ...

spaCy

```
['Belna', 'LOC'],
['Beaune Coiret Frankreich', 'LOC']
```

Flair embeddings:

```
'Beaune', 'S-LOC', 'Frankreich', 'S-LOC',
```

BP embeddings

```
'Belna', 'S-LOC', 'Coiret', 'S-LOC',
'Frankreich', 'S-LOC',
```

```
'Büöen', 'S-LOC',
```

Fragen & Ausblick

Ansätze zur Verbesserung/Weiterentwicklung

- Ausdehnung auf gesamtes Set; Erstellung von GT und Scoring
- Typisierung der Dokumente anhand von Länge, Textarten
- vertiefte Analyse der Flair- vs. BP-Entities

Nachnutzung

- Bereitstellung Skripte (insbes. zum Datenbezug), Daten-Dumps
- Evaluation für die Nutzung der Entities für spezifische Sacherschliessung/Retrieval

Resumé

Können **NER**-Ansätze auf Bernensia-Volltexte erfolgreich angewendet werden – trotz z.T. starker OCR-Fehler?

JA, und tendenziell „praxisreif“.

- entitäten-reiche Textarten können stark profitieren!
- Robustheit gegen OCR-Fehler kann bestätigt werden

Lessons Learned

- Steile, aber tolle Lernkurve ☺ im Oktober zum ersten Mal Python – im März auf dem HPC
- Unterschätze niemals textuelle Daten
- Gilt insbesondere für die nötige Rechenpower

Vielen Dank für die Aufmerksamkeit!

Kathi Woitas
kathi.woitas@students.bfh.ch

Universitätsbibliothek Bern
Digital Scholarship Services