Recent developments in the Diplomata Belgica-project.

Case study: NER applied to a corpus of Middle Dutch charters

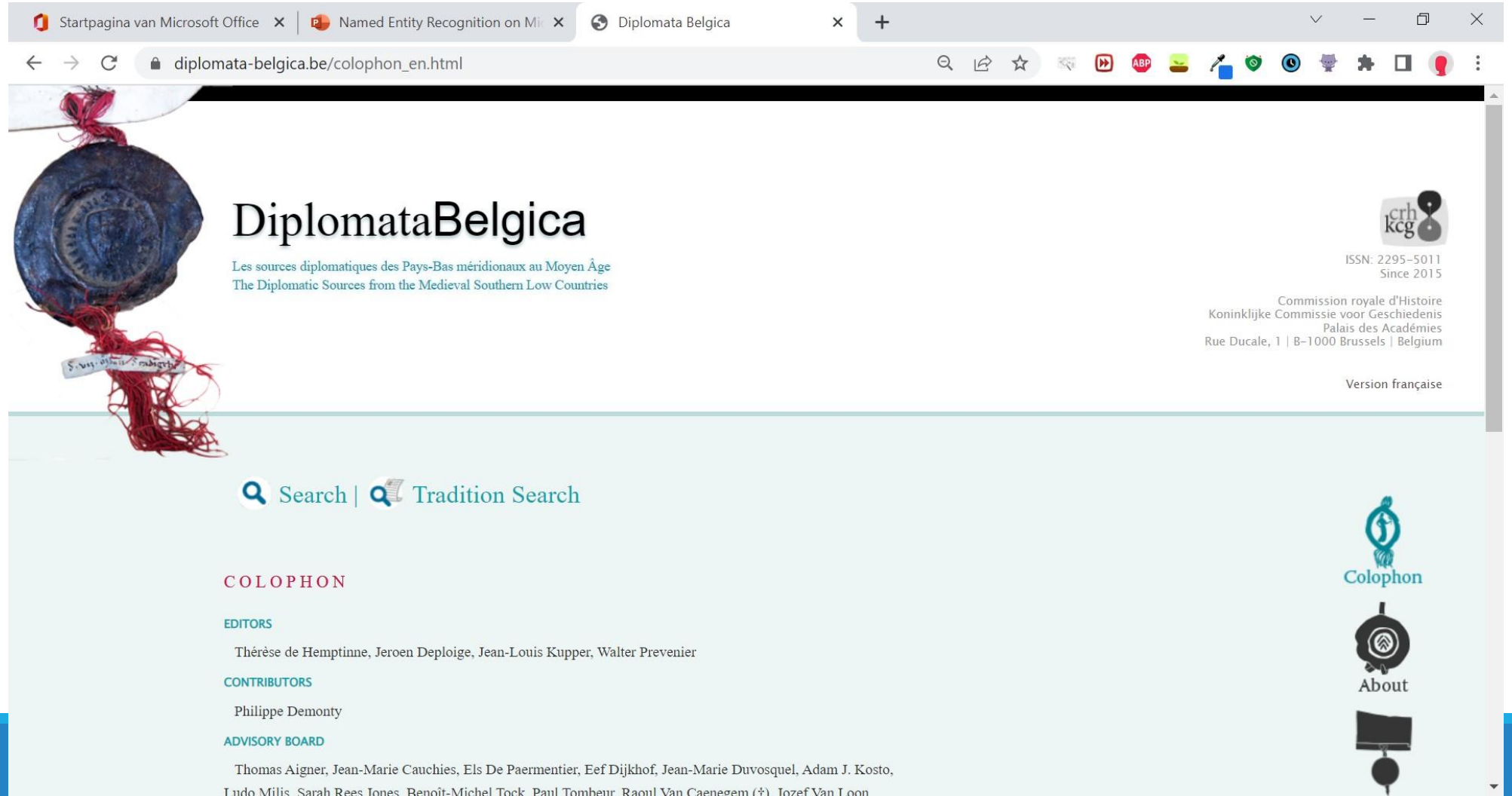
JEROEN DEPLOIGE AND MARIJKE BEERSMANS

GHENT UNIVERSITY AND UNIVERSTY OF ANTWERP

GRAZ, DIDIP CONFERENCE, 28/9/22

Recent developments

2015 - Launch *Diplomata Belgica*



2015 - Launch *Diplomata Belgica*

"*Diplomata Belgica* offers a critical survey of all the diplomatic sources, edited or still unpublished, and issued by both natural persons and legal bodies from the medieval Southern Low Countries. *Diplomata Belgica* covers present day Belgium as well as those areas which belonged historically to the Southern Low Countries but are part now of France (French Flanders, French Hainault), the Netherlands (parts of the provinces of Zeeland, Noord-Brabant, Limburg), the Grand Duchy of Luxembourg, or Germany (parts of the Rhineland).

(…) The database aims at exhaustivity for the period before 1250 and will, in the future, also include late medieval diplomatic materials without striving after completeness."

2015 - Launch *Diplomata Belgica*

- Royal Historical Commission of Belgium in collaboration with Ghent University
- From Thesaurus Diplomaticus (Brepols CD-rom, 1996) to database in open access (2015)
- Two accesses to the data:
  - Charters (context, text, photographs)
  - Material transmission (archival funds, cartularies etc.)

Expanding content, improving data quality

- 2022: ca. 43,000 charters - ca. 20,000 full texts - ca. 8000 photographs
- New corpora
  - Chirographs of Ypres (S. De Valeriola, 6000+ documents)
  - C14NL – Corpus of fourteenth-century non-literary texts in Middle Dutch (KANTL, 600+ documents)
- Improving data quality
  - Toponymical and anthroponymical information
  - User feedback

New User Interface and Search Machine

- User consultation in 2019

- In progress:
  - Sophisticated possibilities of advanced search combined with the flexibility of "elastic search"
  - Better exploitation of georeferencing of "both actors" and place dates
  - Resolving limitations in chronological searches

- Sneak preview:
  - https://watch.screencastify.com/v/R25zhlwrKmvrzCkSxuGA

DH Research / Computational experimentation

- Digital Diplomatics Conference – Naples 2011
  - De Tré G., Deploige J. (2016). “Time Modelling in Digital Humanities : Challenges Posed by the Development of a Database of Medieval Charters.” *IT : INFORMATION TECHNOLOGY*, edited by Manfred Thaller, vol. 58, no. 2, de Gruyter Oldenbourg, 2016, pp. 97–103
  - Billiet C., Van de Weghe N., Deploige J., De Tré G. (2017. “Visualizing and Reasoning with Imperfect Time Intervals in 2-D.” *IEEE TRANSACTIONS ON FUZZY SYSTEMS*, vol. 25, no. 6, Institute of Electrical and Electronics Engineers (IEEE), 2017, pp. 1698–713,

DH Research / Computational experimentation

- Leclercq, E., & Kestemont, M. (2021). "Advances in Distant Diplomatics: A Stylometric Approach to Medieval Charters". *Interfaces: A Journal of Medieval European Literatures*, special cluster ed. J. Deploige, J. De Gussem, Medieval Authorship and Canonicity in the Digital Age,(8): 214–244.
- Aguilar, Sergio Torres , Stutzmann, Dominique (2021). "Named Entity Recognition for French medieval charters". *Workshop on Natural Language Processing for Digital Humanities*, Dec 2021, Helsinki, Finland.
- De Valeriola, Sébastien, Nicolas Ruffini-Ronzani, and Étienne Cuvelier. (2022). “Dealing with the Heterogeneity of Interpersonal Relationships in the Middle Ages. A Multi-Layer Network Approach.” Digital Medievalist, 15(1): 1–28.

Case study: NER applied to a corpus of Middle Dutch charters

Overview

1. Named Entity Recognition
2. Preparation
    1. Data collection
    2. Annotation
3. Training models
    1. Neural networks
    2. Architectures
4. Results
    1. Bi-LSTM
    2. SpaCy
5. Test on out-of-corpus text
6. Conclusion

Named Entity Recognition

Named Entity Recognition (NER)

- Type of **automatic information extraction**
- Named entity ~ rigid designator/proper noun
- Input: ***Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.***
- Output***: [PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .***

Named Entity Recognition (NER)

- "Solved" for modern, high resource languages (near-human performance)

ABANDON ALL HOPE YE WHO ENTER HERE is scrawled in blood red lettering on the side of the Chemical Bank ORG near the corner of Eleventh LOC and First and is in print large enough to be seen from the backseat of the cab asit lurches forward in the traffic leaving Wall Street and just as Timothy Price PERSON notices the words a bus pulls up, the advertisement for Les Misérables PERSON on its side blocking his view, but Price who is with Pierce & Pierce ORG and twenty-six doesn't seem to care because he tells the driver he will give him five dollars MONEY to turn up the radio, "Be My Baby" on WYNN, and the driver, black, not American NORP , does so.

- Ancient languages?
  - Aguilar & Stutzmann (2021)
- **Is NER feasible for Middle Dutch as well?**

Preparation

Data collection

- **Supervised** machine learning
- Data for **training** models and **testing** their performance
- **C14NL-PoS** corpus
  - 155 charters, 50 000 tokens
  - Subcorpus of C14NL
    - Subcorpus of corpus Van Reenen-Mulder
  - All originals
  - Produced locally
  - Raw transcription
  - Enriched with lemma and part-of-speech tags
  - Aldermen charters

```
<w lemma="dat" ana="c410" part="N">Dat</w>
<w lemma="moeten" ana="c214" part="N">moeten</w>
<w lemma="weten" ana="c250" part="N">weten</w>
<w lemma="al" ana="c441" part="N">alle</w>
<w lemma="die" ana="c421" part="N">die</w>
<w lemma="zijn" ana="c204" part="N">sijn</w>
<w lemma="en" ana="c800" part="N">ende</w>
<w lemma="zullen" ana="c214" part="N">zelen</w>
<w lemma="zijn" ana="c250" part="N">sijn</w>
<w lemma="dat" ana="c810" part="N">dat</w>
<w lemma="petrus" ana="c020" part="N">pieter</w>
<w lemma="pantijn" ana="c020" part="N">pantijn</w>
<w lemma="hebben" ana="c213" part="N">heeft</w>
<w lemma="kopen" ana="c273" part="N">ghecocht</w>
<w lemma="en" ana="c800" part="N">ende</w>
<w lemma="wel" ana="c500" part="N">wel</w>
```

Data collection

Annotation

- Entity set: person (PERS), location (LOC), date (DATE), money (MONEY)
- BIO-format

jane	B-PERS
den	I-PERS
smet	I-PERS
alse	O
Prouiserres	O
van	O
der	O
kerken	O

van	O
pamele	B-LOC
v.	B-MONEY
lb	I-MONEY
par	I-MONEY
vlaenderscher	I-MONEY
munten	I-MONEY

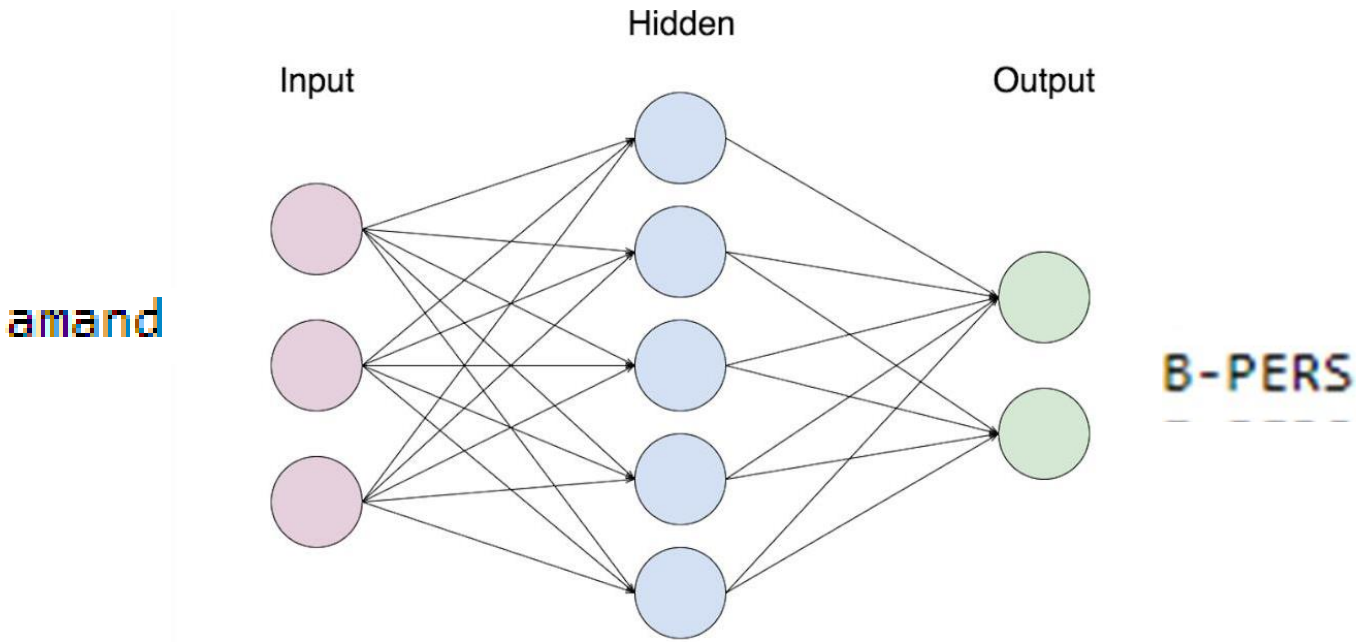
- Manual annotation - some rules
  - Dates are annotated in their entirety
    - *dusentech driehondert een ende zestech de xxvijste dach va hoeymaent* (27th of July 1361)
  - Specific street names or house names are LOC
  - Payments in kind are not MONEY

Train-test split

- Charter level split
- To train the model
  - Training set: 124 charters
  - Development set: 15 charters
- To test performance
  - Test set: 16 charters

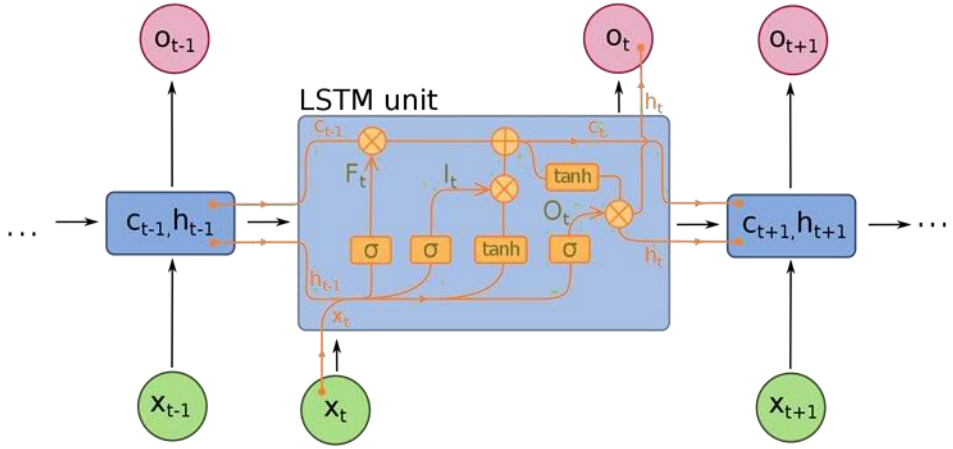
Training

Neural networks



Our models

- Bi-LSTM
  - Made with Keras
  - Layered architecture: embedding, LSTM (bidirectional), output
  - LSTM, a type of neural network that can “remember” previous input
- SpaCy
  - State of the art NLP library
  - Industry oriented
  - Provides a premade pipeline
  - Uses CNN under the hood



Results

Results Bi-LSTM

True										
	B-DATE	27	6	0	0	2	0	0	0	5
	B-LOC	0	48	3	4	1	0	0	0	29
	B-MONEY	0	0	34	0	0	0	7	0	12
	B-PERS	0	1	0	202	0	0	0	4	47
		2								50
					1					8
	I-MONEY	0	0	1	0	0	0	115	0	7
	I-PERS	0	0	0	3	0	0	0	141	80
	O	0	3	8	4	1	0	14	16	4264
		B-DATE	**B-LOC**	**B-MONEY**	**B-PERS**	**I-DATE**	**I-LOC**	**I-MONEY**	**I-PERS**	**O**
					Predicted					

(green = correct, orange = missed, yellow = imagined, red = misclassified)

Results Bi-LSTM

CORPUS LEVEL

	precision	recall	f1-score	support
B-DATE	0.93	0.68	0.78	40
B-LOC	0.83	0.56	0.67	85
B-MONEY	0.74	0.64	0.69	53
B-PERS	0.94	0.80	0.86	254
I-DATE	0.96	0.66	0.78	151
I-LOC	1.00	0.10	0.18	10
I-MONEY	0.85	0.93	0.89	123
I-PERS	0.88	0.63	0.73	224
O	0.95	0.99	0.97	4310
accuracy			0.94	5250
macro avg	0.90	0.67	0.73	5250
weighted avg	0.94	0.94	0.93	5250

ENTITY LEVEL

	precision	recall	f1-score	support
DATE	0.53	0.50	0.51	40
LOC	0.76	0.53	0.62	85
MONEY	0.56	0.62	0.59	53
PERS	0.71	0.62	0.66	254
micro avg	0.67	0.59	0.63	432
macro avg	0.64	0.57	0.60	432
weighted avg	0.68	0.59	0.63	432

Results SpaCy

True										
	B-DATE	29	0	0	0	0	0	0	0	1
	B-LOC	0	61	0	6	0	0	0	1	17
	B-MONEY	0	0	49	0	0	0	1	0	3
	B-PERS	0	0	0	208	0	0	0	3	43
	I-DATE	2	1	0	0	105	0	0	0	43
	I-LOC	0	0	0	0	0	2	0	3	5
	I-MONEY	0	0	1	0	0	0	118	0	4
	I-PERS	0	2	0	2	0	0	0	165	55
	O	2	6	6	9	18	0	19	22	4228
		B-DATE	**B-LOC**	**B-MONEY**	**B-PERS**	**I-DATE**	**I-LOC**	**I-MONEY**	**I-PERS**	**O**
					Predicted					

(green = correct, orange = missed, yellow = imagined, red = misclassified)

Results SpaCy

CORPUS LEVEL

	precision	recall	f1-score	support
B-DATE	0.88	0.72	0.79	40
B-LOC	0.87	0.72	0.79	85
B-MONEY	0.88	0.92	0.90	53
B-PERS	0.92	0.82	0.87	254
I-DATE	0.85	0.70	0.77	151
I-LOC	1.00	0.20	0.33	10
I-MONEY	0.86	0.96	0.90	123
I-PERS	0.85	0.74	0.79	224
O	0.96	0.98	0.97	4310
accuracy			0.95	5250
macro avg	0.90	0.75	0.79	5250
weighted avg	0.94	0.95	0.94	5250

ENTITY LEVEL

	precision	recall	f1-score	support
DATE	0.73	0.60	0.66	40
LOC	0.87	0.72	0.79	85
MONEY	0.82	0.87	0.84	53
PERS	0.87	0.77	0.81	254
micro avg	0.85	0.75	0.80	432
macro avg	0.82	0.74	0.78	432
weighted avg	0.85	0.75	0.80	432

Unseen text (Bi-LSTM)

TRUE

Wie jacob van den houke wydoot brunijnc Robrecht lammertijn artur veyse en jehan abelkin Scepene van der poort van veurne doen te wetene tallen den gonen die deze lre zullen zien of horen leisen Dat jehan de buc de tanne en pieroene zijn wijf camen voor ons Ende ghauen vp wel en wettlike met halme en met ghiften jehan den breden onzen poorte zes sceleghe par. siaers eweliker renten den groten tournoisen ouer twalef peninghe par. of andre munte jnt auenant ligghende onder de stede trechof en de recken die jehans van den recke wa ren daer hanekin nu wonende es te gheldene jehan den brede vors. of den ex bringhe van deze lre te tween payementen jnt jaer dats te wetene dene helt te zinte jehans messe nu eerst

PREDICTED

Wie jacob van den houke wydoot brunijnc Robrecht lammertijn artur veyse en jehan abelkin Scepene van der poort van veurne doen te wetene tallen den gonen die deze lre zullen zien of horen leisen Dat jehan de buc de tanne en pieroene zijn wijf camen voor ons Ende ghauen vp wel en wettlike met halme en met ghiften jehan den breden onzen poorte zes sceleghe par. siaers eweliker renten den groten tournoisen ouer twalef peninghe par. of andre munte jnt auenant ligghende onder de stede trechof en de recken die jehans van den recke waren daer hanekin vranke nu wonende es te gheldene jehan den brede vors. of den ex bringhe van deze lre te tween payementen jnt jaer dats te wetene dene helt te zinte jehans messe nu eerst

Unseen text (SpaCy)

TRUE

Wie jacob van den houke wydoot brunijnc Robrecht lammertijn artur veyse en jehan abelkin Scepene van der poort van veurne doen te wetene tallen den gonen die deze lre zullen zien of horen leisen Dat jehan de buc de tanne en pieroene zijn wijf camen voor ons Ende ghauen vp wel en wettlike met halme en met ghiften jehan den breden onzen poorte zes sceleghe par. siaers eweliker renten den groten tournoisen ouer twalef peninghe par. of andre munte jnt auenant ligghende onder de stede trechof en de recken die jehans van den recke wa ren daer hanekin nu wonende es te gheldene jehan den brede vors. of den ex bringhe van deze lre te tween payementen jnt jaer dats te wetene dene helt te zinte jehans messe nu eerst

PREDICTED

Wie jacob van den houke wydoot brunijnc Robrecht lammertijn artur veyse en jehan abelkin Scepene van der poort van veurne doen te wetene tallen den gonen die deze lre zullen zien of horen leisen Dat jehan de buc de tanne en pieroene zijn wijf camen voor ons Ende ghauen vp wel en wettlike met halme en met ghiften jehan den breden onzen poorte zes sceleghe par. siaers eweliker renten den groten tournoisen ouer twalef peninghe par. of andre munte jnt auenant ligghende onder de stede trechof en de recken die jehans van den recke waren daer hanekin vranke nu wonende es te gheldene jehan den brede vors. of den ex bringhe van deze lre te tween payementen jnt jaer dats te wetene dene helt te zinte jehans messe nu eerst

Conclusion

- The models are not state of the art
  - Aguilar & Stutzmann results

		Middle French (rounded to 2 digits)		Middle Dutch	
		SpaCy	Custom	Spacy	Custom
PERS	Pr	0,94	0,96	0,87	0,71
	Rc	0,96	0,96	0,77	0,62
	f1	0,95	0,96	0,81	0,66
LOC	Pr	0,95	0,96	0,87	0,76
	Rc	0,95	0,95	0,72	0,53
	f1	0,95	0,96	0,79	0,62

Conclusion

- The models are not SOTA, but promising that NER on Middle Dutch is feasible
  - The dataset was small
  - French written language more standardised at this point
- The SpaCy outperforms our own
  - “Sacrificing” precision for recall
  - Better able to detect entity boundaries

Sources

Aguilar, S. T., & Stutzmann, D. (2021). Named Entity Recognition for French medieval charters. Workshop on Natural Language Processing for Digital Humanities. Helsinki.

Centrum voor Teksteditie en bronnenstudie - KANTL. (s.d.). Corpus van veertiende-eeuwse niet-literaire Nederlandse teksten | C14NL. Retrieved june 6, 2022, from https://bouwstoffen.kantl.be.

Clips. (2005, may 8). Language-Independent Named Entity Recognition (I). Retrieved from https://www.clips.uantwerpen.be/: https://www.clips.uantwerpen.be/conll2002/ner/

The future

Pan-European collaboration

Diplomata Belgica courted by…

- Cartae Europae Medii Aevi (CEMA) (Nicolas Perreaux – LaMOP, France)
- Digi-DEEDS (consortium directed by Michael Gervers – Toronto)
- Monasterium / DiDip (Georg Vogeler – Graz / ERC)

Some obstacles, more opportunities

- Problems?
  - Integration or federation?
  - Duplication of data?
  - Loss of search facilities?
  - Intellectual ownership?
  - Regional/national funding?

Some obstacles, more opportunities

- Creative Commons licenses?
  - Photographs
    - **CC BY-NC** This license lets others remix, adapt, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.
  - Texts of the charters
    - **CC BY-NC-SA** This license lets others remix, adapt, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms.
  - Metadata
    - **CC BY-NC-ND** Licensees may copy, distribute, display and perform only verbatim copies of the work, not derivative works and remixes based on it. - UNLESS OTHERWISE AGREED