An empirical study of domain adaptation for named entity recognition on historical documents

Baptiste Blouin

24-06-2021

ENP-China

“ "Elites, Networks and Power in modern urban China" project explores the transformative process of elites in China between 1830 and 1949. It focuses on three main urban areas which were the engines of change in modern China: Beijing/Tianjin, Guangzhou/Hong Kong, and grater Shanghai. “

- Over a century of English and Chinese press to study.
- Use of information extraction.
- The interpretation of the results requires stable performance.

Named Entity Recognition

Elon Musk PERSON apparently wasn't aware that his company SpaceX had a Facebook ORG page. The SpaceX and Tesla PRODUCT CEO has responded to a comment on Twitter GPE calling for him to take down the SpaceX, Tesla and Elon Musk ORG official pages in support of the #deletefacebook movement by first ORDINAL acknowledging he didn't know one existed, and then following up with promises that he would indeed take them down.

He's done just that, as the SpaceX NORP Facebook page is now gone, after having been live earlier today DATE (as you can see from the screenshot included taken at around 12:10 PM ET) TIME .

Possibilities

- Use pre-trained systems on contemporary data.
- Annotate your own data and train a system on it.
- Combining both.

Questions

- How far is the historical domain from the contemporary?
- How much data should be annotated?
- Can we reduce this amount of annotation?
- Can we use unannotated data to facilitate the transfer?
- Does the quality of the documents influence the performance of the systems?

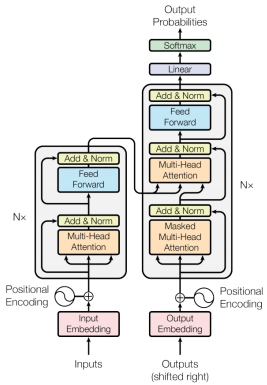
Datasets

Datasets	Train	Dev	Test	Frequency of entity	Entity types	Sources	Time period
CoNLL-03	23,499 |Entity| 204,567 |Token|	5,942 51,578	5,648 46,666	11,6%	4	News	1996-1997
OntoNotes 5.0	81,828 1,088,503	11,066 147,724	11,257 152,728	7.5 %	18	Newswire Broadcast News Broadcast Conversation Web Data	1994-2008
WNUT-17	3,160 62,729	1,250 15,733	1,589 23,394	5.9%	6	Tweets Comments of YouTube	2014-2017
ACE 05	34.669 145,025	4,336 17,763	3,777 14,859	24.1%	7	Newswire Broadcast News Broadcast Conversation Web Data ...	2003-2004
LitBank	29,894 168,787	4,133 20,802	3,425 20,943	17,8%	7	Fiction	1852-1923
HIPE		2,575 26,873	1,301 15,988	9%	5	Newspapers	1798-2018

Table: Data statistics

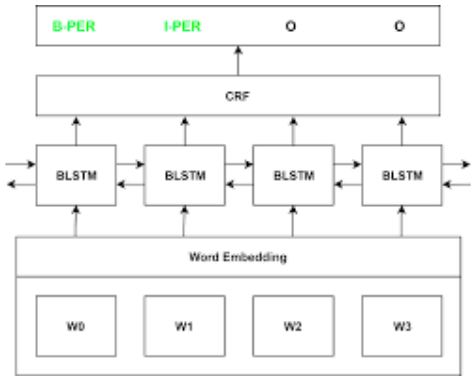
Models

BERT [Devlin et al., 2019]



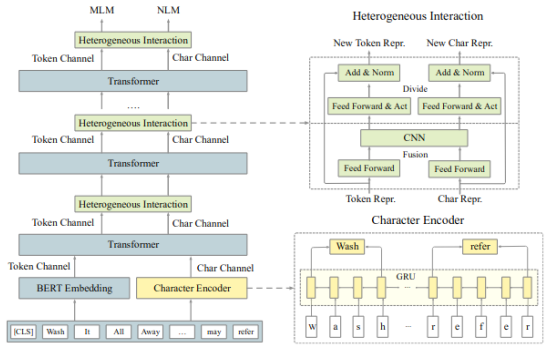
bert-base-uncased

BI-LSTM CRF [Lample et al., 2016] [Akbik et al., 2019]



Common Crawl FastText

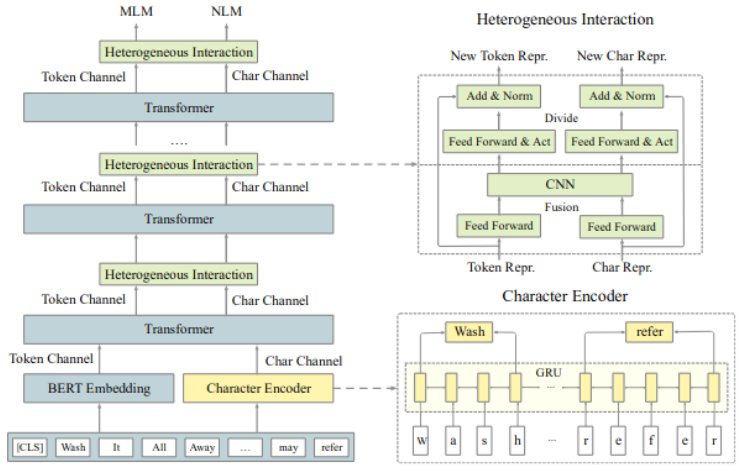
CharBERT [Ma et al., 2020]



charbert-bert-wiki

CharBERT: Character-aware Pre-trained Language Model [Ma et al., 2020]

- The character encoder.
- Heterogeneous interaction.
- Unsupervised character pre-training : Noisy Language Modeling (NLM)



	QNLI		CoNLL-2003 NER		SQuAD 2.0	
Models	Original	Attack	Original	Attack	Original	Attack
BERT	90.7	63.4	91.24	60.79	76.3	50.1
AdvBERT	90.8	75.8	90.68	71.47	76.6	52.4
BERT+WordRec	84.0	76.1	82.52	67.79	63.5	55.2
CharBERT	**91.7**	**80.1**	**91.81**	**76.14**	**78.6**	**56.3**

Experiences

- Why should we annotate and how much data should we annotate?
- How to choose the source corpus to make a coherent transfer to our data?
- Can we facilitate the transfer to the historical domain with embeddings?
- How much do OCR errors bias our systems?

CrossNER: Evaluating Cross-Domain Named Entity Recognition [Liu et al., 2020]

Domain	Unlabeled Corpus	Labeled NER			Entity Categories
		Train	Dev	Test	
Reuteurs	-	14,987	3,466	3,684	person, organization, location, miscellaneous
Politics	9.07M	200	541	651	politician, person, organization, political party, event, election, country, location, miscellaneous
Music	9.82M	100	380	456	music genre, song, band, album, musical artist, musical instrument, award, event, country, location, organization, person, miscellaneous
...	...	...	...	...	...
Artificial Intelligence	287.62K	100	350	431	field, task, product, algorithm, researcher, metrics, university, country, person, organization, location, miscellaneous

Should we annotate? - Experimental setup

- Source corpus : ACE 05
  - Same annotation guideline as LitBank.
  - Annotation guideline different from HIPE.
- Transfer method :
  - Learning on the source corpus - testing on the target corpus.
  - Finetuning of this model according to a variable amount of target data.
    - Compared to learning directly on the target data in variable quantity.
- All three models are used.

Should we annotate? - LitBank

Models / Splits	0	10	50	250	400	1000	3000	6000
BERT	0	0	0	54.28	63.72	72.87	79.04	**80.37**
BERT-ACE	**70.44**	69.39	73.27	75.93	76.78	78.85	80.7	80.91
CharBERT	0	0	30.51	63.97	69.15	74.51	78.81	**80.08**
CharBERT-ACE	**67.38**	68.11	71.00	75.11	76.04	77.56	79.78	80.61
BI-LSTM CRF	0	0	0	24.47	44.41	57.81	68.3	**72.87**
BI-LSTM CRF-ACE	**57.75**	55.88	62.38	67.55	69.38	71.49	74.72	75.48

Table: Results obtained on the LitBank test depending on the system used as well as the amount of LitBank training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

Should we annotate? - LitBank

Models / Splits	0	10	50	250	400	1000	3000	6000
BERT	0	0	0	54.28	63.72	**72.87**	79.04	80.37
BERT-ACE	70.44	69.39	**73.27**	75.93	76.78	78.85	80.7	80.91
CharBERT	0	0	30.51	63.97	69.15	**74.51**	78.81	80.08
CharBERT-ACE	67.38	68.11	71.00	**75.11**	76.04	77.56	79.78	80.61
BI-LSTM CRF	0	0	0	24.47	44.41	**57.81**	68.3	72.87
BI-LSTM CRF-ACE	57.75	55.88	**62.38**	67.55	69.38	71.49	74.72	75.48

Table: Results obtained on the LitBank test depending on the system used as well as the amount of LitBank training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

Should we annotate? - LitBank

Models / Splits	0	10	50	250	400	1000	3000	6000
BERT	0	0	**0**	54.28	63.72	72.87	79.04	80.37
BERT-ACE	70.44	69.39	73.27	75.93	76.78	78.85	80.7	80.91
CharBERT	0	0	**30.51**	63.97	69.15	74.51	78.81	80.08
CharBERT-ACE	67.38	68.11	71.00	75.11	76.04	77.56	79.78	80.61
BI-LSTM CRF	0	0	**0**	24.47	44.41	57.81	68.3	72.87
BI-LSTM CRF-ACE	57.75	55.88	62.38	67.55	69.38	71.49	74.72	75.48

Table: Results obtained on the LitBank test depending on the system used as well as the amount of LitBank training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

Should we annotate? - LitBank

Models / Splits	0	10	50	250	400	1000	3000	6000
BERT	0	0	0	54.28	63.72	72.87	79.04	80.37
BERT-ACE	70.44	69.39	73.27	75.93	76.78	78.85	80.7	**80.91**
CharBERT	0	0	30.51	63.97	69.15	74.51	78.81	80.08
CharBERT-ACE	67.38	68.11	71.00	75.11	76.04	77.56	79.78	**80.61**
BI-LSTM CRF	0	0	0	24.47	44.41	57.81	68.3	72.87
BI-LSTM CRF-ACE	57.75	55.88	62.38	67.55	69.38	71.49	74.72	**75.48**

Table: Results obtained on the LitBank test depending on the system used as well as the amount of LitBank training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

Should we annotate? - HIPE

Models / Splits	0	10	50	250	400	444
BERT	0	0	0.04	42.6	51.46	**52.14**
BERT-ACE	**10.08**	25.17	47.84	54.55	57.37	58.14
CharBERT	0	0	32.82	54.15	57.28	**57.57**
CharBERT-ACE	**13.00**	27.33	46.93	57.85	61.07	61.13
BI-LSTM CRF	0	0	0.35	29.64	37.63	**39.09**
BI-LSTM CRF-ACE	**6.67**	17.7	31.58	44.1	46.75	46.77

Table: Results obtained on the HIPE test depending on the system used as well as the amount of HIPE training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

Should we annotate? - HIPE

Models / Splits	0	10	50	250	400	444
BERT	0	0	0.04	42.6	51.46	**52.14**
BERT-ACE	10.08	25.17	47.84	**54.55**	57.37	58.14
CharBERT	0	0	32.82	54.15	57.28	**57.57**
CharBERT-ACE	13.00	27.33	46.93	**57.85**	61.07	61.13
BI-LSTM CRF	0	0	0.35	29.64	37.63	**39.09**
BI-LSTM CRF-ACE	6.67	17.7	31.58	**44.1**	46.75	46.77

Table: Results obtained on the HIPE test depending on the system used as well as the amount of HIPE training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

Should we annotate? - HIPE

Models / Splits	0	10	50	250	400	444
BERT	0	0	**0.04**	42.6	51.46	52.14
BERT-ACE	10.08	25.17	47.84	54.55	57.37	58.14
CharBERT	0	0	**32.82**	54.15	57.28	**57.57**
CharBERT-ACE	13.00	27.33	46.93	57.85	61.07	**61.13**
BI-LSTM CRF	0	0	**0.35**	29.64	37.63	39.09
BI-LSTM CRF-ACE	6.67	17.7	31.58	44.1	46.75	46.77

Table: Results obtained on the HIPE test depending on the system used as well as the amount of HIPE training used in the case where our systems are already pre-trained on the entire ACE (MODEL-ACE) and in case it is not (MODEL)

CharBERT-ACE x10 : 63.12 - HIPE winner : 63.2

Should we annotate? - Results

- The use of pre-trained models requires a perfect alignment between the datasets.
- Annotation of target data is required.
- Pre-training a system on contemporary data significantly reduces the cost of annotation.
- CharBERT
  - Needs less data to learn on new data.
  - Seems more robust to OCR error.

Which source corpus to choose ? - Experimental setup

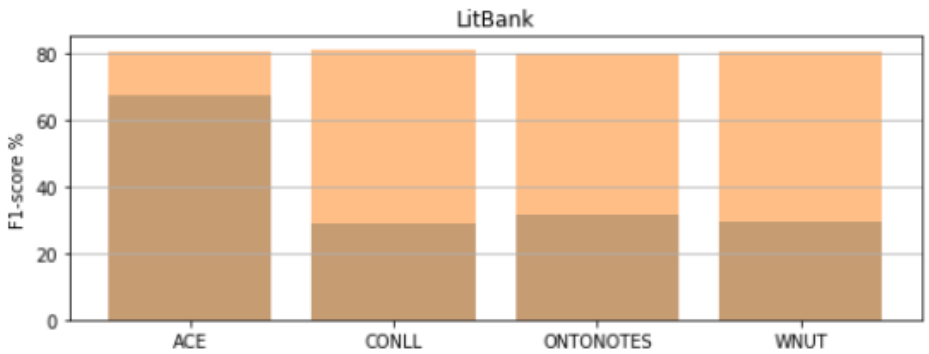
- Questions:
  - How to choose a pre-training corpus ?
  - Is it necessary ?
- Setup :
  - Source corpus :
    - ACE 05
    - CoNNL 03
    - OntoNotes 5
    - WNUT 17
  - Transfer method :
    - Learning on the source corpus - testing on the target corpus. (0%)
    - Finetuning of this model on the entire target data. (100%)
  - Use of CharBERT.

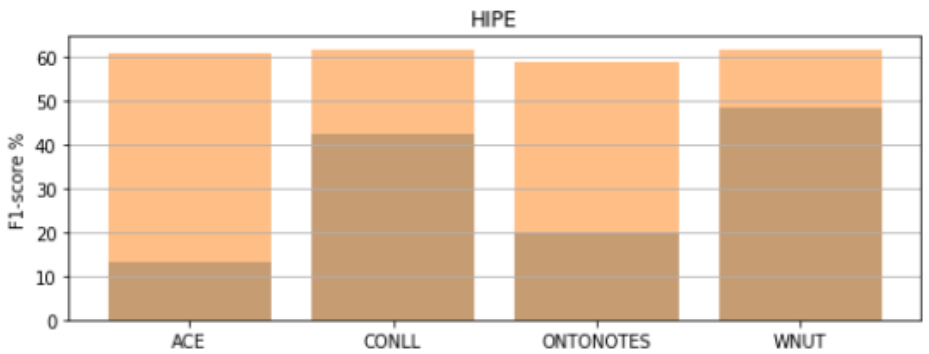
Dataset	CoNLL-03 (4)	OntoNotes 5.0 (18)	WNUT-17 (6)	ACE 05 (6)	LitBank (6)	HIPE (5)
LitBank (6)	3 [76.99%]	5 [98.79%]	2 [76.59%]	6 [**100%**]	6 [100%]	3 [76.99%]
HIPE (5)	3 [92.52%]	5 [**100%**]	3 [57.38%]	3 [92.52%]	3 [92.52%]	5 [100%]

Table: Quantity of entity types shared between datasets. The value in parenthesis represents the number of different entity types for each corpus. The value between brackets is the percentage of entities of shared type in the test set of the target corpus (line).

Which source corpus to choose ?

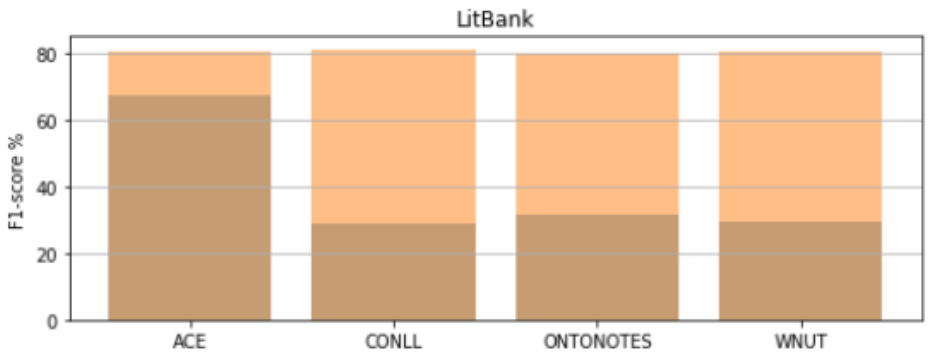
	LitBank		HIPE	
Datasets	0%	100%	0%	100%
ACE 05	**67.38**	80.56	13.00	61.13
CoNLL 03	28.83	80.95	42.43	61.70
OntoNotes 5	31.63	79.73	20.09	58.96
WNUT 17	29.45	80.47	**48.34**	61.60

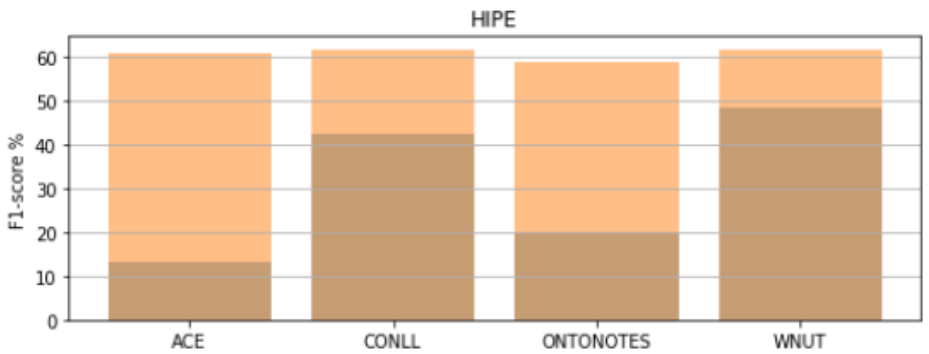




Which source corpus to choose ?

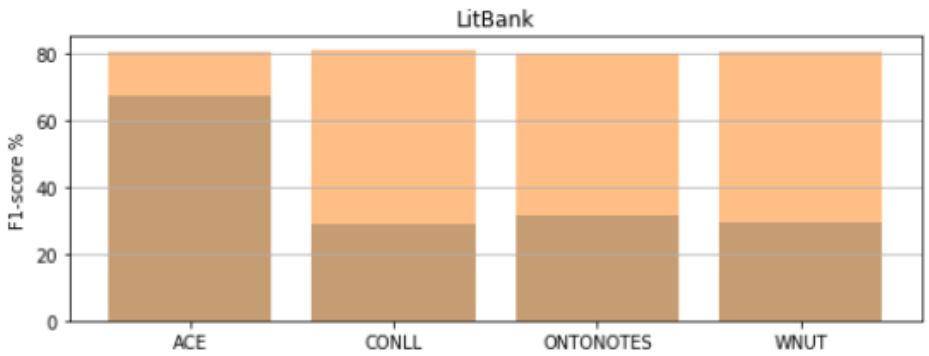
	LitBank		HIPE	
Datasets	0%	100%	0%	100%
ACE 05	67.38	80.56	13.00	61.13
CoNLL 03	28.83	80.95	42.43	61.70
OntoNotes 5	**31.63**	79.73	**20.09**	58.96
WNUT 17	29.45	80.47	48.34	61.60

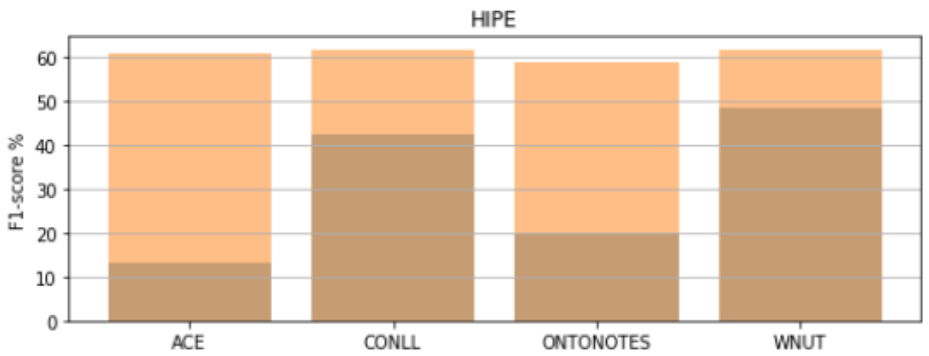




Which source corpus to choose ?

	LitBank		HIPE	
Datasets	0%	100%	0%	100%
ACE 05	67.38	80.56	**13.00**	61.13
CoNLL 03	28.83	80.95	42.43	61.70
OntoNotes 5	31.63	79.73	**20.09**	58.96
WNUT 17	29.45	80.47	48.34	61.60





Which source corpus to choose? - Results

- If we have no annotated target data :
  - Using the same annotation guideline is optimal.
  - Using the same tagset is not enough.
- If we have target annotated data, even in small quantities, the two corpora show their independence from the source data set.

CrossNER: Evaluating Cross-Domain Named Entity Recognition [Liu et al., 2020]

"Results show that emphasizing the partial corpus with specialized entity categories in BERT's domain-adaptive pre-training (DAPT) consistently improves its domain adaptation ability. "

Corpus	Politics (9M)	Science (5M)	Music (10M)	Litera. (9M)	AI (0.3M)	Average
w/o DAPT	68.71	64.94	68.30	63.63	58.88	64.89
DAPT	69.37	66.68	72.05	65.15	61.48	66.95
Integrated	71.44	67.53	74.02	66.57	61.90	68.29

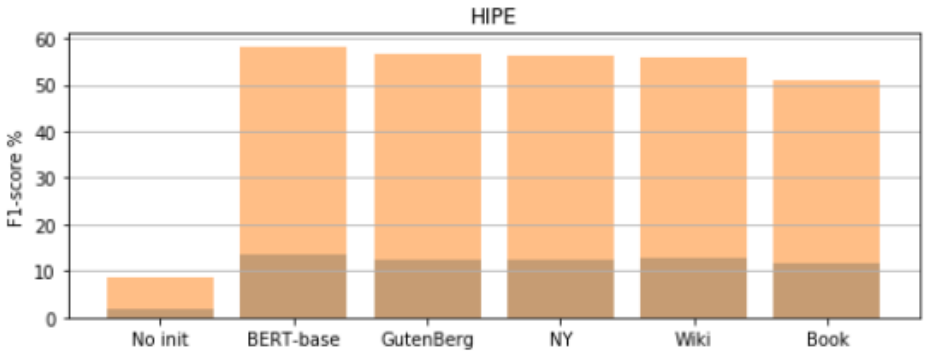
Table: F1-scores of CrossNER proposed methods in three settings and baseline models. Results are averaged over three runs.

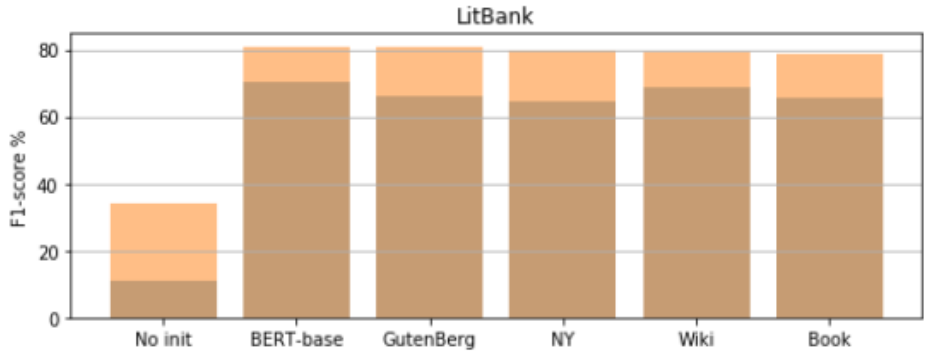
Using unannotated data - Experimental setup

- Questions:
  - Can we use unannotated data to facilitate transfer?
- Setup :
  - BERT finetuning on 2 million sentences (5 epochs):
    - GutenBerg : Source corpus of LitBank.
    - New York Times : Corpus in the same domain as HIPE.
    - BookCorpus : Corpus of the same domain as LitBank + Original corpus of BERT.
    - Wikipédia : BERT's original corpus.
  - Transfer method :
    - Source corpus: ACE 05
    - Learning on the source corpus - testing on the target corpus. (0%)
    - Finetuning of this model on the entire target data. (100%)
  - Use of BERT.

Using unannotated data

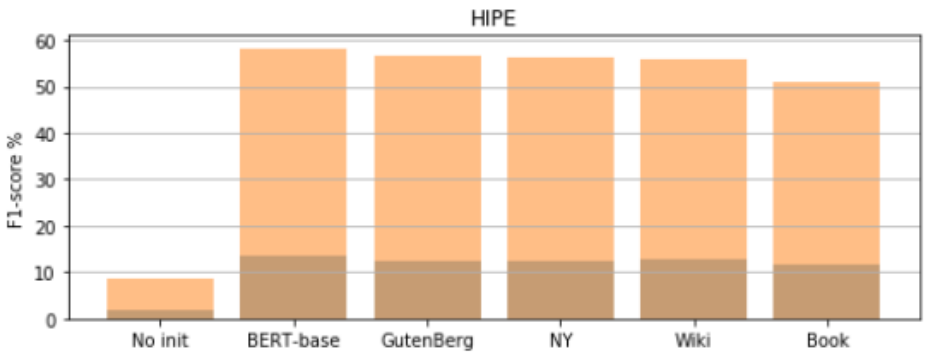
	LitBank		HIPE	
Datasets	0%	100%	0%	100%
No init	10.71	34.02	01.72	08.44
BERT-base	70.44	80.80	13.44	58.14
BookCorpus	65.56	78.97	11.55	51.10
Wikipedia	68.76	79.25	12.88	55.74
NY	64.68	79.61	15.53	56.52
GutenBerg	66.31	80.59	12.53	56.52

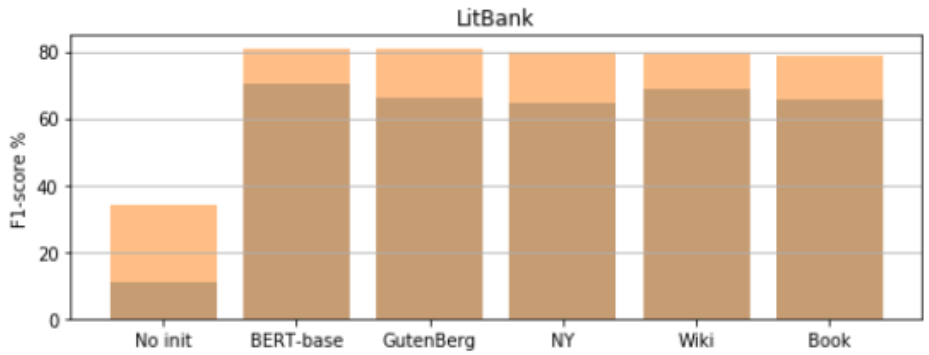




Using unannotated data

	LitBank		HIPE	
Datasets	0%	100%	0%	100%
No init	10.71	34.02	01.72	08.44
BERT-base	70.44	**80.80**	13.44	58.14
BookCorpus	65.56	78.97	11.55	51.10
Wikipedia	68.76	79.25	12.88	55.74
NY	64.68	79.61	15.53	56.52
GutenBerg	66.31	80.59	12.53	56.52
GutenBerg + test LitBank	64.83	**81.57**	12.97	57.87





Are OCR errors a bias? - Experimental setup

- Questions:
  - How much OCR errors bias the processing of historical documents and domain transfer?
  - Does pre-training a system on noisy data facilitates the transfer?
- Setup :
  - Noise generation :
    - Learning the distribution of OCR errors on the IDCAR 17 corpus.
    - Deletion, addition and swapping characters according to a noise level, following this distribution.
  - Transfer method :
    - Learning on the source corpus - testing on the target corpus. (0%)
    - Finetuning of this model on the entire noisy target data. (100%)
    - Learning on the noisy source corpus - testing on the noisy target corpus. (0%)
    - Finetuning of this model on the entire noisy target data. (100%)
  - Use of BERT and CharBERT.

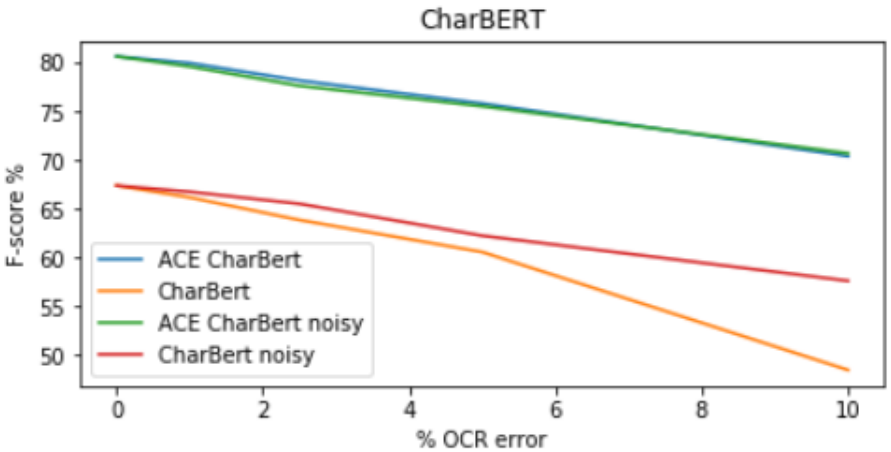
Are OCR errors a bias?

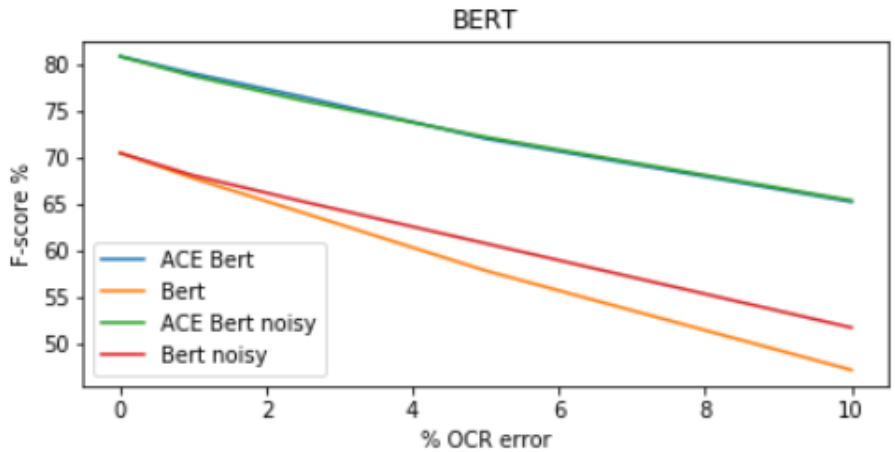
ACE normal - LitBank noisy

	Bert		**CharBERT**	
% noise	0%	100%	0%	100%
0	**70.44**	**80.80**	**67.38**	**80.61**
1	67.73	78.96	66.17	79.92
2.5	64.03	76.45	63.86	78.14
5	57.81	71.97	60.58	75.79
10	**47.16**	**65.20**	**48.50**	**70.39**

ACE noisy - LitBank noisy

	Bert		**CharBERT**	
	0%	100%	0%	100%
0	70.44	80.80	67.38	80.61
1	68.03	78.69	66.74	79.54
2.5	65.23	76.06	65.52	77.60
5	60.73	72.14	61.60	75.50
10	51.70	65.34	57.63	70.69





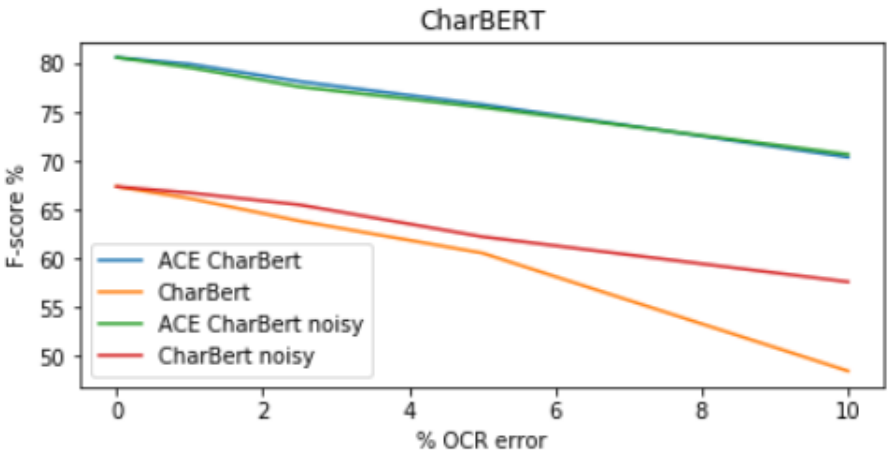
Are OCR errors a bias?

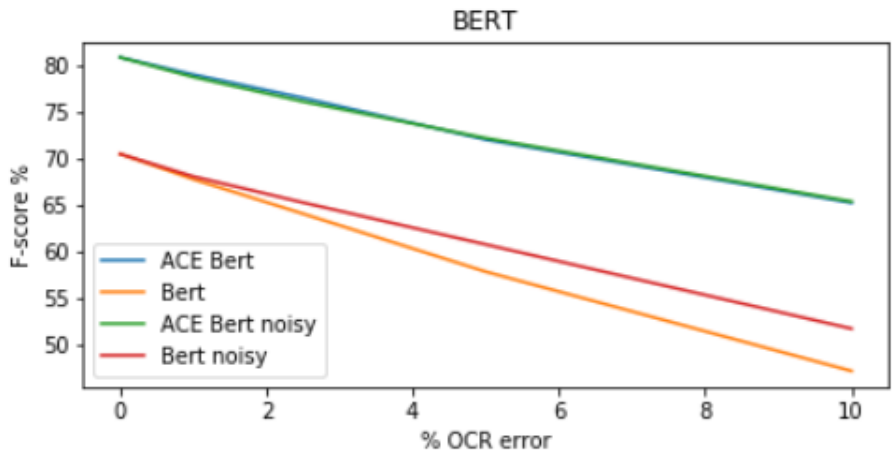
ACE normal - LitBank noisy

	Bert		**CharBERT**	
	0%	100%	0%	100%
% noise				
0	70.44	80.80	67.38	80.61
1	67.73	78.96	66.17	79.92
2.5	64.03	76.45	63.86	78.14
5	57.81	71.97	60.58	75.79
10	47.16	65.20	48.50	70.39

ACE noisy - LitBank noisy

	Bert		**CharBERT**	
	0%	100%	0%	100%
	70.44	**80.80**	**67.38**	**80.61**
	68.03	78.69	66.74	79.54
	65.23	76.06	65.52	77.60
	60.73	72.14	61.60	75.50
	51.70	**65.34**	**57.63**	**70.69**





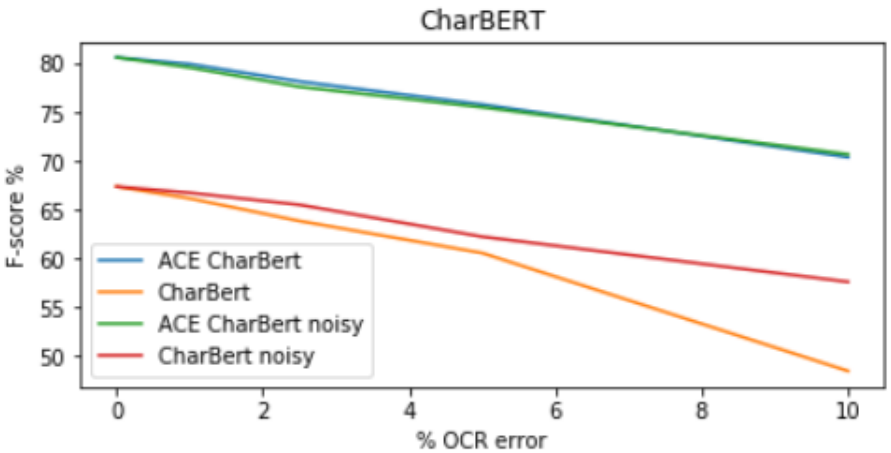
Are OCR errors a bias?

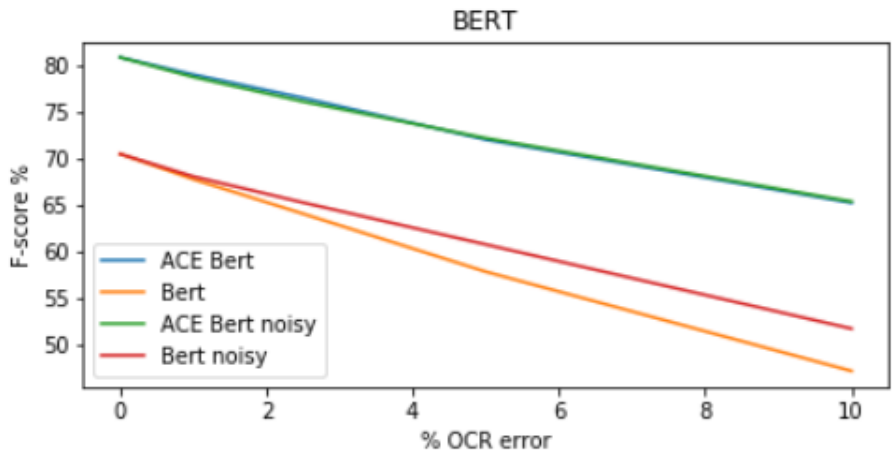
ACE normal - LitBank noisy

	Bert		**CharBERT**	
% noise	0%	100%	0%	100%
0	70.44	80.80	67.38	80.61
1	67.73	78.96	66.17	79.92
2.5	64.03	76.45	63.86	78.14
5	57.81	71.97	60.58	75.79
10	**47.16**	65.20	**48.50**	70.39

ACE noisy - LitBank noisy

	Bert		**CharBERT**	
	0%	100%	0%	100%
0	70.44	80.80	67.38	80.61
1	68.03	78.69	66.74	79.54
2.5	65.23	76.06	65.52	77.60
5	60.73	72.14	61.60	75.50
10	**51.70**	65.34	**57.63**	70.69





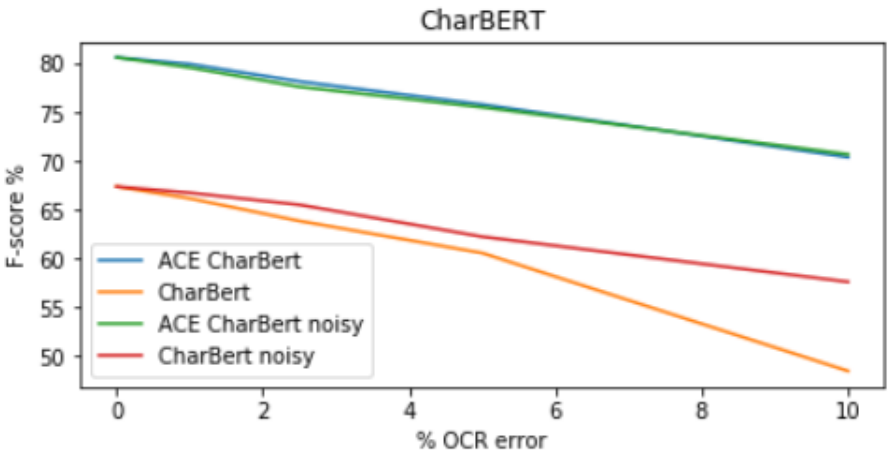
Are OCR errors a bias?

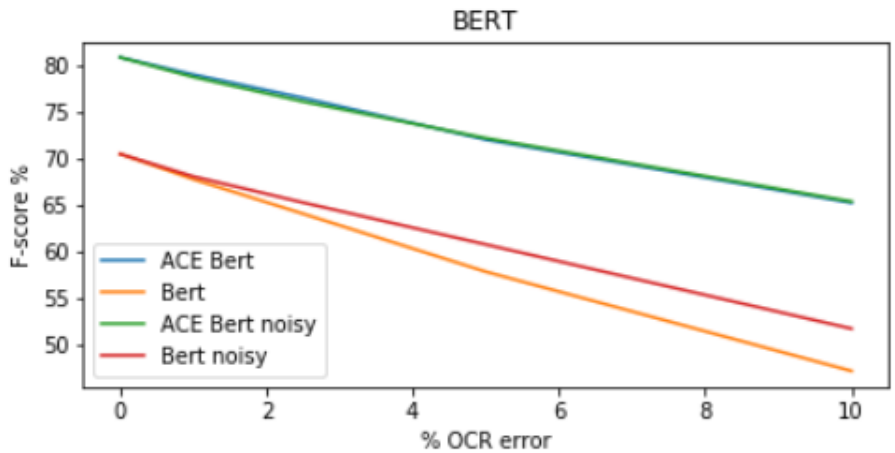
ACE normal - LitBank noisy

	Bert		**CharBERT**	
% noise	0%	100%	0%	100%
0	70.44	80.80	67.38	80.61
1	67.73	78.96	66.17	79.92
2.5	64.03	76.45	63.86	78.14
5	57.81	71.97	60.58	75.79
10	47.16	**65.20**	48.50	**70.39**

ACE noisy - LitBank noisy

	Bert		**CharBERT**	
	0%	100%	0%	100%
	70.44	80.80	67.38	80.61
	68.03	78.69	66.74	79.54
	65.23	76.06	65.52	77.60
	60.73	72.14	61.60	75.50
	51.70	**65.34**	57.63	**70.69**





Are OCR errors a bias? - Results

- OCR errors are a significant bias to transfer approaches.
- CharBERT is more robust than BERT at the noise level in our data.
- Using noisy source data following the same distribution as the target data brings better results in the zero-shot scenario.

	CharBERT	
	0%	100%
% of noise		
0%	**13.00**	61.13
10%	12.44	**62.35**

Table: Using noisy ACE 05 to evaluate HIPE.

Conclusion

- Under certain conditions the use of pre-trained systems is possible to use them on historical data.
- Domain adaptation allows to reduce considerably the annotation cost.
- Noise in the data are a bias to our systems.
- Preparing systems to see target data in advance facilitates transfer to the historical domain.
- CharBERT is more robust to small amounts of data and noise.

Questions?

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019).
Flair: An easy-to-use framework for state-of-the-art nlp.
In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of deep bidirectional transformers for language understanding.
pages 4171–4186.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016).
Neural Architectures for Named Entity Recognition.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., and