Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data

Amir Hazem, **Mérième Bouhandi**, Florian Boudin, Béatrice Daille

LS2N - Université de Nantes, France

May 10, 2022

Outline

1. Introduction
2. BERT for ATE
3. Data Sets
4. Experiments and Results
5. Discussion
6. Conclusion

Automatic Term Extraction (ATE)

- Key component for several NLP applications
- Still exhibits weak results
- Efficient Transformers-based models (BERT) in many NLP tasks
- No systematic evaluation of ATE has been conducted so far
- We run an extensive study on fine-tuning pre-trained BERT models for ATE

We propose 3 Strategies

- BERT embeddings as input features for a biLSTM with a CRF layer
- BERT as a sequence labeling model for NER
- BERT as a binary classifier

Bi-LSTM for Sequence Labeling with BERT Embeddings

- Address ATE as sequence labeling task
- biLSTM are standard to the task of named entity recognition using sequence labeling
- A CRF layer is connected to the LSTM last output layer
- BERT embeddings are used as input of our model to overcome the problem of low training data

BERT as NER system

- ATE as a named entity recognition task
- Each term is seen as a named entity
- Adapt the pre-trained BERT for NER in order to perform ATE
- We consider only one named entity type: a term
- Each term of a given sentence is assigned a label, and all non-terms are assigned O

BERT as Binary Classifier

- BERT has been trained for Next Sentence Prediction (NSP)

Input 2 pairs of sentences
Output predict if the second sentence is subsequent of the first one

- similarly we fine-tune our model by providing:

Input1 2 sentence/term pairs
Input2 2 sentence/non-term pairs
Output predict if the term (non-term) is part of the sentence

BERT as Binary Classifier

- Learn shared features between terms and their contexts the same way it does for subsequent pairs of sentences
- We select 50% of the training pairs where the second sequence is a term, while the other 50% pairs consist of randomly chosen n-grams
- For each positive training pair, we randomly select within the sentence of the same pair an n-gram as non-term to keep a balanced training set

Data Sets

- Lack of data for cross-lingual and cross-domain transfer methods for ATE [schuster-etal-2019]
- TermEval shared task [rigoutsterryn-EtAI:2020:COMPUTERM] offers an appropriate annotation methodology
- Falls within both multilingual and multi-domain scenarios
- Four specialized domains: corruption, dressage, wind energy, and heart failure
- Three languages: English (en), French (fr), and Dutch (nl)

Data Sets

Domain		# Tokens	# Documents	# Term lists
corruption	en	468,711	44	1174
	fr	475,244	31	1217
	nl	470,242	49	1295
dressage	en	102,654	89	1575
	fr	109,572	125	1183
	nl	103,851	125	1546
wind energy	en	314,618	38	1534
	fr	314,681	12	968
	nl	308,742	29	1245
heart failure	en	45,788	190	2585
	fr	46,751	215	2423
	nl	47,888	175	2257

Table: Number of unique tokens and documents per corpus for English (en), French (fr) and Dutch (nl) as well as the size of the evaluation term lists.

Experiments and Results

- We follow the the TermEval shared task evaluation procedure
- Using three domains (corruption, wind energy, and dressage) for training and validation
- Heart failure domain for testing
- For each experiment, reported results are the mean average of 10 runs

Pre-trained Models

English

- bert-base-cased (BERT) [devlin2018bert]
- bert-base-multilingual-cased (BERT-multi)
- roberta-base (RoBERTa) [liu2019roberta] [a]

[a] Additional models such as bert-large-cased and distilroberta-base are presented in the supplementary material

French

- Camembert [2019CamemBERT]
- bert-base-multilingual-cased (BERT-multi)

Dutch

- bert-base-multilingual-cased

Dev Set Experiments: Domain Transfer learning

	English		
Test	Train		
Corp	Equi	Wind	Equi+Wind
RoBERTa	36.16	**39.97**	37.33
BERT (NER)	15.97	22.18	**31.48**
BERT-biLSTM-CRF	20.54	19.01	**20.63**
Equi	Corp	Wind	Corp + Wind
RoBERTa	42.07	**44.99**	42.03
BERT (NER)	38.55	**41.74**	40.81
BERT-biLSTM-CRF	17.65	**19.51**	18.04
Wind	Corp	Equi	Corp + Equi
RoBERTa	**36.16**	33.77	32.91
BERT (NER)	**30.58**	26.89	28.39
BERT-biLSTM-CRF	19.99	20.64	**22.11**

Table: Domain transfer results (F1%) of ATE on the English dataset.

Dev Set: Language transfer learning

Test corpus	Training corpus			
	corp		equi	
	en	en+fr+nl	en	en+fr+nl
corp (en)	-	-	35.09	**36.27**
equi (en)	34.68	**39.29**	-	-
wind (en)	**34.04**	32.89	34.04	**36.01**
	fr	en+fr+nl	fr	en+fr+nl
corp (fr)	-	-	**34.84**	34.14
equi (fr)	27.64	**28.23**	-	-
wind (fr)	24.12	**24.32**	24.12	**26.07**

Test corpus	Training corpus			
	wind		All	
	en	en+fr+nl	en	en+fr+nl
corp (en)	35.56	**37.64**	**35.25**	34.16
equi (en)	40.54	**42.10**	40.15	**41.60**
wind (en)	-	-	34.94	**35.42**
	fr	en+fr+nl	fr	en+fr+nl
corp (fr)	**35.71**	35.21	**35.91**	33.97
equi (fr)	**32.88**	32.33	**31.35**	29.98
wind (fr)	-	-	22.97	**24.12**

Table: Language transfer learning results (F1%) on the validation sets. All the experiments were based on bert-base-multilingual model (BERT-multi) .

Test Set Experiments: English

	English		
	P	R	F1
Baselines			
Features (en)	39.44	29.28	33.61
Features (enfrnl)	14.96	39.20	21.65
Vanilla-biLSTM-CRF	6.84	10.16	8.17
Team results			
TALN-LS2N	34.78	70.87	46.66
RACAI	42.40	40.27	41.31
NYU	43.46	23.64	30.62
e-Termino	34.43	14.20	20.10
NLPLab	21.45	15.59	18.06
Proposed			
BERT (en)	36.31	72.15	**48.21**
BERT (NER)	**57.22**	27.74	37.37
BERT-biLSTM-CRF	24.17	38.32	29.54
BERT-multi (en)	33.67	71.79	45.77
BERT-multi (enfrnl)	33.01	72.67	45.37

Table: Our methods results on the heart failure test set (%) contrasted with the baselines and results of the participants teams on the shared task

Test Set Experiments: French

	French		
	P	R	F1
Baselines			
Features (fr)	48.90	53.47	50.92
Features (enfrnl)	18.71	40.75	25.64
Vanilla-biLSTM-CRF	4.52	11.79	6.53
Team results			
TALN-LS2N	45.17	51.55	48.15
e-Termino	36.33	13.50	19.68
NLPLab	16.07	11.18	13.19
Proposed			
CamemBert	40.11	**70.51**	**51.09**
CamemBert (NER)	**57.51**	25.75	35.57
BERT-biLSTM-CRF	21.13	32.48	25.60
Bert-multi (fr)	36.13	68.11	47.18
BERT-multi (enfrnl)	33.16	69.61	44.91

Table: Our methods results on the heart failure test set (%) contrasted with the baselines and results of the participants teams on the shared task

Test Set Experiments: Dutch

	Dutch		
Baselines			
Features (nl)	32.29	36.07	34.07
Features (enfrnl)	16.45	49.83	24.73
Vanilla-biLSTM-CRF	6.27	9.34	7.50
Team results			
NLPLab	18.9	18.6	18.7
e-Termino	29.0	9.6	14.4
Proposed			
BERT-multi	32.86	**75.47**	45.73
BERT (NER)	**63.75**	42.53	**51.02**
BERT-biLSTM-CRF	22.65	34.62	27.38

Table: Our methods results on the heart failure test set (%) contrasted with the baselines and results of the participants teams on the shared task

Discussion

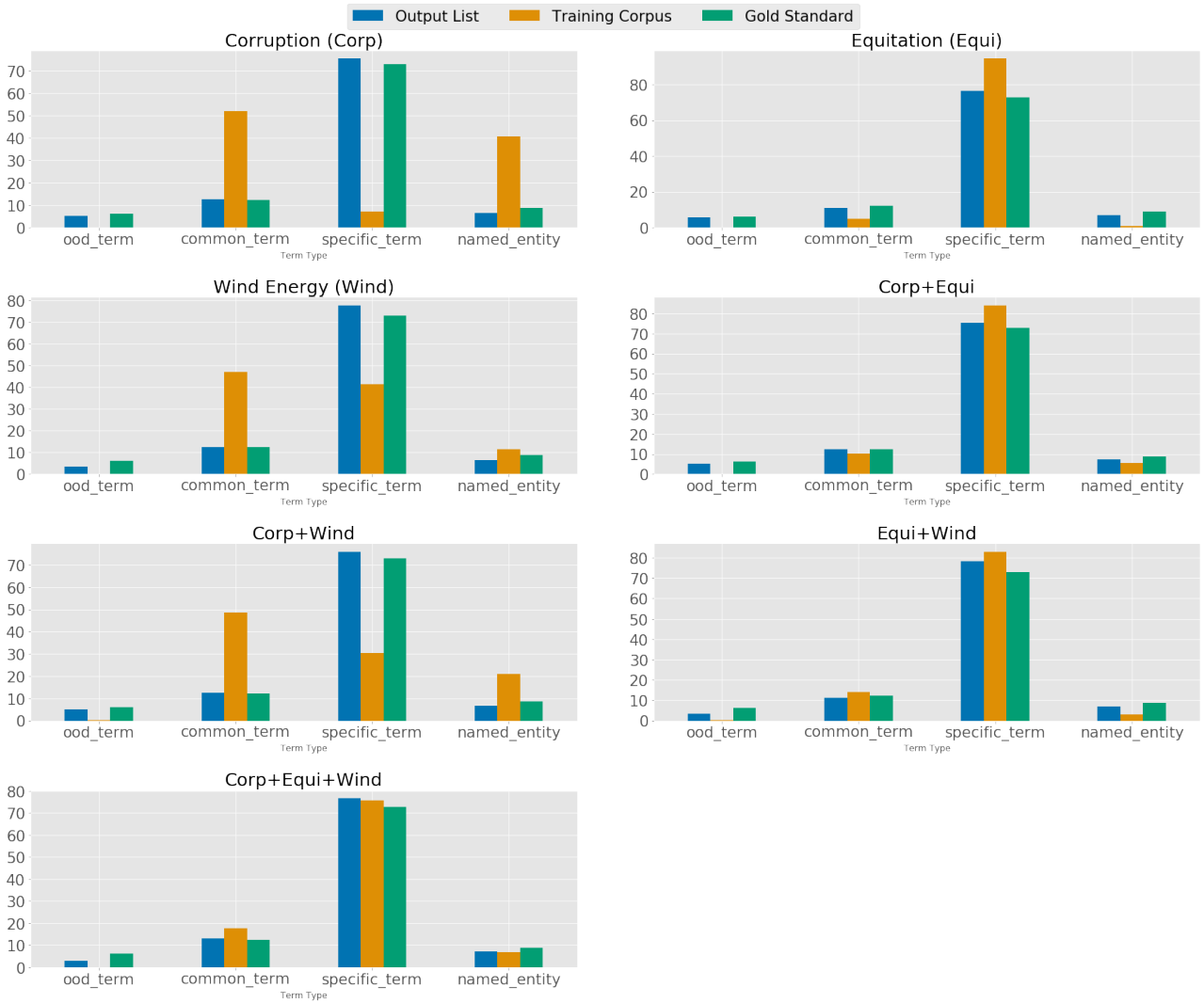
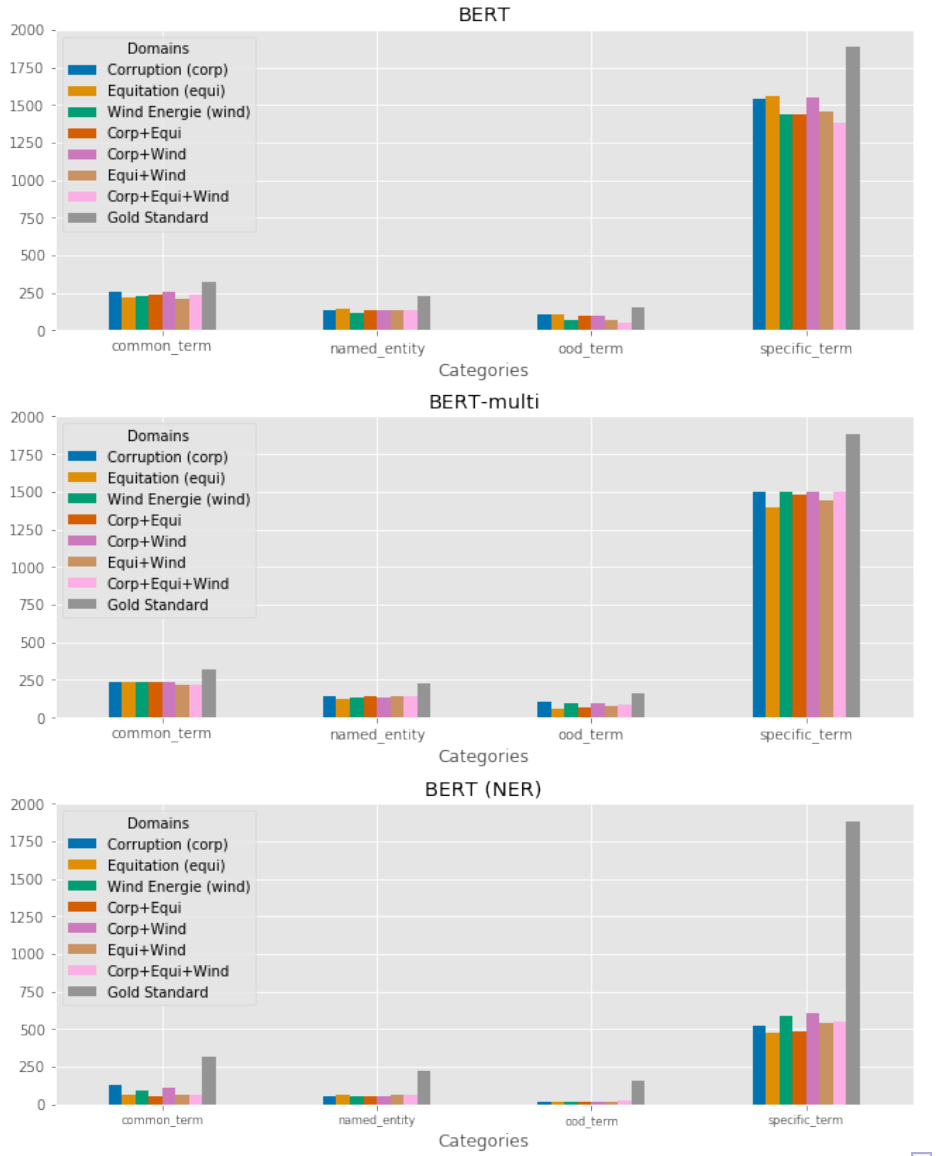


Figure: Proportion (%) of each term type

Discussion: BERT outputs per term type



Discussion: Output lists filtering

	English (BERT)			French (Camembert)			Dutch (BERT-multi)		
	P	R	F1	P	R	F1	P	R	F1
No filtering	34.75	**71.23**	46.71	40.71	**68.81**	51.15	63.76	**42.55**	51.04
Patterns	34.88	70.11	46.58	40.83	67.94	51.01	64.59	42.24	51.08
Specificity	**53.74**	54.64	**54.18**	**57.65**	48.30	52.57	66.21	21.78	32.04
C-Value	48.08	61.02	53.78	55.10	58.78	**56.88**	**83.58**	39.90	**54.01**

Table: Results after filtering down our best models output lists using different filters, improving the results. For the pattern filtering, all terms that fit nominal patterns are kept. For the C-Value and the Specificity, we only keep the first half of the ordered filtered list.

BERT for ATE

- The binary classification system obtained the best results
- Varying the data sets revealed the effectiveness of cross-domain BERT fine-tuning
- Simplicity of our strategies, since neither feature engineering nor pattern design is needed
- Terms share cross-domain marked-contexts captured by BERT

BERT for ATE

- Using several training data sets does not necessarily guarantee better performance
- How to choose the most appropriate training data sets?
- Observing BERT outputs underlined the presence of some inappropriate or ungrammatical term-like candidates
- Post-filtering techniques proved to be remarkably helpful for improving the results

ATE as sequence labeling

- We expected better performance from BERT (NER)
- Only a small proportion of named entities was extracted
- Despite lower F1 scores, BERT (NER) showed acceptable and sometimes competitive results compared to other methods
- Best precision over all the experiments. This is more noticeable for Dutch BERT (NER) as it outperformed the BERT classifier

Conclusion

- We proposed the first systematic study of BERT models for the ATE task on four low resource specialized domains in three languages
- We experimented with BERT as a binary classifier, and as a named entity recognition system, as well as a biLSTM-CRF trained on BERT features
- The obtained empirical results indicate that BERT is able to transfer learning across domains and languages, opening a new promising direction for ATE