Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 648–662
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

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

Amir Hazem, Mériem Bouhandi, Florian Boudin, Béatrice Daille
LS2N - Université de Nantes, France

Abstract

Automatic Term Extraction (ATE) is a key component for domain knowledge understanding and an important basis for further natural language processing applications. Even with persistent improvements, ATE still exhibits weak results exacerbated by small training data inherent to specialized domain corpora. Recently, transformers-based deep neural models, such as BERT, have proven to be efficient in many downstream NLP tasks. However, no systematic evaluation of ATE has been conducted so far. In this paper, we run an extensive study on fine-tuning pre-trained BERT models for ATE. We propose strategies that empirically show BERT's effectiveness using cross-lingual and cross-domain transfer learning to extract single and multi-word terms. Experiments have been conducted on four specialized domains in three languages. The obtained results suggest that BERT can capture cross-domain and cross-lingual terminologically-marked contexts shared by terms, opening a new design-pattern for ATE.

Keywords: ATE, automatic term extraction, terminology, low resource, multilingual.

1. Introduction

Automatic Term Extraction (ATE) is essential to specialized domains knowledge understanding and can further be used as input for other NLP applications, including information retrieval, thesauri construction, machine, and computer-aided translation, etc. The aim of ATE is to identify terminological units in specific domains (Castellví et al., 2001). For that purpose, several methods have been proposed, ranging from linguistic-based (Ananiadou, 1994; Justeson and Katz, 1995), statistical (Schäfer et al., 2015), feature-based (Conrado et al., 2013) as well as hybrid (Basili et al., 1997; Aubin and Hamon, 2006) and deep neural-based (Wang et al., 2016; Hätty and Schulte im Walde, 2018) approaches. Nonetheless, they still exhibit weak results, and manual post-filtering remains necessary to exploit current system outputs (Terryn et al., 2018). Furthermore, feature-based approaches, which, up to now, are considered among the most accurate, require labor-intensive feature engineering with no straightforward transfer learning if a new domain is addressed. The only transfer learning-based method that we are aware of deals with clinical concept detection, and still requires feature engineering (Lv et al., 2014).

Current research on ATE deals with two main problems: (1) the small size of data collections inherent to specialized domains and (2) the limited availability of annotations due to the lack of consensus about the nature of terms (Terryn et al., 2018). The former problem has been addressed in other downstream applications using data selection or data augmentation, as shown in SMT (Moore and Lewis, 2010; Axelrod et al., 2011; Gascó et al., 2012), and NMT (Parcheta et al., 2018) as well as combining specialized and general domain word embeddings as demonstrated in the clinical concept detection (El Boukkouri et al., 2019) and bilingual terminology extraction (Liu et al., 2018) tasks. Nonetheless, these techniques are often difficult to apply for ATE since they all require a reasonable amount of annotated data[1].

The latter problem has also been addressed in other applications using transfer learning, such as, in conversational AI (Schuster et al., 2019) and named entity recognition (Jia et al., 2019; Rahimi et al., 2019) tasks. However, transfer learning is based on mapped representations learned from one sufficient annotated source data and targeted another less-resourced data. In ATE, proper and sufficient source annotations are not available, making such transfer learning difficult even to consider. To overcome the annotation gap occurring in ATE, (Judea et al., 2014) proposed an unsupervised data generation approach. However, it relies on linguistic features that need to be defined beforehand.

Over the last few years, a new family of deep neural approaches impulsed by transformers has brought forward significant improvements in many downstream applications (Vaswani et al., 2017; Devlin et al., 2018; Raffel et al., 2019). BERT, for instance, can learn bidirectional context representations from a considerable amount of general domain data that can be fine-tuned for a specific task.

These approaches remarkably capture semantic and syntactic information (Rogers et al., 2020; Reif et al., 2019), and supersede classical approaches on many NLP tasks. Based on the assumption that terms share terminologically-marked contexts (Meyer, 2001), we put greater stress on this hypothesis and propose to investigate its validity on cross-domain and cross-lingual scenarios by experiencing the usage of BERT for ATE on four specialized domains in three different languages. We believe our work provides three important contribu-

[1] Data augmentation-based SMT, and NMT systems often require additional annotated bi-texts for training or improving language models. In clinical concept detection and bilingual term extraction, embedding models are used as features for deep neural approaches, such as biLSTM, but these methods also require enough annotated data to be trained on.

tions: 1) The first extensive study of BERT for ATE; 2) The proposition of efficient and straightforward BERT-based strategies (as a binary classifier, as a sequence labeling method, and as features for a BiLSTM-CRF) that can be further improved by post-filtering; 3) The empirical verification that the terminologically-marked contexts assumption holds in both the cross-domain and cross-lingual scenarios thanks to transfer learning.

2. Related Work

Several approaches for ATE have been proposed so far, ranging from the early linguistic-based (Bourigault, 1992; Justeson and Katz, 1995) to the recent deep neural-based (Wang et al., 2016; Hätty and Schulte im Walde, 2018). We divide their classification into supervised versus unsupervised methods.

2.1. Unsupervised Methods

Unsupervised methods usually involve two steps: (i) the identification of term-like units and (ii) the filtering and sorting of the extracted units that may not be terms using several rules (Daille, 2017).

The term-like units are often sequences of noun chunks that are selected during the identification step according to their part-of-speech and their morpho-syntactic patterns (Justeson and Katz, 1995), as well as general phrase rules involving phrase borders (Bourigault, 1992), samples of annotated texts and lists of seed terms (Jacquemin, 2001). Bourigault (1992), for instance, introduced a two steps-based method: (1) analysis: in which a set of rules of frontier marker identification (e.g., punctuation marks, verbs, pronouns, etc.) is defined and (2) parsing: in which the maximal-length noun phrases are determined, and sub-group candidates are extracted based on grammatical structure rules. Justeson and Katz (1995) used the linguistic properties of technical terminology. Terms can also be collected using the lexical expansion of seed term lists and nominal phrases that include seed terms (Jacquemin, 2001; Burgos Herera, 2014).

In the filtering step, all the collected term-like sequences are usually ranked according to termhood and unithood (Kageura and Umino, 1996) criteria, which can be measured by frequency, association measures (e.g., mutual information, log-likelihood), specificity measures (Ahmad et al., 1994), etc. Filtering can also be done by removing nested sequences (incorrect, incomplete, or expansions that are not part of the base term). This is usually done using several measures, such as the C-value (Frantzi and Ananiadou, 2000), contextual filtering under the assumption that terms are gregarious, the NC-value (Frantzi and Ananiadou, 2000) or by exploiting the fact that terms, contrary to regular noun phrases, share terminologically-marked contexts (Meyer, 2001; Condamines, 2002). In this work, we assume that terms do share strongly marked contexts that BERT is able to capture, and we push this hypothesis further to cross-domain and cross-lingual scenarios.

2.2. Supervised Methods

Prior works on ATE used different learning algorithms and feature sets (Conrado et al., 2013; Judea et al., 2014; Hätty et al., 2017). (Hatzivassiloglou V, 2001), for instance, presented an automated system for assigning protein and gene class labels to biological terms in free texts. They explored three established learning techniques: Bayesian learning (Duda and Hart, 1973), decision tree using C4.5 implementation (Quinlan, 1993), and inductive rule learning using RIPPER implementation (Cohen, 1996). Takeuchi and Collier (2005) used support vector machines to extract Bio-medical entities, while Zhang and Fang (2010) used Conditional Random Fields (CRF) and syntactic features. These methods, among others, require quite a high number of features to be set. Conrado et al. (2013), for instance, explored Naive Bayes and decision tree using 17 linguistic, statistical, and hybrid features, such as frequency, TF-IDF, POStags, etc. They also used two additional features from corpora that belong to another domain Finally, Hätty et al. (2017) used a random forest classifier, trained on several termhood measures, and recursively applied to the components. To overcome the lack of training data, Judea et al. (2014) proposed an unsupervised training data augmentation step prior to using a binary classifier and a CRF trained on different types of features. Even if their approach is language-independent, it has only been tested on English and still relies on linguistic features that need to be defined beforehand.

More recently, several deep neural approaches have also been applied to ATE. Wang et al. (2016) used two deep neural classifiers: a CNN that learns terms representation using multiple filters to capture sub-gram information and an LSTM that captures the meaning of a term from the sequential combination of its constituents. Amjadian et al. (2016) leveraged local and global embeddings to encapsulate the meaning of the term for the classification step, although they only work with unigram terms. Few neural-based methods regard ATE as a sequence labeling task (Han et al., 2018). In clinical concept detection, which can be seen as a sub-task of ATE, several methods have, however, been proposed (Lv et al., 2014; El Boukkouri et al., 2019). Finally, Šajatović et al. (2019) evaluated 16 state-of-the-art methods at the corpus and document level, showing that there is no best-performing single method for corpus-level ATE.

3. BERT for ATE

We address ATE by introducing three different strategies. The first consists in using BERT as a binary classifier, while the second uses BERT as a sequence labeling model for Named Entity Recognition (NER). Finally, we propose a feature-based strategy which consists in using BERT embeddings as input features for a bidirectional LSTM (biLSTM) with a Conditional Random Fields (CRF) layer.

3.1. BERT as Binary Classifier

BERT is a supervised learning model[2] that has been trained on the Masked Language Model (MLM) objective and Next Sentence Prediction (NSP). For NSP, the model takes as input pairs of sentences and learns to predict if the second sentence is the subsequent sentence of the first one. We similarly fine-tune our model, providing sentence/term pairs and sentence/non-term pairs instead of sentence pairs. In a similar fashion to NSP, 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[3]. Therefore, 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. Class balancing is often a vital setup for classifiers' efficiency. Even if some existing techniques do reduce the impact of unbalanced classes, we only deal with balanced training in this work. The intuition behind this step is that BERT will learn shared features between terms and their contexts the same way it does for subsequent pairs of sentences. Moreover, for each correct input pair, the term is present in both sequences, which we expect to be advantageous to learn more context about terms during the masked LM phase. Upon testing, all the n-grams of a sentence are used to provide sentence/term-like pairs.

3.2. BERT as NER system

We also address ATE as a named entity recognition task where each term is seen as a named entity. Therefore, we can easily adapt the pre-trained BERT for NER in order to perform ATE. This procedure can be seen as an over-simplification of the actual NER task where we consider only one named entity type (a term), whereas, in the classical task, we consider several named entities. Similarly to NER, we use the IOB scheme (I indicates that the token is inside an entity, O indicates that a token is outside the entity, and B indicates that the token begins the entity).

Since the BERT NER is already pre-trained following the IOB scheme, we chose the ORG tag for all the experiments[4]. Hence, each term of a given sentence is assigned by B-ORG or I-ORG labels, and all non-terms are assigned by O-ORG.

3.3. Bi-LSTM for Sequence Labeling with BERT Embeddings

We finally address ATE as sequence labeling task. Bidirectional LSTM (biLSTM) are standard to the task of named entity recognition using sequence labeling. A Conditional Random Fields (CRF) layer is connected to the LSTM last output layer, enabling it to efficiently use neighboring tag information to better model the conditional probability of the output state sequence. Weights in vanilla LSTM can either be initialized with random values or with third-party word embeddings: here, BERT embeddings are used as input of our model to overcome the problem of low training data.

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 1: 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.

4. Data Sets

Up to now, cross-lingual and cross-domain transfer methods for ATE have been hindered by the lack of multilingual data sets annotated according to the same guidelines (Schuster et al., 2019). However, the recently released data set from the first TermEval shared task (Rigouts Terryn et al., 2020) offers an appropriate annotation methodology that falls within both multilingual and multi-domain scenarios. Therefore, in our experiments and evaluation, we use the TermEval data sets, which consist of four specialized domains: corruption, dressage, wind energy, and heart failure in three languages: English (en), French (fr), and Dutch (nl). Table 1 depicts the number of documents and tokens for each corpus as well as the test lists size.

5. Experiments and Results

We follow the evaluation procedure defined in the TermEval shared task (Rigouts Terryn et al., 2020) which consists in using three domains (corruption, wind energy, and dressage) for training and validation and the heart failure domain for testing. We first report the results on the development sets used for cross-validation and fine-tuning, and then the obtained results on the heart failure test set. For each experiment, reported results are the mean average of 10 runs.

5.1. Pre-trained Models

As pre-trained models can behave differently according to the addressed tasks (Liu et al., 2019), we evaluate several BERT models that we have fine-tuned on

[2] While various binary classifiers have been applied to ATE (see the Related Work section), to our knowledge, the only work that addressed ATE using BERT comes from the TermEval shared task (Rigouts Terryn et al., 2020)

[3] The n-gram size is set empirically.

[4] Within the IOB tags (PER, ORG, MISC, and LOC), no significant difference in performance has been observed.

	Corp			**Equi**			**Wind**		
English	P	R	F1	P	R	F1	P	R	F1
BERT	29.46	49.45	36.79	30.76	71.61	**42.94**	21.87	74.03	33.73
BERT-multi	25.47	58.21	35.25	27.87	71.94	40.15	22.92	74.51	**34.94**
RoBERTa	26.85	**62.53**	**37.33**	29.17	**75.59**	42.03	20.93	**77.08**	32.91
BERT (NER)	**52.92**	22.46	31.48	**54.45**	32.63	40.81	**35.48**	23.66	28.39
BERT-multi (NER)	34.09	11.51	17.21	13.41	14.67	14.01	11.14	18.58	13.93
RoBERTa (NER)	31.91	10.22	15.48	49.39	28.44	36.11	32.48	23.08	26.98
BERT-biLSTM-CRF	22.61	18.98	20.63	20.65	16.03	18.05	21.04	23.31	22.11
French	P	R	F1	P	R	F1	P	R	F1
BERT-multi	27.53	53.03	35.91	21.17	60.77	31.35	13.82	68.99	22.97
CamemBERT	29.73	**62.61**	**40.25**	24.38	**65.89**	**35.53**	15.63	**72.51**	**25.71**
BERT-multi (NER)	39.17	11.59	17.89	42.83	20.20	27.45	23.22	25.93	24.5
CamemBERT (NER)	**42.61**	12.33	19.13	**48.01**	23.33	31.40	**24.09**	24.48	24.28
BERT-biLSTM-CRF	18.78	16.21	17.40	18.53	19.08	19.76	17.20	20.72	18.79
Dutch	P	R	F1	P	R	F1	P	R	F1
BERT-multi	23.70	**68.59**	**35.21**	30.39	**63.85**	41.10	15.85	**66.55**	25.52
BERT-multi (NER)	**40.93**	15.68	22.67	**65.25**	32.79	**43.65**	**37.41**	38.31	**37.85**
BERT-biLSTM-CRF	20.51	21.75	21.11	21.17	18.90	19.97	21.70	15.28	17.93

Table 2: ATE scores (%) in terms of Precision (P), Recall (R) and F1-score (F1). For each validation corpus, the two remaining ones were used for fine-tuning. We contrast 3 usages of BERT : as a binary classifier, BERT-NER and BERT embeddings used with a biLSTM-CRF. We also contrast BERT's multi-lingual version as well as RoBERTa for English and Camembert for French.

the specialized corpora. The tested models for English were: bert-base-cased (BERT) (Devlin et al., 2018), bert-base-multilingual-cased (BERT-multi), and roberta-base (RoBERTa) (Liu et al., 2019)[5]. For French, we used Camembert (Martin et al., 2019) and bert-base-multilingual-cased (BERT-multi). Finally, for Dutch, we only considered bert-base-multilingual-cased.

5.2. Dev Set Experiments

5.2.1. Domain Transfer Learning Evaluation

The domain transfer learning procedure consists of fine-tuning BERT on a given domain and testing on another domain. Having three validation data sets (corruption (Corp), dressage (Equi), and wind energy (Wind)), and several BERT models, we subdivided the domain transfer evaluation into two experiments: 1) fixing the training data sets while varying BERT models and 2) fixing the models while varying the training data sets.

Table 2 reports the obtained results when varying BERT models using BERT as a binary classifier, BERT as NER, and BERT with biLSTM-CRF. The contrasted models are BERT and its multilingual version (BERT-multi), as well as RoBERTa for English and CamemBERT for French. For each specialized test set, we fine-tuned on the combination of the remaining data sets, which consists of providing BERT with both domains as one single data set. If this procedure may seem counter-intuitive, we expect that BERT can capture correlations between terms across domains by assuming that they have terminologically-marked contexts.

Overall, better results were obtained in terms of the F1 score when using BERT as a binary classifier. However, BERT (NER) showed competitive results and was sometimes better for Dutch and always better in terms of precision (P). When comparing BERT models used for binary classification, we see that the results are more contrasted for English and that no single model performs the best in each domain. For French, it is clear that Camembert is the best performing model in terms of the F1 score. When it comes to BERT (NER), we observe in some cases a considerable drop in the results for BERT-multi (NER) and Roberta (NER) for the Corp domain. This drop could be linked to the lower number of named entities in the Corp data set compared to the other corpora. Conversely, BERT-multi (NER) outperformed BERT-multi for Dutch on Equi and Wind test sets. Finally, we observe that BERT-biLSTM-CRF obtained way lower results, suggesting that using BERT embeddings as features for a biLSTM-CRF is less effective than using BERT as a classifier. This is not surprising as we already know that biLSTM models often require a massive amount of training data, which is definitely not the case in our data sets.

Based on the previous experiment's best performing models, we report in Table 3 the obtained results by varying the training data sets. Training on the English Wind corpus and testing on Corp, for instance, shows better results (+2.64% with RoBERTa) than combining Wind with Equi. On the contrary, better results are

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

	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**

	French		
Test	Train		
Corp	Equi	Wind	Equi + Wind
CamemBERT	34.87	39.07	**40.25**
CamemBERT (NER)	18.71	**19.99**	19.13
BERT-biLSTM-CRF	**18.04**	16.66	17.40
Equi	Corp	Wind	Corp + Wind
CamemBERT	32.73	32.22	**35.53**
CamemBERT (NER)	**34.09**	31.62	31.40
BERT-biLSTM-CRF	18.96	19.21	**19.76**
Wind	Corp	Equi	Corp + Equi
CamemBERT	26.78	24.57	**25.71**
CamemBERT (NER)	**30.35**	22.54	24.28
BERT-biLSTM-CRF	18.40	17.97	**18.79**

Table 3: Domain transfer results (F1%) of ATE on the English and French data sets (for the sake of clarity, Dutch results are put in the supplementary material).

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**
	nl	en+fr+nl	nl	en+fr+nl
corp (nl)	-	-	29.91	**34.09**
equi (nl)	37.52	**41.56**	-	-
wind (nl)	26.11	**28.03**	26.11	**28.44**
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**
	nl	en+fr+nl	nl	en+fr+nl
corp (nl)	35.52	**38.89**	35.21	**35.3**
equi (nl)	**41.42**	41.36	41.10	39.99
wind (nl)	-	-	25.52	**29.79**

Table 4: Language transfer learning results (F1%) on the validation sets. All the experiments were based on bert-base-multilingual model (BERT-multi) .

obtained for BERT (NER) and BERT-biLSTM-CRF when using both Equi and Wind corpora. However, these observations do not hold for all the experiments and vary depending on the chosen data sets and the used models. Overall, contrasting the results suggest that transfer learning for ATE is domain sensitive and that a careful choice of training data sets is undoubtedly needed for better performance.

5.2.2. Language Transfer Evaluation

The language transfer experiment, shown in Table 4, questions the usefulness of adding other languages for fine-tuning. Hence, given the English corruption test set for instance (annotated as corp (en)), we fine-tune the multilingual BERT model (BERT-multi) on each of the two remaining data sets separately (equi or wind) as well as their combination (noted as All), which simultaneously represent cross-lingual and cross-domain evaluation. As can be seen, adding other languages often increases the results, and when it is not the case, the F1 score difference remains very low. Improvements are more noticeable for the Dutch test sets, where we obtained an F1 score of 41.56% versus F1 of 37.52% without the use of multiple languages when compared to a fine-tuning using the Corp data set only. Nonetheless, we notice that, in some cases, adding more languages decreases the performance, as it can be noticed when testing on Wind (en) and fine-tuning on Corp. Indeed, the F1 score dropped from 34.04% to 32.89% when adding French and Dutch languages to English (en+fr+nl). Finally, some improvements can also be observed when using both cross-lingual and cross-domain combination.

5.3. Test Set Experiments

5.3.1. Baselines

In contrast to our proposed approaches, we implemented a feature-based baseline that exploits eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) for classification [6]. We also implemented a Vanilla-biLSTM-

[6] More than 25 features were tested. We retained the best-performing ones: Frequency, TF-IDF, C-value, Termhood, and Specificity.

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

CRF to contrast with our BERT-biLSTM-CRF[7]. The results are reported in Table 5, alongside the results of the participating teams[8] on the TermEval shared task (Rigouts Terryn et al., 2020). The teams performed various approaches: e-Termino used a rule-based approach, RACAI used a combination of statistical approaches based on several features, NLPLab used a biLSTM based on pre-trained Glove and NYU used chunking, statistical and search-based scores. Finally, the best performing team (TALN-LS2N) used BERT as a binary classifier which is similar to our approach from a data representation perspective but differs from it did not exploit cross-lingual transfer, and 2) they only performed a combination of all training data for fine-tuning. On the contrary, we investigate several domain combinations and show that more cross-domain data does not necessarily lead to better performance.

5.3.2. Results

Table 5 shows the obtained results on the heart failure test set. We first report the results for our cross-domain feature-based approach, noted Features (en) for English and its cross-lingual version, noted Features (enfrnl) for all languages as well as our Vanilla-biLSTM-CRF baseline. We also list the official results of the shared task for all the participating teams. Finally, we report the results of our proposed approaches using BERT, BERT (NER), as well as BERT-biLSTM-CRF.

Overall, our achieved results on the test sets follow the observed results obtained on the development sets. Indeed, BERT for English, CamemBERT for French, and BERT-multi for Dutch obtained the best F1 scores, while BERT (NER) obtained the best precision. Our methods outperformed the baselines as well as the participating teams with a large margin for Dutch. Using cross-lingual transfer learning did not improve the results compared to BERT-multi (en) and BERT-multi (enfrnl). It also noticeably weakened the results for French. This can also be observed for the Features (enfrnl) baseline, where the results drop significantly. Finally, even if the BERT-BiLSTM-CRF is behind the classifier based-BERT model, it shows, however, noticeably better results than the Vanilla-biLSTM CRF, which highlights the substantial contribution of BERT embeddings as input features for the biLSTM for this task.

6. Analysis and Discussion

Table 6 represents the proportion of heart failure term types for BERT for different domain transfer configurations. Regardless of the training sets, we see that specific and common terms are the most captured elements[9]. However, the overall scores and the high percentage of false positives indicate that there is still a big room for improvement. It should also be remarked that the named entities or out-of-domain terms categories are less represented in the training sets; making them harder to spot upon testing. The models' training was

[7] It should be noted that this work focuses on BERT-based strategies, and these baselines were implemented solely for comparison's sake.

[8] Not all the teams submitted outputs for all languages.

[9] See Table 5 in the supplementary material for details.

	Domains						
BERT	Corp	Equi	Wind	Corp+Equi	Corp+Wind	Equi+Wind	Corp+Equi+Wind
Specific Terms	**81.67**	**76.47**	**82.58**	**76.36**	**82.04**	**77.48**	73.44
Common Terms	81.19	71.15	69.27	73.98	80.56	65.83	**74.60**
Named Entities	59.73	53.09	63.71	61.06	58.84	58.40	57.96
OOD Terms	68.15	42.03	71.33	63.05	64.96	42.03	33.75

Table 6: Percentage of each term type in the true positives for BERT output list on the English heart failure test set with respect to transfer learning on different domains. The information about the other models can be found in the supplementary material.

	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 7: 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.

carried out directly on the terms, but we could go a step further by selecting a finer granularity (specific, common, out-off-domain, named entity). However, these classes were extremely unbalanced. While it was not the concern of this work, future work will adapt the models to exploiting each class's particularities. Finally, many classic ATE approaches apply post-filtering to the collected term-like sequences. In the same way, we applied to our models outputs three main automatic filtering techniques that are: pattern filtering where all terms that fit nominal patterns were kept; C-Value and the Specificity measures where only the first half of the ordered lists were kept. Table 7 reports the obtained results of each filtering technique. We first see that pattern filtering did not improve the results, suggesting that BERT's outputs are morpho-syntactically correct. We also see that C-Value filtering significantly improves the results for each model, while Specificity filtering F1 scores are less important for French and drastically weaker for Dutch. However, for English, the F1 is improved by 7.47%.

Among the three proposed BERT-based approaches, the binary classification system obtained the best results. Varying the data sets revealed the effectiveness of cross-domain BERT fine-tuning and lent support the assumption that terms share cross-domain marked-contexts captured by BERT. This finding opens a new interesting line of research for ATE since the lack of training data can be reduced by using other specialized domains. Nonetheless, the results also revealed that using several training data sets does not necessarily guarantee better performance. We still have an unanswered question: how to choose the most appropriate training data sets? Regardless of the empirical improvements, it is worth highlighting the simplicity of our strategies, since neither feature engineering nor pattern design is needed. Besides that, 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.
Our second strategy considered ATE as a sequence labeling problem. While named entities are present in all corpora, we expected better performance from BERT (NER). However, only a small proportion of named entities was extracted[10]. Overall, and despite lower F1 scores, BERT (NER) showed acceptable and sometimes competitive results compared to other methods. It also obtained the best precision over all the experiments. This is more noticeable for Dutch BERT (NER) as it outperformed the BERT classifier. Finally, the larger proportion of extracted specific terms can be explained by their more extensive representation in the training sets or by the fact that they may share more definite marked contexts. We show that BERT can be used for ATE, replacing classic methods to identify term-like units more accurately, and it can be followed by standard post-filtering techniques, further improving its results.

7. Conclusion

In this paper, 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 do-

[10] See Table 5 in the supplementary material for details.

mains and languages, opening a new promising direction for ATE.

8. Bibliographical References

Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1994). What is a term? the semi-automatic extraction of terms from text. *Translation Studies: An Interdiscipline*, pages 267–278. John Benjamins.

Amjadian, E., Inkpen, D., Paribakht, T., and Faez, F. (2016). Local-global vectors to improve unigram terminology extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 2–11, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Ananiadou, S. (1994). A methodology for automatic term recognition. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In Tapio Salakoski, et al., editors, *Advances in Natural Language Processing*, pages 380–387, Berlin, Heidelberg. Springer Berlin Heidelberg.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 355–362, Edinburgh, Scotland, UK.

Basili, R., De Rossi, G., and Pazienza, M. T. (1997). Inducing terminology for lexical acquisition. In *Second Conference on Empirical Methods in Natural Language Processing*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING 1992 Volume 3: The 15th International Conference on Computational Linguistics*.

Burgos Herera, D. A. (2014). *Towards an Image-Term Co-occurrence Model for Multilingual Terminolog Alignment and Cross-Language Image Indexing*. Ph.D. thesis, Universitat Pompeu Fabra.

Castellví, C. M., Estopa, R., and Palatresi, J. (2001). Automatic term detection: A review of current systems. In *Recent Advances in Computational Terminology*, page 53–88. Amster-dam/Philadelphia. John Benjamins, 07.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Cohen, W. (1996). Learning trees and rules with set-valued features. *Proceedings of AAAI'96, IAAI'96*, 08.

Condamines, A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*, 8(1):141–162.

Conrado, M., Pardo, T., and Rezende, S. (2013). A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 16–23, Atlanta, Georgia, June. Association for Computational Linguistics.

Daille, B. (2017). *Term variation in specialised corpora: characterisation, automatic discovery and applications.*, volume 19 of *Terminology and Lexicography Research and Practice series*. John Benjamins Publishing Company, August.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duda, R. O. and Hart, P. E. (1973). Pattern classification and scene analysis. In *A Wiley-Interscience publication*.

El Boukkouri, H., Ferret, O., Lavergne, T., and Zweigenbaum, P. (2019). Embedding strategies for specialized domains: Application to clinical entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 295–301, Florence, Italy, July. Association for Computational Linguistics.

Frantzi, K. and Ananiadou, S. (2000). The C-Value/NC-Value domain independent method for multi-word term extractionAlgorithmsBased on Error Counting. *Journal of Natural Language Processing*, 6(3):145–179.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 152–161, Avignon, France.

Han, X., Xu, L., and Qiao, F. (2018). Cnn-bilstm-crf model for term extraction in chinese corpus. In Xiaofeng Meng, et al., editors, *Web Information Systems and Applications*, pages 267–274, Cham. Springer International Publishing.

Hätty, A. and Schulte im Walde, S. (2018). Fine-grained termhood prediction for German compound terms using neural networks. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 62–73, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Hätty, A., Dorna, M., and Schulte im Walde, S. (2017). Evaluating the reliability and interaction of recursively used feature classes for terminology extraction. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–121, Valencia, Spain, April. Association for Computational Linguistics.

Hatzivassiloglou V, Duboué PA, R. A. (2001). Disambiguating proteins, genes, and rna in text: a machine learning approach. *Bioinformatics*, 17(1):97–106.

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. Cambridge: MIT Press.

Jia, C., Liang, X., and Zhang, Y. (2019). Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy, July. Association for Computational Linguistics.

Judea, A., Schütze, H., and Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 290–300, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289.

Liu, J., Morin, E., and Saldarriaga, P. (2018). Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2855–2866, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lv, X., Guan, Y., and Deng, B. (2014). Transfer learning based clinical concept extraction on data from multiple sources. *Journal of Biomedical Informatics*, 52:55 – 64. Special Section: Methods in Clinical Research Informatics.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv e-prints*, Nov.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins.

Moore, R. C. and Lewis, W. D. (2010). Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 220–224, Uppsala, Sweden.

Parcheta, Z., Sanchis-Trilles, G., and Casacuberta, F. (2018). Data selection for nmt using infrequent n-gram recovery. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT'18)*, pages 219–227, Alacant, Spain.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Rahimi, A., Li, Y., and Cohn, T. (2019). Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy, July. Association for Computational Linguistics.

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. (2019). Visualizing and measuring the geometry of bert. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 8594–8603. Curran Associates, Inc.

Rigouts Terryn, A., Hoste, V., Drouin, P., and Lefever, E. (2020). Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France, May. European Language Resources Association.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works.

Šajatović, A., Buljan, M., Šnajder, J., and Dalbelo Bašić, B. (2019). Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154, Florence, Italy, August. Association for Computational Linguistics.

Schäfer, J., Rösiger, I., Heid, U., and Dorna, M. (2015). Evaluating noise reduction strategies for terminology extraction. In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence (TIA 2015), Granada, Spain.*, pages 123–131, November.

Schuster, S., Gupta, S., Shah, R., and Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Takeuchi, K. and Collier, N. (2005). Bio-medical entity extraction using support vector machines. *Artifi-*

cial Intelligence in Medicine, 33(2):125 – 137. Information Extraction and Summarization from Medical Documents.

Terryn, A. R., Hoste, V., and Lefever, E. (2018). A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, R., Liu, W., and McDonald, C. (2016). Featureless domain-specific term extraction with minimal labelled data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 103–112, Melbourne, Australia, December.

Zhang, X. and Fang, A. C. (2010). An ate system based on probabilistic relations between terms and syntactic functions. In *Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT)*.

9. Appendix

Additional information is presented in this supplementary material. We first present models settings and the detailed statistics of the evaluation term lists of each domain and for each language. Then, we present additional BERT models evaluations through figures and extended tables.

9.1. Settings

9.1.1. BERT Settings

For the fine-tuning phase of BERT, we used the simpletransformers [11] library and its default parameters setting. For BERT as a binary classifier, we used only 1 epoch for fine-tuning, while for BERT as NER task, we used 4 epochs. Nonetheless, no significant impact on the results has been observed when using more epochs.

9.1.2. BiLSTM-CRF Hyperparameters

As for our biLSTM-CRF, we choose 150 hidden units for the BiLSTM, and 100 in the ReLU activated hidden layer, depending on the validation set performance. In a similar fashion to others who used word embeddings like GloVe (Pennington et al., 2014) or fastText (Bojanowski et al., 2017) as their input, we use the last layer of pre-trained bert-base-cased (EN), CamemBERT (FR) and bert-base-multilingual-cased-dutch (NL) as BERT embeddings. The model was built using Keras and trained with a batch size of 64, for 10 epochs.

9.1.3. Infrastructure and Average Runtime

Our models are trained using a GeForce RTX 2080 GPU, for about 6 minutes per run for BERT, 4 minutes for BERT (NER) and 20 minutes for the BiLSTM-CRF.

Domain		# ST	# CT	# NE	# OOD	Total
corruption	en	278	644	248	6	1174
	fr	300	678	236	5	1217
	nl	310	731	249	6	1295
dressage	en	780	309	421	71	1575
	fr	705	238	221	26	1183
	nl	1026	333	153	41	1546
wind energy	en	781	296	444	14	1534
	fr	444	308	195	21	968
	nl	577	342	305	21	1245
heart failure	en	1885	320	228	158	2585
	fr	1714	505	147	59	2423
	nl	1561	450	182	66	2257

Table 8: Number of unique terms per corpus in three languages. Numbers are given for each type of terms: ST for Specific Terms, CT for Common Terms, NE for Named Entities, and OOD for Out of Domain terms. Total covers all the terms.

9.2. Results

In this section, we provide extensive and more comprehensive tables and experiments that we were unable to add to the main paper due to a lack of space. In Table 9, we show ATE scores for each validation corpus. In addition to the BERT models presented in the paper, we also contrast our results with extra models. Tables 10 and 11 show cross-domain and cross-lingual results in more details, on the validation sets as well as on the test set. Table 1 shows for each term type in the output list, its proportion in the gold standard, and corpus used for training for the heart failure English test set. These precise percentages can be found in Table 12. We can see that even if a term type is sometimes less represented in the training corpus, BERT (base and multi) often do a good job retrieving it. Figure 2 suggests that transfer learning for ATE is domain sensitive and that a careful choice of training data sets is undoubtedly needed for better performance since more training data does not always rhyme with better performances.

[11] https://github.com/ThilinaRajapakse/simpletransformers

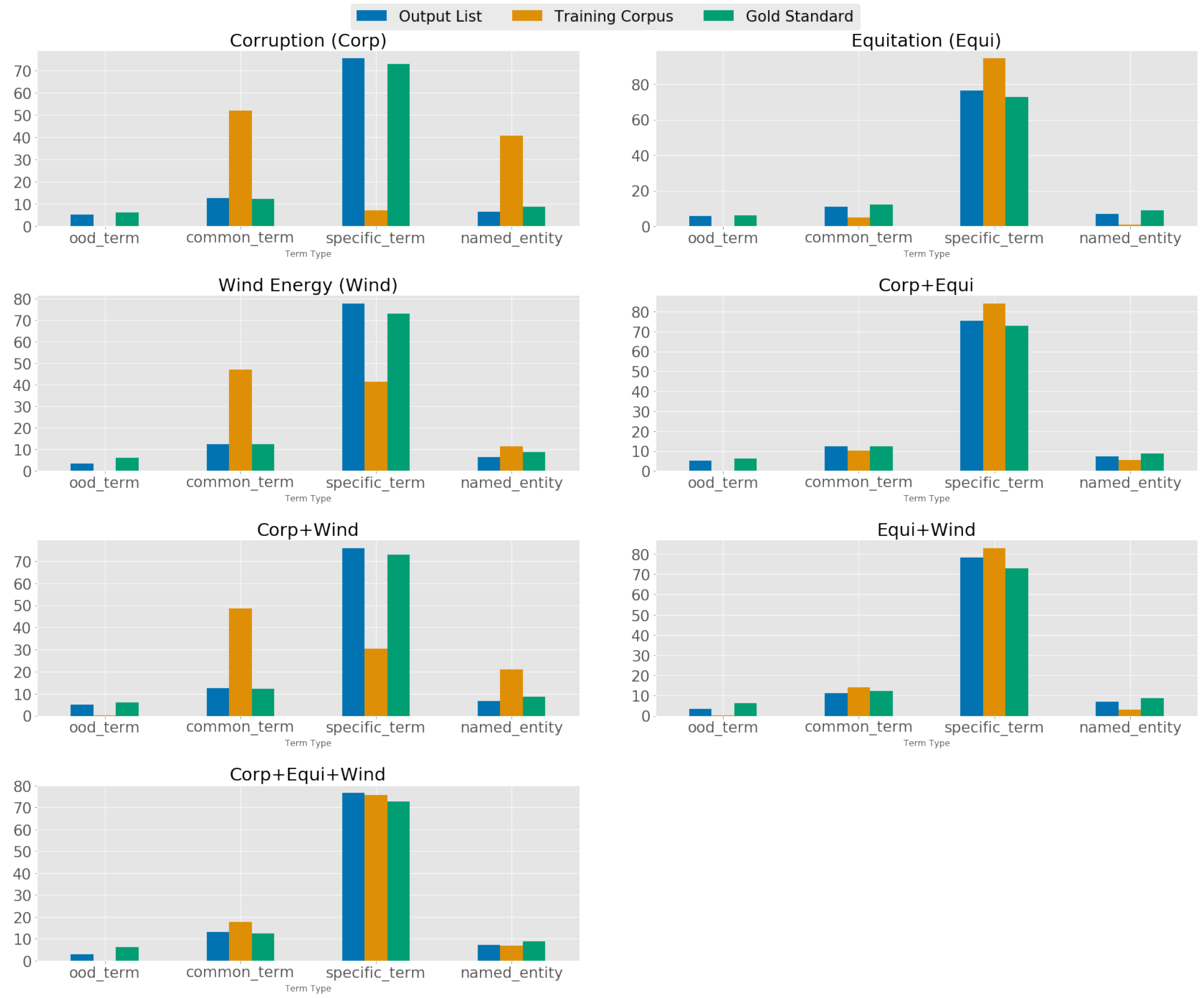

Figure 1: Proportion (%) of each term type in the output list, the gold standard, and corpus used for training for the heart failure English test set. The proportion of out-of-domain terms in the training corpus is so small, it often looks like there are none. These precise percentages can be found in Table12.

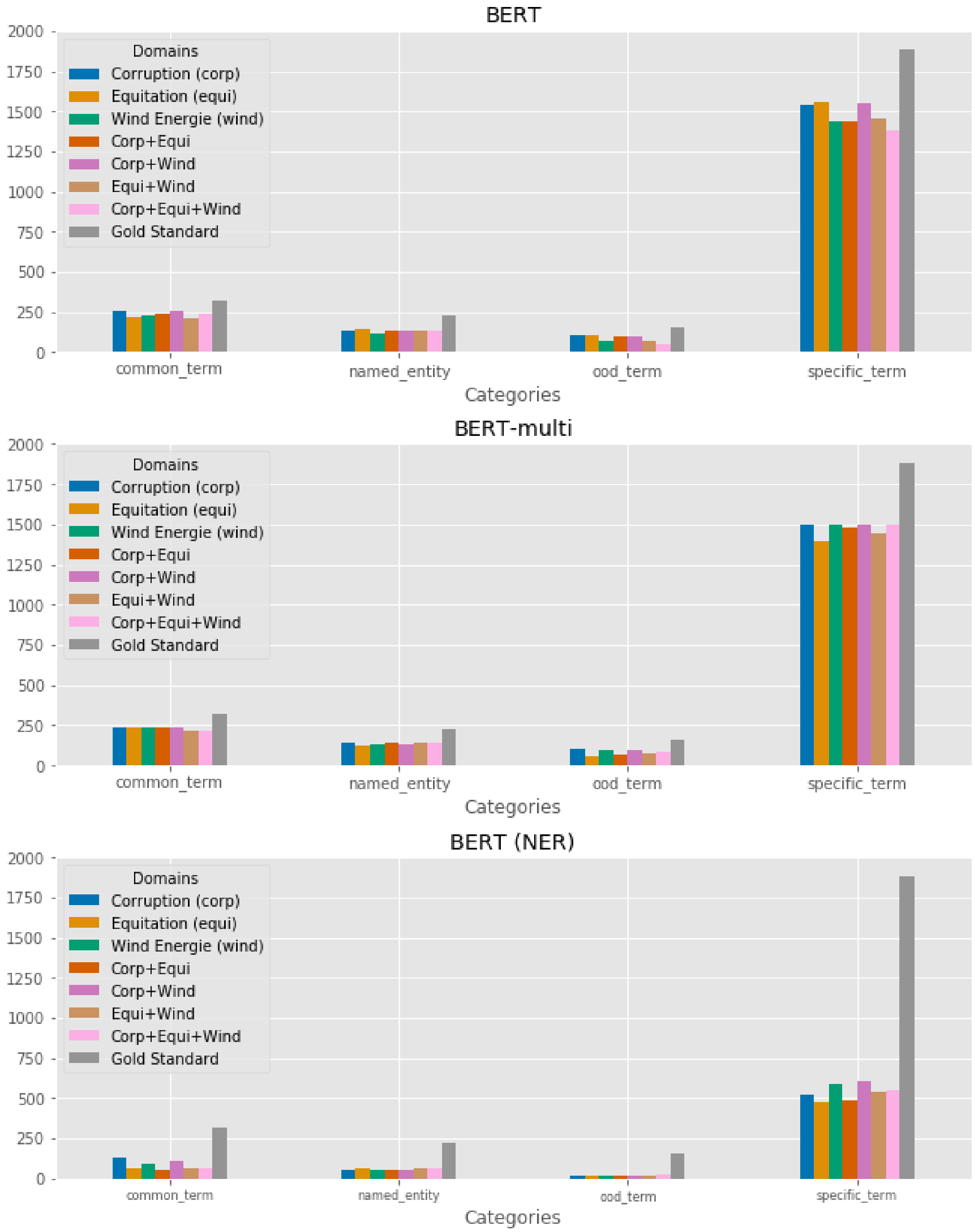

Figure 2: Illustration of BERT, BERT-multi and BERT (NER) models outputs, with regards to term types for the heart failure test sets and for all combinations of the training sets.

	Corp			**Equi**			**Wind**		
English	P	R	F1	P	R	F1	P	R	F1
BERT	29.46	49.45	36.79	30.76	71.61	42.94	21.87	74.03	33.73
distilbert-base-cased	29.07	51.4	37.08	31.09	72.15	**43.42**	22.32	73.06	34.16
BERT-multi	25.47	58.21	35.25	27.87	71.94	40.15	22.92	74.51	**34.94**
distilbert-base-multi-cased	24.54	59.29	34.68	27.46	72.15	39.76	22.56	74.16	34.56
bert-large-cased	28.21	52.53	36.42	31.11	70.5	43.02	21.86	74.87	33.63
RoBERTa	26.85	**62.53**	**37.33**	29.17	**75.59**	42.03	20.93	**77.08**	32.91
BERT (NER)	**52.92**	22.46	31.48	**54.45**	32.63	40.81	**35.48**	23.66	28.39
BERT-multi (NER)	34.09	11.51	17.21	13.41	14.67	14.01	11.14	18.58	13.93
RoBERTa (NER)	31.91	10.22	15.48	49.39	28.44	36.11	32.48	23.08	26.98
distilroberta-base	26.79	55.5	35.89	27.29	73.52	39.77	19.99	75.12	31.56
BERT-biLSTM-CRF	22.61	18.98	20.63	20.65	16.03	18.05	21.04	23.31	22.11
French	P	R	F1	P	R	F1	P	R	F1
BERT-multi	27.53	53.03	35.91	21.17	60.77	31.35	13.82	68.99	22.97
distilbert-base-multi-cased	25.11	56.06	34.62	19.53	62.47	29.75	13.21	67.9	22.09
CamemBERT	29.73	**62.61**	**40.25**	24.38	**65.89**	**35.53**	15.63	**72.51**	**25.71**
BERT-multi (NER)	39.17	11.59	17.89	42.83	20.20	27.45	23.22	25.93	24.5
CamemBERT (NER)	**42.61**	12.33	19.13	**48.01**	23.33	31.40	**24.09**	24.48	24.28
BERT-biLSTM-CRF	18.78	16.21	17.40	18.53	19.08	19.76	17.20	20.72	18.79
Dutch	P	R	F1	P	R	F1	P	R	F1
BERT-multi	23.70	**68.59**	**35.21**	30.39	**63.85**	41.10	15.85	**66.55**	25.52
distilbert-base-multi-cased	21.69	69.82	33.09	28.32	64.99	39.39	15.61	65.54	25.19
BERT-multi (NER)	**40.93**	15.68	22.67	**65.25**	32.79	**43.65**	**37.41**	38.31	**37.85**
BERT-biLSTM-CRF	20.51	21.75	21.11	21.17	18.90	19.97	21.70	15.28	17.93

Table 9: ATE scores (%) in terms of Precision (P), Recall (R) and F1-score (F1). For each validation corpus, the two remaining ones were used for fine-tuning. We contrast 3 usages of BERT : as a binary classifier, BERT-NER and BERT embeddings used with a biLSTM-CRF. We also contrast BERT's multi-lingual version as well as RoBERTa for English and Camembert for French (This table is similar to Table 2 of the main paper with extra models such distilbert and Roberta-large).

	English		
Test	Train		
Corp	Equi	Wind	Equi+Wind
BERT	35.33	**38.01**	36.79
BERT-multi	35.09	**35.56**	35.25
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
BERT	39.93	**44.21**	42.94
BERT-multi	38.93	**40.54**	40.15
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
BERT	33.45	33.34	**33.73**
BERT-multi	34.04	**35.59**	34.94
RoBERTa	**36.16**	33.77	32.91
BERT (NER)	**30.58**	26.89	28.39
BERT-biLSTM-CRF	19.99	20.64	**22.11**

	French		
Test	Train		
Corp	Equi	Wind	Equi + Wind
BERT-multi	34.84	35.71	**35.91**
CamemBERT	34.87	39.07	**40.25**
CamemBERT (NER)	18.71	**19.99**	19.13
BERT-biLSTM-CRF	**18.04**	16.66	17.40
Equi	Corp	Wind	Corp + Wind
BERT-multi	27.64	**32.88**	31.35
CamemBERT	32.73	32.22	**35.53**
CamemBERT (NER)	**34.09**	31.62	31.4
BERT-biLSTM-CRF	18.96	19.21	**19.76**
Wind	Corp	Equi	Corp + Equi
BERT-multi	**24.12**	23.37	22.97
CamemBERT	**26.78**	24.57	25.71
CamemBERT (NER)	**30.35**	22.54	24.28
BERT-biLSTM-CRF	18.40	17.97	**18.79**

	Dutch		
Test	Train		
Corp	Equi	Wind	Equi + Wind
BERT-multi	29.91	**35.52**	35.21
BERT-multi (NER)	**32.42**	27.62	22.67
BERT-biLSTM-CRF	19.09	20.22	**21.11**
Equi	Corp	Wind	Corp + Wind
BERT-multi	37.52	**41.42**	41.1
BERT-multi (NER)	**46.36**	41.51	43.65
BERT-biLSTM-CRF	19.14	**21.20**	19.97
Wind	Corp	Equi	Corp + Equi
BERT-multi	**26.11**	24.8	25.52
BERT-multi (NER)	36.74	35.86	**37.85**
BERT-biLSTM-CRF	16.29	**19.83**	17.93

Table 10: Domain transfer results (F1%) of ATE on the English, French and Dutch data sets (This table is the extension of Table 3 of the main paper

	English						
Test	Train						
hf	Corp	Equi	Wind	corp+equi	corp+Wind	equi+wind	All
BERT	**43.25**	**45.10**	**46.21**	**45.40**	**45.35**	**48.21**	**46.42**
BERT-multi (en)	40.62	43.39	43.70	43.39	42.91	45.22	45.77
BERT-multi (enfrnl)	40.92	43.81	44.71	43.62	44.11	46.29	45.37
RoBERTa	41.75	44.87	44.6	44.15	43.58	44.71	43.35
BERT-biLSTM-CRF	24.28	26.57	28.27	25.22	26.92	29.93	29.26

Table 11: Heart failure test results (F1%) using cross-domain and cross-lingual (enfrnl) training. Equi+Wind means that we fine-tuned BERT on the combination of equitation and wind energy data sets. BERT-multi (enfrnl) means that the training was conducted using three languages.

	Domains						
Occ. in Training Set	Corp	Equi	Wind	Corp+Equi	Corp+Wind	Equi+Wind	Corp+Equi+Wind
Specific Terms	7.11	**94.39**	41.50	**84.04**	30.30	**82.86**	**75.65**
Common Terms	**52.00**	4.73	**46.95**	10.34	**48.59**	13.93	17.56
Named Entities	40.85	0.82	11.40	5.56	20.98	3.12	6.71
OOD Terms	0.02	0.04	0.14	0.04	0.10	0.06	0.06
BERT	Corp	Equi	Wind	Corp+Equi	Corp+Wind	Equi+Wind	Corp+Equi+Wind
Specific Terms	**81.67**	**76.47**	**82.58**	**76.36**	**82.04**	**77.48**	73.44
Common Terms	81.19	71.15	69.27	73.98	80.56	65.83	**74.60**
Named Entities	59.73	53.09	63.71	61.06	58.84	58.40	57.96
OOD Terms	68.15	42.03	71.33	63.05	64.96	42.03	33.75
BERT-NER	Corp	Wind	Equi	Corp+Equi	Corp+Wind	Equi+Wind	Corp+Equi+Wind
Specific Terms	27.93	**31.01**	25.33	**25.65**	32.12	**28.67**	**29.47**
Common Terms	**40.12**	28.21	18.80	17.86	**33.85**	21.36	21.31
Named Entities	23.89	25.22	**28.31**	25.22	25.66	26.54	27.87
OOD Terms	12.10	12.10	10.82	11.46	12.10	12.10	15.28
BERT-multi	Corp	Equi	Wind	Corp+Equi	Corp+Wind	Equi+Wind	Corp+Equi+Wind
Specific Terms	**79.60**	**79.39**	73.92	**78.81**	**79.81**	**76.84**	**79.76**
Common Terms	73.66	73.04	**74.29**	73.66	73.04	66.77	68.65
Named Entities	61.94	59.29	53.53	60.61	56.63	50.31	61.06
OOD Terms	64.96	59.23	38.21	44.58	61.78	50.31	52.22

Table 12: Percentage of each term type in the true positives list for BERT, BERT-NER and BERT-multi outputs on the English heart failure test set in respect to transfer learning on different domains. The top part of the table shows (in %) how well each type is represented in the training data.