How different is different? Systematically identifying distribution shifts and their impacts in NER datasets

Xue Li (x.li3@uva.nl)
University of Amsterdam
Paul Groth (p.t.groth@uva.nl)
University of Amsterdam

Research Article

Keywords:

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

How different is different? Systematically identifying distribution shifts and their impacts in NER datasets

Xue Li[1] and Paul Groth[1]

[1]Informatics Institute, University of Amsterdam, Science Park, Amsterdam, 1098 XH, Netherlands.

Contributing authors: x.li3@uva.nl; p.t.groth@uva.nl;

Abstract

When processing natural language, we are frequently confronted with the problem of distribution shift. For example, using a model trained on a news corpus to subsequently process legal text exhibits reduced performance. While this problem is well-known, to this point, there has not been a systematic study of detecting shifts and investigating the impact shifts have on model performance for NLP tasks. Therefore, in this paper, we detect and measure two types of distribution shift, across three different representations, for 12 benchmark Named Entity Recognition datasets. We show that both input shift and label shift can lead to dramatic performance degradation. For example, fine-tuning on a wide spectrum dataset (Ontonotes) and testing on an email dataset (Cerec) that shares labels leads to a 63-points drop in F1 performance. Overall, our results indicate that the measurement of distribution shift can provide guidance to the amount of data needed for fine-tuning and whether or not a model can be used "off-the-shelf" without subsequent fine-tuning. Finally, our results show that shift measurement can play an important role in NLP model pipeline definition.

Keywords: Distribution Shift, Named Entity Recognition

1 Introduction

Differences between training and inference distributions are a common occurrence in the field of Natural Language Processing (NLP). This difference can be observed, for instance, when the input data undergoes changes over time or when a model is employed on data from a new domain. This is known as *distribution shift* (Quionero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009). Consider the following example from Named Entity Recognition (NER):

Fig. 1 An example of distribution shift in Named Entity Recognition.

The example shows two phenomena. First, the entities in the training example tend to be relatively well-known entities (e.g. EU), which are highly probable to be present in the data sources utilized by pre-trained language models (Devlin, Chang, Lee, & Toutanova, 2019; Lee et al., 2019) that are widely used for NLP tasks. Conversely, the entities in the inference example are unique to a particular domain (e.g. IETF). Second, the labels in the training example differ from the ones in the inference example. This is because entities from different domains possess different types, such as "Organization" versus "Protocol", and variations in labelling for the same type, such as "Location" and "Place". These phenomena embody two common shifts in NLP: input distribution shifts and label distribution shifts (Quionero-Candela et al., 2009). While identifying changes of types for the same entity mentions across domains is infeasible, in this work we primarily focus on category shift in the label space (Lekhtman, Ziser, & Reichart, 2021). Despite the substantial body of literature on measuring domain similarity(Dai, Karimi, Hachey, & Paris, 2019), detecting *when* a shift occurs remains a challenging task in the field. This task is known as *shift detection*.

A key area where shift detection is useful is *domain adaptation*, which aims at adapting a model in the presence of distribution shifts (Csurka, 2017). One of the common supervised approaches to achieve adaptation is fine-tuning deep neural networks (Vu et al., 2020). While fine-tuning can be effective, there is still a cost, such as determining the required amount of additional data for fine-tuning. To inform this decision, shift detection methods are frequently employed in other areas that employ machine learning (Cobb & Van Looveren, 2022; Kulinski, Bagchi, & Inouye, 2020; Rabanser, Günnemann, & Lipton, 2019). This work frequently adopts statistical hypothesis testing as an underlying principled approach to the problem (Cobb & Van Looveren, 2022; Rabanser et al., 2019). Statistical two-sample testing is a methodology for determining whether the distribution of the training data p is equivalent to the distribution of the test data q. While this approach has been explored for computer vision tasks involving high-dimensional data, it has seen limited application to NLP

Hence, to better inform these decisions and quantify the potential impact of distribution shifts, this paper undertakes a systematic investigation of shifts across benchmark corpora using statistical tests, which have been widely adopted for shift detection in the context of other machine learning tasks. In this work, we specifically focus on the NER task and detect distribution shifts across 12 different datasets that are representative of various domains. We use word frequency and sentence-level representations to characterize input distributions, and label frequency to characterize label distributions. Appropriate statistical tests are identified for each representation and employed to detect and quantify shifts. We then investigate the impact of domain shift in both the input and label space on performance in the supervised setting. We establish a relationship between the shift distance and the performance degradation. These results provide insights into what statistical test one needs to perform to make such a determination.

Summarizing, the contributions of this paper are as follows:

- The systematic measurement of distribution shift between 12 NER benchmark datasets covering multiple domains.
- Empirical evidence that shifts impact performance.
- Evidence that sentence-based representations provide better information for shift detection.

2 Related Work

Distribution shifts are prominent in real-world applications (Engstrom, Tran, Tsipras, Schmidt, & Madry, 2019; Michel, 2021; Recht, Roelofs, Schmidt, & Shankar, 2019; Specia et al., 2020), leading to growing interest in detecting them for machine learning tasks (Cobb & Van Looveren, 2022; Kulinski et al., 2020; Rabanser et al., 2019).

Shift types

In the broader landscape of machine learning, Wiles et al. (2022) conducted a fine-grained analysis of distribution shifts, classifying them as spurious correlation, low data drift, and unseen data shift. Additionally, they evaluated 19 different methods on both synthetic and real-world datasets for vision tasks.

The Use of Statistical Tests

The use of statistical tests for dataset shift detection was brought to the fore by Rabanser et al. (2019). In their work, they developed a dataset shift detection framework which contains a dimensionality reduction component and a two-sample-testing component. They investigated multiple combinations of methods for each component, and tested on artificially generated covariates and label distribution shifts. Recently, based on two-sample tests for shift

Corpus	Year	Document source	Domain	# Types	Category
GUM	2017	Wiki-family	Various	11	Wiki data
Wikigold	2009	Wikipedia text	Various	4	Wiki data
BTC	2016	Twitter	Mainstream news	3	Informal text
W-NUT17	2017	User-generated text	Various	6	Informal text
CEREC	2005	Informal emails	Work	4	informal text
AnEM	2012	Anatomical text	Anatomy	11	Specific field
i2b2-06	2006	Clinical text	Biomedical	7	Specific field
SEC-Filings	2015	Electronic filings	Finance	4	Specific field
SciERC	2018	Scientific abstracts	Scientific	6	Specific field
Re3d	2018	Documents related to defense and security analysis	Conflict in Syria and Iraq	10	Specific field
CoNLL-03	2003	Reuters news	Mainstream news	4	News
OntoNotes	2007-2012	Magazine, news, web, tele, etc	Various	18	General

Table 1 List of annotated datasets for English NER from different domains.

detection, Cobb and Van Looveren (2022) developed a general drift detection framework borrowing machinery from causal inference. The framework is used to deal with the situation when the inference data are not expected to form an i.i.d. sample from the historical data distribution.

Domain similarity

Within the field of NLP, researchers have explored various methods for measuring domain similarity in the context of domain adaptation including using target vocabulary covered rate and language model perplexity(Dai et al., 2019). However, these methods work well under the assumption that there are sufficient data from the source and target distribution. Therefore, in our work we adopt non-parametric statistical hypothesis testing framework to detect shift without knowing the actual parameters of the population.

Shift detection in NLP

Within NLP, Arora, Huang, and He (2021) focused on out-of-distribution texts and two approaches for detection. Shifts are categorized into *background* shift and *semantic* shift. Model calibration and density estimation are investigated for shift detection across 14 pairs of natural language understanding datasets. Comparing density estimation methods and calibration methods. We investigate different types of shifts than these works.

Given the importance of shift detection, a number of datasets have been developed (Koh et al., 2021; Malinin et al., 2022), however, they are not for the widely used task of NER.

Our work adds to this existing literature. First, we employ widely used labelled NER datasets and compare not only changes in fields (e.g. science to finance) but also changes in text style (e.g. news style text to social media style text). Second, we test the impact of representation choice on shift detection. Lastly, we provide new evidence for the performance impact of distribution shifts on task performance.

3 Methodology

Our methodology consists of the following steps: data collection; representation choice; statistical hypothesis testing and shift impact measurement. For

Corpus	Sample Size	# Types	Entity Types
GUM	3424	11	Organization, Person, Location, Event, Abstract, Object, Time, Substance, Plant, Quantity, Animal
Wikigold	1688	4	Organization, Person, Location, Miscellaneous
BTC	9318	3	Organization, Person, Location
W-NUT17	5591	6	Organization, Person, Location, Group, Product, Creativework
CEREC	2031	4	Organization, Person, Location, Digits
AnEM	4423	11	Multi-tissue_structure, Organism_substance, Organism_subdivision, Organ, Cellular_component, Cell, Immaterial_anatomical_entity, Tissue, Pathological_formation, Anatomical_system, Developing_anatomical_structure
i2b2-06	40280	7	Person, Location, ID, Date, Phone, Age
SEC-Filings	1435	4	Organization, Person, Location, Miscellaneous
SciERC	2687	6	Material, OtherScientificTerm, Generic, Method, Task, Metric
Re3d	2687	10	Organization, Person, Location, Temporal, Nationality, Quantity, Weapon, Money, MilitaryPlatform, DocumentReference
CoNLL-03	17350	4	Organization, Person, Location, Miscellaneous
OntoNotes	17760	18	Organization, Location, Person, Work_of_Art, Cardinal, Event, NORP, Date, FAC, Quantity, Ordinal, Time, Product, Percent, Money, Law, Language

Table 2 List of NER datasets with corresponding entity types. Sample size is shown in number of sentences.

data collection, we acquire datasets from different domains. Domains are characterized by their language usage arising from the style employed to the use of language particular to given field usage. For all datasets, both the space of input text and the space of labels are considered. In terms of representations, two types of representations are used for the input and one for labels. Statistical hypothesis testing appropriate for each representation is used to detect distribution shifts. The calculated statistics are then used to measure the extent of a shift. Lastly, the impact of each shift on model performance is ascertained. We now walk through each of these steps in detail.

3.1 Data collection

We collected 12 datasets from different domains covering news, social media, encyclopedic content, finance, science, emails, and business. Table 1 shows the list of datasets with the published year, document source, domains and the number of entity types. Table 2 shows the list of datasets and their entity types. Every dataset is treated as both source dataset and target dataset, resulting in 78 pairs of datasets. We group the datasets into five categories, which we now describe in-turn.

Wiki data

GUM (Zeldes, 2017) (the Georgetown University Multilayer Corpus) is collected and expanded as part of the curriculum of a course. The current corpus contains texts from public wikis (e.g. Wikinews, Wikivoyage, wikiHow, Wikipedia) as well as social media sites (e.g. Reddit, Youtube). Example types include *event*, *time*, *animal* and *abstract*. **Wikigold** (Balasuriya, Ringland, Nothman, Murphy, & Curran, 2009) is a gold-standard NER dataset sourced from Wikipedia. Wikigold has standard types such as *person* and *organization*.

Informal text

Formal texts such as in news and Wikipedia are normally verified by multiple people sometimes even experts. Hence, the majority of text has correct grammar and spelling. In comparison, user-generated informal data such as social media texts, often contain less formal language usage characterized by slang, poor grammar, misspellings, the use of satire, etc. **BTC** (Derczynski, Bontcheva, & Roberts, 2016)(Broad Twitter Corpus) is a NER dataset where the source data is from Twitter that not only has tweets on general topics but also on specific topics such as disasters. BTC includes 3 types: *person*, *location* and *organization*. **WNUT17** (Derczynski, Nichols, van Erp, & Limsopatham, 2017) is a NER dataset where the text sources are Reddit, Twitter, YouTube and StackExchange comments. WNUT17 focuses especially on emerging and rare entities. The dataset contains 6 types, including *creative*, *corporation* and *product* besides common types. **CEREC** (Dakle & Moldovan, 2020) is a large-scale corpus for entity resolution in email conversations. The emails are taken from the first large public corpus the Enron Email Corpus (Klimt & Yang, 2004) which contains emails of 150 employees of the Enron Corporation. Cerec contains standard types such as *person* and *digits* type.

Specific fields

AnEm (Ohta, Pyysalo, Tsujii, & Ananiadou, 2012) is a corpus annotated with species-independent anatomical entity mentions. The texts are from academic papers from the biomedical scientific literature. AnEm contains 11 domain-specific types such as *organ*, *cell* and *organism_substance*. **i2b2** (Uzuner, Luo, & Szolovits, 2007) is a corpus that contains unstructured clinical notes from the Research Patient Data Registry at Partners Healthcare. The dataset consists of 8 types such as *hospital*, *phone* and *doctor*. **SEC-filings** (Salinas Alvarado, Verspoor, & Baldwin, 2015) (U.S. Security and Exchange Commission filings) is a randomly selected and manually annotated finance dataset. The texts are from public-domain financial reports. The dataset includes standard types from CoNLL, i.e. *organization*, *person*, *location* and *misc*. **SciERC** (Luan, He, Ostendorf, & Hajishirzi, 2018) is a dataset that includes annotations for scientific entities in 500 scientific abstracts from AI conferences and workshop proceedings. The dataset focuses on scientific related types including *material*, *method* and *task*. **Re3d** (Dstl & Laboratory, n.d.) was constructed from documents that are relevant to the defence and security analysis domain, specifically, focusing on the topic of the conflict in Syria and Iraq. It includes domain-specific types such as *weapons* and *military platform*.

News

CoNLL-03 (Tjong Kim Sang & De Meulder, 2003)[1] is a dataset where the texts are taken from the Reuters news stories from 1996 to 1997. It contains the standard types including *person*, *location*, *organization* and *misc*.

[1] We use only the English data.

General

OntoNotes (Weischedel, Hovy, Marcus, & Palmer, 2017)[2] is a large annotated corpus that consists of various genres of texts including news, conversational telephone speech, weblogs, newsgroups, broadcast, and talk shows). OntoNotes include a large variety of types (18) including common types and less common ones such as *money* and *percent*.

Even though the datasets are grouped into five categories, there is still overlap. Wiki-based datasets and OntoNotes or CoNLL belong to different categories, but they might share many similar general entities. This is because common entities in the news are highly likely to have Wikipedia pages. Intuitively, the "similarity" between datasets in the wiki group should be larger. Conversely, the "similarity" between the domain-specific financial dataset SEC and the news dataset CoNLL should be smaller. We introduce methods to statistically quantify the distance between datasets in the following sections.

For all datasets, we preprocess them as follows. Duplicates are removed to prevent overfitting. Labels are unified across datasets. Different datasets use different labels to refer to the same type. Hence, to better compare performance, we unify the labels with the same semantic meanings. For example, 'person' and 'PER' will be unified under the same label.

3.2 Shift detection and measurement

We use statistical testing to determine and measure shifts between datasets. Formally, given a labeled source domain data $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)\} \sim p$ and labeled target domain data $\{(\boldsymbol{x'}_1, \boldsymbol{y'}_1), ..., (\boldsymbol{x'}_n, \boldsymbol{y'}_n)\} \sim q$, shift detection determines whether p equals q. The null hypothesis is $H_0 : p = q$ and the alternative hypothesis $H_0 : p \neq q$. The statistical values are used as shift measurements. Both shifts occurring in the input distribution $p(\boldsymbol{x})$ and the label distribution $p(\boldsymbol{y})$ are investigated.

We now discuss the representations we use for the datasets and the corresponding statistical tests we employ.

3.2.1 Representation for input space

We investigate two different representations for the input space.

Word frequencies: in this setting, $\boldsymbol{x}$ represents the frequency of each word. The underlying assumption is that the occurrences of words within a dataset indicate how important a word is. The word frequency distribution over the vocabulary represents the dataset.

Distributional representation: In this setting, each instance of $\boldsymbol{x}$ is an n-dimensional vector representation of a sentence within a dataset. Sentence-BERT(Reimers & Gurevych, 2019) is used to encode each sentence. The idea behind sentence-BERT is that semantically similar sentences are closer in vector space (Reimers & Gurevych, 2019). The data points in this n-dimensional space are the distribution for each dataset.

[2]Similiar to CONLL, only English data is used.

3.2.2 Representation for label space

Within the NER task, datasets from different domains have different types of entities. We use category counts as our label distribution $p(\boldsymbol{y})$. Among different domains, the most general types include Person, Organization, Location and Miscellaneous. As mentioned in subsection 3.1, we unify labels with the same semantic meanings. We note that very field-specific datasets will have a different label space than more general datasets.

Following recent work on distribution shifts, for label space, we formulate the problem as one of *unseen data shift* where some attribute values are unseen under p but are seen under q (Wiles et al., 2022). For example, the type *Method* might have zero observation in many datasets such as in CoNLL and Wikigold, but it will have many observations in dataset SciERC. However, it does not necessarily mean that there are no entities that have the type Methods in the Wikigold dataset, but due to specific data generation processes, those entities are not annotated. We see this as an outcome of different sampling processes. We assume all datasets share a common label set $\boldsymbol{A}^l$ and some labels in the set are unseen in p but are seen in q due to systematic sampling error.

3.2.3 Statistical Hypothesis Testing

For each type of representation, a different statistical test is necessary, which we now detail. Shift decisions are reported based on the significant level. By default, we use .05 as the significant threshold for all tests. Furthermore, we use this statistical testing as a means to measure distribution shift and draw a connection between shift and performance.

Chi-Squared Test

For frequency distributions of input and label count distribution, each sample $\boldsymbol{x}_n$ is one categorical value that represents word occurrence in the domain. We adopt Pearson's Chi-Squared test, a parametric test for determining if two frequency distributions are the same. The crucial underlying assumption is that a corpus is modelled as a sequence of independent Bernoulli trials. The relevant statistic χ^2 can be computed as:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_i is the observed value for category i and E_i is the expected value for category i. All word occurrences below 5 are filtered out.

There has been a long debate if the chi-squared test, or statistical testing in general, should be applied for corpus linguistics (Gries, 2005). However, it is still widely employed within the literature (Rabanser et al., 2019). Given that the distribution shift literature also employs chi-squared testing, we also make use of it here.

We employ two data processing procedures while using Chi-Squared tests. First, before applying the Chi-Squared test to data, we implement a normalization procedure to ensure that both the observed and expected values are on the same scale (Underhill & Bradfield, 2013). This normalization enhances the robustness of the test to different sample sizes. Second, by design, label distribution may contain a considerable number of zeros for certain categories. Since Chi-Squared test is not viable when dividing by zero, we added a small constant ($1e-5$) to each category to ensure that we obtain results without changing the numerical meaning of the results [3].

Another potential test for this sort of distribution is the Kolmogorov-Smirnov (KS) two-sample test. However, this test fits the cumulative distribution which requires values to be sorted. Sorting items in a vocabulary is not meaningful.

Maximum Mean Discrepancy (MMD)

For multi-dimensional representations obtained from sentence-BERT, we employ MMD (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012), a non-parametric kernel-based two-sample test to determine if two samples are drawn from two different distributions p and q. MMD tries to calculate the L_2 distance between the mean embeddings μ_p and μ_q of the distributions in a reproducing kernel Hilbert space $\mathcal{F}$ as:

$$\mathbf{MMD}^2(P,Q) = < \mu_p, \mu_p > -2 < \mu_p, \mu_q > + < \mu_q, \mu_q > .$$

Empirically, we use the unbiased estimate of the squared MMD statistic:

$$\mathbf{MMD}^2 = \frac{1}{m^2-m}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) + \frac{1}{n^2-n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\kappa(\boldsymbol{x}'_i,\boldsymbol{x}'_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\kappa(\boldsymbol{x}_i,\boldsymbol{x}'_j)$$

where the kernel is computed with a squared exponential function $\kappa(\boldsymbol{x},\tilde{\boldsymbol{x}}) = e^{-\frac{1}{\sigma}|\boldsymbol{x}-\tilde{\boldsymbol{x}}|^2}$. σ is the median distance between points (Gretton et al., 2012).

3.3 Impact on Performance

The last step in our method is to detect how the shift affects model performance. Our hypothesis is that *as the degree of distribution shift increases, so does the likelihood that a model makes an error and hence the degree of this error will also increase.* As one of the most widely used large-scale pre-trained language models, we use BERT (Devlin et al., 2019) to measure performance. Specifically, we measure the effect of shifts from $p(x)$ and shifts from $p(y)$ on performance.

[3] This approach is inspired by the methods used to address the divide by zero problem in multi-class logistic regression in machine learning

Our baseline NER model is obtained by taking BERT and fine-tuning it on the CoNLL dataset. We first test the model performance on all datasets excluding CoNLL. We then fine-tune the baseline model on each dataset and test on all other datasets.

Specifically, each dataset is split into a training set and an inference set following an 80:20 ratio. Then we pair any two of the datasets and use the training set of the first as the source domain and the inference set for the second as the target domain. We fine-tune the original baseline model on the source training set and evaluate on the target inference set. Fine-tuning is performed for 10 epochs. Similar to the original BERT paper, we use a batch size of 32 and a learning rate of 5e-5. We train and test our model on GPU GeForce 1080Ti with 256 GB memory. The average runtime for fine-tuning one dataset is multiple hours depending on the data size. Fine-tuning and inference for the total 12 datasets takes about 20 hours. We do not add additional parameters to the baseline model, and hence the number of trainable parameters is 110 million. Our code for both testing and evaluation is available as supplementary material.

To draw a connection between distribution shifts and performance degradation, we calculate the correlation between the measurement $shift$ and the performance difference $perf_{ab}$ between any two datasets a and b.

3.4 Experimental Setup

We conduct various experiments under different setups. For shift detection, we verify the validity of the tests on the sampled datasets of the same corpus. If the results indicate no shift detected, this implies that the testing has effectively validated that the two distributions are the same. We subsequently apply tests to all pairs of datasets.

As illustrated in Table 2, the datasets exhibit varying sample sizes. To mitigate the potential impact of the size differences, we uniformly sample a subset of 900 samples from each dataset and perform all experiments. We then investigate if varying sample sizes would affect the testing results. Furthermore, we employ identical tests and performance measures on the original full-sized datasets, the results of which are included in the Appendix A.

For the performance measures, BERT is pre-trained on a particular scope of texts and may favor datasets from certain domains. To address this potential bias, we utilize both BERT-base and BioBERT-base (Lee et al., 2019) and compare their respective performance outcomes. The complete results are provided in the Appendix A.

4 Results and Discussion

We now present the results of applying the method detailed above. We begin with an analysis of the input datasets to verify our hypothesis about the shift between distributions representing shift between domains.

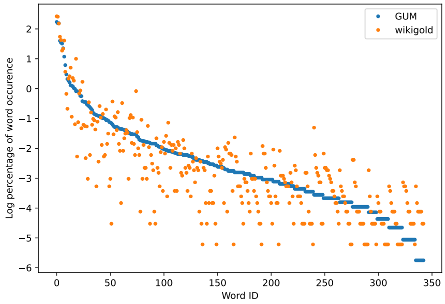

Fig. 2 GUM and Wikigold.

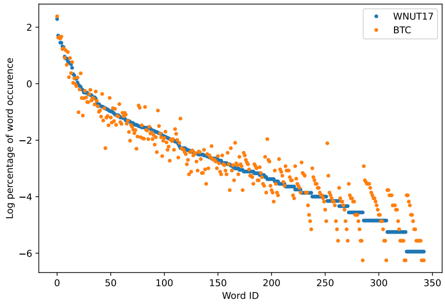

Fig. 3 WNUT-17 and BTC.

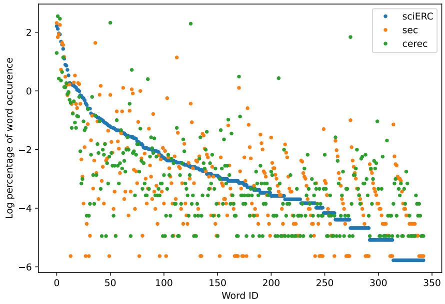

Fig. 4 SciERC, SEC and CEREC.

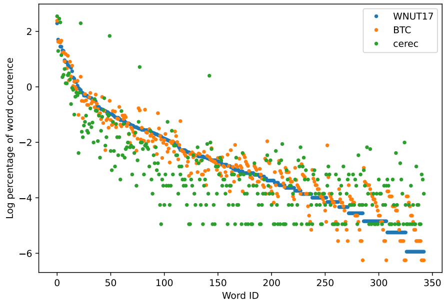

Fig. 5 WNUT-17, BTC and CEREC.

4.1 Datasets analysis

Figures 2 to 5 show the word frequency plots on selected pairs of datasets. WNUT-17 and BTC both include text from Twitter, and we see that word frequency is similar across both datasets. Conversely, in the case of SciERC, SEC and Cerec datasets, which more distinctly represent different domains, we observe greater dispersion within their respective word frequency distributions. Intuitively, these results suggest that word frequency distributions can serve as an indicator of a domain.

4.2 Hypothesis testing

Considering space limitations, Tables 3, 5 and 6 display only a subset of results for hypothesis testing with corresponding distributions. We selected two dataset pairs wherein the source and target distribution are from the same dataset. Then we selected the top 5 and bottom 5 pairs in ascending order. All tables are selected following the same style. All tests are conducted on both sampled datasets and full datasets, and complete results for all dataset pairs are available in Appendix A.

Source data	Target data	Number of samples from test					
		5	50	200	500	1,000	2,000
AnEM	AnEM	-0.2542	-0.0281	-0.0073	-0.0029	-0.0015	-0.0007
BTC	BTC	-0.2906	-0.0286	-0.0070	-0.0028	-0.0014	-0.0007
GUM	WNUT17	0.6140⋆	0.1566⋆	0.0964⋆	0.0619⋆	0.0425	0.0271
GUM	wikigold	0.4367⋆	0.2047⋆	0.1266⋆	0.0828⋆	0.0615⋆	0.0294
BTC	WNUT17	0.2581⋆	0.0503⋆	0.0509⋆	0.0466	0.0353	0.0311
conll	wikigold	0.2646⋆	0.1069⋆	0.0669⋆	0.0385	0.0479	0.0348
WNUT17	wikigold	0.4361⋆	0.0789⋆	0.0532⋆	0.0297	0.0283	0.0406
...	...	...	...	...	...	...	...
ontonotes	sec	0.3163⋆	0.3097⋆	0.2311⋆	0.1569⋆	0.1379⋆	0.1438⋆
i2b2	sec	0.2510⋆	0.1378⋆	0.1387⋆	0.1521⋆	0.1336⋆	0.1462⋆
re3d	sec	0.2838⋆	0.1956⋆	0.1589⋆	0.1482⋆	0.1455⋆	0.1533⋆
sec	sciERC	0.1811⋆	0.1640⋆	0.1612⋆	0.1690⋆	0.1497⋆	0.1536⋆
BTC	sec	0.1342⋆	0.1646⋆	0.1832⋆	0.1801⋆	0.1551⋆	0.1586⋆

Table 3 MMD statistics for selected pair of full-sized datasets with a different number of samples. Ordered by the distance between pairs of datasets with 2000 samples. Sign ⋆ indicates there is a shift detected.

Source data	Target data	Number of samples from test					
		5	50	200	500	900	
sec	sec	-0.2881	-0.0275	-0.0069	-0.0027	-0.0015	
AnEM	AnEM	-0.2951	-0.0294	-0.0073	-0.0029	-0.0016	
BTC	WNUT17	-0.0088	0.0085	0.0073	0.0063	0.0068	
GUM	wikigold	0.0476	0.0160	0.0194	0.0209	0.0209	
conll	wikigold	0.0105	0.0263	0.0276	0.0264	0.0274	
ontonotes	GUM	0.0989⋆	0.0408	0.0318	0.0307	0.0294	
GUM	WNUT17	0.0744⋆	0.0432	0.0402	0.0369	0.0359	
...	...	...	...	...	...	...	...
BTC	sec	0.1085⋆	0.1495⋆	0.1571⋆	0.1534⋆	0.1438⋆	
i2b2	sciERC	0.1871⋆	0.1556⋆	0.1424⋆	0.1439⋆	0.1440⋆	
re3d	sec	0.0855⋆	0.1543⋆	0.1587⋆	0.1558⋆	0.1502⋆	
i2b2	sec	0.1229⋆	0.1608⋆	0.1619⋆	0.1599⋆	0.1506⋆	
sec	sciERC	0.1556⋆	0.1613⋆	0.1658⋆	0.1568⋆	0.1521⋆	

Table 4 MMD statistics for selected pair of sampled datasets (900 samples) with a different number of samples. Ordered by the distance between pairs of datasets with 900 samples. Sign ⋆ indicates there is a shift detected.

4.2.1 Chi-squared testing for input distribution

Table 5 shows the chi-squared testing on both sampled dataset pairs and original-sized dataset pairs. The distance between the same distribution is also

Source data	Target data	Statistics	Shift Decision	Source data	Target data	Statistics	Shift Decision
conll	conll	0.0000		conll	conll	0.0000	
cerec	cerec	0.0000		cerec	cerec	0.0000	
BTC	WNUT17	96.0458		GUM	BTC	75.6658	
GUM	WNUT17	123.5303		GUM	WNUT17	81.6450	
GUM	BTC	127.4870		BTC	WNUT17	93.8928	
ontonotes	WNUT17	144.4873		ontonotes	WNUT17	98.6004	
ontonotes	re3d	149.6937		ontonotes	BTC	124.5910	
...	...	...	...	...	...	...	...
cerec	ontonotes	4726.8631	♣	WNUT17	sec	2590.0214	♣
conll	sciERC	6930.8975	♣	ontonotes	sciERC	2608.5222	♣
cerec	sciERC	7169.8736	♣	conll	sciERC	3051.1489	♣
conll	AnEM	7186.4441	♣	BTC	sec	4274.0432	♣
cerec	i2b2	7927.0406	♣	cerec	sciERC	4769.4978	♣

Table 5 Chi-squared statistics on full-sized datasets (left) and sampled datasets (900 samples)(right) in ascending order. ♣ indicates there is a shift detected.

reported as a sanity check. When the source and target distributions are equivalent, the testing indicates that no shift is detected. This indicates that the testing is capable of identifying when two distributions are identical.

For the full-sized datasets, the left table in Table 5 reveals that among the 78 dataset pairs, 13 pairs are detected with shifts. Meanwhile, the right table indicates that for the sampled datasets, out of the same 78 pairs, 22 pairs are detected to have shifts. These results suggest that the test is more sensitive to identifying shifts when there is a smaller sample size. On closer inspection of the dataset pairs, we observe that out of the 13 shift-detected full-sized pairs, 11 pairs are also detected in the sampled datasets, which reaches an approximately 84.6% agreement. Additionally, the full results presented in (Table 5) reveal that a higher Chi-Squared value does not necessarily imply the detection of a shift. For instance, while the OntoNotes and i2b2 datasets have high Chi-Squared values, no shift is detected. This outcome could arise due to the data samples being non-representative of the full distribution, thereby resulting in the test's inability to make a confident conclusion.

Analyzing these results, we note that BTC and WNUT-17 datasets have the smallest distance, which is inline with the frequency plots (Figure 3). On the other hand, the BTC dataset and SEC finance dataset have the furthest distance, which, as expected, reflects that these two datasets have very different text styles. One surprising outcome is GUM and SciERC which have a relatively small distance using this representation while being from what appear to be different domains. These examples illustrate that this test can quantify the distance between datasets.

4.2.2 MMD testing for input distribution

For the distributional representations, we apply MMD with a different number of samples (i.e. embedded sentences) from n = [5, 50, 200, 500, 1000, 2000]. Tables 3 and 4 shows the results of these tests. The table is ordered by the scores generated using 2000 samples. Again, we measure the difference between a dataset and itself as a sanity check. Negative results that are close to zero from an unbiased MMD test indicate a small distance.

Table 3 shows the results from MMD on full-sized datasets and Table 4 shows the results on sampled datasets. Both results show a similar trend where the tests are more sensitive to shifts with smaller data sizes.

CoNLL, a widely used benchmark dataset in NER, is surprisingly far from other datasets in the distance measured by the chi-squared test. However, with MMD tests, the distance is fairly close. This is an indication that sentence-level representation provides more information than word-frequency representation.

4.2.3 Label distribution

To detect category shift in label distribution, we utilized the Chi-squared test, as detailed in subsection 3.4. This testing was performed on both the sampled and full-sized datasets, and the results are presented in Table 6. Results reveal a significant difference between the datasets that share the same categories and those that have different categories.

Datasets that are focused on specialized fields typically contain more specific labels. Consequently, the dissimilarity between these datasets and those from general domains is greater. For example, while the input shift between BTC and WNUT17 may be small, the label shift is relatively significant due to their distinct label spaces. For the NER task, generalizing model performance to datasets that have distinct categories is more challenging, as evidenced in the following section.

Source data	Target data	Statistics	Shift Decision
conll	conll	0.00	
cerec	cerec	0.00	
conll	wikigold	0.04	
BTC	sec	0.07	
BTC	wikigold	0.51	
BTC	WNUT17	0.84	
BTC	re3d	1.22	
...	...	...	...
conll	sciERC	311849153.00	♣
i2b2-06	sciERC	320519601.10	♣
BTC	sciERC	459742109.25	♣
sec	sciERC	555622639.19	♣
cerec	sciERC	699845821.45	♣

Source data	Target data	Statistics	Shift Decision
conll	conll	0.00	
cerec	cerec	0.00	
conll	wikigold	0.04	
BTC	sec	0.10	
BTC	wikigold	0.52	
BTC	WNUT17	0.84	
BTC	re3d	1.11	
...	...	...	...
conll	sciERC	308299988.60	♣
i2b2-06	sciERC	315852394.93	♣
BTC	sciERC	451014301.26	♣
sec	sciERC	546583192.69	♣
cerec	sciERC	675866015.85	♣

Table 6 Chi-squared testing for label distributions of full-sized datasets (left) and sampled datasets (900 samples) (right).

4.3 Performance measurement

As noted in subsection 3.4, we conducted four sets of experiments, including fine-tuning models on both original-sized and sampled datasets with 900 samples using both BERT-base and BioBERT-base models. The complete results can be seen in Appendix A.

Tables 7 to 9 present the micro-averaged F1 performance of the models trained and tested on the sampled datasets with BERT, full-sized datasets with BERT, and full-sized datasets with BioBERT, respectively. Each row in

Table 7 Micro-average F1 score when the model is fine-tuned on the source dataset (the row) and tested on the target dataset (the column). Fine-tuning uses BERT-base model with 10 epochs. All datasets are sampled datasets with 900 samples.

	conll	cerec	ontonotes	i2b2-06	re3d	wikigold	SEC	GUM	BTC	sciERC	WNUT17	AnEM	average f1
conll	**0.66**	0.25	0.24	0.22	0.24	0.44	0.07	0.04	0.25	0.00	0.30	0.00	**0.22**
cerec	0.26	**0.70**	0.08	0.15	0.21	0.24	0.04	0.17	0.10	0.00	0.13	0.00	0.20
ontonotes	0.18	0.01	**0.22**	0.03	0.09	0.02	0.01	0.01	0.04	0.00	0.09	0.00	0.15
i2b2-06	0.14	0.11	0.04	**0.69**	0.04	0.11	0.01	0.01	0.17	0.00	0.09	0.00	0.14
re3d	0.17	0.18	0.12	0.00	**0.54**	0.21	0.06	0.11	0.15	0.00	0.18	0.00	0.13
wikigold	0.52	0.20	0.23	0.11	0.25	**0.62**	0.08	0.08	0.20	0.00	0.22	0.00	0.13
SEC	0.08	0.08	0.05	0.02	0.04	0.12	**0.90**	0.02	0.06	0.00	0.07	0.00	0.13
GUM	0.11	0.12	0.03	0.02	0.13	0.12	0.02	**0.27**	0.08	0.00	0.07	0.00	0.13
BTC	0.41	0.17	0.15	0.17	0.20	0.31	0.04	0.05	**0.66**	0.00	0.21	0.00	0.13
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.30**	0.00	0.00	0.12
WNUT17	0.21	0.16	0.06	0.17	0.07	0.19	0.00	0.03	0.14	0.00	**0.37**	0.00	0.12
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.30**	0.11

Table 8 F1 scores on full-sized datasets. Fine-tuning uses the BERT-base model with 10 epochs.

	conll	cerec	ontonotes	i2b2-06	wikigold	WNUT17	GUM	re3d	SEC	BTC	sciERC	AnEM	average f1
conll	**0.90**	0.35	0.31	0.13	0.62	0.26	0.12	0.33	0.10	0.31	0.00	0.00	**0.29**
cerec	0.35	**0.90**	0.13	0.10	0.31	0.14	0.21	0.19	0.08	0.20	0.00	0.00	0.25
ontonotes	0.28	0.14	**0.93**	0.05	0.26	0.14	0.08	0.17	0.06	0.17	0.00	0.00	0.23
i2b2-06	0.27	0.20	0.12	**0.99**	0.33	0.06	0.05	0.22	0.02	0.16	0.00	0.00	0.22
wikigold	0.57	0.25	0.30	0.10	**0.88**	0.13	0.13	0.30	0.13	0.20	0.00	0.00	0.21
WNUT17	0.43	0.32	0.29	0.15	0.44	**0.70**	0.10	0.24	0.09	0.27	0.00	0.00	0.21
GUM	0.20	0.17	0.05	0.01	0.23	0.04	**0.68**	0.28	0.04	0.06	0.00	0.00	0.21
re3d	0.19	0.25	0.15	0.04	0.19	0.11	0.14	**0.61**	0.06	0.20	0.00	0.00	0.21
SEC	0.22	0.14	0.19	0.05	0.25	0.09	0.05	0.12	**0.93**	0.21	0.00	0.00	0.21
BTC	0.59	0.35	0.32	0.14	0.62	0.24	0.14	0.34	0.15	**0.87**	0.00	0.00	0.20
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.57**	0.00	0.19
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.79**	0.19

Table 9 F1 scores on full-sized datasets. Fine-tuning uses the BioBERT-base model with 10 epochs.

	conll	cerec	ontonotes	i2b2-06	wikigold	WNUT17	GUM	re3d	SEC	BTC	sciERC	AnEM	average f1
conll	**0.88**	0.32	0.26	0.11	0.58	0.21	0.10	0.29	0.36	0.23	0.00	0.00	**0.28**
cerec	0.30	**0.87**	0.11	0.06	0.30	0.12	0.20	0.20	0.31	0.15	0.00	0.00	0.25
ontonotes	0.23	0.14	**0.92**	0.10	0.22	0.12	0.07	0.15	0.07	0.15	0.00	0.00	0.23
i2b2-06	0.33	0.22	0.17	**1.00**	0.36	0.09	0.09	0.26	0.23	0.16	0.00	0.00	0.23
wikigold	0.49	0.25	0.29	0.10	**0.84**	0.11	0.11	0.28	0.38	0.15	0.00	0.00	0.21
WNUT17	0.40	0.24	0.25	0.14	0.43	**0.61**	0.09	0.23	0.19	0.25	0.00	0.00	0.21
GUM	0.15	0.16	0.04	0.03	0.21	0.03	**0.63**	0.26	0.04	0.06	0.00	0.00	0.21
re3d	0.15	0.21	0.11	0.03	0.16	0.07	0.13	**0.53**	0.09	0.11	0.00	0.00	0.20
SEC	0.21	0.19	0.15	0.11	0.18	0.08	0.05	0.10	**0.90**	0.21	0.00	0.00	0.20
BTC	0.50	0.33	0.32	0.14	0.58	0.22	0.12	0.30	0.31	**0.85**	0.00	0.00	0.20
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.55**	0.00	0.19
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.77**	0.19

the table indicates the dataset the model is fine-tuned on. Correspondingly, the columns indicate the dataset which the fine-tuned model is tested upon. The last column reports the average F1 scores of all the test results, which represents the generalization ability of the model when fine-tuned on a specific dataset. The rows are ordered by this average F1 score.

Table 8 reveals that even though BTC and WNUT17 contain texts from the same domain, the model's generalization ability on WNUT17 decreases significantly when fine-tuned on BTC, as there is a significant label category shift between these two datasets.

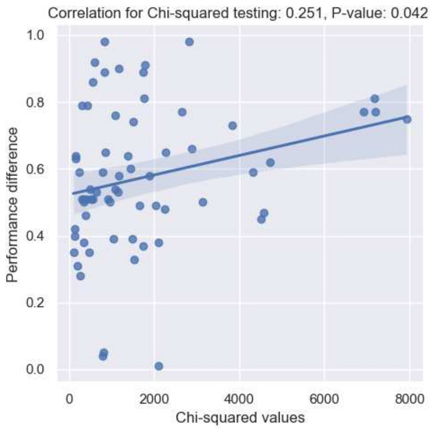

Fig. 6 Plots for Chi-squared measures with word frequency input distribution and performance difference. Linear regression model fitted.

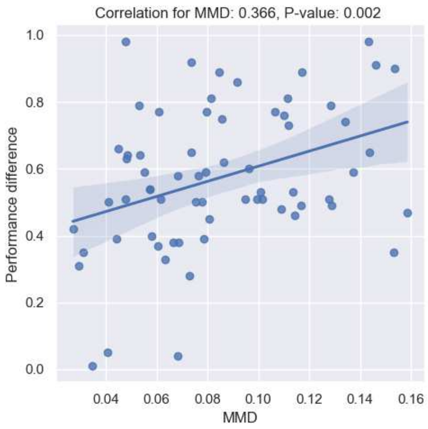

Fig. 7 Plots for MMD measures with sentence-level input distributions and performance difference. Linear regression model fitted.

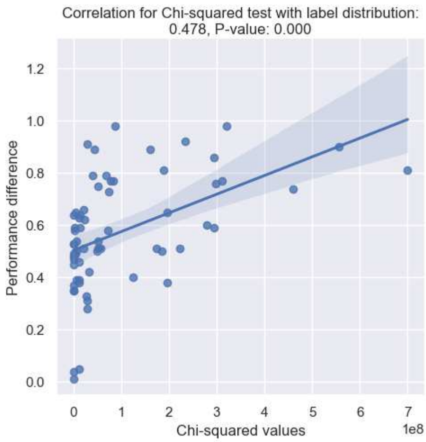

Fig. 8 Plots for Chi-squared measures with label distribution and performance difference. Linear regression model fitted.

Comparing Table 7 and Table 8, we observe that when we control the number of data samples, the average F1 scores tend to decrease. However, the overall rankings of the datasets are similar between the two tables, except for WNUT17 and Re3d. Using a subset of the original dataset reduces the generalization ability significantly, indicating the additional data samples in the original data set help improve the generalization. Conversely, for dataset Re3d, the generalization ability increases while the number of samples decreases, indicating that the additional data samples in the dataset harm the generalization.

Due to mutually exclusive sets of categories, we encounter many zero F1 scores on AnEM dataset and SciERC dataset. Even though fine-tuning helps improve the performance on the same dataset, the generalization ability is low, indicating that category shift has a significant impact on the performance.

Comparing Table 8 and Table 9, we observe that the fine-tuning performance is slightly impacted by the text on which these language models were pre-trained. However, the average F1 score rankings remain the same.

4.4 Correlation

To investigate further how shifts impact model performance, we report the correlation between the testing statistics and performance differences in figs. 6 to 8. Assuming there are datasets D_a and D_b. $Perf_{ab}$ indicates the performance difference on D_a and D_b when the model is fine-tuned on source dataset D_s where $s \in \{D_i \mid i = \{1, 2, ..., 12\}\}$. $Shift_{ab}$ is the distance between D_a and D_b with regarding to each statistical test. The correlation is calculated between $perf$ and $shift$.

Based on the presented plots, it is evident that the label category shift shows the most statistically significant correlation (P < .0001) with model performance. This finding suggests that category shift can serve as a reliable indicator of model performance in a supervised setting when evaluating in a new domain. With respect to input distribution shift, while the word frequency distribution's correlation with model performance is the lowest, it is still significant (P = .042). The MMD results reveal a moderately strong correlation (P = 0.002). This indicates that in an unsupervised setting, MMD testing with sentence-level representation distribution can be used to estimate model performance when transferring between domains.

5 Conclusion

In this work, we investigated input data and label distribution shifts across 12 benchmark NER datasets. We compared two different types of representations for input shifts. We systematically measured the shifts using the lens of statistical testing. We measured performance differences by fine-tuning BERT models and calculating the correlation between shifts and performance.

The results show that both word frequency distribution and sentence-level distributional representations are useful for ascertaining shift. Changing between domains results in measurable differences in distribution shifts. Results show that label shift correlates more significantly with performance degradation than input shifts for NER. However, there is still a correlation between input shift and performance degradation. Here, sentence-level representations provide more signals for the relation between distribution shift and performance.

Based on these results, we believe that shift detection and the measurement of distribution shifts can play important roles in tackling NLP tasks, especially for new and low-resource domains. In particular, when applying a model to a new domain, or as data changes, the measures detailed above can help researchers and practitioners decide whether the expense of gathering new annotated data and subsequent fine-tuning is warranted. In the future, we hope that distribution shift measurement can become part of widely used NLP paradigms such as crowd-sourcing and active learning.

Acknowledgement

This work is funded by the Dutch Research Council (NWO) through grant MVI.19.032. CRediT [4] author statement is as follows. Xue Li: Conceptualization, Methodology, Software, Writing - Original Draft. Paul Groth: Conceptualization, Supervision, Writing - Reviewing and Editing.

References

Arora, U., Huang, W., He, H. (2021, November). Types of Out-of-Distribution Texts and How to Detect Them. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10687–10701). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved 2022-07-18, from https://aclanthology.org/2021.emnlp-main.835 10.18653/v1/2021.emnlp-main.835

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., Curran, J.R. (2009, August). Named entity recognition in Wikipedia. *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (people's web)* (pp. 10–18). Suntec, Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W09-3302

Cobb, O., & Van Looveren, A. (2022, August). *Context-Aware Drift Detection.* arXiv. Retrieved 2022-09-20, from http://arxiv.org/abs/2203.08644 (arXiv:2203.08644 [cs, stat]) 10.48550/arXiv.2203.08644

Csurka, G. (2017). Domain adaptation for visual applications: A comprehensive survey. *CoRR*, *abs/1702.05374*. Retrieved from http://arxiv.org/abs/1702.05374 https://arxiv.org/abs/1702.05374

Dai, X., Karimi, S., Hachey, B., Paris, C. (2019). Using similarity measures to select pretraining data for NER. *CoRR*, *abs/1904.00585*. Retrieved from http://arxiv.org/abs/1904.00585 https://arxiv.org/abs/1904.00585

Dakle, P.P., & Moldovan, D. (2020). CEREC: A corpus for entity resolution in email conversations. *Proceedings of the 28th*

[4] https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement

international conference on computational linguistics. International Committee on Computational Linguistics. Retrieved from https://doi.org/10.18653%2Fv1%2F2020.coling-main.30 10.18653/v1/2020.coling-main.30

Derczynski, L., Bontcheva, K., Roberts, I. (2016, December). Broad Twitter corpus: A diverse named entity recognition resource. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 1169–1179). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclanthology.org/C16-1111

Derczynski, L., Nichols, E., van Erp, M., Limsopatham, N. (2017, September). Results of the WNUT2017 shared task on novel and emerging entity recognition. *Proceedings of the 3rd workshop on noisy user-generated text* (pp. 140–147). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-4418 10.18653/v1/W17-4418

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423 10.18653/v1/N19-1423

Dstl, D.S., & Laboratory, T. (n.d.). *Dstl/re3d: Relationship and entity extraction evaluation dataset.* Retrieved from https://github.com/dstl/re3d

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A. (2019, 09–15 Jun). Exploring the landscape of spatial robustness. K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 1802–1811). PMLR. Retrieved from https://proceedings.mlr.press/v97/engstrom19a.html

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A. (2012, mar). A kernel two-sample test. *J. Mach. Learn. Res.*, *13*(null), 723–773.

Gries, S.T. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on kilgarriff. , *1*(2), 277–294. Retrieved 2022-10-19, from https://doi.org/10.1515/cllt.2005.1.2.277

doi:10.1515/cllt.2005.1.2.277

Klimt, B., & Yang, Y. (2004). The enron corpus: A new dataset for email classification research. J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Machine learning: Ecml 2004* (pp. 217–226). Berlin, Heidelberg: Springer Berlin Heidelberg.

Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., … Liang, P. (2021, July). WILDS: A Benchmark of in-the-Wild Distribution Shifts. *Proceedings of the 38th International Conference on Machine Learning* (pp. 5637–5664). PMLR. Retrieved 2022-09-20, from https://proceedings.mlr.press/v139/koh21a.html (ISSN: 2640-3498)

Kulinski, S., Bagchi, S., Inouye, D.I. (2020). Feature shift detection: Localizing which features have shifted via conditional distribution tests. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 19523–19533). Curran Associates, Inc.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, *abs/1901.08746*. Retrieved from http://arxiv.org/abs/1901.08746 https://arxiv.org/abs/1901.08746

Lekhtman, E., Ziser, Y., Reichart, R. (2021, November). DILBERT: Customized pre-training for domain adaptation with category shift, with an application to aspect extraction. *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 219–230). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.20 10.18653/v1/2021.emnlp-main.20

Luan, Y., He, L., Ostendorf, M., Hajishirzi, H. (2018, October-November). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3219–3232). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1360 10.18653/v1/D18-1360

Malinin, A., Band, N., Ganshin, Alexander, Chesnokov, G., Gal, Y., … Yangel, B. (2022, February). *Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks.* arXiv. Retrieved 2022-08-19, from http://arxiv.org/abs/2107.07455 (arXiv:2107.07455 [cs, stat]) 10.48550/arXiv.2107.07455

Michel, P. (2021). Learning neural models for natural language processing in the face of distributional shift. *ArXiv*, *abs/2109.01558*.

Ohta, T., Pyysalo, S., Tsujii, J., Ananiadou, S. (2012, 01). Open-domain anatomical entity mention detection. (p. 27-36).

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D. (2009). Dataset shift in machine learning..

Rabanser, S., Günnemann, S., Lipton, Z. (2019). Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V. (2019, 09–15 Jun). Do ImageNet classifiers generalize to ImageNet? K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 5389–5400). PMLR. Retrieved from https://proceedings.mlr.press/v97/recht19a.html

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, *abs/1908.10084*. Retrieved from http://arxiv.org/abs/1908.10084 https://arxiv.org/abs/1908.10084

Salinas Alvarado, J.C., Verspoor, K., Baldwin, T. (2015, December). Domain adaption of named entity recognition to support credit risk assessment. *Proceedings of the australasian language technology association workshop 2015* (pp. 84–90). Parramatta, Australia. Retrieved from https://aclanthology.org/U15-1010

Specia, L., Li, Z., Pino, J., Chaudhary, V., Guzmán, F., Neubig, G., ... Li, X. (2020, November). Findings of the WMT 2020 shared task on machine translation robustness. *Proceedings of the fifth conference on machine translation* (pp. 76–91). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.4

Tjong Kim Sang, E.F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003* (pp. 142–147). Retrieved from https://aclanthology.org/W03-0419

Underhill, L., & Bradfield, D. (2013). *INTROSTAT (statistics textbook)* (Unpublished doctoral dissertation).

Uzuner, ö., Luo, Y., Szolovits, P. (2007, 09). Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, *14* (5), 550-563. Retrieved from https://doi.org/10.1197/jamia.M2444 https://arxiv.org/abs/https://academic.oup.com/jamia/article-pdf/14/5/550/2136261/14-5-550.pdf
10.1197/jamia.M2444

Vu, T., Wang, T., Munkhdalai, T., Sordoni, A., Trischler, A., Mattarella-Micke, A., . . . Iyyer, M. (2020). Exploring and predicting transferability across NLP tasks. *CoRR*, *abs/2005.00770*. Retrieved from https://arxiv.org/abs/2005.00770 https://arxiv.org/abs/2005.00770

Weischedel, R.M., Hovy, E.H., Marcus, M.P., Palmer, M. (2017). Ontonotes : A large training corpus for enhanced processing..

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Cemgil, T. (2022). A FINE-GRAINED ANALYSIS ON DISTRIBUTION SHIFT. , 15.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, *51*(3), 581–612.

http://dx.doi.org/10.1007/s10579-016-9343-x

Appendix A Appendix

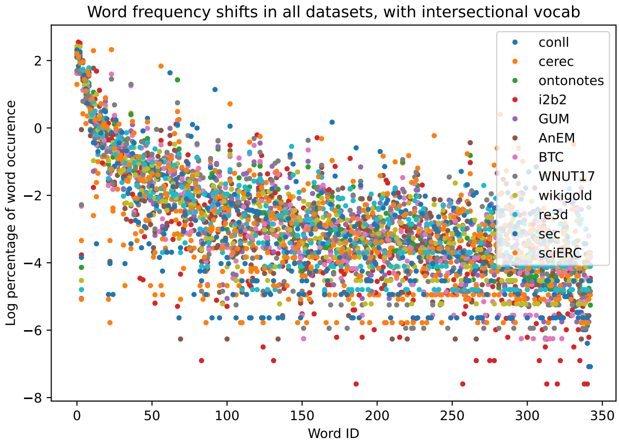

Fig. 1 Frequency plots across all datasets with an intersectional vocabulary where the vocabulary is an intersection of all vocabularies.

Table 1 MMD statistics for all combination without repetition of datasets with different number of samples. The table is ordered by the distance between pairs of datasets when the number of samples are 2000. The star signs show the results are statistically significant and there are shifts.

Source data	Target data	Number of samples from test					
		5	50	200	500	1,000	2,000
AnEM	AnEM	-0.2542	-0.0281	-0.0073	-0.0029	-0.0015	-0.0007
BTC	BTC	-0.2906	-0.0286	-0.0070	-0.0028	-0.0014	-0.0007
GUM	GUM	-0.2521	-0.0277	-0.0071	-0.0029	-0.0015	-0.0007
WNUT17	WNUT17	-0.2521	-0.0289	-0.0073	-0.0030	-0.0015	-0.0007
cerec	cerec	-0.2776	-0.0286	-0.0072	-0.0029	-0.0014	-0.0007
conll	conll	-0.2653	-0.0284	-0.0073	-0.0029	-0.0014	-0.0007
i2b2	i2b2	-0.2520	-0.0284	-0.0071	-0.0028	-0.0014	-0.0007
ontonotes	ontonotes	-0.2587	-0.0254	-0.0067	-0.0029	-0.0014	-0.0007
sciERC	sciERC	-0.2890	-0.0284	-0.0071	-0.0028	-0.0014	-0.0007
wikigold	wikigold	-0.2695	-0.0282	-0.0072	-0.0029	-0.0015	-0.0008
sec	sec	-0.2831	-0.0273	-0.0068	-0.0027	-0.0014	-0.0009
re3d	re3d	-0.2600	-0.0279	-0.0071	-0.0029	-0.0015	-0.0015
GUM	WNUT17	0.6140*	0.1566*	0.0964*	0.0619*	0.0425	0.0271
GUM	wikigold	0.4367*	0.2047*	0.1266*	0.0828*	0.0615*	0.0294
BTC	WNUT17	0.2581*	0.0503*	0.0509*	0.0466	0.0353	0.0311
conll	wikigold	0.2646*	0.1069*	0.0669*	0.0385	0.0479	0.0348
WNUT17	wikigold	0.4361*	0.0789*	0.0532*	0.0297	0.0283	0.0406
GUM	AnEM	0.4231*	0.1762*	0.0975*	0.0727*	0.0607*	0.0411
conll	WNUT17	0.4148*	0.0678*	0.0402	0.0338	0.0532*	0.0443
conll	GUM	0.3968*	0.1629*	0.1158*	0.0870*	0.0934*	0.0450
i2b2	AnEM	0.2739*	0.0860*	0.0557*	0.0535*	0.0461	0.0477
AnEM	wikigold	0.2688*	0.1235*	0.0717*	0.0520*	0.0503*	0.0478
ontonotes	re3d	0.2496*	0.2194*	0.1538*	0.0500	0.0535*	0.0480
cerec	WNUT17	0.3337*	0.0627*	0.0509*	0.0467	0.0465	0.0484
ontonotes	GUM	0.5314*	0.3557*	0.2142*	0.1115*	0.0890*	0.0530*
ontonotes	WNUT17	0.4052*	0.2008*	0.1389*	0.0457	0.0423	0.0534*
cerec	GUM	0.3492*	0.1526*	0.1075*	0.0874*	0.0824*	0.0552*
ontonotes	BTC	0.2600*	0.2338*	0.1792*	0.0792*	0.0686*	0.0572*
conll	cerec	0.1973*	0.0861*	0.0626*	0.0508*	0.0738*	0.0574*
GUM	BTC	0.3451*	0.1934*	0.1547*	0.1223*	0.0994*	0.0582*
cerec	wikigold	0.2564*	0.1174*	0.0783*	0.0634*	0.0615*	0.0605*
conll	AnEM	0.2566*	0.1009*	0.0656*	0.0606*	0.0812*	0.0607*
AnEM	WNUT17	0.4262*	0.1042*	0.0554*	0.0449	0.0487	0.0617*
conll	BTC	0.1510*	0.0878*	0.0817*	0.0761*	0.0871*	0.0632*
ontonotes	wikigold	0.3596*	0.1709*	0.1677*	0.0701*	0.0650*	0.0666*
BTC	wikigold	0.2147*	0.1211*	0.1080*	0.0824*	0.0687*	0.0682*
i2b2	wikigold	0.3163*	0.1125*	0.0753*	0.0686*	0.0642*	0.0683*
conll	ontonotes	0.3720*	0.2402*	0.1244*	0.0697*	0.0970*	0.0686*
GUM	re3d	0.4959*	0.2415*	0.1646*	0.1108*	0.1087*	0.0728*
i2b2	GUM	0.4385*	0.1880*	0.1217*	0.1121*	0.0940*	0.0735*
conll	i2b2	0.2461*	0.0980*	0.0729*	0.0724*	0.0836*	0.0737*
conll	re3d	0.3366*	0.1395*	0.0453	0.0541*	0.0952*	0.0753*
cerec	BTC	0.0892*	0.0911*	0.0928*	0.0907*	0.0865*	0.0763*
GUM	sciERC	0.3731*	0.1781*	0.1375*	0.1213*	0.1008*	0.0778*
WNUT17	re3d	0.3548*	0.0967*	0.0740*	0.0483	0.0650*	0.0784*
wikigold	re3d	0.3274*	0.1211*	0.0898*	0.0638*	0.0760*	0.0793*
i2b2	WNUT17	0.4524*	0.0822*	0.0651*	0.0710*	0.0656*	0.0797*
BTC	re3d	0.2138*	0.1317*	0.1071*	0.0878*	0.0910*	0.0807*
cerec	AnEM	0.2287*	0.1215*	0.0747*	0.0754*	0.0846*	0.0815*
ontonotes	AnEM	0.3626*	0.2827*	0.1698*	0.0845*	0.0846*	0.0845*
cerec	i2b2	0.1851*	0.0734*	0.0571*	0.0755*	0.0769*	0.0858*
cerec	ontonotes	0.3310*	0.2643*	0.1853*	0.0952*	0.0933*	0.0865*
wikigold	sciERC	0.2372*	0.1347*	0.1095*	0.1055*	0.0981*	0.0918*
AnEM	BTC	0.1894*	0.1244*	0.1082*	0.1025*	0.1010*	0.0948*
WNUT17	sciERC	0.3590*	0.0939*	0.0878*	0.0880*	0.0867*	0.0964*
AnEM	re3d	0.3190*	0.1483*	0.0982*	0.0819*	0.0999*	0.0994*
cerec	re3d	0.2920*	0.1363*	0.1039*	0.0878*	0.1045*	0.1008*
AnEM	sciERC	0.2466*	0.1329*	0.0922*	0.1032*	0.1069*	0.1016*
conll	sciERC	0.2141*	0.1163*	0.1087*	0.1114*	0.1292*	0.1065*
cerec	sec	0.1203*	0.1248*	0.1239*	0.1231*	0.1081*	0.1091*
i2b2	BTC	0.2184*	0.1072*	0.1113*	0.1160*	0.1094*	0.1102*
cerec	sciERC	0.1908*	0.1145*	0.1050*	0.1144*	0.1175*	0.1116*
ontonotes	i2b2	0.3943*	0.2750*	0.1961*	0.1162*	0.1077*	0.1119*
wikigold	sec	0.2417*	0.1701*	0.1441*	0.1281*	0.1045*	0.1135*
GUM	sec	0.3692*	0.2313*	0.1893*	0.1633*	0.1321*	0.1144*
conll	sec	0.1941*	0.1385*	0.1251*	0.1226*	0.1294*	0.1167*
ontonotes	sciERC	0.3224*	0.2789*	0.2021*	0.1281*	0.1205*	0.1173*
AnEM	sec	0.2065*	0.1559*	0.1331*	0.1321*	0.1207*	0.1277*
i2b2	re3d	0.3472*	0.1175*	0.1160*	0.1131*	0.1236*	0.1286*
WNUT17	sec	0.3779*	0.1311*	0.1277*	0.1186*	0.1059*	0.1289*
BTC	sciERC	0.1668*	0.1354*	0.1548*	0.1544*	0.1449*	0.1342*
re3d	sciERC	0.2966*	0.1667*	0.1414*	0.1280*	0.1416*	0.1374*
i2b2	sciERC	0.2644*	0.1371*	0.1261*	0.1480*	0.1416*	0.1435*
ontonotes	sec	0.3163*	0.3097*	0.2311*	0.1569*	0.1379*	0.1438*
i2b2	sec	0.2510*	0.1378*	0.1387*	0.1521*	0.1336*	0.1462*
re3d	sec	0.2838*	0.1956*	0.1589*	0.1482*	0.1455*	0.1533*
sec	sciERC	0.1811*	0.1640*	0.1612*	0.1690*	0.1497*	0.1536*

Table 2 MMD statistics for sampled datasets. The star signs show the results are statistically significant and there are shifts.

Source data	Target data	Number of samples from test				
		5	50	200	500	900
sec	sec	-0.2881	-0.0275	-0.0069	-0.0027	-0.0015
AnEM	AnEM	-0.2951	-0.0294	-0.0073	-0.0029	-0.0016
BTC	BTC	-0.2925	-0.0290	-0.0073	-0.0029	-0.0016
WNUT17	WNUT17	-0.2969	-0.0292	-0.0073	-0.0029	-0.0016
cerec	cerec	-0.2560	-0.0284	-0.0071	-0.0029	-0.0016
conll	conll	-0.2993	-0.0291	-0.0073	-0.0029	-0.0016
i2b2	i2b2	-0.2879	-0.0283	-0.0071	-0.0028	-0.0016
ontonotes	ontonotes	-0.2936	-0.0290	-0.0073	-0.0029	-0.0016
re3d	re3d	-0.2971	-0.0284	-0.0071	-0.0028	-0.0016
sciERC	sciERC	-0.2749	-0.0285	-0.0071	-0.0029	-0.0016
wikigold	wikigold	-0.3013	-0.0297	-0.0074	-0.0030	-0.0016
GUM	GUM	-0.2953	-0.0298	-0.0075	-0.0030	-0.0017
BTC	WNUT17	-0.0088	0.0085	0.0073	0.0063	0.0068
GUM	wikigold	0.0476	0.0160	0.0194	0.0209	0.0209
conll	wikigold	0.0105	0.0263	0.0276	0.0264	0.0274
ontonotes	GUM	0.0989*	0.0408	0.0318	0.0307	0.0294
GUM	WNUT17	0.0744*	0.0432	0.0402	0.0369	0.0359
conll	GUM	0.0300	0.0396	0.0381	0.0366	0.0365
GUM	BTC	0.0817*	0.0451	0.0407	0.0396	0.0386
ontonotes	wikigold	0.0561*	0.0477	0.0426	0.0402	0.0405
GUM	AnEM	0.0866*	0.0454	0.0407	0.0426	0.0432
conll	ontonotes	0.0369	0.0513*	0.0486	0.0449	0.0433
ontonotes	BTC	0.0564*	0.0466	0.0437	0.0458	0.0447
AnEM	wikigold	0.0518*	0.0490	0.0474	0.0457	0.0465
conll	BTC	0.0324	0.0456	0.0471	0.0468	0.0465
ontonotes	WNUT17	0.0509*	0.0559*	0.0491	0.0483	0.0476
WNUT17	wikigold	0.0247	0.0482	0.0486	0.0480	0.0482
conll	WNUT17	0.0526*	0.0544*	0.0530*	0.0483	0.0487
BTC	wikigold	0.0276	0.0509*	0.0480	0.0494	0.0496
i2b2	AnEM	0.0540*	0.0530*	0.0540*	0.0520*	0.0508*
ontonotes	re3d	-0.0277	0.0566*	0.0521*	0.0548*	0.0541*
cerec	GUM	0.2320*	0.0587*	0.0552*	0.0538*	0.0561*
conll	cerec	0.1734*	0.0489	0.0529*	0.0538*	0.0573*
cerec	WNUT17	0.1661*	0.0657*	0.0564*	0.0544*	0.0581*
conll	AnEM	0.0456	0.0683*	0.0621*	0.0593*	0.0586*
cerec	BTC	0.1639*	0.0619*	0.0587*	0.0577*	0.0609*
cerec	wikigold	0.1416*	0.0632*	0.0585*	0.0585*	0.0617*
cerec	ontonotes	0.2162*	0.0707*	0.0615*	0.0604*	0.0638*
GUM	re3d	0.0550*	0.0637*	0.0625*	0.0644*	0.0641*
ontonotes	AnEM	0.0987*	0.0796*	0.0682*	0.0678*	0.0668*
i2b2	wikigold	0.0259	0.0762*	0.0689*	0.0688*	0.0692*
conll	re3d	0.0248	0.0789*	0.0750*	0.0719*	0.0713*
conll	i2b2	0.0752*	0.0833*	0.0762*	0.0748*	0.0729*
wikigold	re3d	0.0392	0.0786*	0.0746*	0.0722*	0.0730*
GUM	sciERC	0.1532*	0.0803*	0.0730*	0.0732*	0.0740*
AnEM	WNUT17	0.0984*	0.0787*	0.0792*	0.0779*	0.0768*
i2b2	GUM	0.1222*	0.0838*	0.0761*	0.0784*	0.0775*
AnEM	BTC	0.0854*	0.0827*	0.0807*	0.0816*	0.0806*
BTC	re3d	0.0494	0.0774*	0.0802*	0.0835*	0.0815*
cerec	AnEM	0.2092*	0.0827*	0.0802*	0.0803*	0.0830*
cerec	i2b2	0.0626*	0.0823*	0.0807*	0.0822*	0.0849*
wikigold	sciERC	0.1274*	0.0896*	0.0858*	0.0850*	0.0862*
WNUT17	re3d	0.0492	0.0950*	0.0911*	0.0882*	0.0882*
i2b2	WNUT17	0.0917*	0.0956*	0.0887*	0.0899*	0.0900*
ontonotes	sciERC	0.1322*	0.1029*	0.0968*	0.0947*	0.0945*
i2b2	BTC	0.0628*	0.0988*	0.0933*	0.0954*	0.0948*
ontonotes	i2b2	0.1087*	0.1157*	0.0961*	0.0964*	0.0970*
AnEM	sciERC	0.1553*	0.0960*	0.0943*	0.0955*	0.0973*
conll	sciERC	0.1178*	0.1059*	0.0987*	0.0969*	0.0976*
cerec	re3d	0.1926*	0.1039*	0.0980*	0.0957*	0.0996*
AnEM	re3d	0.0714*	0.0990*	0.1008*	0.1020*	0.1016*
WNUT17	sciERC	0.1567*	0.1101*	0.1051*	0.1045*	0.1044*
cerec	sec	0.2222*	0.1078*	0.1146*	0.1121*	0.1052*
ontonotes	sec	0.1088*	0.1167*	0.1196*	0.1154*	0.1066*
BTC	sciERC	0.1675*	0.1150*	0.1074*	0.1087*	0.1080*
cerec	sciERC	0.2926*	0.1140*	0.1080*	0.1065*	0.1110*
GUM	sec	0.1057*	0.1218*	0.1251*	0.1207*	0.1114*
conll	sec	0.0858*	0.1108*	0.1224*	0.1202*	0.1121*
wikigold	sec	0.0725*	0.1210*	0.1281*	0.1226*	0.1141*
AnEM	sec	0.1234*	0.1400*	0.1393*	0.1364*	0.1277*
i2b2	re3d	0.0872*	0.1392*	0.1314*	0.1300*	0.1290*
re3d	sciERC	0.0952*	0.1353*	0.1330*	0.1329*	0.1333*
WNUT17	sec	0.1214*	0.1535*	0.1582*	0.1517*	0.1429*
BTC	sec	0.1085*	0.1495*	0.1571*	0.1534*	0.1438*
i2b2	sciERC	0.1871*	0.1556*	0.1424*	0.1439*	0.1440*
re3d	sec	0.0855*	0.1543*	0.1587*	0.1558*	0.1502*
i2b2	sec	0.1229*	0.1608*	0.1619*	0.1599*	0.1506*

Table 3 Chi-squared statistics for all combinations without repetition of datasets. The table is ordered by the chi-squared value following ascending order. All word occurrences that below 5 are filtered for the effective usage of Chi-suared testing. The symbol ♣ indicates there are shifts detected.

Source data	Target data	Statistics	Shift decision
conll	conll	0.0000	
cerec	cerec	0.0000	
ontonotes	ontonotes	0.0000	
i2b2	i2b2	0.0000	
GUM	GUM	0.0000	
AnEM	AnEM	0.0000	
BTC	BTC	0.0000	
WNUT17	WNUT17	0.0000	
wikigold	wikigold	0.0000	
re3d	re3d	0.0000	
sec	sec	0.0000	
sciERC	sciERC	0.0000	
BTC	WNUT17	96.0458	
GUM	WNUT17	123.5303	
GUM	BTC	127.4870	
ontonotes	WNUT17	144.4873	
ontonotes	re3d	149.6937	
GUM	wikigold	191.4220	
wikigold	re3d	232.6331	
GUM	re3d	257.9313	
ontonotes	GUM	292.8625	
AnEM	re3d	302.5694	
AnEM	wikigold	337.4599	
ontonotes	wikigold	346.6776	
GUM	sciERC	349.3116	
GUM	sec	385.1260	
AnEM	WNUT17	401.3902	
i2b2	re3d	426.1499	
re3d	sec	471.7018	
ontonotes	BTC	479.0829	
AnEM	BTC	500.5096	
wikigold	sciERC	544.9597	
AnEM	sec	548.6586	
i2b2	GUM	584.2967	
wikigold	sec	642.1604	
BTC	wikigold	787.5949	
re3d	sciERC	789.3532	
WNUT17	wikigold	794.3760	
ontonotes	AnEM	820.9558	
i2b2	AnEM	823.5791	
ontonotes	sec	853.5255	
AnEM	sciERC	913.3419	
GUM	AnEM	961.5095	
conll	WNUT17	1047.2785	
i2b2	BTC	1086.9625	
conll	cerec	1089.8079	
cerec	re3d	1148.7631	
i2b2	wikigold	1164.4519	
sec	sciERC	1171.7814	
cerec	WNUT17	1378.3443	
WNUT17	sciERC	1434.5123	
WNUT17	re3d	1478.7925	
BTC	sciERC	1510.2569	
conll	BTC	1524.6080	
WNUT17	sec	1656.9017	
cerec	wikigold	1727.7312	♣
ontonotes	sciERC	1728.9888	
cerec	AnEM	1757.9515	♣
i2b2	sec	1777.9979	
cerec	BTC	1879.3686	
conll	sec	2027.9428	
conll	ontonotes	2092.5698	
conll	wikigold	2100.3001	
cerec	sec	2240.8336	♣
conll	i2b2	2257.3186	
i2b2	WNUT17	2653.3429	
i2b2	sciERC	2808.0538	♣
conll	GUM	2877.0674	
conll	re3d	3124.3117	♣
ontonotes	i2b2	3838.2474	
cerec	GUM	4310.9531	♣
BTC	re3d	4513.7231	♣
BTC	sec	4568.5178	♣
cerec	ontonotes	4726.8631	♣
conll	sciERC	6930.8975	♣
cerec	sciERC	7169.8736	♣
conll	AnEM	7186.4441	♣

Table 4 Chi-squared statistics for sampled datasets. The symbol ♣ indicates there are shifts detected.

Source data	Target data	Statistics	Shift decision
conll	conll	0.0000	
cerec	cerec	0.0000	
ontonotes	ontonotes	0.0000	
i2b2	i2b2	0.0000	
GUM	GUM	0.0000	
AnEM	AnEM	0.0000	
BTC	BTC	0.0000	
WNUT17	WNUT17	0.0000	
wikigold	wikigold	0.0000	
re3d	re3d	0.0000	
sec	sec	0.0000	
sciERC	sciERC	0.0000	
GUM	BTC	75.6658	
GUM	WNUT17	81.6450	
BTC	WNUT17	93.8928	
ontonotes	WNUT17	98.6004	
ontonotes	BTC	124.5910	
ontonotes	re3d	131.3946	
AnEM	BTC	158.8629	
GUM	wikigold	162.4105	
i2b2	wikigold	172.9666	
AnEM	WNUT17	189.5393	
wikigold	sciERC	191.4669	
AnEM	re3d	195.1509	
GUM	re3d	207.6499	
i2b2	GUM	211.5066	
wikigold	re3d	213.8132	
GUM	AnEM	214.1968	
GUM	sciERC	244.2752	
ontonotes	wikigold	249.6848	
i2b2	AnEM	258.3698	
conll	AnEM	258.8873	
ontonotes	AnEM	283.8807	
AnEM	wikigold	292.7534	
conll	BTC	295.2595	
i2b2	BTC	296.2821	
re3d	sec	316.5017	
i2b2	re3d	320.1426	
wikigold	sec	327.4453	
ontonotes	GUM	345.1330	
ontonotes	i2b2	357.7485	
GUM	sec	401.7539	
re3d	sciERC	404.8397	
WNUT17	re3d	407.1329	
AnEM	sec	433.0472	
conll	re3d	433.0808	
i2b2	WNUT17	490.4596	
sec	sciERC	501.0867	
cerec	WNUT17	603.9794	
AnEM	sciERC	635.9219	
ontonotes	sec	716.1215	
BTC	wikigold	748.7217	
conll	sec	760.0368	
conll	WNUT17	778.9208	
cerec	re3d	815.4335	♣
conll	GUM	832.8918	
cerec	wikigold	836.1332	♣
WNUT17	wikigold	861.2727	
cerec	i2b2	875.3262	♣
i2b2	sec	881.3150	♣
cerec	AnEM	892.4618	♣
i2b2	sciERC	905.3239	♣
cerec	GUM	1089.9833	♣
conll	i2b2	1425.7854	♣
cerec	sec	1428.6658	♣
BTC	sciERC	1643.8628	♣
cerec	BTC	1795.8086	♣
conll	cerec	1810.5222	♣
cerec	ontonotes	1961.2929	♣
WNUT17	sciERC	2016.5294	♣
conll	wikigold	2145.5425	♣
conll	ontonotes	2348.6941	♣
BTC	re3d	2480.3993	♣
WNUT17	sec	2590.0214	♣
ontonotes	sciERC	2608.5222	♣
conll	sciERC	3051.1489	♣
BTC	sec	4274.0432	♣

Table 5 Label distribution Chi-squared testing statistics for all combinations without repetition of datasets. The table is ordered by the test value in ascending order.

Source data	Target data	Statistics	Shift Decision
conll	conll	0.00	
cerec	cerec	0.00	
ontonotes	ontonotes	0.00	
i2b2-06	i2b2-06	0.00	
GUM	GUM	0.00	
AnEM	AnEM	0.00	
BTC	BTC	0.00	
WNUT17	WNUT17	0.00	
wikigold	wikigold	0.00	
re3d	re3d	0.00	
sec	sec	0.00	
sciERC	sciERC	0.00	
conll	wikigold	0.04	
BTC	sec	0.07	
BTC	wikigold	0.51	
BTC	WNUT17	0.84	
BTC	re3d	1.22	
conll	sec	2.03	
wikigold	sec	4.24	
cerec	sec	326074.25	♣
cerec	re3d	745315.20	♣
cerec	wikigold	781788.25	♣
cerec	WNUT17	841992.77	♣
re3d	sec	1310152.93	♣
cerec	GUM	2564429.92	♣
cerec	BTC	3257665.19	♣
WNUT17	sec	5112738.62	♣
ontonotes	sec	5223820.29	♣
conll	cerec	5780109.23	♣
conll	re3d	6382202.06	♣
conll	WNUT17	7210082.54	♣
WNUT17	re3d	11686258.40	♣
ontonotes	re3d	11940158.70	♣
WNUT17	wikigold	12258168.10	♣
GUM	sec	12466422.81	♣
ontonotes	wikigold	12524494.36	♣
ontonotes	WNUT17	13489000.18	♣
wikigold	re3d	13583255.59	♣
conll	GUM	21959515.63	♣
AnEM	sec	22315222.59	♣
cerec	ontonotes	22872211.25	♣
conll	BTC	27895788.79	♣
GUM	re3d	28494675.69	♣
i2b2-06	sec	29859890.76	♣
GUM	wikigold	29889167.66	♣
GUM	WNUT17	32190918.71	♣
ontonotes	GUM	41083014.26	♣
ontonotes	AnEM	43451712.45	♣
GUM	AnEM	50349808.09	♣
AnEM	re3d	51006213.20	♣
cerec	i2b2-06	51746456.14	♣
ontonotes	BTC	52188909.35	♣
AnEM	wikigold	53502390.51	♣
AnEM	WNUT17	57622585.01	♣
i2b2-06	re3d	68251165.27	♣
i2b2-06	wikigold	71591287.74	♣
ontonotes	i2b2-06	74740292.02	♣
i2b2-06	WNUT17	77104499.80	♣
conll	AnEM	84469257.53	♣
i2b2-06	AnEM	86817785.05	♣
GUM	BTC	124546587.29	♣
ontonotes	sciERC	160417887.32	♣
AnEM	sciERC	174040643.86	♣
GUM	sciERC	185884729.65	♣
cerec	AnEM	189564270.53	♣
conll	ontonotes	195857558.82	♣
conll	i2b2-06	197142187.31	♣
AnEM	BTC	222941642.42	♣
i2b2-06	GUM	234834691.92	♣
WNUT17	sciERC	279748960.06	♣
wikigold	sciERC	293835211.87	♣
re3d	sciERC	294915993.80	♣
i2b2-06	BTC	298317122.09	♣
conll	sciERC	311849153.00	♣
i2b2-06	sciERC	320519601.10	♣
BTC	sciERC	459742109.25	♣
sec	sciERC	555622639.19	♣

Table 6 Label distribution Chi-squared testing statistics for sampled datasets (900 samples).

Source data	Target data	Statistics	Shift Decision
conll	conll	0.00	
cerec	cerec	0.00	
ontonotes	ontonotes	0.00	
i2b2-06	i2b2-06	0.00	
GUM	GUM	0.00	
AnEM	AnEM	0.00	
BTC	BTC	0.00	
WNUT17	WNUT17	0.00	
wikigold	wikigold	0.00	
re3d	re3d	0.00	
sec	sec	0.00	
sciERC	sciERC	0.00	
conll	wikigold	0.04	
BTC	sec	0.10	
BTC	wikigold	0.52	
BTC	WNUT17	0.84	
BTC	re3d	1.11	
conll	sec	1.47	
wikigold	sec	2.60	
cerec	sec	334179.62	♣
cerec	wikigold	839683.36	♣
cerec	re3d	843668.88	♣
cerec	WNUT17	917624.79	♣
re3d	sec	1066254.19	♣
cerec	GUM	2900540.31	♣
cerec	BTC	3687180.00	♣
WNUT17	sec	4377314.31	♣
ontonotes	sec	4780033.07	♣
conll	cerec	5836613.50	♣
conll	re3d	6491937.00	♣
conll	WNUT17	7061035.89	♣
WNUT17	wikigold	10998734.44	♣
WNUT17	re3d	11050923.20	♣
GUM	sec	11530780.73	♣
wikigold	re3d	11602120.49	♣
ontonotes	wikigold	12010632.37	♣
ontonotes	re3d	12067622.20	♣
ontonotes	WNUT17	13125498.34	♣
AnEM	sec	20244050.26	♣
conll	GUM	22319385.77	♣
cerec	ontonotes	26018437.52	♣
i2b2-06	sec	26465192.08	♣
conll	BTC	28372520.38	♣
GUM	wikigold	28973014.22	♣
GUM	re3d	29110489.58	♣
GUM	WNUT17	31662383.01	♣
ontonotes	GUM	41488682.25	♣
ontonotes	AnEM	44092822.88	♣
AnEM	wikigold	50866559.82	♣
GUM	AnEM	50979958.03	♣
AnEM	re3d	51107918.60	♣
ontonotes	BTC	52740634.40	♣
AnEM	WNUT17	55588158.29	♣
cerec	i2b2-06	59943853.08	♣
i2b2-06	wikigold	66498217.78	♣
i2b2-06	re3d	66813748.00	♣
i2b2-06	WNUT17	72670797.12	♣
ontonotes	i2b2-06	74879460.31	♣
conll	AnEM	85589611.44	♣
i2b2-06	AnEM	87686295.03	♣
GUM	BTC	127225200.96	♣
ontonotes	sciERC	158825547.52	♣
AnEM	sciERC	173009643.23	♣
GUM	sciERC	183633507.68	♣
cerec	AnEM	187632540.01	♣
conll	i2b2-06	195573212.31	♣
conll	ontonotes	200209544.04	♣
AnEM	BTC	223363307.19	♣
i2b2-06	GUM	229706748.40	♣
WNUT17	sciERC	279425297.10	♣
i2b2-06	BTC	292004448.24	♣
wikigold	sciERC	294289564.04	♣
re3d	sciERC	305681552.91	♣
conll	sciERC	308299988.60	♣
i2b2-06	sciERC	315852394.93	♣
BTC	sciERC	451014301.26	♣
sec	sciERC	546583192.69	♣

Table 7 F1 scores on sampled datasets (900 samples). Fine-tuning uses the BERT-base model with 10 epochs.

	conll	cerec	ontonotes	i2b2-06	re3d	wikigold	SEC	GUM	BTC	sciERC	WNUT17	AnEM	average f1
conll	**0.66**	0.25	0.24	0.22	0.24	0.44	0.07	0.04	0.25	0.00	0.30	0.00	**0.22**
cerec	0.26	**0.70**	0.08	0.15	0.21	0.24	0.04	0.17	0.10	0.00	0.13	0.00	0.20
ontonotes	0.18	0.01	**0.22**	0.03	0.09	0.02	0.01	0.01	0.04	0.00	0.09	0.00	0.15
i2b2-06	0.14	0.11	0.04	**0.69**	0.04	0.11	0.01	0.01	0.17	0.00	0.09	0.00	0.14
re3d	0.17	0.18	0.12	0.00	**0.54**	0.21	0.06	0.11	0.15	0.00	0.18	0.00	0.13
wikigold	0.52	0.20	0.23	0.11	0.25	**0.62**	0.08	0.08	0.20	0.00	0.22	0.00	0.13
SEC	0.08	0.08	0.05	0.02	0.04	0.12	**0.90**	0.02	0.06	0.00	0.07	0.00	0.13
GUM	0.11	0.12	0.03	0.02	0.13	0.12	0.02	**0.27**	0.08	0.00	0.07	0.00	0.13
BTC	0.41	0.17	0.15	0.17	0.20	0.31	0.04	0.05	**0.66**	0.00	0.21	0.00	0.13
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.30**	0.00	0.00	0.12
WNUT17	0.21	0.16	0.06	0.17	0.07	0.19	0.00	0.03	0.14	0.00	**0.37**	0.00	0.12
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.30**	0.11

Table 8 F1 scores on full-sized datasets. Fine-tuning uses the BERT-base model with 10 epochs.

	conll	cerec	ontonotes	i2b2-06	wikigold	WNUT17	GUM	re3d	SEC	BTC	sciERC	AnEM	average f1
conll	**0.90**	0.35	0.31	0.13	0.62	0.26	0.12	0.33	0.10	0.31	0.00	0.00	**0.29**
cerec	0.35	**0.90**	0.13	0.10	0.31	0.14	0.21	0.19	0.08	0.20	0.00	0.00	0.25
ontonotes	0.28	0.14	**0.93**	0.05	0.26	0.14	0.08	0.17	0.06	0.17	0.00	0.00	0.23
i2b2-06	0.27	0.20	0.12	**0.99**	0.33	0.06	0.05	0.22	0.02	0.16	0.00	0.00	0.22
wikigold	0.57	0.25	0.30	0.10	**0.88**	0.13	0.13	0.30	0.13	0.20	0.00	0.00	0.21
WNUT17	0.43	0.32	0.29	0.15	0.44	**0.70**	0.10	0.24	0.09	0.27	0.00	0.00	0.21
GUM	0.20	0.17	0.05	0.01	0.23	0.04	**0.68**	0.28	0.04	0.06	0.00	0.00	0.21
re3d	0.19	0.25	0.15	0.04	0.19	0.11	0.14	**0.61**	0.06	0.20	0.00	0.00	0.21
SEC	0.22	0.14	0.19	0.05	0.25	0.09	0.05	0.12	**0.93**	0.21	0.00	0.00	0.21
BTC	0.59	0.35	0.32	0.14	0.62	0.24	0.14	0.34	0.15	**0.87**	0.00	0.00	0.20
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.57**	0.00	0.19
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.79**	0.19

Table 9 F1 scores on sampled datasets (900 samples). Fine-tuning uses the BioBERT-base model with 10 epochs.

	conll	cerec	ontonotes	i2b2-06	SEC	re3d	wikigold	sciERC	GUM	BTC	WNUT17	AnEM	average f1
conll	**0.58**	0.18	0.22	0.03	0.14	0.23	0.45	0.00	0.08	0.20	0.16	0.00	**0.19**
cerec	0.25	**0.77**	0.08	0.09	0.26	0.17	0.20	0.00	0.19	0.08	0.10	0.00	0.19
ontonotes	0.03	0.00	**0.26**	0.00	0.02	0.02	0.02	0.00	0.00	0.04	0.04	0.00	0.14
i2b2-06	0.15	0.11	0.02	**0.66**	0.07	0.07	0.08	0.00	0.01	0.11	0.04	0.00	0.13
SEC	0.23	0.21	0.06	0.20	**0.90**	0.07	0.19	0.00	0.02	0.33	0.14	0.00	0.13
re3d	0.16	0.22	0.14	0.01	0.07	**0.50**	0.20	0.00	0.13	0.09	0.09	0.00	0.12
wikigold	0.49	0.16	0.20	0.07	0.18	0.25	**0.57**	0.00	0.08	0.16	0.12	0.00	0.12
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.36**	0.00	0.00	0.00	0.00	0.12
GUM	0.09	0.16	0.04	0.01	0.05	0.15	0.09	0.00	**0.26**	0.07	0.04	0.00	0.12
BTC	0.34	0.19	0.17	0.07	0.27	0.16	0.28	0.00	0.06	**0.59**	0.16	0.00	0.12
WNUT17	0.16	0.14	0.04	0.07	0.20	0.07	0.15	0.00	0.01	0.08	**0.19**	0.00	0.11
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.25**	0.10

Table 10 F1 scores on full-sized datasets. Fine-tuning uses the BioBERT-base model with 10 epochs.

	conll	cerec	ontonotes	i2b2-06	wikigold	WNUT17	GUM	re3d	SEC	BTC	sciERC	AnEM	average f1
conll	**0.88**	0.32	0.26	0.11	0.58	0.21	0.10	0.29	0.36	0.23	0.00	0.00	**0.28**
cerec	0.30	**0.87**	0.11	0.06	0.30	0.12	0.20	0.20	0.31	0.15	0.00	0.00	0.25
ontonotes	0.23	0.14	**0.92**	0.10	0.22	0.12	0.07	0.15	0.07	0.15	0.00	0.00	0.23
i2b2-06	0.33	0.22	0.17	**1.00**	0.36	0.09	0.09	0.26	0.23	0.16	0.00	0.00	0.23
wikigold	0.49	0.25	0.29	0.10	**0.84**	0.11	0.11	0.28	0.38	0.15	0.00	0.00	0.21
WNUT17	0.40	0.24	0.25	0.14	0.43	**0.61**	0.09	0.23	0.19	0.25	0.00	0.00	0.21
GUM	0.15	0.16	0.04	0.03	0.21	0.03	**0.63**	0.26	0.04	0.06	0.00	0.00	0.21
re3d	0.15	0.21	0.11	0.03	0.16	0.07	0.13	**0.53**	0.09	0.11	0.00	0.00	0.20
SEC	0.21	0.19	0.15	0.11	0.18	0.08	0.05	0.10	**0.90**	0.21	0.00	0.00	0.20
BTC	0.50	0.33	0.32	0.14	0.58	0.22	0.12	0.30	0.31	**0.85**	0.00	0.00	0.20
sciERC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.55**	0.00	0.19
AnEM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.77**	0.19