Exploring a Unified Sequence-To-Sequence Transformer for Medical Product Safety Monitoring in Social Media

Shivam Raval[1] * **Hooman Sedghamiz**[1] * **Enrico Santus**[1] **Tuka Alhanai**[2]
Mohammad Ghassemi[3] **Emmanuele Chersoni**[4]
[1] DSIG - Bayer Pharmaceuticals, New Jersey, USA
[2] New York University, Abu Dhabi, UAE
[3] Michigan State University, Michigan, USA
[4] The Hong Kong Polytechnic University, Hong Kong
sbraval@asu.edu, {hooman.sedghamiz,enrico.santus}@bayer.com
tuka.alhanai@nyu.edu, ghassem3@msu.edu
emmanuele.chersoni@polyu.edu.hk

Abstract

Adverse Events (AE) are harmful events resulting from the use of medical products. Although social media may be crucial for early AE detection, the sheer scale of this data makes it logistically intractable to analyze using human agents, with NLP representing the only low-cost and scalable alternative.

In this paper, we frame AE Detection and Extraction as a sequence-to-sequence problem using the T5 model architecture and achieve strong performance improvements over competitive baselines on several English benchmarks (F1 = 0.71, 12.7% relative improvement for AE Detection; Strict F1 = 0.713, 12.4% relative improvement for AE Extraction). Motivated by the strong commonalities between AE-related tasks, the class imbalance in AE benchmarks and the linguistic and structural variety typical of social media posts, we propose a new strategy for multi-task training that accounts, at the same time, for task and dataset characteristics. Our multi-task approach increases model robustness, leading to further performance gains. Finally, our framework shows some language transfer capabilities, obtaining higher performance than Multilingual BERT in zero-shot learning on French data.

1 Introduction

Before market release, drugs are regularly tested for safety and effectiveness in clinical trials. However, since no clinical trial is large enough to find all potential Adverse Events (AEs) on a wide and diverse range of population, Pharmacovigilance continuously monitors the market to timely intervene, in case unexpected AEs are discovered.

According to multiple sources (Sen, 2016; Alatawi and Hansen, 2017), AEs are systematically under-reported in official channels. A growing number of patients, though, talk about them on social platforms like Twitter and health forums, sharing medical conditions, treatment reviews, side effect descriptions and so on. These outlets contain crucial information for Pharmacovigilance, but the sheer scale of this data – velocity, volume, variety – makes manual exploration prohibitively expensive. For this reason, Natural Language Processing (NLP) technologies represent the only low-cost and scalable alternative.

In recent years, the research community approached this problem by promoting thematic workshops and shared tasks, such as the Social Media Mining For Health Applications (SMM4H) (Weissenbacher et al., 2019; Klein et al., 2020), as well as by creating resources, such as CADEC (Karimi et al., 2015). Despite these efforts, the automatic detection of AEs from social outlets has still to face major challenges: i) posts containing AEs are rare compared to other posts (i.e. **rare signal and imbalanced data**); ii) text typologies largely differ across media (i.e. **text length and structure**); iii) **informal and figurative language** is dominant, often containing slang, idioms, sarcasm and metaphors; iv) datasets contain broad **differences in the annotations**, sometimes focusing only on the symptom mentions and others times including temporal, locative and intensity modifiers; v) annotated resources for model fine-tuning are only available for a small set of languages (i.e. **cross-lingualism**).

Most of these challenges have led the research community to develop end-to-end solutions for each task, missing the benefit of performing multi-

* equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3534–3546
November 7–11, 2021.

ple related tasks with the same model (i.e. transfer learning). In this paper, we aim to tackle the above-mentioned challenges all at once by framing the AE detection and AE extraction tasks as generative sequence-to-sequence (seq-to-seq) problems, to be addressed with a single architecture, namely T5 (Raffel et al., 2019). In previous studies, the T5 architecture showed high flexibility in dealing with text from different domains and typologies, even in knowledge-intensive tasks (Petroni et al., 2021). Furthermore, the T5 architecture is capable of incrementally learning new tasks with few or no labels (Xue et al., 2020).

In the following paragraphs, not only we show that T5 outperforms strong baselines on multiple English benchmarks (F1 = 0.71, 10.94% relative improvement for AE detection; Strict F1 = 0.713, 12.46% relative improvement for AE extraction), but, to fully unleash its potential and thereby address the above-mentioned challenges (i.e., small, varied and imbalanced data), we introduce a novel multi-task/data training framework that efficiently handles task complexity, data imbalance and textual differences, further improving over the state-of-the-art results. Assessed in multiple cross-textual and (zero-shot learning) cross-lingual AE detection and AE extraction settings, T5 shows robustness and improves over all the competitive baselines, despite being simpler – in terms of number of layers and parameters – than the competitors.

To summarize, our contributions are: i) we use T5 for framing AE detection and extraction as a sequence-to-sequence problem, obtaining strong performance on multiple tasks and datasets; ii) we describe a new approach for balancing data across tasks and datasets in a multi-task setting, which leads to F1-score improvements on all benchmarks; iii) we test our model in a crosslingual transfer (English to French) scenario, showing that it outperforms Multilingual BERT in zero-shot learning.

2 Related Work

Early efforts in automated Pharmacovigilance have targeted Electronic Health Records (EHR) to detect evidence of AEs (Uzuner et al., 2011; Jagannatha et al., 2019). However, not all AEs lead to clinical visitations: many users prefer to discuss their experiences with drugs on the Internet, and this fact led to a growing interest in the automatic detection of adverse events from social media platforms.

Some of the early machine learning systems for AE detection from social media data used a combination of various classifiers along with word embeddings as features (Sarker and Gonzalez, 2015; Nikfarjam et al., 2015; Daniulaityte et al., 2016; Metke-Jimenez and Karimi, 2016).

After the introduction of challenges such as Social Media Mining for Health Applications (SMM4H) (Weissenbacher et al., 2018, 2019) and CADEC (Karimi et al., 2015), most works focused on neural networks (Sarker et al., 2018; Minard et al., 2018). With the development of attention mechanism (Vaswani et al., 2017), Transformer-based language models such as BERT (Devlin et al., 2019) and its biomedical (e.g., BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019) and PubMedBERT (Gu et al., 2020)) and non-biomedical variants (e.g., SpanBERT (Joshi et al., 2020)) obtained state-of-the-art performance in AE detection (Weissenbacher et al., 2019; Klein et al., 2020; Portelli et al., 2021a,b).

Models like BERT and its variants can be described as encoder-only: in order to carry out a specific task, a decoder has to be followed by task-specific trainable network, most often in the form of a linear layer. Recent developments in NLP led to the introduction of models such as T5 (Raffel et al., 2019), which is an encoder-decoder Transformer architecture. In a series of studies, T5 and its variants have shown performance gain on various datasets and applications (Raffel et al., 2019; Xue et al., 2020), despite being smaller in terms of parameters. The prefix training approach adopted by T5 allows users to fine-tune on various tasks concurrently, creating a single model that can incrementally learn while being capable of performing different tasks simultaneously. To our knowledge, the present contribution is the first to frame AE detection and extraction as generative problems.

3 Methods

3.1 The T5 Model

We employ T5, a pre-trained encoder-decoder transformer proposed by Raffel et al. (2019). This model maps a vector sequence of n input words represented by $\mathbf{X_{1:n}} = \mathbf{x_1}, \cdots, \mathbf{x_n}$ to an output sequence of $\mathbf{Y_{1:m}} = \mathbf{y_1}, \cdots, \mathbf{y_m}$ with an *a-priori* unknown length of m, with a conditional probability defined as:

$$p_{\theta_{model}}(\mathbf{Y}_{1:m}|\mathbf{X}_{1:n}) \quad (1)$$

The architecture of the model is very similar to

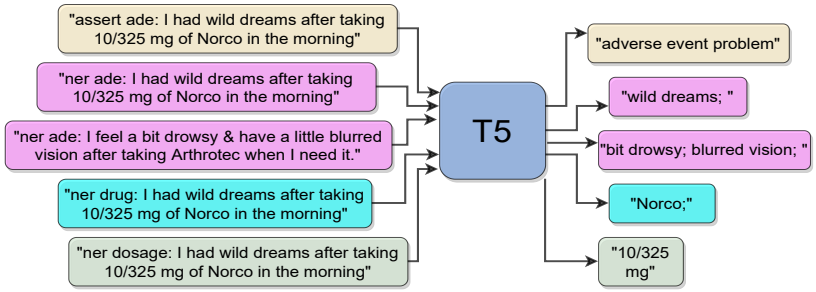

Figure 1: Diagram of our sequence-to-sequence framework, which is a fine-tuned T5 model for four prefix: "assert ade:" (yellow box) for detection task, "ner ade:" (pink box) for the task of extracting AE's , "ner drug:" (blue box) for extracting drug mentions and "ner dosage:" (green box) for extracting drug dosage information from the input.

the original Transformer proposed by Vaswani et al. (2017). An input sequence is first passed to the encoder which consists of self-attention followed by feed-forward layers. The encoder maps the input to a sequence of embeddings that go through normalization and drop-out layers. The decoder attends to the output of the encoder using several attention layers. The self-attention layers, instead, employ masking to make the decoder only attend to the past tokens, in an auto-regressive manner:

$$p_{\theta_{decoder}}(\mathbf{Y}_{1:m}) = \prod_{i=1}^{m} p_{\theta_{decoder}}(\mathbf{y_i}|\mathbf{Y}_{0:i-1}) \quad (2)$$

where $p_{\theta_{decoder}}(\mathbf{y_i}|\mathbf{Y}_{0:i-1})$ is the probability distribution of the next token $\mathbf{y}_i$. Finally, the output of the decoder passes through a SoftMax layer over the vocabulary. Raffel et al. (2019) proposed to add a prefix in front of the input sequence to inform the model about which task to perform (e.g. summarization, question answering, classification etc.; see Figure 1).The model was trained on the Colossal Clean Crawled Corpus (C4), a massive corpus (about 750 GB) of web-extracted and cleaned text.

3.1.1 Pre-Training and Pre-Finetuning

Raffel et al. (2019) explored a wide range of architectures and pre-training objectives, finding that encoder-decoder models generally outperform decoder-only language models, and that a BERT-style denoising objective – where the model is trained to recover masked words in the input – works best. Moreover, the best variant of their system made use of an objective that corrupts contiguous spans of tokens, similarly to the span corruption strategy introduced for the SpanBERT model (Joshi et al., 2020).

The resulting model was then pre-finetuned on a variety of tasks from the following sources: the GLUE (Wang et al., 2018) and the SuperGLUE (Wang et al., 2019) benchmarks for natural language understanding, the abstractive summarization data by Hermann et al. (2015) and Nallapati et al. (2016), the SQUAD question answering dataset (Rajpurkar et al., 2016) and the WMT translation benchmarks for translation from English to French, from English to German and from English to Romanian. The tasks were all treated as a single task in the sequence-to-sequence format, by concatenating all the datasets together and appending the task-specific prefixes to the instances.

T5 comes in versions, *small* (60 million parameters), *base* (220 million parameters), *large* (770 million parametrs), *3B* (3 billion parameters) and *11B* (11 billion parameters). In the paper we will use the term T5 to either refer to the architecture or to the T5-Base, as opposed to T5-Small, which will always be mentioned as such.

3.2 Seq-to-Seq AE-related Tasks

Given an input sequence of words $\mathbf{X_{1:n}} = \mathbf{x_1}, \cdots, \mathbf{x_n}$ that potentially contains drug, dosage and AE mentions, we frame the AE detection (i.e. binary classification) and extraction (i.e. span detection) tasks as seq-to-seq problems, further fine-tuning T5 to generate $\mathbf{Y_{1:m}} = \mathbf{y_1}, \cdots, \mathbf{y_m}$, where Y is either the classification label or the text span with the AE. By selecting the prefixes (see Table 1), we train T5 on all these tasks (see Figure 1).

Prefix	Task Definition	Task Type
assert ade	Contains AE or not	CLS (binary)
ner ade	Extract AE span	NER (span)
ner drug	Extract drug span	NER (span)
ner dosage	Extract drug dosage span	NER (span)

Table 1: Prefix and task definition. AE assertion is binary classification (CLS), while the remaining tasks are Name Entity Recognition (NER).

For the AE detection, as it is a binary classification task, we have chosen the prefix "assert ade:" and the labels i) "adverse event problem" (i.e., positive) and ii) "health ok" (i.e., negative). Usually, to train Named Entity Recognition (NER) systems, the input data is transformed into standard Inside–Outside–Beginning (IOB) format and individual tokens are classified in one of the IOB tags. However, the T5 model can utilize the direct span as a generation target. If multiple spans can be extracted, they can be provided to the system separated by a semicolon or other special characters. For our experiments on language transfer, we simply apply the "assert ade: " to data in a different language (i.e., French). The model will automatically leverage the knowledge acquired during the *pre-finetuning* in the machine translation task. Tasks and definitions are summarized in Table 1.

3.3 Multi-Task and Multi-Dataset Fine-Tuning

Generative models like T5 can be easily trained on multiple tasks. However, multi-task learning poses challenges as models may overfit or underfit, depending on the task difficulty, the label distribution and the variability across tasks and datasets (Arivazhagan et al., 2019). In Raffel et al. (2019), proportional mixing and temperature scaling training strategies were adopted to address the data balance across tasks. In this work, we extend these strategies to a **multi-dataset scenario**, in which tasks are trained on multiple datasets containing heterogeneous data. This scenario is typical in AE detection, where data comes from medical blogs, forums, tweets and other social media outlets, each of which carries specific writing styles as well as different textual structures and lengths. The annotation scheme may differ too across datasets, with some schemes focused on the symptoms only, while others including also the temporal, manner and intensity modifiers.

We assume a multinomial probability distribution θ_t over the fine-tuning task t, given that the fine-tuning task t itself is comprised of dataset(s) d. We define M_d as the number of samples of dataset d and ρ_d the probability of drawing an example from d during training.

In *proportional mixing*, we intuitively sample in proportion to the dataset size. Therefore, the probability of drawing a sample from task t is computed as $\theta_t = \frac{min(\gamma_t, N_t)}{\sum_t min(\gamma_t, N_t)}$, where N_t corresponds to the number of samples available for task t across all datasets, computed as $N_t = \sum_d M_d$. The probability of drawing from dataset d is similarly estimated as $\rho_d = \frac{min(\gamma_d, M_d)}{\sum_t min(\gamma_d, M_d)}$. For the sake of algorithm re-utilization, these parameters γ_d and γ_t are introduced because, even with proportional mixing, large datasets may still dominate the training. These parameters are meant to limit the impact of such large datasets and they have been set to 2^{14} as in the original paper (Raffel et al., 2019).

Temperature scaling has also been shown to boost multi-task training performance (Raffel et al., 2019; Goodwin et al., 2020). It was used for Multilingual BERT, to make sure that the model had sufficient training on low-resource languages (Devlin et al., 2019). To implement scaling with a temperature $\mathcal{T}$, the mixing rate for each task and dataset is raised to the power of $1/\mathcal{T}$, and then the rates are re-normalized so that they sum to 1. Therefore, initially, the probabilities are computed with temperature scaling, respectively, as $\theta_t = \frac{\sqrt[\mathcal{T}]{\theta_t}}{\sum_t \sqrt[\mathcal{T}]{\theta_t}}$ (for the probability of drawing from task t) and as $\rho_d = \frac{\sqrt[\mathcal{T}]{\rho_d}}{\sum_d \sqrt[\mathcal{T}]{\rho_d}}$ (for the probability of drawing from dataset d). We set T as 2 as it is the best reported value for temperature scaling strategy demonstrated in Raffel et al. (2019) and Goodwin et al. (2020).

To assess the value of using multi-dataset sampling, in our experiments we will compare the original proportional mixing and temperature scaling by Raffel et al. (2019) with our approach.

4 Experimental Settings

General figures for all the datasets are reported in Table 2, while more detailed textual statistics are available in Appendix A. More details about the training can be found in Appendix B.

4.1 Datasets

SMM4H This dataset was introduced for the Shared Tasks on AE in the Workshop on Social Media Mining for Health Applications (SMM4H) (Weissenbacher et al., 2018). The dataset is composed of Twitter posts, typically short, informal texts with non-standard ortography, and it contains annotations for both detection (i.e., Task 1, classification) and extraction (i.e., Task 2, NER) of AEs. The number of samples differs from the original dataset as many tweets vanished, due to deletion or access restriction in the platform. Splits are stratified, to maintain an equal ratio of positive and negative examples (see Table 2).

CADEC CADEC contains 1,250 medical forum posts annotated with patient-reported AEs. In this dataset, texts are long and informal, often deviating from English syntax and punctuation rules. Forum posts may contain more than one AE. For our goals, we adopted the training, validation, and test splits proposed by Dai et al. (2020) (see Table 2).

ADE corpus v2 This dataset (Gurulingappa et al., 2012) contains case reports extracted from MEDLINE and it was used for multi-task training, as it contains annotations for all tasks in Table 1, i.e. drugs, dosage, AE detection and extraction. Splits are stratified, to maintain an equal ratio of positive and negative examples (see Table 2).

WEB-RADR This dataset is a manually curated benchmark based on tweets. We used it exclusively to test the performance of the multi-task models, as it was originally introduced only for testing purposes (Dietrich et al., 2020) (see Table 2).

Dataset	Total	Positive	Negative
SMM4H Task 1 (AE Detection)	15,482	1,339	14,143
Train (80%)	12,386	1,071	11,315
Validation (10%)	1,548	134	1,414
Test (10%)	1,548	134	1,414
SMM4H Task 2 (AE Det., AE & Drug Extr.)	2,276	1300	976
Train (60%)	1,365	780	585
Validation (20%)	455	260	195
Test (20%)	456	260	196
CADEC (AE Det., AE & Drug Extr.)	1,250	1,105	145
Train (70%)	875	779	96
Validation (15%)	187	163	24
Test (15%)	188	163	25
ADE Corpus v2 (AE Detection)	23,516	6,821	16,695
Train (60%)	14,109	4,091	10,018
Validation (20%)	4,703	1,365	3,338
Test (20%)	4,704	1,365	3,339
ADE Corpus v2 (AE Extraction)	6,821	6,821	0
Train (60%)	4,091	4,091	0
Validation (20%)	1,365	1,365	0
Test (20%)	1,365	1,365	0
ADE Corpus v2 (Drug Extraction)	7,100	7,100	0
Train (60%)	4,260	4,260	0
Validation (20%)	1,420	1,420	0
Test (20%)	1,420	1,420	0
ADE Corpus v2 (Drug Dosage Extraction)	279	0	0
Train (60%)	167	0	0
Validation (20%)	56	0	0
Test (20%)	56	0	0
WEB-RADR (AE Detection & Extraction)			
Test	57,481	1,056	56,425
SMM4H-French (AE Detection)			
Test	1,941	31	1,910

Table 2: Dataset Statistics and Splits.

SMM4H-French The SMM4H French Dataset contains a total of 1,941 samples out of which 31 samples belong to AE (positive) class and 1,910 samples have the label Non-AE (negative class). This dataset is only used for testing the zero-shot transfer (see Table 2).

4.2 Settings

AE Detection We train and test T5 and the baselines (see 4.3.1) on the SMM4H Task 1 dataset. We then assess the robustness of T5 and the best performing baseline on the test sets of CADEC, ADE Corpus v2 and WEB-RADR.

AE Extraction We train and test T5 and the baselines (see 4.3.2) on the SMM4H Task 2 dataset. We then assess the robustness of T5 and the best performing baseline by testing them (trained on either SMM4H Task 2 or CADEC) on the test sets of SMM4H Task 2, CADEC, ADE Corpus v2 and WEB-RADR.

Multi-Task Learning We train T5-Base on all the training sets for all tasks, using proportional mixing or temperature scaling both with the original multi-task (see 4.3.3) and with our proposed multi-task and multi-dataset approach, and we evaluate the resulting models on the available test sets.

Language Transfer We train T5 and the Multilingual BERT (see 4.3.4) on the SMM4H Task 1 English dataset, and then we test it in a zero-shot learning setting on the SMM4H-French dataset.

4.3 Baselines

4.3.1 AE Detection

Our baselines are five pre-trained BERT variants with a classification head fine-tuned for AE detection. A weighted cross-entropy loss function is used for all of them to adjust for class imbalance.
BioBERT (Lee et al., 2020) was built upon the original BERT and further pre-trained on PubMed abstracts. We used BioBERT v1.1, which was reported to perform better in biomedical tasks.
BioClinicalBERT (Alsentzer et al., 2019) was pre-trained on MIMIC III dataset containing Electronic Health Records (EHR) of ICU patients.
SciBERT (Beltagy et al., 2019) was pre-trained on 1.14 million papers, randomly selected from semantic scholar, with an 18-82 ratio between computer science and biomedical papers.
PubMedBERT (Gu et al., 2020) was pre-trained

from scratch on PubMed abstracts, without building upon the vocabulary of the original BERT.
SpanBERT (Joshi et al., 2020) adopts a different pre-training objective from BERT. This model is trained by masking full contiguous spans instead of single words or subwords, which allows it to encode span-level information.

4.3.2 AE Extraction Task Baselines

For the AE EXTRACTION task, we use the four models described in Portelli et al. (2021a), namely BERT, BERT+CRF, SpanBERT, and SpanBERT+CRF. The authors reported state-of-the-art performance with the SpanBERT models on SMM4H, and their implementation is publicly available at https://github.com/ailabUdineGit/ADE.

4.3.3 Multi-Task Learning

For Multi-Task Learning, we use as baseline the T5 model fine-tuned with the original training strategies by Raffel et al. (2019), which balance across tasks (TB, task balancing) but do not account for multi-dataset learning (DB, dataset balancing). We refer to them as $T5_{TB\text{-}PM}$ for proportional mixing and $T5_{TB\text{-}TS}$ for temperature scaling. We refer to our approach, which accounts also for the multi-dataset learning, as $T5_{TDB\text{-}PM}$ for proportional mixing and $T5_{TDB\text{-}TS}$ for temperature scaling.

4.3.4 Language Transfer

As a baseline for Language Transfer, we use Multilingual BERT (the uncased version), which was pre-trained on monolingual corpora in 102 languages (Devlin et al., 2019). The model was fine-tuned by adding a classification head on the top to perform AE Detection in a zero-shot setting.

4.4 Metrics

We adopt the same metrics of the SMM4H competition. [1] For the AE Detection (i.e., the *assert ade* prefix) we use precision, recall, and F1-score for the positive (AE) class. For the AE Extraction (i.e., the *ner ade*, *ner drug*, *ner dosage* prefixes) we use both Strict and Partial Match F1-Score (Weissenbacher et al., 2019; Klein et al., 2020). The same AE Detection and AE Extraction metrics have also been used in the Multi-Task setting and in the Language Transfer settings.

[1] https://competitions.codalab.org/competitions/20798

5 Results and Analysis

5.1 AE Detection

Table 3 summarizes precision, recall and F1 score obtained by T5-Small, T5-Base and the baselines on the SMM4H Task 1 test set.

Model	Precision	Recall	F1
BioBERT	55.5	63.1	59.0
BioClinicalBERT	68.3	59.7	63.7
SciBERT	68.8	55.9	61.7
PubMedBERT	59.7	61.9	60.8
SpanBERT	55.0	73.1	62.8
T5-Small	58.1	65.0	61.3
T5-Base	**68.8**	**73.7**	**71.1**

Table 3: Precision, Recall and F1-Score for the positive AE class in the SMM4H Task 1 test set

T5-Small obtains competitive performance, lagging slightly behind the performance of some BERT variants, while T5-Base outperforms all the other approaches, with a 12.7% relative F1-score improvement over the best baseline, BioClinicalBERT (the improvement for the McNemar test is significant at $p < 0.001$). It should also be noticed that the two versions of T5, together with SpanBERT, improve over the Recall of the other BERT variants. The result seems to comply with the report by Portelli et al. (2021a,b), who found that models relying on span-based objectives had increased recall in the task, probably because they are better at identifying longer AE spans that would otherwise go undetected.

Model\ Test set	SMM4H Task 2	CADEC	ADE Corpus v2	WEB-RADR
BioClinical BERT	82.5	90.1	28.6	32.3
T5-Base	**88.0**	**93.7**	**31.7**	**35.8**

Table 4: F1-Score for T5-Base and BioClinicalBERT trained on SMM4H Task 1 and tested on all datasets.

Table 4 provides the results for the model generalization evaluation that we run for T5 and the best baseline. In this evaluation, we train the systems on SMM4H Task 1 and test on the other datasets (i.e. SMM4H Task 2, CADEC, ADE Corpus V2 and WEB-RADR), which differ from the training set in terms of linguistic features, text structures, text lengths and even annotation schemes. Both models obtain high performance on SMM4H Task 2 and CADEC, despite their textual differences. The large linguistic difference of the ADE Corpus v2

	SMM4H Task 2		CADEC	
Architecture	Partial F1	Strict F1	Partial F1	Strict F1
BERT	66.1	55.9	77.7	65.2
BERT+CRF	68.1	59.5	77.2	64.4
SpanBERT	66.7	59.2	79.2	67.2
SpanBERT + CRF	70.1	63.4	**79.4**	67.6
T5-Small	70.7	66.1	75.6	65.7
T5-Base	**75.1**	**71.3**	79.1	**69.8**
Dai et al. (2020)	-	-	-	69.0

Table 5: Partial and Strict F1 score for the AE Extraction task on SMM4H Task 2 and CADEC. For CADEC, we also report the current SOTA model by Dai et al. (2020).

(i.e., MEDLINE case reports) explains instead the drastic drop in performance for both systems in this dataset, in which T5-Base still performs better than the baseline. WEB-RADR also proves to be a challenging benchmark for its extreme class imbalance, but our system still achieves an F1-score around 0.36 for the positive class, while BioClinicalBERT is performs than the T5-Base.

5.1.1 Qualitative Analysis on SMM4H

In order to better understand the model performance, we picked some samples from the SMM4H Task 1 test dataset to compare between captured and non-captured AE and analyze the reason behind the miss-classification. In few cases, the model has problems identifying non-standardized acronyms, for example the input "really bad RLS from <drug name>", is classified as non-AE by the model compared to its original label as an AE. The model is not able to understand the meaning behind RLS, which denotes Restless Leg Syndrome in this scenario. We observed that if the RLS is changed to nightmares, headache or restless leg syndrome, the model recognizes the input as an AE. The model is able to capture most of the AE unusual references such as "<drug name> burns like thousand suns", "<drug name> was a joke", "<drug name> tastes like battery acid". Yet we found some cases in which the model failed. For example, the inputs "stomach feels like a cement mixer after taking <drug name>" was classified as non-AE. In this case, "cement mixer" is used in a figurative way to refer to the fact that the stomach is not well or it is churning. Once we replace this figurative image with a term such as churning, the model correctly classifies the sample as AE.

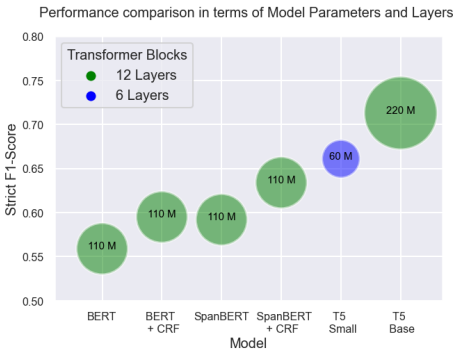

Figure 2: Performance in the AE Extraction task, with the number of layers and parameters for each system.

5.2 AE Extraction

Table 5 summarizes the results for the AE Extraction task for T5 and the baselines trained on SMM4H Task 2, including the scores for a recent SOTA system on CADEC (Dai et al., 2020). It can be seen that both T5 models outperform all the baselines on the SMM4H data, while on the longer and more structured CADEC texts the SpanBERT architectures are more competitive for the partial F1-score. On the other hand, our best model still retains a better performance for the Strict F1 metric, suggesting that it is more accurate in detecting the boundaries of the AE span. T5-Base also outperforms the system by Dai et al. (2020).

Model	SMM4H Task 2	CADEC	ADE Corpus v2	WEB-RADR
Trained on SMM4H Task 2				
SpanBERT + CRF	70.1 (63.4)	15.7 (2.8)	24.6 (15.1)	18.9 (7.3)
T5-Base	**75.1 (71.3)**	**24.4 (20.5)**	**38.9 (29.5)**	**36.2 (13.9)**
Trained on CADEC				
SpanBERT + CRF	35.4 (28.6)	79.4 (67.6)	31.2 (24.8)	20.1 (7.9)
T5-Base	**57.9 (51.6)**	79.1 (69.8)	**50.3 (43.7)**	**30.4 (18.8)**

Table 6: Partial (strict) F1-scores for T5-Base and SpanBERT+CRF trained on SMM4H Task 2 and CADEC and evaluated on all datasets.

In order to further evaluate the system general-

Text Statistics	BERT	BERT+CRF	SpanBERT	SpanBERT+CRF	T5-Base
Dale Chall Readability^{+}	8.34	8.15	8.20	8.32	**9.44**
Automated Readability^{+}	8.42	8.37	8.35	8.51	**10.37**
Flesch Reading Ease^{-}	62.73	63.57	62.60	62.92	**53.18**

Table 7: Text Statistics metric to evalute the quality of span generated by models trained on SMM4H Task 2 dataset ($^{+}$ represents higher score is better and $^{-}$ means lower score is better)

Task	Model/Dataset	SMM4H Task 1	SMM4H Task 2	CADEC	ADE Corpus v2	WEB-RADR	Avg. Score
assert ade	$T5_{TB\text{ -PM}}$	55.2	**91.5**	92.7	91.7	31.9	72.6
	$T5_{TDB\text{ - PM}}$	**67.9**	88.5	**98.7**	91.7	**37.4**	**76.8**
	$T5_{TB\text{ -TS}}$	65.3	83.5	91.1	90.9	36.1	73.3
	$T5_{TDB\text{ - TS}}$	**69.4**	**89.4**	**98.7**	**91.5**	**37.3**	**77.2**
ner ade	$T5_{TB\text{ -PM}}$	-	75.7 (71.8)	46.5 (39.9)	58.4 (53.4)	38.6 (15.1)	54.8 (45.0)
	$T5_{TDB\text{ - PM}}$	-	75.7 (71.8)	**74.4 (64.0)**	**59.7 (55.9)**	**38.7 (15.8)**	**62.1 (51.8)**
	$T5_{TB\text{ -TS}}$	-	75.3 (70.2)	45.2 (38.4)	59.7 (56.1)	38.9 (15.6)	54.7 (45.0)
	$T5_{TDB\text{ - TS}}$	-	**75.7 (71.1)**	**75.3 (66.0)**	**60.3 (56.7)**	**39.1 (15.8)**	**62.6 (52.4)**
ner drug	$T5_{TB\text{ -PM}}$	-	**92.3 (92.3)**	88.7 (88.7)	79.4 (79.0)	-	86.8 (86.6)
	$T5_{TDB\text{ - PM}}$	-	90.3 (90.3)	**92.4 (91.8)**	**82.2 (82.0)**	-	**88.3 (88.0)**
	$T5_{TB\text{ -TS}}$	-	88.2 (88.1)	88.4 (88.1)	80.2 (79.8)	-	85.6 (85.3)
	$T5_{TDB\text{ - TS}}$	-	**91.8 (91.8)**	**94.1 (93.4)**	**83.1 (82.8)**	-	**89.6 (89.3)**
ner dosage	$T5_{TB\text{ -PM}}$	-	-	-	73.2 (67.8)	-	73.2 (67.8)
	$T5_{TDB\text{ - PM}}$	-	-	-	**78.5 (71.4)**	-	**78.5 (71.4)**
	$T5_{TB\text{ -TS}}$	-	-	-	76.7 (71.4)	-	76.7 (71.4)
	$T5_{TDB\text{ - TS}}$	-	-	-	**78.5 (71.4)**	-	**78.5 (71.4)**

Table 8: F1-scores for the multi-task setting. Task Balancing (TB) is compared to our Task and Dataset Balancing (TDB) approach, with PM = Proportional Mixing and TS = Temperature Scaling. F1 of the positive class is reported for AE Detection (the *assert ade* row), while *partial (strict)* F1 is reported for the Extraction tasks.

ization capability, we test on all the AE Extraction datasets both T5-Base and SpanBERT+CRF (best baseline), after training them on SMM4H Task 2 and on CADEC. In Table 6, it can be seen that T5-Base has better generalization than the baseline on all datasets, with F1-scores that are 10 points higher or more. Training on CADEC generalizes better (with the only exception of the partial metric for WEB-RADR), while systems trained on the SMM4H perform poorly on the other benchmarks.

Fig. 2 compares the baselines and the T5 performance in AE extraction, in terms of number of layers/parameters. The plot suggests that the model parameters and the number of layers are not the factors for the T5 models performance gain, e.g. T5-Small has almost half the number of parameters (60 million) and half of the layers (6 layers) of BERT and its variants and it still performs better.

5.2.1 Analysis of Extracted Spans on SMM4H

We employ some commonly used text statistics to assess the spans extracted by the T5 model. Table 7 compares three text statistics metrics for the model trained on the SMM4H Task 2 dataset. The higher scores obtained by T5 in the Dale Chall Readability (Chall and Dale, 1995) and Automated Readability index (Smith and Senter, 1967) suggest this model is able to generate a higher percentage of AE spans with rare terms. The lower Flesch Reading score (Kincaid et al., 1975), instead, indicates that the model generates spans that are more readable.

5.3 Multi-Task Learning

Table 8 includes the scores on all the test sets for the multi-task T5 models, trained either with the original or with our proposed strategy (see 3.3).

In AE Detection, our $T5_{TDB}$ approach always outperforms the original $T5_{TB}$ by a large margin (5.8% relative improvement for PM and 5.3% for TS), except for the Proportional Mixing case in SMM4H Task 2.Margins are smaller in ADE Corpus V2 and WEB-RADR. Looking at the comparison between TS and PM for T5, the former is better in the SMM4H subsets and comparable in all the others, globally obtaining a higher average score.

Our training approach improves both partial and

strict F1-scores on the AE Extraction task, where the models are tested on all datasets except for SMM4H Task 1, which does not have AE Extraction annotations (13.3% relative improvement for PM and 14.4% for TS). In all datasets, our training strategies obtain equal or superior performance for both partial and strict F1 scores, with large gains on CADEC and more marginal gains on SMM4H Task 2, ADE Corpus v2 and WEB-RADR. TS is again preferable to PM, obtaining a higher average score. The results for the Drug and Dosage tasks are similar: in Drug Extraction, SMM4H Task 2 confirms to be more challenging for $T5_{\text{TDB-PM}}$ ($T5_{\text{TB-PM}}$ outperforms it by 2 points), while $T5_{\text{TDB-TS}}$ outperforms its counterpart. In all the other settings, the Task and Dataset Balancing approaches score higher than Task Balancing-only ones.

Overall, our approaches consistently achieve gains in the multi-task setting, independently from task type (i.e. Detection or Extraction) and annotation scheme. TS proves to be superior to PM in all tasks, even though it may lag slightly behind PM in some datasets.

5.4 Cross-lingual Transfer

As a final evaluation, we tested the ability of T5-Base and Multilingual BERT to generalize the AE Detection task to a new language, i.e. French. Notice that the SMM4H French data proved to be challenging, due to the extreme class imbalance (Klein et al., 2020). It can be seen in Table 9 that T5-Base obtains higher F1-score, specifically thanks to a higher precision. Multilingual BERT, instead, shows higher recall. Overall, the T5-Base performance in zero-shot learning is encouraging, and further improvements are likely to come with few shot learning or with more targeted strategies for multilingual training.

Architecture	Zero-Shot		
	Precision	Recall	F1
Multilingual BERT	10.2	**32.2**	15.5
T5-Base	**17.9**	22.6	**20.0**

Table 9: Metrics for Multilingual BERT and T5-Base on zero-shot learning on SMM4H-French.

6 Conclusions

In order to address several typical challenges of the healthcare domain (small, imbalanced and highly variable datasets, cross-lingual data), we proposed to treat AE Detection and AE/Drug/Dosage Extraction tasks as sequence-to-sequence problems, adapting the T5 architecture and improving over all the baselines in both the Detection and the Extraction tasks. To maximize the benefit of multi-task and multi-dataset learning, we introduced a new training strategy that extends Raffel et al. (2019), showing that our approach accounts for multiple and diverse datasets and leads to consistent improvements over the original T5 proposal. Finally, the model also shows some language transfer abilities in the zero shot setting, leaving the door open for future experiments to extend our training framework towards multilinguality (Xue et al., 2020).

Acknowledgments

We would like to thank the reviewers and the chairs for their insightful reviews and suggestions.

References

Yasser M Alatawi and Richard A Hansen. 2017. Empirical Estimation of Under-reporting in the US Food and Drug Administration Adverse Event Reporting System (FAERS). *Expert Opinion on Drug Safety*, 16(7):761–767.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the NAACL Workshop on Clinical Natural Language Processing*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, Yonghui Wu, and Google AI. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *arXiv preprint arXiv:1907.05019*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint arXiv:1903.10676*.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An Effective Transition-based Model for Discontinuous NER. In *Proceedings of ACL*.

Raminta Daniulaityte, Lu Chen, Francois R Lamy, Robert G Carlson, Krishnaprasad Thirunarayan, and Amit Sheth. 2016. "When 'Bad' Is 'Good'": Pdentifying Personal Communication and Sentiment in Drug-related Tweets. *JMIR Public Health and Surveillance*, 2(2):e162.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Juergen Dietrich, Lucie M Gattepaille, Britta Anne Grum, Letitia Jiri, Magnus Lerch, Daniele Sartori, and Antoni Wisniewski. 2020. Adverse Events in Twitter-development of a Benchmark Reference Dataset: Results from IMI WEB-RADR. *Drug Safety*, pages 1–12.

Rudolf Flesch and Alan J Gould. 1949. *The Art of Readable Writing*, volume 8. Harper New York.

Travis R Goodwin, Max E Savery, and Dina Demner-Fushman. 2020. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. In *Proceedings of EMNLP*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779*.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-related Adverse Effects from Medical Case Reports. *Journal of Biomedical Informatics*, 45(5):885–892.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. *arXiv preprint arXiv:1506.03340*.

Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). *Drug Safety*, 42(1):99–111.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A Corpus of Adverse Drug Event Annotations. *Journal of Biomedical Informatics*, 55:73–81.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog count and Flesch reading ease Formula) for Navy Enlisted Personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, et al. 2020. Overview of the Social Media Mining for Health Applications (# SMM4H) Shared Tasks at COLING 2020. In *Proceedings of the COLING Social Media Mining for Health Applications Workshop & Shared Task*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2016. Concept Identification and Normalisation for Adverse Drug Event Discovery in Medical Forums. In *BMDID@ ISWC*. Citeseer.

Anne-Lyse Minard, Christian Raymond, and Vincent Claveau. 2018. IRISA at SMM4H 2018: Neural Network and Bagging for Tweet Classification. In *Proceedings of the EMNLP Workshop on Social Media Mining for Health Applications*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive Text Summarization Using Sequence-to-sequence RNNs and Beyond. In *Proceedings of CONLL*.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from Social Media: Mining Adverse Drug Reaction Mentions Using Sequence Labeling with Word Embedding Cluster Features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. KILT: A Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of NAACL*.

Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. 2021a. BERT Prescriptions to Avoid Unwanted Headaches: A Comparison of Transformer Architectures for Adverse Drug Event Detection. In *Proceedings of EACL*.

Beatrice Portelli, Daniele Passabi, Edoardo Lenzi, Giuseppe Serra, Enrico Santus, and Emmanuele Chersoni. 2021b. Improving Adverse Drug Event Extraction with SpanBERT on Different Text Typologies. In *Proceedings of the AAAI International Workshop on Health Intelligence (W3PHIAI)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQUAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of EMNLP*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and Systems for Medication-related Text Classification and Concept Normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 Shared Task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Abeed Sarker and Graciela Gonzalez. 2015. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-corpus Training. *Journal of Biomedical Informatics*, 53:196–207.

Sukanta Sen. 2016. Consumer Reporting of Adverse Drug Reactions: A Current Perspective. *International Journal of Green Pharmacy (IJGP)*, 10(03).

Edgar A Smith and RJ Senter. 1967. Automated Readability Index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-purpose Language Understanding Systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the ACL Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In *Proceedings of the ACL Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the Social Media Mining for health (SMM4H) Shared Tasks at EMNLP 2018. In *Proceedings of the EMNLP Workshop on Social media Mining for Health Applications Workshop & Shared Task*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A Massively Multilingual Pre-trained Text-to-text Transformer. *arXiv preprint arXiv:2010.11934*.

A Dataset Textual Statistics

Table 10 presents textual statistics to show the difference in type of datasets with respect to their input sequence length, target (extraction) span sequence length and other parameters. It can observed that the input sequence length is relatively short for the SMM4H and for WEB-RADR datasets, while CADEC and ADE corpus datasets tend to include longer texts. The Flesch reading ease score (Flesch and Gould, 1949) indicates the readability of the sentence, with lower values representing that the text is difficult to understand for the average reader. The ADE corpus datasets have the lowest Flesch reading score, as the text is adopted from MEDLINE and contains more medical terms, while Twitter data (SMM4H, WEB-RADR) and the health forum (CADEC) datasets contain a lower amount of scientific terminology and are typically made of shorter texts, with a lower degree of syntactic complexity.

B Training Details

All the experiments have been performed on the top of Hugging-face's Python package (Wolf et al., 2019). [2] The code for the models implemented in the paper is available at https://github.com/shivamraval98/MultiTask-T5_AE

B.1 AE Detection

The baseline BERT models for AE detection were trained on one NVIDIA Tesla V100 16 GB GPU and it takes the model approximately 30 minutes to execute for all epochs. The hyperparameters used for baseline models are detailed in Table 11.

Model	Epoch	Batch Size	Warm-up Steps
BioBERT	3	32	400
BioClinicalBERT	5	40	500
SciBERT	5	40	400
PubMedBERT	5	40	300
SpanBERT	3	40	400

Table 11: Hyperparameters for AE Detection baselines. The learning rate and weight decay was kept constant with values $5e-05$ and 0.01 respectively

The T5 models were trained using a cluster of four NVIDIA Tesla V100 16 GB GPU, with 80 batch size per GPU and 10 epochs for T5-Small, and 16 batch size per GPU and 7 epochs for T5-Base. The learning rate for the both the t5 models was set to $1e-04$. The input and the generated sequence length were set to 130 and 20, respectively, with exponential length penalty set to 2 for the generated sequence. For the rest of the hyperparameters, we used the default values in the library.

The T5-Small model approximately takes 3-5 minutes per epoch while T5-Base executes for 7-10 minutes per epoch in the aforementioned cluster environment setting.

B.2 AE Extraction

The hyperparameters for the baseline models (BERT, BERT+CRF, SpanBERT and SpanBERT+CRF) of AE extraction were set as described in Portelli et al. (2021a). The hyperparame-

[2] https://github.com/huggingface/transformers

Dataset	Avg. Seq Length	Avg. Span Length (AE, Drug or Dosage)	Avg. Stopwords in span	Avg. Freq. of AE per sample	Unique AE words	% of AE Samples	Unique Drug Mentions	Flesch Reading Ease Score
SMM4H Task 1 (AE Detection)	98.9	-	-	-	-	8.6	-	64.7
SMM4H Task 2 (AE Det., AE & Drug Extr.)	108.8	9.1	0.2	1	1108	57.1	69	62.1
CADEC (AE Det., AE & Drug Extr.)	459.4	16.1	2.4	6	2303	89.0	320	69.1
ADE Corpus v2 (AE Detection)	132.5	-	-	-	-	28.9	-	23.2
ADE Corpus v2 (AE Extraction)	152.1	18.5	0.1	1	2662	100	-	13.6
ADE Corpus v2 (Drug Extraction)	152.3	10.8	0	-	-	100	1251	14.3
ADE Corpus v2 (Drug Dosage Extraction)	163.4	8.5	0	-	-	100	-	23.6
WEB-RADR (AE Detection & Extraction)	106.3	16.5	1.1	2	2037	1.8	-	61.3
SMM4H French (AE Detection)	142.4	-	-	-	-	1.6	-	-

Table 10: Comparison of the AE datasets according to different textual statistics.

ter setting for the T5-Small and T5-Base for both SMM4H Task 2 and CADEC dataset is presented in Table 12 and the default values were utilized for the rest of the hyperparameters.

Model	ISL	OSL	BS	EP	LR	Time
SMM4H Task 2 AE Extraction						
T5-Small	130	20	80	10	1e-4	5
T5-Base	130	20	64	7	1e-4	7
CADEC AE Extraction						
T5-Small	512	150	64	25	1e-3	10
T5-Base	512	150	32	20	1e-3	20

Table 12: Hyperparameters for T5-Small and T5-Base when trained on SMM4H and CADEC AE Extraction Task (ISL = Input Sequence Length, OSL = Output Sequence Length, BS = Batch Size (over all GPU's), EP = Epoch, LR = Learning Rate, Time = Training Time in mins per epoch).

B.3 Multi-Task Training

The Multi-Task Training was performed on T5-Base by combining all the training sets and experimenting for the originally proposed Task Balancing (TB) approach, and for our proposed task plus multi-dataset balancing (TDB) strategy for proportional mixing (PM) and temperature scaling (TS). The same hyperparameters were utilized for all settings with batch size 8, learning rate $1e-04$, input sequence length 512 and output sequence length 150. Temperature value was kept to be 2 for the temperature scaling method. For every multi-task setting, it took the model approximately 60 minutes to train for one epoch in the 4 GPU cluster computing environment setting.

B.4 Cross-Lingual Transfer

Multilingual BERT was trained using the four GPU cluster setting with batch size 256 over all GPU's for 7 epochs. The learning rate was set as $5e-05$ with 0 warmup steps and 0.01 weight decay. The T5-Base model trained on English SMM4H Task 1 AE Detection dataset was utilized to perform zero-shot on SMM4H French Dataset.