CE-005

機械翻訳テキストを用いた低資源言語特定分野向けＢＥＲＴの事前学習
Pretraining Language- and Domain-Specific BERT on Automatically Translated Text

石垣 達也 [1)]　上原 由衣 [1)]　トピチ ゴラン [1)]　高村 大也 [1)]
Tatsuya Ishigaki　Yui Uehara　Goran Topić　Hiroya Takamura

概要

専門的なテキストを対象とした言語処理タスクにおいて、SciBERT など特定分野に特化した事前学習言語モデルの有効性が知られている。しかし、英語以外の言語の細分化された分野については獲得可能なデータ量が不足し、BERT の学習が難しい。本研究では BERT の事前学習に必要な大規模コーパスが存在しない設定を想定し、英語等の資源豊富な言語から機械翻訳を用い獲得した専門的なテキストを、事前学習に用いる枠組みを提案する。本研究では、特に前処理での語彙の獲得手法、事前学習時に使用するデータなどを変化させた様々な BERT を翻訳テキストを用いて学習し、比較する。日本語の材料化学分野のテキストに着目した実験より、翻訳テキストを用い学習した様々な BERT は、日本語 Wikipedia を用いて学習した従来の BERT よりも一貫して良い性能を示すことを確認した。この結果は、従来困難であった低資源言語および細分化された専門分野向けの BERT が学習できる可能性を示唆するものである。

1　はじめに

SciBERT [7] などの科学分野のテキストから事前学習された BERT を用いることで、生物医学でのエンティティ抽出 [15]、化学分野の関係抽出 [9] など、専門的なテキストを対象とする言語処理タスクにおいて性能向上が報告されている。このような背景から、医学 [1]、生物医学 [12]、金融 [3]、材料化学 [7] など多くのより細分化された専門分野のテキストを用いて学習された BERT が公開されている。専門的なテキストは分野が細分化するにつれて獲得が難しくなり、特に英語以外の言語では深刻な問題となる。その一方、英語以外の言語でも社内文書などを対象に、事前学習モデルを用いて高品質なテキスト解析を行う需要は高い。

そこで、本研究では資源が豊富な英語で記述された専門テキストを、対象言語に機械翻訳し、BERT の事前学習に用いる枠組みを提案する。本稿では特に大規模な専門テキストの獲得が難しい対象として、日本語の材料科学分野に着目し、提案枠組みの効果を検証する。図 1 に提案枠組みの概要を示す。提案する専門分野向け BERT は、人間によって英語で記述された論文テキストを機械翻訳を用い日本語に変換したコーパスから学習する。翻訳データを用いる提案枠組みは非常に単純であり、多くの言語および分野に対し適用可能な利点がある。本稿では特に、翻訳データを用いた BERT の学習について、特に 2 点に着目し様々な学習手法を比較する: 1) 翻訳データを活用することで日本語 Wikipedia を用いた一般的な BERT を用いるよりも良い性能を示すか、2) BERT 学習のための前処理において、専門テキストを用いてサブワード辞書を構築すると性能が向上するか。

翻訳テキストを用いて学習した様々な BERT を、日本語の材料科学分野のエンティティ抽出および関係抽出タスクにおいて性能評価する実験を行う。この実験より、翻訳テキストを用いて学習した BERT は従来の日本語 Wikipedia を用いて学習した一般分野向けの BERT より一貫して良い性能を示すことがわかった。さらに、日本語 Wikipedia と翻訳テキストを混合したコーパスから BERT を学習する場合においては、前処理のサブワード辞書構築において、専門語彙を含める方策がより有効であることがわかった。

本稿の貢献は以下である: 1) 専門分野向けの BERT を翻訳テキストを用いて学習する新たな枠組みを提案する、2) 翻訳テキストを用い学習した BERT は一貫して日本語 Wikipedia のみを用いて学習した BERT よりも良い性能を示すことを確認した、3) 日本語の材料化学分野向けの BERT を公開する[1)]。

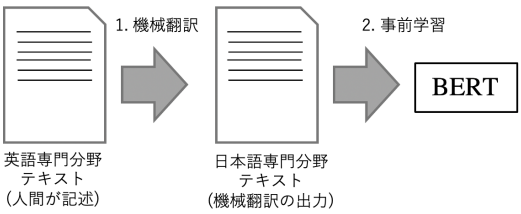

図 1　翻訳テキストを用いて特定言語および分野に特化した BERT を学習する枠組み。

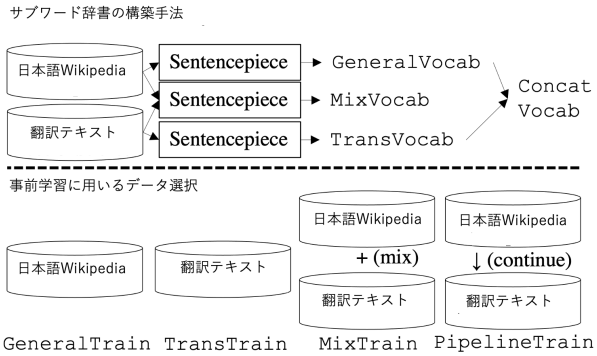

図 2　サブワード辞書の構築手法 (上半分) と事前学習に用いるデータ (下半分) の概要。

2　関連研究

事前学習の手法はこれまでに多く提案されている。一般的な BERT は Wikipedia など一般分野のテキストのみを用いて学習される [6]。一方、SciBERT は事前学習に専門分野のテキストを用いることで専門的なテキストに対する解析性能の向上を報告している [5]。専門分野向けの BERT の学習手法としては、他にも一般的なテキストで学習したのち、さらに専門分野のテキストを用い追加の事前学習を行う手法が存在する [20, 12, 22]。さらに、一般テキストと専門テキストを混合したコーパスを作成し BERT を学習する手法が考えられ、これは多言語 BERT は [6] のアイデアに近い。

1)　産総研人工知能研究センター

1)　https://kirt.airc.aist.go.jp/resources

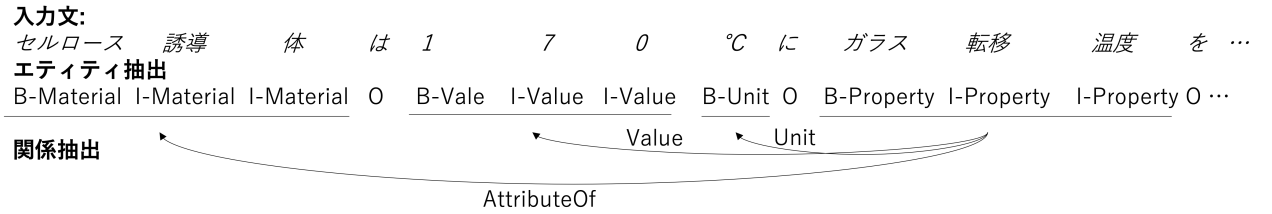

図 3　材料科学分野におけるエンティティ抽出および関係抽出タスク。

事前学習に使用するデータの使用法のみならず、単語埋め込み層に入力するサブワードを定義した辞書をいかに構築するかという点も問題となる。辞書の構築手法の観点から、一般分野のテキストのみから獲得する手法 [12, 6]、専門分野のテキストのみ用いる手法 [5]、一般テキストおよび専門テキストからそれぞれ辞書を構築し 2 つの辞書の和集合をとる手法 [20] に分けられる。本稿では異なる学習データの使用法および辞書の構築手法を組み合わせる。こうして学習した様々な BERT を下位タスクにおいて用い、性能を比較する。

提案する枠組みは、機械翻訳を用いたデータ拡張の研究とも関連する [4, 19]。機械読解 [21] やフェイクニュース判定 [2] などの下位タスクの学習データ作成に用いられるデータ拡張では、資源豊富な言語のラベル付きのデータを機械翻訳を用いて低資源言語に変換する手法が採用される。本研究では下位タスクの学習ではなく、事前学習に用いるラベルなしデータを獲得する点が異なる。

3　手法

本節では翻訳テキストを用い BERT を事前学習する手法について述べる。事前学習で用いるデータの使用法およびサブワード辞書の構築手法について順に説明する。

3.1　事前学習に用いるデータ

事前学習には日本語 Wikipedia に加え、専門分野のテキストとして英語文献を翻訳したデータを用いる。日本語の材料科学分野では、事前学習に用いる大規模コーパスの獲得が難しい。一方、英語では例えば Web of Science[2] に収録された論文データなど、大規模な材料科学分野のテキストが入手可能である。そこで、本研究では Web of Science に収録された材料化学分野の文献を含む論文誌として “DSSHPSH” and “ESCI” とラベル付けされたものを選択した。さらに材料化学分野の論文にフィルタリングするため、これらの論文誌の収録論文のうち “Materials Science” のタグが付与された論文を収集した。これらの収録論文は 1965 年から 2018 年の間に記述され、例外として 1900 年に記述された古い論文も含む。収集した論文から要旨のみを抽出し、自然言語処理の専門家でなくとも使用できる翻訳システムという点を重視し、Amazon Translate を用いて英語から日本語へと翻訳した[3]。最終的に 2,501,178 本の論文から要旨を翻訳した 21,115,139 文を専門分野のテキストとして獲得した。さらに、日本語の Wikipedia の 2020 年 4 月のバージョンに含まれる 1,197,647 記事から 21,584,456 文を抽出し、一般分野のテキストコーパスも獲得した。これら 2 種類のデータを用いて、いくつかの BERT を学習し比較する。

3.2　サブワード辞書の構築と事前学習手法

BERT 学習の既存手法はサブワード辞書の構築手法と事前学習に用いるデータの観点から少なくとも 2 種類に分類できる。サブワード辞書の構築手法には図 2 の上半分に示す 4 手法が考えられる。`GeneralVocab` はサブワード分割を一般分野のテキストのみを Sentencepiece 等の既存アルゴリズム [10] で学習するのに対し、`TransVocab` は翻訳テキストのみを用いる。Delvin らの BERT [6] は前者を採用しており、SciBERT [5, 7] は後者を採用している。さらに、`MixVocab` は一般分野と専門分野のテキストを混合しサブワード分割を学習する。これは多言語 BERT [6] の手法を分野特化型 BERT に適用したものと捉えることができる。`ConVocab` は一般分野および専門分野のテキストそれぞれからサブワード辞書を構築し、2 つの和集合を取ったものである。これは exBERT [20] の手法に近い。

事前学習に用いるデータ選択の観点からは、図 2 の下半分に示す 4 つの手法を考える。`GeneralTrain` uses only the general texts [6]. `TransTrain` は翻訳テキストのみを用いて学習する手法である。`MixTrain` は一般分野のテキストおよび翻訳テキストを混合し学習する。`PipelineTrain` は一般分野のテキストで BERT のパラメータを学習し、次に翻訳テキストを用いて学習を継続する。サブワード辞書の構築手法とデータ選択の観点から合計 10 の組み合わせ手法が考えられ、これらを比較する実験を行う。

4　実験

構築する事前学習モデルは日本語の材料化学分野でのエンティティ抽出および関係抽出タスクで評価する。

4.1　性能評価タスクとモデル

性能評価に用いる日本語材料化学分野のエンティティ抽出および関係抽出の例を図 3 に示す。エンティティ抽出は入力テキストから 4 種類のエンティティを抽出する。具体的には、1) セルロース誘導体などの物質名、2) ガラス転移温度 (Tg) などの物性、3) 数値、4) 単位である。物性については当該分野において応用上重要である Tg および弾性率にのみ着目する設定を扱う。関係抽出は抽出されたエンティティ間の関係をラベル付けするタスクである。例えば、Tg はセルロース誘導体という物質の持つ物性であるため、2 つのエンティティ間には “AttributeOf” のラベルを付与する。エンティティ抽出および関係抽出タスクにおいて用いるラベル定義を、それぞれ表 3 および表 4 に示す。

エンティティ抽出器および関係抽出器はそれぞれ別にラベルに対する交差エントロピー損失を用いて学習する．エンティティ抽出器において、解析対象の文はまず

2) https://www.webofscience.com
3) AmazonTranslateは2020年1月のバージョンを使用した。https://aws.amazon.com/translate/

	事前学習データとサブワード辞書			エンティティ (Tg)			関係 (Tg)			エンティティ (弾性率)			関係 (弾性率)		
1) GeneralTrain and TransTrain				P	R	F	P	R	F	P	R	F	P	R	F
1	General	-	GeneralVocab	88.27	90.28	89.15	80.68	76.88	78.47	92.64	93.15	92.87	77.99	78.51	78.36
2	Translated	-	TransVocab	90.67	92.45	91.49	80.23	78.75	79.17	93.43	94.59	94.00	78.32	79.57	78.96
2) MixTrain															
3	Mix	-	GeneralVocab	90.73	91.46	91.00	81.78	77.56	79.33	93.73	94.48	94.08	79.61	80.47	79.68
4	Mix	-	TransVocab	90.79	92.38	91.54	81.10	78.61	79.65	93.45	94.83	94.12	79.69	79.68	80.71
5	Mix	-	MixVocab	91.34	92.14	91.67	81.26	79.07	79.85	93.78	95.01	94.38	79.02	79.03	79.98
6	Mix	-	ConVocab	90.43	92.22	91.26	81.03	79.42	80.01	94.11	95.04	94.56	79.58	79.64	80.55
3) PipelineTrain															
7	General	Translated	GeneralVocab	91.10	92.08	91.54	81.49	78.49	79.75	93.62	94.69	94.13	79.60	80.43	79.60
8	General	Translated	TransVocab	90.93	91.40	91.10	81.23	79.35	79.99	**94.21**	94.22	94.18	79.77	**80.75**	79.86
9	General	Translated	MixVocab	90.57	**92.47**	91.46	81.68	78.47	79.75	93.59	94.66	94.11	78.92	79.54	79.33
10	General	Translated	ConVocab	90.85	92.12	91.40	80.10	79.57	79.55	93.60	95.25	94.40	79.15	80.00	79.86

表 1　日本語材料化学分野でのエンティティ抽出および関係抽出の性能。データ選択の観点から上から 3 つのカテゴリに分かれる。最も上には一般分野、専門分野いずれかのデータを用いて事前学習する手法 (GeneralTrain および TransTrain)、2 つ目のカテゴリには一般テキストおよび専門テキストの混合データを用いる手法 (MixTrain)、3 つ目のカテゴリには一般テキストで学習し専門テキストで追加の事前学習を行う手法 (PipelineTrain) である。例えば、モデル 9 は一般テキストでまず学習され、次に翻訳テキストで学習され、サブワード辞書の構築には混合データを用いた (MixVocab) ことを示す。

	Text (1st)	Text (2nd)	Vocab.	関係抽出 (Tg)			関係抽出 (弾性率)		
1) GeneralTrain and TransTrain				P	R	F	P	R	F
1	General	-	GeneralVocab	71.04	70.30	70.35	71.04	71.86	71.00
2	Translated	-	TransVocab	71.82	73.39	72.31	72.61	72.66	71.65
2) MixTrain									
3	Mix	-	GeneralVocab	73.49	72.05	72.49	72.35	74.39	73.02
4	Mix	-	TransVocab	73.80	73.38	73.41	71.81	75.13	73.12
5	Mix	-	MixVocab	73.26	73.10	72.89	71.91	74.70	72.91
6	Mix	-	ConcatVocab	72.28	74.27	73.04	73.05	74.86	73.42
3) PipelineTrain									
7	General	Translated	GeneralVocab	72.56	72.67	72.40	72.25	74.54	73.04
8	General	Translated	TransVocab	73.42	73.59	73.19	73.89	74.49	73.80
9	General	Translated	MixVocab	73.09	72.73	72.61	72.00	73.88	72.58
10	General	Translated	ConcatVocab	72.35	74.16	73.01	72.75	74.46	72.74

表 2　エンティティ抽出器の予測エンティティを関係抽出器で用いた場合の性能。

エンティティラベル
B-Material
B-Property
B-Value
B-Unit
I-Material
I-Property
I-Value
I-Unit
O

表 3　エンティティ抽出でのラベル定義。

関係ラベル
AttributeOf
Value
Unit
Abbreviation
Synonym
Conjunction
Other

表 4　関係抽出でのラベル定義。

MeCab [11] を用いて形態素に分割され、さらに 3.2 節で述べた手法でサブワードに分割される。サブワード系列の先頭に文頭を表す特別なトークン [CLS] を付与する。各サブワードは BERT により読み込まれ、各サブワードに対する分散表現を得る。各サブワードに対する分散表現は 1 層の隠れ状態を持つ多層パーセプトロンを用い、選択するラベル候補の数と同じ、大きさ 9 のベクトルに変換される。各次元の値が候補ラベルに対するスコアとみなされ、スコア最大のラベルを出力する。学習時には、事前学習済み BERT で重みを初期化したモデルを、正解ラベルが付与されたデータセットを用いてファインチューニングする。

関係抽出器はエンティティの対を入力とし、7 つの候補ラベルに対しスコアを計算する。Trieu ら [18] と同様に、Sohrab らの手法 [17] で計算したエンティティに対する分散表現に加え、エンティティの種類ラベルに対する分散表現も獲得し、これらを組み合わせ最終的なエンティティ表現とする。2 つのエンティティ表現を入力とし、4 つの素性ベクトルを結合し分類器の入力とする。具体的には、1) head エンティティの表現、2) tail エンティティの表現、3) 2 つの表現の要素積 [16, 13]、4) 入力文の [CLS] トークンの表現の結合ベクトルである。結合ベクトルは多層パーセプトロンによる分類器で大きさ 7 のベクトルへと変換される。大きさ 7 は候補ラベルの数と等しく、各次元の値を各候補ラベルへのスコアとみなす。最終的に、スコア最大となるラベルを出力する。

4.2　学習設定

データ: エンティティ抽出および関係抽出器の学習のためのデータセットについて述べる。材料化学分野においては、研究者によってガラス転移温度 (以後 Tg と呼ぶ)

や弾性率など着目する物性が異なる。そこで、2 種類のアノテーションデータを作成し性能評価実験を行う。1 つ目のデータは Tg に着目しアノテーションしたデータ、2 つ目は弾性率に着目し作成したデータである。材料化学分野の専門家がエンティティおよび関係ラベルのアノテーションを行った。Tg についてのデータセットは高分子学会論文誌に含まれる 106 報に対し、アノテーションを行ったもので 14,189 文が含まれる。弾性率に着目したデータセットは、同様の論文誌から 100 報抽出し 12,864 文にアノテーションを行ったものである。このデータセットは公開が予定されている。

学習パラメータ: サブワード辞書の構築には SentencePiece [10] を用いた。ハイパーパラメータである語彙サイズの設定には、`GeneralVocab`, `TransVocab` および `MixVocab` に対し 32,104 とした。`ConVocab` は `GeneralVocab` と `TransVocab` の和集合となり、語彙サイズはは 49,858 であった。比較する 10 種類の BERT は、すべて 30 エポック分の学習を行い、最適化器には Adam [8] を初期学習率 10^{-4} の設定で用いた。学習には NVIDIA V100 を 4 本用い、分散学習した。エンティティおよび関係抽出の学習では、160 エポック分の学習を初期学習率を 10^{-5} に設定した Adam で行い、F1 スコアが最大となるモデルを選択し評価に用いる。

5 結果

表 1 に比較対象である 10 の BERT のエンティティ抽出および関係抽出タスクでの性能を示す。評価指標には Sohrab ら [17] にならい、テキストスパンを対象に計算した精度、再現率および F1 スコアのマクロ平均を用いる。これらの値は 5 分割交差検定を行い計算した。表は上から下にかけて 3 つに分割されている。上から順に 1) 一般分野のテキストもしくは専門分野のテキストのみで学習する BERT (`GeneralTrain` および `TransTrain`)、2) 2 種類のテキストを混合したコーパスから学習した BERT (`MixTrain`)、3) まず一般分野のコーパスで学習し専門分野のコーパスでさらに追加の事前学習を行った BERT (`PipelineTrain`) である。関係抽出器の評価においては、正解のエンティティを入力とした [5]。正解エンティティではなく予測エンティティを用い関係抽出器を用いた実験についても行った。結果を表 2 に示す。後述するが、この場合も正解エンティティを与える設定と似た傾向を示した。

翻訳テキストで学習した BERT は下位タスクの性能向上に寄与するか？翻訳テキストを用いて学習したモデル (表 1 のモデル 2 から 10) は、一般分野のテキストのみで学習したベースライン手法 (モデル 1) よりもモデル 2 の Tg の精度 (P) 以外のすべての指標において良い性能を示した。具体的な値をみると、一般分野のテキストのみで学習したモデル 1 の F1 スコアは Tg に関するアノテーションデータにおいて 89.15 であるのに対し、専門分野テキストのみで学習したモデル 2 の F1 スコアは翻訳データであるにも関わらず 91.49 であり、2.34 の向上がみられた。Tg のデータでの関係抽出タスクにおいても同様に、0.7 (78.47 から 79.17) の向上が見られており、翻訳テキストの活用が一貫して性能向上に寄与した。弾性率に関するアノテーションデータでの実験でも同様の傾向を確認した。

翻訳データから学習したサブワード辞書は下位タスクの性能向上に寄与するか？混合データから学習した BERT (`MixTrain`) において、使用するサブワード辞書を変更した場合の結果をみる。Tg に対するエンティティ抽出において、専門分野のテキストから学習したサブワード分割を用いるモデル 4 から 6 はそれぞれ 91.54、91.67 および 91.26 と一般分野のテキストからサブワード辞書を構築 (`GeneralVocab`) するモデル 3 の 91.00 よりも高い F1 スコアを得た。この結果より、混合データを用いる設定において分野特化したサブワード辞書の効果が確認できる。Tg の関係抽出、弾性率のエンティティ抽出や関係抽出でも同様の傾向であった。一方、一般分野のテキストでまず学習し専門分野のテキストで追加の事前学習を行う `PipelineTrain` では、一般分野のテキストでサブワード分割を学習した場合であっても F1 スコアが各タスクにおいて 91.54, 79.75, 94.13, and 79.60 と高い。専門分野のテキストをサブワード辞書の構築に用いる手法 (モデル 8 から 10) は一般分野のテキストを用いる手法と性能差がない。よって、`PipelineTrain` を用いる場合、サブワード辞書の構築は必ずしも専門分野のテキストから行わずとも良い性能が期待できる特徴がある。Wikipedia などの一般テキストから学習された既存の BERT を、単純に専門分野の翻訳テキストでさらに事前学習すれば良いため、この特徴は実応用上好ましい。

6 おわりに

本稿では、専門分野向けの BERT を英語から機械翻訳したテキストで学習する新たな枠組みについて調査した。実験より、機械翻訳した専門分野テキストの活用が BERT の下位タスクでの性能向上に寄与することがわかった。非常にシンプルで他分野、他言語、BERT の派生モデルへの応用可能性も高い利点がある。今後は材料化学分野の情報抽出タスクにとどまらず、他分野、他言語、BART [14] など言語生成モデルでの効果を検証したい。

参考文献

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[2] Maaz Amjad, Grigori Sidorov, and Alisa Zhila. Data augmentation using machine translation for fake news detection in the Urdu language. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, pages 2537–2542, Marseille, France, May 2020.

[3] Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. In *arXiv preprint (1908.10063, 2019)*, 2019.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations (ICLR 2015)*, pages 1–15, 2014.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP2019)*, pages 3615–3620, Hong Kong, China, 2019. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2019)*, pages 4171–4186, Minnesota, USA, 2019. Association for Computational Linguistics.

[7] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction, 2021.

[8] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[9] Jens Kringelum, Sonny Kim Kjærulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. Chemprot-3.0: a global chemical biology diseases mapping. *Database J. Biol. Databases Curation*, 2016, 2016.

[10] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, 2018. Association for Computational Linguistics.

[11] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, pages 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746, 2019.

[13] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP2017)*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[15] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[16] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP2018)*, pages 3219–3232, Brussels, Belgium, 2018. Association for Computational Linguistics.

[17] Mohammad Golam Sohrab, Anh-Khoa Duong Nguyen, Makoto Miwa, and Hiroya Takamura. mgsohrab at WNUT 2020 shared task-1: Neural exhaustive approach for entity and relation recognition over wet lab protocols. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 290–298, Online, November 2020. Association for Computational Linguistics.

[18] Hai-Long Trieu, Thy Thy Tran, Khoa N A Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. DeepEventMine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917, 06 2020.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS2017)*, pages 5998–6008, 2017.

[20] Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online, November 2020. Association for Computational Linguistics.

[21] Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. Fast and accurate reading comprehension by combining self-attention and convolution. In *Proceedings of Sixth International Conference on Learning Representations (ICLR2018)*, Vancouver, Canada, 2018.

[22] Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP2020)*, pages 5461–5468, Online, 2020. Association for Computational Linguistics.