生成された読影所見の自動評価に向けた固有表現認識とモダリティ推定

田川 裕輝 [1]　西埜 徹 [1]　谷口 元樹 [1]　谷口 友紀 [1]　大熊 智子 [1]
若宮 翔子 [2]　荒牧 英治 [2]
[1] 富士ゼロックス株式会社　[2] 奈良先端科学技術大学院大学
{tagawa.yuki, nishino.toru, motoki.taniguchi, taniguchi.tomoki, ohkuma.tomoko}@fujixerox.co.jp
{wakamiya, aramaki}@is.naist.jp

1　はじめに

膨大な電子カルテ文書を日々扱う医療分野において自然言語処理に対するニーズは大きい．我々は放射線科に焦点を当てて，自然言語処理による医師の負担軽減を目指している．放射線科の読影医は毎日，読影所見と呼ばれる膨大な量の放射線画像の診断所見 (図 1) を作成している．そのため，読影所見を自動かつ高精度で生成することができれば，医師の負担軽減につながる．

テキストを生成する場合，事実との整合性や必要な情報に対する網羅性が重要となる．そのため，生成テキストの評価にはテキストの表層的な情報だけでなく，内容に基づいた評価をしなければならない．しかし，テキスト生成分野で広く利用されている BLEU などの N-gram の一致に基づいた手法で評価すると，正解テキストと生成されたテキストとの間で内容が異なる場合でも，表層が類似しているだけで高い評価値となることが報告されている [1]．そのため，Wiseman ら [2] はバスケットボールのボックススコアからの試合の要約生成において，正解テキストと生成テキストのそれぞれに対して，選手名等の固有表現 (NE) や得点数等の数値データ，それらを対応付ける属性から構成されるタプルを抽出し，抽出されたタプルの一致に基づいた評価指標で生成テキストを評価している．この評価指標により，BLEU では検出できないエラーを検出して人手評価に近い評価が出来るようになったことを報告している．

読影所見の自動生成においてもこのような内容に基づいた評価をするためにはテキスト中の NE の適合率や網羅率，NE 間の関係性等を考慮した評価指標が必要となる．生成された所見を評価する場合には、病名や病変などの重要な NE の脱落や，病気や病変が確認された部位などを正確に評価しなければならない．このような所見内容を直接評価する指標が確立できると，その評価値を最適化する生成モデルを訓練して質の高い所見を自動生成することが期待できる．

本研究では生成された読影所見の自動評価に向けて，読影所見に対する固有表現認識 (NER) とモダリティ推定を行う．NER とは病変や病名などテキスト中の NE を認識する技術である．モダリティ推定とは病変や病名が実際に生じているのか否かの事実性を判別する技術である．これらの NER とモダリティ推定の結果を利用し，生成された所見の評価を行う．

医療テキストに対する NER の研究の多くは英語を対象としたもの [3, 4] であり，日本語を対象とした研究は少ない．Yano [5] は日本語の病歴要約 [6] を対象に病名認識と病名のモダリティ推定に取り組んでいる．また，荒牧ら [7] は日本語の症例報告を対象に病名に対して，NE とモダリティのアノテーションを行い，大規模な症例報告データ構築とそれを用いた病名認識器を実現している．一方で，我々の目的は病名認識器の構築そのものではなく，自動生成された読影所見に対して，内容に基づいた詳細な評価を行うことである．本研究の貢献は以下の 3 つである．(1) 病歴要約 [6] や症例報告 [7] と比べ，より詳細な 8 種類の NE ラベルと 3 種類のモダリティラベルを設計し，読影所見コーパスを作成する．(2)(1) のコーパスを用いて系列ラベリングタスクで基本的なモデルである CRF，BiLSTM-CRF [8]，BERT [9]，BERT-CRF の精度の比較，出力のエラー分析を行う．(3)(2) で構築された NER とモダリティ判定を用いて NE に基づいた評価指標で生成所見の評価実験を実施し，自動評価指標の ROUGE，BLEU と比べ，人手評価との相関が高いことを確認する．

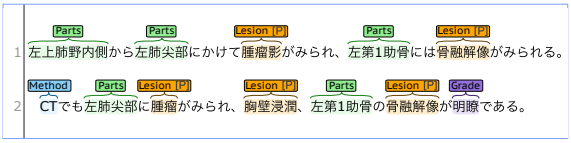

図 1: アノテーション済み読影所見の一例

2　データ

我々は共同研究先である X 病院から提供された肺結節に関する読影所見を対象に NER とモダリティ推定のためのラベルを設計した．表 1 に示す 8 種類の NE ラベルを定義し，2,080 件の読影所見に対してアノテーションを行った．モダリティラベルは Lesion，Disease に対して，実際に患者が患っているか (P; Positive)，患っていないか (N; Negative)，患っている疑いや可能性があるか (S; Suspicious) の 3 種類を用意した．アノテーションされた読影所見の例を図 1 に示す．“腫瘤影”という病変情報の直後に“みられ”という患者が病変を患っていることを断定する表現があるため，“腫瘤影”にはモダリティとして P ラベルが付与されている．

データセットの統計量を表 2 に示す．読影所見の特徴として，読影医は放射線画像から読み取れる病変の有無などの所見のみを記述し，何らかの病気を否認するような診断を記述することは少ない．そのため，Lesion-S と Disease-N の出現回数がその他のラベルと比べて少なくなっている．

3　モデル

NER では所見を入力し，所見中の各トークンに対する正解ラベルの系列を出力するようにモデルを学習させる．モダリティ推定では出力ラベル系列を Lesion-P のように，NE ラベルにモダリティラベルを結合したものを新たにラベルとして用意し，NER とモダリティ推定を同時に解く．モデルには系列ラベリングタスクで基本的な CRF，BiLSTM-CRF [8]，BERT [9]，BERT-CRF を用いる．

ラベル	説明
Parts	主には部位を示す表現．“中間層”，“末端”，“底部”，“一部”などの画像中の位置を表す名詞も含まれる．
Lesion	画像情報を客観的に観察して得られる病変情報“すりガラス影”，“結節”等．所見にはそれが“認める”，“認めない”など断定的な表現で記述されている．
Disease	観察された病変から診断される病名“肺炎，“肺癌”など．“疑われる”など断定的ではない表現で記述される．
Time	具体的な日付や“前回”，“年一度”などの表現が含まれる．
Method	“単純 CT”，“PET”，“造影剤”，“dynamic study”などの検査に使用される器材や検査方法．
Change	“著変”，“増大”など，前回の診断との比較から得られた観察結果．主にサ変動詞語幹．
Numeric	数詞と助数詞による具体的な寸法や“一部”，“全体”，“最大”などの数量表現．
Grade	“明瞭な”，“わずかに”など程度を表す修飾語．状態や形状を含む．

表 1: NE ラベルとその説明

NE	モダリティ	件数
Parts	-	4,644
Time	-	2,136
Method	-	983
Change	-	1,905
Numeric	-	2,364
Grade	-	1,733
Lesion	Positive	4,178
	Negative	657
	Suspicious	200
Disease	Positive	825
	Negative	141
	Suspicious	1,547
所見数		2,080
所見の平均文数		2.78
所見の平均サブワード数		59.23
所見の平均ラベル数		10.25

表 2: 作成したデータセットの統計量

3.1 CRF

CRF は系列ラベリングタスクを解く際に利用される識別モデルである．入力系列 $X = x_0, x_1, ..., x_N$ に対する出力ラベル系列 $Y = y_0, y_1, ..., y_N$ が与えられたとき，CRF で計算される条件付き確率 $P(Y|X)$ は以下の式で定義される．

$$P(Y|X) = \frac{1}{Z}\exp(\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k^{\mathsf{T}} f_k(y_{i-1}, y_i, X)), \quad (1)$$

f_k は $i-1$ と i 番目のラベルと入力系列 X に対する素性ベクトルである．また，λ_k は素性ベクトル f_k に対する重みパラメータである．Z は正規化項である．λ_k は $P(Y|X)$ の対数尤度を最大化するように学習される．入力系列 X に対する最適なラベル系列 y^* は以下の式で求める．

$$y^* = \underset{Y}{\operatorname{argmax}} P(Y|X). \quad (2)$$

3.2 BiLSTM-CRF

BiLSTM-CRF [8] はまずトークン埋め込み層により，入力トークン x_i の one-hot ベクトル o_i に対する埋め込みベクトル v_i を獲得する．次に，双方向の LSTM を用いて，トークン x_i の隠れ状態 h_i を獲得する．

$$v_i = o_i T, \quad (3)$$

$$\overrightarrow{h_i} = \mathrm{LSTM_f}(v_i, \overrightarrow{h_{i-1}}), \quad (4)$$

$$\overleftarrow{h_i} = \mathrm{LSTM_b}(v_i, \overleftarrow{h_{i-1}}), \quad (5)$$

$$h_i = \overrightarrow{h_i} \oplus \overleftarrow{h_i}, \quad (6)$$

ここで，$T \in \mathbb{R}^{vocab \times m}$ はトークン埋め込みテーブルであり，$vocab$ は学習データ中の全トークン数，m はトークン埋め込みベクトルの次元数である．また，$\mathrm{LSTM_f}$ と $\mathrm{LSTM_b}$ はそれぞれ順方向と逆方向の LSTM を表し，$\overrightarrow{h_i}$，$\overleftarrow{h_i}$ は d 次元のベクトルである．また，$\oplus$ はベクトルの結合を表す．

その後，CRF によりラベル系列を出力する．

$$e_i = h_i W, \quad (7)$$

$$\mathrm{Score}(X, Y) = \sum_{i=0}^{N} Trans_{y_i, y_{i+1}} + \sum_{i=1}^{N} e_{i, y_i}, \quad (8)$$

$$P(Y|X) = \frac{\exp(\mathrm{Score}(X, Y))}{\sum_{\tilde{y} \in \tilde{Y}} \exp(\mathrm{Score}(X, \tilde{y}))}, \quad (9)$$

ここで，$W \in \mathbb{R}^{2d \times l}$ は重み行列であり，l は全ラベル数である．$Trans \in \mathbb{R}^{l \times l}$ はラベル遷移行列であり，$Trans_{y_i, y_{i+1}}$ は i 番目のラベル y_i から $i+1$ 番目のラベル y_{i+1} に遷移するスコアを表す．e_{i, y_i} は i 番目のトークンのラベル y_i に対するスコアである．また，入力系列 X に対して，あり得る全てのラベル系列 $\tilde{Y}$ で正規化する．学習時は $P(Y|X)$ の対数尤度を最大化するように学習する．入力系列 X に対する最適なラベル系列 y^* は以下の式で求める．

$$y^* = \underset{Y}{\operatorname{argmax}}\, \mathrm{Score}(X, Y). \quad (10)$$

3.3 BERT，BERT-CRF

BERT [9] は Transformer という自己注意機構を備えたモデルを複数層重ねたモデルであり，大規模なテキストで事前学習し，各タスクに応じて再学習することで，様々なタスクで高い精度を記録している．我々は BERT に対して各ラベルに対するスコアを出力する全結合層を追加し，再学習を行う．再学習時は出力した各ラベルに対するスコアと正解ラベルに対する SoftmaxCrossEntropy を誤差とし，最小化するように学習する．

BERT-CRF は上述した BERT の最終層に CRF を追加したモデルである．BiLSTM-CRF と同じように，BERT で入力系列をエンコードし，得られたベクトル表現から CRF により，正解ラベル系列を出力するように再学習する．

4 実験

4.1 実験設定

出力ラベル系列には BIO 方式[1]を採用した．CRF の特徴量としては前後 2 トークンの表層を利用した．BiLSTM-CRF はトークン埋め込みベクトルの次元数を 128，$\mathrm{LSTM_f}$，$\mathrm{LSTM_b}$ の隠れ層のサイズを 256 とした．BERT，BERT-CRF は日本語 Wikipedia で事前学習した Whole Word Masking モデル[2]を利用し，再学習した．実装には Transformers[3]を利用した．また，全モデルにおいて，事前学習モデルと同じサブワードレベルのトークナイザーを利用し，所見をトークン化した．精度評価には 10 分割交差検定を適用し，各交差において開発データに対する Micro-F1 が最も高い epoch のモデルを採用した．F1 値の算出方法は CoNLL-2003 と同じ方法を利用した．

[1]BIO 方式では NE の開始を表す Begin，同一種の NE の継続を表す Inside，どの NE にも当てはまらない Outside を利用する．

[2]http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル

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

	CRF	BiLSTM-CRF	BERT	BERT-CRF
Parts	0.951	0.946	0.932	**0.953**
Time	0.963	0.968	0.967	**0.978**
Method	0.936	0.946	0.936	**0.953**
Change	**0.961**	0.954	0.956	0.960
Numeric	**0.967**	0.962	0.960	0.964
Grade	0.789	0.780	0.815	**0.840**
Lesion	0.918	0.917	0.895	**0.937**
Disease	0.925	0.924	0.912	**0.934**
Macro-F1	0.926	0.925	0.922	**0.940**
Micro-F1	0.931	0.929	0.922	**0.942**

表 3: 10 分割交差検定における NER の F1 値

	CRF	BiLSTM-CRF	BERT	BERT-CRF
Lesion-P	0.909	0.919	0.914	**0.931**
Lesion-N	0.854	**0.902**	0.873	0.889
Lesion-S	0.453	0.522	0.465	**0.533**
Disease-P	0.730	0.804	0.787	**0.834**
Disease-N	0.594	**0.709**	0.576	0.687
Disease-S	0.830	0.875	0.866	**0.891**
Macro-F1	0.728	0.788	0.747	**0.794**
Micro-F1	0.907	0.914	0.910	**0.929**

表 4: 10 分割交差検定におけるモダリティ推定の F1 値

4.2 実験結果

NER の精度を表 3 に示す．精度を比較すると，BERT-CRF が Macro，Micro−F1 ともに最も高い値となった．一方で，BERT が最も低い値となった．エラー分析の結果，BERT は学習データに存在しないラベル系列を出力しており，Precision の低下が確認された．例えば，B-Lesion の後に I-Disease など一貫性の無いラベル系列を出力していた．BERT のラベル出力層は全結合層であり，ラベルを出力する際に，直前に予測したラベルの情報を考慮していないため，このような問題が起きたと考えられる．それに対して，テキストのエンコード能力が高いとされる BERT に，ラベル間の遷移を扱える CRF[4]をラベル出力層として利用した BERT-CRF が最も高い精度となった．

モダリティ推定の精度を表 4 に示す．表 3 に示した NER の結果と同じく，BERT-CRF が最も高い精度となった．一方で，CRF の精度が大きく低下していることが確認できる．CRF はモダリティを判断するための根拠にあたる表現が病変や病名の表現と距離が離れている場合に予測に失敗していた．CRF のウィンドウサイズを調整することで対応することも考えられるが，その他のラベルの予測精度への影響や経験則に頼らざるを得ないという問題もある．それに対して，LSTM や BERT などではウィンドウサイズの指定は不要であり，CRF と比べて高い精度となった．

4.3 考察

モデルが予測を間違えた NE の多くがトークナイズ時にサブワード化された NE であった．実際，テストデータ中の NE のうちトークナイズ時にサブワード化されたものは約 13%であるが，テスト時に予測に失敗した NE のうちサブワード化された NE は約 55%と割合に大きな差が見られた．読影所見の特徴として病変や病名は英語で記述されることがあり，細かい粒度でサブワードに分割されるため，質の高いベクトル表現を獲得できなかったと考えられる．例えば，“PureGGO”という病変表現は“P_ure_G_GO”[5]のように分割される．このような場合，“PureGGO”または“Pure_GGO”のようにして扱うのが適切と思われるが，事前学習コーパスにこのような英単語が頻出するわけではないため，細かい粒度でサブワードに分割されてしまう．この問題を緩和するためには再学習時に扱うテキストと同じ分野のテキストでサブワードを構築し，事前学習を行うことが挙げられる．実際に，BioBERT [3] や ClinicalBERT [4] など大規模な Biomedical 分野のテキストで事前学習することで当該分野のテキストを対象とした NER の精度向上が報告されている．また，ドメインが限られたコーパスに対して NER を行う際に，単純に BPE [10] で語彙を構築するのではなく，専門用語辞書を利用して，辞書制約付きの BPE [11] で語彙を構築することで精度が向上するという報告もある．このような技術を適用することで更なる精度向上が期待される．

[4]CRF のラベル遷移行列に対して，学習データに存在しない系列の要素には初期値として-10000 を入力した．

[5] _はサブワードの開始を意味する．

5 生成テキストの評価

本研究の目的は生成された所見を内容に基づいて評価することである．N-gram の一致に基づいた評価指標では書かれている内容が異なる場合でも，表層が類似しているだけで，評価値が高くなるという問題が報告されている [1]．そこで，我々は学習させたモダリティ推定モデルの出力を利用し，NE の一致に基づいた自動評価を行う．

5.1 評価方法

評価対象として，放射線画像に対する所見タグ群から読影所見を生成するモデル [12] が出力したテキストを用いた．生成モデルには Seq2Seq に注意機構を加えたモデルを利用し，105 件の所見を生成した．生成された所見に対して，アノテータによる人手評価，ROUGE，BLEU を利用した自動評価，NE による自動評価を行い，人手評価値とのピアソンの積率相関係数を算出した．以下ではそれぞれの評価方法について説明する．

5.1.1 アノテータによる人手評価

人手評価では生成所見と正解所見を比較したとき，生成モデルが所見タグと適合した内容の所見を生成できているか (Precision)，所見タグの内容を網羅した所見を生成できているか (Recall) を加味した評価を行った．まず，アノテータには入力となった所見タグ群とその所見タグ群から生成された所見が与えられる．次に，アノテータは所見タグ群とそれに対応する生成所見中の表現との間のマッチングをとる．例えば，入力の 4 個の所見タグのうち，3 個の所見タグに関する内容が正確に生成所見に記述されていた場合，再現率は 3/4=0.75 となる．同じようにして，適合率も計算し，F1 値を計算した．この F1 値を人手評価値とした．詳しくは西埜ら [12] の研究を参照されたい．

5.1.2 自動評価

自動評価指標として要約生成や機械翻訳などで広く利用される ROUGE と BLEU を利用する．それぞれを簡単に説明すると，ROUGE は正解テキストに含まれている N-gram を生成テキストがどの程度網羅しているかを表す指標であり，BLEU は生成テキストに含まれている N-gram が正解テキストにどの程度出現しているかを表す指標である．本研究では ROUGE-1，ROUGE-2，ROUGE-L，SentBLEU を利用し，正解所見と生成所見を評価した．実装には SumEval [6]を利用した．

[6]https://github.com/chakki-works/sumeval

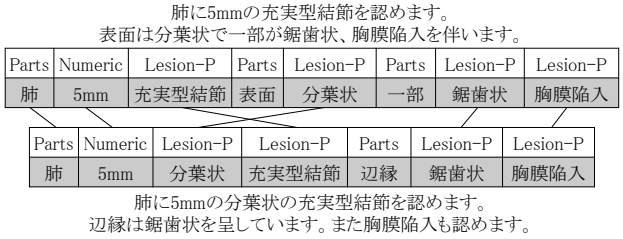

図 2: 評価方法の例

5.1.3 NE による自動評価

NE に基づいた評価では正解所見と生成所見に対して認識された NE の一致に基づいた F1 値を計算する．F1 値は以下の式で計算する．

$$P = \frac{ExactMatchCount}{PredNEs}, \tag{11}$$

$$R = \frac{ExactMatchCount}{GoldNEs}, \tag{12}$$

$$F1 = \frac{2PR}{P+R}. \tag{13}$$

$ExactMatchCount$ とは生成所見と正解所見との間で一致した NE の数である．$PredNEs$，$GoldNEs$ とはそれぞれ生成，正解所見から認識された NE の数である．

評価方法の例を図 2 に示す．まず，正解所見と生成所見に対し，上述したモデルを利用して NE ラベルとモダリティラベルを認識する．その後，認識した NE 間で一致を計算する．この例では 6 個の NE がラベルと表層ともに一致している．そのため，P=6/7，R=6/8 となり，F1=0.8 となる．また，NE 認識モデルには最も高い精度であった BERT－CRF を利用した．

5.2 評価結果

正解所見と生成所見のペアに対して，ROUGE-1, ROUGE-2, ROUGE-L，SentBLEU，NE に基づいた評価手法のそれぞれで評価値を算出し，人手評価値との相関係数を計算した．相関係数を表 5 に示す．NE を利用した評価指標が人手評価との相関係数が最も高いことが確認できる．ROUGE や BLEU などの N-gram の一致に基づいた手法では“内部が空洞になっています”と“内部に空洞を認めます”という 2 つの表現に対して，同じ意味を表現しているにも関わらず，N-gram の一致が少ないため，低い評価値を出力していた．一方で，NE に基づいた手法では“(内部, Parts)”，“(空洞，Lesion-P)”に汎化した上で評価できるため，人手評価と近い評価が可能となった．

5.3 NE に基づいた評価指標の課題

NE に基づいた評価指標に課題はいくつかある．1 つ目は言い換えに対応できない点である．例えば，“充実型結節”は“充実性結節”や“充実型の結節”と様々な表層で記述されるが，現状の評価指標では NE 間のアライメントに文字列間の完全一致を利用しているため，NE の言い換えに対応できない．最近では，BERT から得られるベクトル表現を利用し，要約生成や機械翻訳などの評価において，言い換えに対応できない問題を緩和する評価手法 [1] が提案されている．本研究においても，再学習させた BERT のベクトル表現や編集距離などを利用して，言い換えの問題に対応する必要がある．

ROUGE-1	ROUGE-2	ROUGE-L	SentBLEU	NE
0.256	0.236	0.233	0.242	**0.310**

表 5: 人手評価と各評価指標との相関係数

2 つ目は NE 間の関係性を考慮していない点である．例えば，“**左肺上葉**に不整形陰影，**左肺下葉**に結節を認めます”，と“**左肺下葉**に不整形陰影，**左肺上葉**に結節を認めます”は明らかに異なることを記述しているが，現状の評価指標では違いを評価値に反映できない．正確に評価するためには Parts と Lesion，Disease 間の関係など NE 間の関係性を加味した評価指標が必要である．そのため，現在は NE 間の関係のアノテーションも進めている．

6 おわりに

本研究では生成された読影所見の自動評価に向けた所見の構造化のために，読影所見に対する NER とモダリティ推定を行った．8 種類の NE ラベルと 3 種類のモダリティラベルを設計し，データを作成した．また，CRF，BiLSTM-CRF，BERT，BERT-CRF の 4 つのモデルで精度評価を行った．評価の結果，BERT-CRF が NER，モダリティ推定ともに高い精度を示した．一方で，エラー分析の結果，モデルが予測を間違えた NE の多くがトークナイズ時にサブワード化された NE であることを確認した．この問題を解決するため，大規模な医療テキストでサブワードを構築し，事前学習する方法などが考えられる．

また，訓練したモデルを利用し，実際に生成された読影所見の評価も行った．ROUGE と BLEU の評価値と比べ，NE に基づいた評価指標の方が，人手評価と相関係数が高いことが確認できた．今後は，言い換えに対応できない問題や NE 間の関係性を考慮できない問題に取り組む予定である．

参考文献

[1] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of EMNLP-IJCNLP*, 2019.

[2] Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of EMNLP*, 2017.

[3] 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. *Bioinformatics*, 2019.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of Clinical NLP Workshop*, 2019.

[5] Ken Yano. Neural disease named entity extraction with character-based bilstm+ crf in japanese medical text. *CoRR*, 2018.

[6] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, and Eiji Aramaki. Overview of the ntcir-10 mednlp task. Citeseer, 2013.

[7] 荒牧英治, 若宮翔子, 矢野憲, 永井宥之, 岡久太郎, 伊藤薫. 病名アノテーションが付与された医療テキスト・コーパスの構築. 自然言語処理, 2018.

[8] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *CoRR*, 2015.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, 2019.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, 2016.

[11] 浦澤合, 関根裕人, 乾孝司, 岩倉友哉. 文書からの化合物名抽出のためのサブワード有効性調査. 人工知能学会全国大会論文集, 2019.

[12] 西埜徹, 桃木陽平, 谷口友紀, 田川裕輝, 谷口元樹, 大熊智子, 中村佳児. 不均衡性を考慮した深層強化学習による読影レポート生成. 言語処理学会第 26 回年次大会, 2020.