スパン間の類似性に基づく事例ベース構造予測

大内 啓樹 [1,2]　鈴木 潤 [2,1]　小林 颯介 [2,3]　横井 祥 [2,1]　栗林 樹生 [2,4]　乾 健太郎 [2,1]
[1] 理化学研究所　[2] 東北大学　[3] 株式会社 Preferred Networks　[4]Langsmith 株式会社

hiroki.ouchi@riken.jp
{jun.suzuki,sosk,yokoi,kuribayashi,inui}@ecei.tohoku.ac.jp

1 はじめに

ニューラルネットを用いたモデルによって，構造予測タスクの精度は飛躍的に向上した．その一方で，精度向上と引き換えに，人間によるモデルの予測根拠の解釈が困難なものとなった [8]．解釈性の高い予測は，人間の意思決定や人間-機械間のインタラクションを促進する役割を果たすため，実用上極めて重要な研究課題である [14]．しかしながら，解釈性の高い予測過程を備えたモデルを用いて既存のニューラルモデルと同等の精度を達成することは一般的に困難であり，精度と解釈性を両立する構造予測モデルの構築は未だに成功に至っていない．

本研究の目的は，既存のニューラルモデルと同等の精度を保ちつつ，より解釈性の高いモデルを構築することである．その取り組みとして，**事例ベース学習** (*instance-based learning*) [1] の枠組みに着目する．事例ベース学習は事例間の類似度 (あるいは距離) を学習する機械学習手法の一つであり，解析対象の事例と類似度の高い学習事例のラベルをそのまま出力することを想定している．したがって，分類器ベースのモデルとは異なり，予測に使われた事例を根拠として示すことができる．このような解釈性の高い予測過程にもかかわらず，構造予測タスクに事例ベースの手法からアプローチした研究はほとんど存在しない．

本研究では，構造予測タスクの事例ベースモデルで高精度を達成する鍵は「スパンの特徴ベクトルの学習」であると考える．スパンとは一単語以上から構成される言語単位である．近年の構造予測タスクでは，分類器の学習を通してスパンの特徴ベクトルも学習することによって高精度を達成している [15, 16, 7]．これらの分類器ベースの学習とは異なり，「同じラベルを持つスパン同士が特徴ベクトル空間上で近づくように学習する」手法を提案する．この学習法によって，スパン間の類似性に基づく事例ベースのラベル予測を実現する．3つの構造予測タスク (固有表現抽出，階層型固有表現抽出，統語チャンキング) の評価実験を通して提案手法の有効性を示す．本研究の主な貢献は以下の二点である．

貢献 1: 事例ベース学習の枠組みでスパン特徴ベクトルを学習する初の試み．

貢献 2: 高精度を保ちつつ解釈性の高い構造予測モデルを構築可能であることの実証．

2 関連研究

分類器ベースのニューラルモデルは予測のブラックボックス性 [6] という共通の技術的課題を持つ．近年多くの研究者が，このブラックボックスに対する分析手法や解釈を与える手法の開発に取り組んでいる [14, 9]．本研究では，所与のブラックボックスモデルの解釈性向上をめざす方向性とは異なり，最初から高い解釈性を備えたモデルを構築することを目的とする．

現在のようにニューラルモデルが成功を収める以前，事例ベース学習は多くの構造予測タスクに利用されてきた [11, 12, 2]．しかし，分類器ベースのニューラルモデルの台頭に伴い，事例ベースのモデルに関する研究報告は減少している．ひとつの例外として，BIO タグ付与モデルにおいて事例ベースの学習を導入した Wiseman and Stratos (2019) の研究 [20] がある．事例ベース学習の枠組みで各単語の特徴ベクトルを学習したが，分類器ベースのモデルと比べ精度が低下する結果となった (4 節の表 2 を参照)．その主要因は，ラベリングする単位はスパンであるにも関わらず， 彼らの BIO モデルでは単語単位で独立に類似性を学習する点にあると考えられる (解析単位の不一致)．本研究では，単語単位のラベリングとは異なり，スパン単位のラベリングという方向性で事例ベース構造予測に取り組む．

3 手法

本節では，既存のスパン分類アプローチに基づく構造予測について述べてから，提案手法について詳述する．

3.1 既存手法: スパン分類アプローチ

スパン分類アプローチ [15, 16] では，T 単語からなる入力文 $\boldsymbol{x} = (w_1, w_2, \cdots, w_T)$ の可能なスパンを列挙し，各スパン $s = (a, b)$[*1] にラベル $y \in \mathcal{Y}$ を付与する．例として固有表現抽出を考える．

$$\begin{array}{cccccc} \text{Franz}_1 & \text{Kafka}_2 & \text{is}_3 & \text{a}_4 & \text{novelist}_5 \\ [& \texttt{PER} &] & & \end{array}$$

ここで Franz Kafka は 1 単語目から 2 単語目までのスパンであるため $s = (1, 2)$ と表す．このスパンに対して PER (人間) という固有表現ラベルを予測できれば正

[*1] a と b は文内の単語インデックスを表す: $(1 \leq a \leq b \leq T)$.

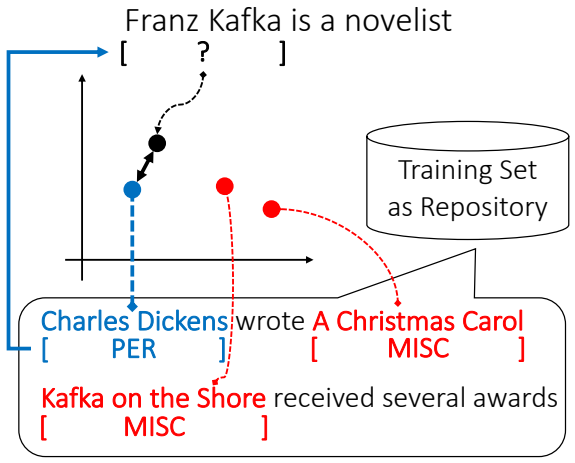


図 1: 事例ベーススパンモデルの概略． 解析対象スパン Franz Kafka を検索クエリとし，学習データ中の類似スパン Charles Dickens のラベル PER を付与している．スパン間の類似度はベクトル空間上で計算する．

解となる．予測時は，入力文 $\boldsymbol{x}$ の可能なスパン集合 $s \in \mathcal{S}(\boldsymbol{x})$ の各スパンに対して多クラス分類を行う．上記の例文の可能なスパン集合は次の通りである: $\mathcal{S}(\boldsymbol{x}) = \{(1,1),(1,2),(1,3),\cdots,(2,2),(2,3),\cdots(5,5)\}$．ここで，固有表現ではないスパンに対しては null ラベルを付与する．例えば，上記の例文の a novelist のスパン $s = (4,5)$ には null を付与する．これは，系列ラベリングで用いられる BIO タグセットの O タグと同様の役割を担っていると解釈できる．したがって，ラベル集合 $\mathcal{Y}$ は各データセットの定めるラベルセット $\mathcal{Y}^{\text{data}}$ と null からなる: $\mathcal{Y} = \mathcal{Y}^{\text{data}} \cup \{\texttt{null}\}$．一般的に，各スパン s にラベル y を割り当てる確率は softmax を用いて以下のようにもとめられる:

$$P(y|s) = \frac{\exp(\mathbf{w}_y \cdot \mathbf{h}_s)}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{w}_{y'} \cdot \mathbf{h}_s)} \ . \tag{1}$$

ここで，各ラベル $y \in \mathcal{Y}$ と紐づく重みベクトル $\mathbf{w}_y$ と各スパンベクトル $\mathbf{h}_s$ の内積から確率が計算される．モデルパラメータは negative log-likelihood の最小化を通して学習されることが多い．このようなモデルを本稿では**分類器ベース**と呼ぶ．

3.2 提案手法: 事例ベーススパンモデル

事例ベーススパンモデルは，解析対象のスパンをクエリとし，特徴ベクトル空間上で類似度が高い (あるいは距離が近い) スパンを学習データから検索することによってラベルを付与する (図 1)．この予測を高精度に行うには，「同一のラベルを持つ事例同士が近くに位置する」ように特徴ベクトル空間を学習する必要がある．本研究では，まず，検索クエリとなるスパン s_i の近傍事例として学習データ中のスパン s_j が選ばれる確率 (*stochastic nearest neighbors* [4]) を定義する．[*2]

$$P(s_j|s_i,\mathcal{D}) = \frac{\exp(\text{sim}(\mathbf{h}_{s_i},\mathbf{h}_{s_j}))}{\sum_{s_k \in \mathcal{S}(\mathcal{D}):i \neq k} \exp(\text{sim}(\mathbf{h}_{s_i},\mathbf{h}_{s_k}))} \ , \tag{2}$$
$$P(s_i|s_i,\mathcal{D}) = 0 \ .$$

ここで，学習データ $\mathcal{D}$ 中のすべての文の可能なスパン集合 $\mathcal{S}(\mathcal{D}) = \{s \in \mathcal{S}(\boldsymbol{x}) \mid (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}\}$ が検索対象となる．$\text{sim}(\mathbf{h}_{s_i},\mathbf{h}_{s_j})$ は両スパンベクトル間の類似度である．この確率をもとに，スパン s_i に正解ラベル y_i を割り当てる確率を求める．

$$P(y_i|s_i) = \sum_{s_j \in \mathcal{S}(\mathcal{D}):y_i = y_j} P(s_j|s_i,\mathcal{D}) \ .$$

ここでは，ラベル $y_i = y_j$ を満たす学習データ中のスパン s_j との近傍確率 (式 2) の合計値を計算している．損失関数として negative log-likelihood を用いる．

$$\mathcal{L} = -\sum_{s_i \in \mathcal{S}(\mathcal{D})} \log P(y_i|s_i) \ .$$

推論時は周辺確率が最大となるラベル $\hat{y}$ を出力する．

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y|s) \ .$$

近傍確率の計算効率化　式 2 の近傍確率 $P(s_j|s_i,\mathcal{D})$ は学習データ $\mathcal{D}$ 中のすべてのスパンに依存する．そのため，計算時間・使用メモリ両面において極めてコストが高く，実際の計算機上で学習を行うことは困難である．そこで，学習データ $\mathcal{D}$ からランダムにサンプリングされた K 文 ($\{\boldsymbol{x}^{(k)}\}_{k=1}^{K}$) を用いることによって計算コストを削減する．このシンプルな計算コスト削減法によって，モデルの学習を単一 GPU で 24 時間以内に終えることが可能になった．

4 実験

4.1 実験設定

データ　3 つの構造予測タスクの標準ベンチマークデータを用いて提案手法を評価する．

1. 固有表現抽出: The CoNLL-2003 dataset [18]
2. 階層型固有表現抽出: The GENIA dataset [5]
3. 統語チャンキング: The CoNLL-2000 dataset [17]

比較モデル　分類器ベースのニューラルモデルとの比較実験を通して，提案手法による事例ベースモデルが効果的に学習できているかを検証する．比較モデルとして，構造予測タスクで高精度を達成している分類器ベー

[*2] 元論文の stochastic nearest neighbors は，各事例の線形変換先の特徴ベクトル空間上におけるユークリッド距離に基づいて定義される．本研究では各事例を (i) 任意の変換先の特徴ベクトル空間上での (ii) 任意の類似度・距離尺度に基づいて確率を定義するため，元論文の一般形と見なすことができる．

表 1: 分類器・事例ベースモデルの比較実験結果．各セルは F_1 値と標準偏差を示している．太字は統計的に有意 (Permutation test: $p < 0.05$) に高い F_1 値を示す．

タスク	分類器ベース	事例ベース
固有表現抽出	90.65 ±0.25	90.70 ±0.06
階層型固有表現抽出	73.76 ±0.35	**74.20** ±0.16
統語チャンキング	94.82 ±0.07	94.83 ±0.08

ススパンモデル (3.1 節) を用いる．公正な比較のため，両モデルのエンコーダ部分 (単語埋め込み[*3]，文字レベル CNN，LSTM) は同じアーキテクチャを用いる．

スパン特徴ベクトルの計算　式 1・2 のスパン特徴ベクトル $\mathbf{h}_s$ として LSTM-minus に基づくベクトル [19] を用いる．まず，入力文 $\boldsymbol{x} = (w_1, w_2, \cdots, w_T)$ を単語埋め込みと文字レベル CNN を用いてベクトル系列 $\mathbf{w}_{1:T} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T)$ に変換し，双方向 LSTM を用いて $\overrightarrow{\mathbf{h}}_{1:T}$ と $\overleftarrow{\mathbf{h}}_{1:T}$ を得る[*4]．これらのベクトルから，各スパン $s = (a, b)$ の両端に基づく特徴ベクトル $\mathbf{h}_s^{\text{lstm}}$ を計算する．固有表現抽出には $\mathbf{h}_s^{\text{lstm}} = [\overrightarrow{\mathbf{h}}_b - \overrightarrow{\mathbf{h}}_{a-1}; \overleftarrow{\mathbf{h}}_a - \overleftarrow{\mathbf{h}}_{b+1}]$ を用いて，階層型固有表現と統語チャンキングには $\mathbf{h}_s^{\text{lstm}} = [\overrightarrow{\mathbf{h}}_b - \overrightarrow{\mathbf{h}}_{a-1}; \overleftarrow{\mathbf{h}}_a - \overleftarrow{\mathbf{h}}_{b+1}; \overrightarrow{\mathbf{h}}_a + \overrightarrow{\mathbf{h}}_b; \overleftarrow{\mathbf{h}}_a + \overleftarrow{\mathbf{h}}_b]$ を用いる．[*5]このベクトルを重み行列 $\mathbf{W}$ で線形変換したものをスパン特徴ベクトルとする: $\mathbf{h}_s = \mathbf{W}\,\mathbf{h}_s^{\text{lstm}}$．分類器ベースモデルでは，このスパン特徴ベクトル $\mathbf{h}_s$ と各ラベル y に紐づく重みベクトル $\mathbf{w}_y$ との内積に基づいて確率分布を計算する (式 1)．事例ベースモデルでは，式 2 のスパン間の類似度としてスパン特徴ベクトル間の内積に基づいて確率分布を計算する: $\text{sim}(\mathbf{h}_{s_i}, \mathbf{h}_{s_j}) = \mathbf{h}_{s_i} \cdot \mathbf{h}_{s_j}$．

モデルのセットアップ　事例ベースモデルの学習時は，各ミニバッチ (サイズ = 8) ごとに $K = 50$ 文をランダムサンプリングして用いる．テスト時は，入力文と類似度の高い $K = 50$ 文を学習データからサンプリングして用いる．この類似度計算には，各文の単語埋め込みの平均ベクトル[*6]を計算し，それらのコサイン類似度を用いる．その他のハイパーパラメータは Ma and Hovy (2016) [10] の設定にしたがう．

4.2　定量評価

この実験では，異なるランダムシードを用いて独立に学習した 5 つのモデルの平均 F_1 スコアを報告する．

分類器ベースモデルとの比較　表 1 は各テストデータに対する F_1 値を示している．固有表現抽出と統語チャンキングでは，両モデルとも同等 (統計的有意差なし) の F_1 値を記録した．階層型固有表現抽出では，事例ベースモデルは分類器ベースモデルを上回る (統計的に有意に高い)F_1 値を達成した．これらの結果は，我々が提案した事例ベース学習によってスパン特徴ベクトルが有効に学習されていることを示唆している．

[*3] 100 次元の GloVe [13] を用いる．

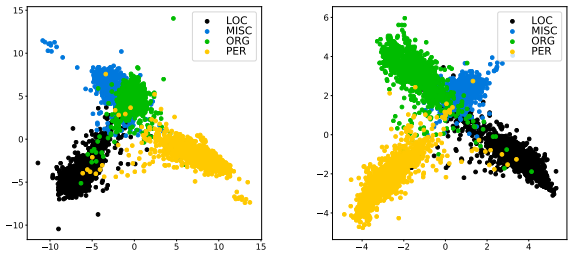
[*4] $\overrightarrow{\mathbf{h}}_t$ と $\overleftarrow{\mathbf{h}}_t$ は前向きと後向きの LSTM の隠れ層を表す．

[*5] 予備実験の結果，スパンの両端を足したベクトルを加えたほうが，階層型固有表現と統語チャンキングでは精度が顕著に高かったため，固有表現抽出とは異なるベクトルを用いる．

[*6] 各文 $\boldsymbol{x} = (w_1, \cdots, w_T)$ に対して $\frac{1}{T}\sum_t \mathbf{w}_t^{\text{glove}}$ とする．

表 2: 既存研究 Wiseman and Stratos (W&S) [20] との比較．各セルは Flat NER の F_1 値と，分類器・事例ベースモデル間の F_1 値の差分 (diff) を示している．

	分類器ベース	事例ベース	diff
W&S [20]	90.76	89.94	−0.82
本研究	90.65	90.70	+0.05



(a) 分類器ベースモデル．　(b) 事例ベースモデル．

図 2: スパン特徴ベクトル空間の可視化．

既存研究との比較　使用している単語埋め込みなどが異なり，直接の比較はできないが，参考のため，表 2 に既存研究 Wiseman and Stratos [20] との F_1 値の比較結果 (CoNLL-2003 評価セット) を示す．彼らは BERT [3] を用いた BIO タグ付与アプローチに基づくモデルを採用している．彼らの事例ベース BIO モデルは分類器ベース BIO モデルと比べて F_1 値が 1 ポイント近く低下している．一方，我々の事例ベーススパンモデルは分類器ベーススパンモデルと同等の F_1 値を保っている．この結果から，高精度を保ちつつ解釈性の高い事例ベースモデルが構築可能であることがわかった．

4.3　定性評価

我々は「同一ラベルを持つスパン同士がベクトル空間上で近くに位置する」ような学習手法を考案した．この意図が実現されているかを，実際に学習された特徴ベクトル空間と近傍の実例の分析を通して確かめる．

スパン特徴ベクトル空間の可視化　図 2 は，固有表現抽出 (CoNLL-2003 開発データ) における分類器・事例ベースモデルのスパン特徴ベクトル $\mathbf{h}_s$(式 1 または 2) を主成分分析で二次元にマップしたものである．これを見ると，分類器ベースのスパン特徴ベクトルは図の中央で重なり合っている部分が多く，特に `ORG` のスパン (緑) と正解ラベル `MISC` のスパン (青) が重なり合っている．一方，事例ベースのスパン特徴ベクトルは重なり合っている部分が比較的少なく，4 つのクラスタに分かれている．これらの結果から，提案手法である事例ベース学習によってスパン特徴ベクトルが効果的に学習できていることがわかった．

表 3: スパン近傍検索の実例．CoNLL-2003 開発データ中の Phil Simmons というエンティティをクエリとし，学習データから近傍 5 件を検索した．

	クエリ	... all-rounder [Phil Simmons] took four for ...
	分類器ベースモデル	
1	null	... [Former England captain Will Carling] ...
2	null	... England [captain Will Carling] along ...
3	PER	... ([A. Giles] 50 ; B. Julian 4-66 , C. Lewis ...
4	null	... [Will Carling along with Jeremy Guscott] ...
5	PER	... with [Jeremy Guscott] , Rory Underwood ...
	事例ベースモデル	
1	PER	... Yorkshire captain [David Byas] completed ...
2	PER	[Ian Botham] began his test career in 1977 ...
3	PER	... paceman [Darren Gough] polishing off ...
4	PER	... [Michael Tucker] homered and drove in ...
5	PER	... Australian captain [Greg Chappell] with ...

表 4: 事例ベースモデルの予測誤り例．CoNLL-2003 開発データ中の Air France の正解ラベルは ORG (組織) であるが，誤って LOC (場所) を付与している．

	クエリ	... spokesman for [Air France] 's pro-Socialist ... 予測:LOC 正解:ORG
1	LOC	... [Colombia] turned down American 's ...
2	LOC	... competition involving [Scotland] , Wales , ...
3	LOC	... deal signed in [Nigeria] 's capital Abuja ...
4	LOC	... general strike in the West Bank and [Gaza] .
5	LOC	... on its way to [Romania] via the former ...

近傍検索の実例　表 3 は，Phil Simmons をクエリとした際の近傍 5 つの検索結果である．分類器ベースモデルでは，エンティティではないスパンと PER (人間) ラベルを持つスパンが混在している．一方，事例ベースモデルでは PER (人間) ラベルを持つスパンが一貫して近傍を占めていることがわかる．[*7]この傾向は他の多くのケースでも同様に観測された．これらの分析結果から，同一ラベルを持つスパンが近くに位置していることを確認できた．また，この一例のように，スパンレベルの検索システムや多様なアプリケーションへの応用可能性も示唆された．

事例ベースモデルの予測誤り例　表 4 は，正解ラベルが ORG (組織) である Air France に対して，誤って LOC (場所) と予測した例を示している．この例のように，組織名に場所名が含まれる事例や，場所名と同一の組織名で表される事例[*8]に対する誤りが比較的多い傾向が見られた．また，Air France の近傍事例として LOC のエンティティが占めていることが誤りの一因となっている．このように，学習事例のどれを根拠として誤った予測をしたのかが容易に分析できる点は，本提案手法の重要な利点であると考えられる．

[*7] クエリの Phil Simmons はクリケット選手であり，近傍事例もクリケット選手が多かった (David Byas, Ian Botham, Darren Gough, Greg Chappell).

[*8] 例えば，Seattle Mariners は Seattle と表記されることも多く，正解は ORG であるが誤って LOC と予測する例が見られる．

5　おわりに

本研究ではスパン特徴ベクトルの事例ベース学習手法を提案した．3 つの構造予測タスクの実験を通して，分類器ベースのニューラルモデルと同等の精度を保ちつつ，より解釈性の高い事例ベースモデルを構築することに成功した．本研究の貢献は，近年盛んに研究されている言語モデルを用いて精度を改善する方向性とは直交するものである．したがって，これからの方向性として，本提案手法を言語モデルで強化した構造予測モデルにも適用することが挙げられる．また，定性評価で示したように，学習したスパン特徴ベクトルを多様なアプリケーションへ応用する方向性も考えられる．

謝辞　本研究は JSPS 科研費 19K20351, 19H04162 の助成を受けたものです．

参考文献

[1] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, Vol. 6, No. 1, pp. 37–66, 1991.
[2] Walter Daelemans and Antal Van den Bosch. *Memory-based language processing*. Cambridge University Press, 2005.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
[4] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Ruslan R Salakhutdinov. Neighbourhood components analysis. In *Proceedings of NIPS*, pp. 513–520, 2005.
[5] J-D Kim, Tomoko Ohta, Yuka Tateisi, Jun' ichiTsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, Vol. 19, No. suppl_1, pp. i180–i182, 2003.
[6] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of ICML*, pp. 1885–1894, 2017.
[7] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of EMNLP*, pp. 188–197, 2017.
[8] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of EMNLP*, pp. 107–117, 2016.
[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of NIPS*, pp. 4765–4774, 2017.
[10] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*, pp. 1064–1074, 2016.
[11] Makoto Nagao. *A framework of a mechanical translation between Japanese and English by analogy principle*. Elsevier Science Publishers, 1984.
[12] Joakim Nivre, Johan Hall, and Jens Nilsson. Memory-based dependency parsing. In *Proceedings of CoNLL*, pp. 49–56, Boston, Massachusetts, USA, 2004.
[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP)*, pp. 1532–1543, 2014.
[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*, pp. 1135–1144, 2016.
[15] Mohammad Golam Sohrab and Makoto Miwa. Deep exhaustive model for nested named entity recognition. In *Proceedings of EMNLP*, pp. 2843–2849, 2018.
[16] Mitchell Stern, Jacob Andreas, and Dan Klein. A minimal span-based neural constituency parser. In *Proceedings of ACL*, pp. 818–827, 2017.
[17] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 shared task chunking. In *Proceedings of CoNLL*, 2000.
[18] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*, pp. 142–147, 2003.
[19] Wenhui Wang and Baobao Chang. Graph-based dependency parsing with bidirectional LSTM. In *Proceedings of ACL*, pp. 2306–2315, 2016.
[20] Sam Wiseman and Karl Stratos. Label-agnostic sequence labeling by copying nearest neighbors. In *Proceedings of ACL*, pp. 5363–5369, July 2019.