Vol.2017-NL-233 No.5
2017/10/24

含意関係認識コーパスの偏りによる性能評価への影響

土屋 雅稔[1,a)]

概要：現在、英語の含意関係認識コーパスとして広く用いられている Stanford Natural Language Inference (SNLI) コーパスの仮説文には，語彙の大きな偏りがある．そのため，仮説文を参照することなく，仮説文のみを用いて含意関係ラベルを推定することが，66%の精度で可能となっている．本稿では，この偏りについての分析を報告すると共に，この偏りが深層学習に対して与える影響について述べる．

Performance Impact Caused by Hidden Bias of Natural Language Inference Corpus

MASATOSHI TSUCHIYA[1,a)]

1. はじめに

ニューラルネットワーク (NN) を含む機械学習的な手法を適用する場合，学習データの品質は，非常に重要な問題の 1 つである．エラーを含まない学習データは存在しないが，通常のエラーは正規分布に従うことが期待されるため，十分な分量の学習データによって克服することが可能である．しかし，学習データのエラーが，正規分布には従わず，バイアスを持っているような場合には，大きな問題となり得る．例えば，Reidsma ら [1] は，対話行為コーパスにおけるアノテーションエラーが一定のバイアスを持っている場合，ベイジアンネットワークの学習結果が偏ることを報告している．NN は，学習データ全体を総当りにメモ化することも可能であるような膨大なパラメータ空間を有する [2]．そのため，このようなバイアスを持つエラーが NN に対して与える影響は，他の機械学習的な手法よりも，より深刻になることが予想される．

本稿では，大規模な含意関係認識タスク用コーパスを対象として，その品質をチェックするための新しい経験的な手法を提案する．提案手法は，2 段階からなる．第 1 段階では，調査対象コーパスに何らかのバイアスが存在しない場合であれば，確実に棄却されるような異常な仮説を設定する．本稿では，そのような異常な仮説を設定するために，含意関係認識タスクそのものの定義に注目する．含意関係認識タスクは，前提文と仮説文の文対を対象として，その分対の関係を分類するタスクとして定義されている [3], [4], [5], [6]．したがって，前提文が不要であり，仮説文の情報のみから分類が可能であるという仮説は，含意関係認識タスクの定義と相容れない．そのため，このような仮説は，バイアスを含まない含意関係認識用コーパスにおいては確実に棄却されると期待される．第 2 段階では，この異常な仮説が棄却されるか否かを，機械学習モデルを用いて検討する．本稿では，機械学習モデルとして，Naive Bayes モデルを用いた．なお，この提案手法は，統計的仮説検定の考え方から大きな示唆を受けてはいるが，統計的仮説検定とは異なり，第 2 段階の機械学習モデルの選択と設計に自由度があるため，異常な仮説が常に棄却されるとは限らない問題点がある．

本稿では，提案手法の有効性を検討するため，Stanford Natural Language Inference (SNLI) コーパス [7] と Sentences Involving Compositional Knowledge (SICK) コーパス [8] を対象とする実験結果を示す．SICK コーパスを対象とする実験結果は，設定した異常な仮説を棄却したのに対して，SNLI コーパスを対象とする実験結果は，仮説を棄却できなかった．この結果から，SNLI コーパスには，前提文によってコンテキスト情報が与えられていない場合であっ

[1] 豊橋技術科学大学 情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology
[a)] tsuchiya@imc.tut.ac.jp

ⓒ 2017 Information Processing Society of Japan

s_1	Two boys are swimming in the pool.
s_2	Two girls are playing basketball.
s_3	Two women are swimming in the pool.
s_h	Two children are swimming in the pool.

図 1 含意関係認識タスク例

ても，仮説文のみから含意関係ラベルが推定できるようなバイアスが含まれていることが示唆される．さらに本稿では，先行研究によって提案されている 2 種類の含意関係認識用 NN モデルを対象として，このバイアスが，どのような性能上の影響を与えているかについて述べる．

2. 提案手法

2.1 異常な仮説の設定

最初に，調査対象コーパスにバイアスが存在しない状態を仮定し，この仮定が満たされている場合には，確実に棄却されるような異常な仮説を設定する．そのような仮説を設定するため，本稿では，調査対象コーパスのタスク定義に注目する．

含意関係認識についてのワークショップ SemEval[6] では，含意関係認識を，前提文と仮説文が与えられた時，その文対に対して，当該文対の関係を表す 3 種類の含意関係ラベル (含意 · 中立 · 矛盾) を付与するタスクと定義している．図 1 に，含意関係認識タスクの具体例を示す．前提文 s_1 と仮説文 s_h が与えられた時，この文対の関係は，前提文 s_1 によって与えられる文脈情報に基づいて，含意と分類される．前提文 s_2 と仮説文 s_h の関係は中立，前提文 s_3 と仮説文 s_h の関係は矛盾と分類される．これらの具体例から明らかに，文 s_h の含意関係ラベルは，前提文によって文脈情報が与えられない限り，決めることができない．逆に考えると，前提文が与えられていないにも関わらず，単独の仮説文から含意関係ラベルを推定することが可能であるならば，そのようなコーパスは，含意関係認識タスクのコーパスとして不適切である．

このような観察に基づき，本稿では，調査対象とする含意関係認識コーパスにおいて隠れたバイアスが存在することを検証するための異常な仮説として，以下を定義する．

仮説 前提文による文脈情報が与えられていない状態で，仮説文のみから含意関係ラベルを決定できる．

この仮説は，明らかに，含意関係認識タスクの自然な直感に反する異常な仮説であり，正常な含意関係認識コーパスにおいては，確実に棄却されることが期待される．逆に考えると，このような異常な仮説が棄却されないようなコーパスは，少なくとも，含意関係認識タスクのコーパスとしては不適切なバイアスが含まれていると考えられる．

2.2 含意関係ラベル推定モデル

提案手法の第 2 段階は，第 1 段階で定義した異常な仮説が棄却されるか否かを，機械学習的手法を用いて検討する段階である．本稿では，Wang ら [9] によって，次式のように定義された Naive Bayes モデルを用いる．

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^{n} P(x_i|y), \tag{1}$$

ここで，y は含意関係ラベル，x_i は素性である．本稿では，仮説文に含まれる全ての単語 unigram を，出現頻度に基づく足切りなども行わずに，素性として用いる．この含意関係ラベル推定モデルが，調査対象コーパスの含意関係ラベルの出現比よりも有意に高い精度を達成できる場合には，第 1 段階で定義した異常な仮説が棄却されていない．

3. 含意関係認識コーパスの構築手順

本節では，本稿の調査対象とする SNLI コーパスと SICK コーパスの構築手順の概要について述べる．

3.1 SNLI コーパス

SNLI コーパスの構築手順は，大まかに 3 つのステップからなる．第 1 ステップは，Flickr30k コーパス [10] から，前提文を収集するステップである．Flickr30k コーパスは，30k 個の写真に対してクラウドソーシングによって付与された約 160k 個の表題からなるコーパスである．これらの表題は，写真そのものを撮影者とは関係がない別の作業者によって作成されているため，写真の情景をそのまま説明する文となっている傾向がある．

第 2 ステップは，Amazon Mechanical Turk (AMT) を用いて，仮説文を収集するステップである．SNLI コーパスの作成者は，AMT の作業者に対して 1 つの前提文を提示し，1 つの前提文に対して含意する仮説文，中立の仮説文，矛盾する仮説文の計 3 文を人手で作成するように依頼した．このように依頼することにより，SNLI コーパスにおいては，含意関係ラベルはほぼ均等に出現することが保証されている (表 2).

第 3 ステップは，作成されたデータの品質を保証するための多数決である．第 1 ステップおよび第 2 ステップによって収集された前提文と仮説文および含意関係ラベルの 3 つ組を，AMT の複数の作業者に提示し，作業者間でデータの妥当性についての合意が得られた 3 つ組のみをコーパスに含めている．

3.2 SICK コーパス

SICK コーパスの構築手順は，大まかに 4 つのステップからなる．第 1 ステップでは，Flickr8k コーパス [11] および SemEval-2012 STS MSR-Video Descriptions データセット[*1]

[*1] http://www.cs.york.ac.uk/semeval-2012/

表 1 含意関係認識コーパスの比較

	SNLI	SICK
学習セット文対数	55k	4,500
開発セット文対数	10k	500
テストセット文対数	10k	4,927
前提文平均単語長	14.1	9.8
仮説文平均単語長	8.3	9.5
学習セット語彙数	36,427	2,178
テストセット語彙数	6,548	2,188
テストセット未知語率	0.25%	0.32%

から，ランダムに文を選択する．これらの文は全て，英語で記述された情景の説明文である．第 2 ステップでは，第 1 ステップで得られた文を正規化する．第 3 ステップでは，第 2 ステップで得られた文に対して意味的に類似した文や意味的に反対の文を，いくつかの人手で作成された規則を用いて作成する．規則としては，類義語や反義語への単語の置換，否定表現の削除などがある．最後の第 4 ステップでは，第 3 ステップで作成された文対を対象として，AMT の作業者に依頼して，含意関係ラベルを付与する．

3.3 SNLI コーパスと SICK コーパスの比較

SNLI コーパスと SICK コーパスの類似点は，2 つの観点から説明することができる．第 1 に，両コーパスは，先述の通り，情景に対する英語説明文から作成されているという点で類似している．そのため，両コーパスの前提文および仮説文の平均単語長はよく似通っている (表 1)．第 2 に，両コーパスは，語彙の点からも，きわめて類似している．SNLI コーパスの学習セットに含まれる語彙を既知語と見なして，SICK コーパスのテストセットの未知語率を求めると，0.05% という極めて低い値が得られた．このような低い未知語率は，SICK コーパスのテストセットは，SNLI コーパスの学習セットと，ほぼ同じ語彙からなることを示している．

それに対して，SNLI コーパスと SICK コーパスは 2 つの点で異なる．第 1 の相違点は，文の作成手順である．SNLI コーパスの仮説文は，人間の作業者によって作成されているのに対して，SICK コーパスの前提文と仮説文は，原型となる文から人手作成された規則に基づいて自動生成されている．作業手順から明らかに，仮説文を作成する SNLI コーパスの作業者は，作業前に提示された前提文に起因するバイアスから逃れることはできない．第 2 の相違点は，表 2 に示す通り，SNLI コーパスの含意関係ラベル出現比がほぼ均等であるのに対して，SICK コーパスの含意関係ラベル出現比は均等ではない．この相違点もまた，両コーパスの構築手順の相違によるものである．

task6/index.php?id=data

4. 実験

表 3 に，前提文を参照することなく，仮説文のみを用いて含意関係ラベルを推定した結果を示す．Naive Bayes モデルを実装するためのフレームワークとしては，`scikit-learn`[12] を用いた．SNLI コーパスの仮説文のみを用いて学習した含意関係ラベル推定モデルは，63.3%の精度を達成した．これは，SNLI コーパスの含意関係ラベル出現比 (表 2) から期待されるチャンスレシオ 34.3%よりも明らかに有意に高い精度である．それに対して，SICK コーパスの仮説文のみを用いて学習した含意関係ラベル推定モデルは，56.7%の精度しか達成できなかった．これは，SICK コーパスの含意関係ラベル出現比 (表 2) から期待されるチャンスレシオ 56.7%とほとんど変わらない精度である．

図 2 は，SNLI コーパスによって学習した含意関係ラベル推定モデルと，SICK コーパスによって学習した含意関係ラベル推定モデルとの違いを明らかに示している．図 2 の左側の混同行列は，SNLI コーパスによって学習した含意関係ラベル推定モデルから得られた行列であり，右側の混同行列は，SICK コーパスによって学習した含意関係ラベル推定モデルから得られた行列である．右側の混同行列から明らかに，SICK コーパスによって学習したモデルは，個々の仮説文に対して適切な含意関係ラベルを推定するのではなく，単に最も多数回出現しているラベル (中立) を出力している．それに対して，SNLI コーパスによって学習したモデルは，個々の仮説文に対して適切な含意関係ラベルを推定 · 出力しようとしている．

これらの結果は，SICK コーパスに対しては，先に定義した異常な仮説が棄却されていることを示している．同時に，SNLI コーパスに対しては，先に定義した異常な仮説が棄却されておらず，SNLI コーパスには仮説文のみから含意関係ラベルを推定できるようなバイアスが含まれていることを示唆している．

5. 検討

4 節で述べた通り，SNLI コーパスについては，2 節で定義した異常な仮説が棄却されない．したがって，SNLI コーパスには，何らかのバイアスが存在していると考えられる．SNLI コーパスは，既に含意関係認識タスクの学習データとして，多数の研究に利用されているため [7], [14], [15], [16], [17], [18], [19], [20], [21], [22] ，このバイアスによる影響の大きさを検討することは重要である．本節では，この影響の大きさについて検討する．

5.1 SNLI コーパスの経験的分割

SNLI コーパスのテストセットを，SNLI コーパスの学習

ⓒ 2017 Information Processing Society of Japan

表 2 含意関係ラベルの分布

	SNLI コーパス			SICK コーパス		
	学習セット	開発セット	テストセット	学習セット	開発セット	テストセット
含意	183,416 (33.4%)	3,329 (33.8%)	3,368 (34.3%)	1,299 (28.9%)	144 (28.8%)	1,414 (28.7%)
中立	182,764 (33.3%)	3,235 (32.9%)	3,219 (32.8%)	2,536 (56.4%)	282 (56.4%)	2,793 (56.7%)
矛盾	183,187 (33.4%)	3,278 (33.3%)	3,237 (33.0%)	665 (14.8%)	74 (14.8%)	720 (14.6%)
計	549,367	9,842	9,824	4,500	500	4,927

推定ラベル	正解ラベル		
	含意	中立	矛盾
含意	2275	644	706
中立	508	1976	563
矛盾	585	599	1968

(a) SNLI コーパス

推定ラベル	正解ラベル		
	含意	中立	矛盾
含意	3	3	2
中立	1411	2790	718
矛盾	0	0	0

(b) SICK コーパス

図 2 含意関係ラベル推定結果の混同行列

表 3 仮説文のみを用いた含意関係ラベルの推定結果

コーパス	精度
SNLI コーパス	63.3%
SICK コーパス	56.7%

表 4 含意関係ラベル推定モデルによる SNLI コーパスの分割

	E_e	H_e
含意	2,275 (36.6%)	1,093 (30.3%)
中立	1,976 (31.8%)	1,243 (34.5%)
矛盾	1,968 (31.6%)	1,269 (35.2%)
計	6,219 (63.3%)	3,605 (36.7%)

セットに基づいて作成された含意関係ラベル推定モデルを用いて，2 つのサブセットに分割する．第 1 のサブセットは，含意関係ラベル推定モデルによって推定されたラベルが正解ラベルと一致する文対からなるサブセットである．このサブセット E_e を，(含意関係ラベルが経験的に) **推定可能サブセット**と呼ぶ．第 2 のサブセットは，第 1 のサブセットの補集合であり，含意関係ラベル推定モデルによって推定されたラベルが正解ラベルと一致しない文対からなるサブセットである．このサブセット H_e を，(含意関係ラベルが経験的に) **推定困難サブセット**と呼ぶ．

表 4 に，推定可能サブセットと推定困難サブセットを分類した結果を示す．SNLI コーパスのテストセットに含まれる文対の内，63.3%は推定可能サブセット E_e に分類され，残りは推定困難サブセット H_e に分類される．表 4 より，推定可能サブセット E_e と推定困難サブセット H_e の含意関係ラベルの出現比は，大きく異なってはいないため，含意関係ラベルの種別は，推定可能 · 推定困難の違いに対して大きな影響を与えていないと考えられる．

5.2 含意関係認識用 NN モデル

本稿では，2 つの含意関係認識用 NN モデルを対象として，SNLI コーパスに存在するバイアスの影響を検討する．

第 1 のモデルは，Bowman ら [7] によって提案され，Mou ら [13] によって性能評価がなされたモデルである．以後，本稿では，このモデルを並列 LSTM モデルと呼ぶ．並列 LSTM モデルは，次の式によって定義される．

$$
\begin{aligned}
\boldsymbol{h}_{p,i} &= \mathrm{LSTM}_p(W_e x_{p,i} + W_{hp}\boldsymbol{h}_{p,i-1}) \\
\boldsymbol{h}_{h,i} &= \mathrm{LSTM}_h(W_e x_{h,i} + W_{hh}\boldsymbol{h}_{h,i-1}) \\
\boldsymbol{l}_1 &= \tanh(W_1[\boldsymbol{h}_{p,|\boldsymbol{x_p}|}, \boldsymbol{h}_{h,|\boldsymbol{x_h}|}] + B_1) \\
\boldsymbol{l}_2 &= \tanh(W_2\boldsymbol{l}_1 + B_2) \\
\boldsymbol{l}_3 &= \tanh(W_3\boldsymbol{l}_3 + B_3) \\
\boldsymbol{y} &= \mathrm{softmax}(\boldsymbol{l}_3)
\end{aligned}
$$

第 2 のモデルは，Rocktaschel ら [14] によって提案されたモデルである[*2]．以後，本稿では，このモデルを直列 LSTM モデルと呼ぶ．直列 LSTM モデルは，次の式によって定義される．

$$
\begin{aligned}
\boldsymbol{h}_{p,i} &= \mathrm{LSTM}_p(W_e x_{p,i} + W_{hp}\boldsymbol{h}_{p,i-1}) \\
\boldsymbol{h}_{h,0} &= \mathrm{LSTM}_h(W_{hh}\boldsymbol{h}_{p,|\boldsymbol{x_p}|}) \\
\boldsymbol{h}_{h,i} &= \mathrm{LSTM}_h(W_e x_{h,i} + W_{hh}\boldsymbol{h}_{h,i-1}) \\
\boldsymbol{l} &= \tanh(W_l\boldsymbol{h}_{h,|\boldsymbol{x_h}|} + B_l) \\
\boldsymbol{y} &= \mathrm{softmax}(\boldsymbol{l})
\end{aligned}
$$

5.3 含意関係認識用 NN モデルに対するバイアスの影響評価

表 5 は，並列 LSTM モデルおよび直列 LSTM モデルに対して，SNLI コーパスを用いて学習とテストを行なった結果を示す．両モデルともに，テストセット全体および推定可能サブセット E_e に対しては高精度を達成しているにも関わらず，推定困難サブセット H_e に対しては，非常に精度が劣化していることが分かる．このような大きな劣化は，テストセット全体に対して得られた両モデルの高精度

[*2] Rocktaschel ら [14] は，前提文を処理する LSTM と仮説文を処理する LSTM の間に attention を考慮したモデルも提案しているが，本稿では議論の簡単さのために，attention なしのモデルを用いる．

表 5 含意関係認識用 NN モデルの性能

モデル	先行研究における結果	$E_e \cup H_e$	E_e	H_e
並列 LSTM モデル	76.3% [13]	76.8%	87.8%	57.8%
直列 LSTM モデル	80.9% [14]	81.4%	90.1%	65.6%

表 6 仮説文の全単語を未知語シンボルに置換えた場合の含意関係認識用 NN モデルの性能

モデル	$E_e \cup H_e$	E_e	H_e
並列 LSTM モデル	54.1%	66.0%	33.7%
直列 LSTM モデル	48.6%	56.7%	34.7%

は，そのかなりの部分が，推定可能サブセット E_e による底上げによるものであることを示唆している．

表 6 は，SNLI コーパスの学習セットによって訓練された 2 つの NN モデルに対して，前提文の全単語を未知語シンボルに置換した場合の性能を示す．前提文の全単語を未知語シンボルに置換することにより，前提文によって与えられるコンテキスト情報はほぼ全てが失われる (厳密には，前提文の単語長の情報のみは残る)．よって，仮に，両モデルが，前提文によって与えられるコンテキスト情報に基づいて仮説文の含意関係ラベルを推定しているならば，大きな性能劣化が観察されるはずである．しかし，両モデルともに，推定可能サブセット E_e に対しては，表 4 に示されたチャンスレシオ 36.8%よりも明らかに高い精度を達成している．この結果は，両モデルが，推定可能サブセット E_e に対しては，前提文によるコンテキスト情報を参照せず，仮説文の情報のみに基づいて動作するモデルとなっていることを示唆している．このような両モデルの振る舞いは，両モデルの提案者が本来期待している振る舞いとは大きく異なっていると考えられる．

6. 結論

本稿では，大規模な含意関係認識タスク用コーパスを対象として，その品質をチェックするための新しい経験的な手法を提案した．提案手法は，2 つの段階からなる．第 1 段階は，仮に調査対象コーパスにバイアスが含まれていない場合には確実に棄却されると考えられる異常な仮説を設定する段階であり，第 2 段階は，機械学習モデルを用いて当該仮説が棄却されるか否かを検討する段階である．本稿では，この提案手法を用いて，SICK コーパスに対してはバイアスが含まれるという仮説は棄却されること，同時に，SNLI コーパスに対してはバイアスが含まれるという仮説が棄却できないことを示した．

また本稿では，このバイアスが，SNLI コーパスに基づいて学習した含意関係認識用 NN モデルに対して，大きな影響を与えていることを示した．特に，前提文の単語を未知語シンボルに置換する実験を通じて，SNLI コーパスに基づいて学習した含意関係認識用 NN モデルは，実際には前提文を参照しておらず，仮設文のみを参照して動作していることを示した．この結果は，隠れたバイアスを含む学習データに基づいて学習した NN モデルは，その NN の設計者がまったく意図しない動作をするモデルとなっている可能性があることを示唆している．

今後の課題としては，本提案手法を別のタスクのコーパスに対して適用することを考えている．

Acknowledgments

A part of this research was supported by JSPS KAKENHI Grant No. 15K12097. I would like to express my sincere appreciation to Dr. Mitsuo Yoshida and Dr. Adam Meyers for their valuable comments.

参考文献

[1] Reidsma, D. and Carletta, J.: Reliability Measurement Without Limits, *Computational Linguistics*, Vol. 34, No. 3, pp. 319–326 (online), DOI: 10.1162/coli.2008.34.3.319 (2008).

[2] Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O.: Understanding deep learning requires rethinking generalization, *Proceedings of The International Conference on Learning Representations (ICLR2016)*, (online), available from ⟨http://arxiv.org/abs/1611.03530⟩ (2017).

[3] Condoravdi, C., Crouch, D., de Paiva, V., Stolle, R. and Bobrow, D. G.: Entailment, intensionality and text understanding, *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning* (Hirst, G. and Nirenburg, S., eds.), pp. 38–45 (online), available from ⟨http://www.aclweb.org/anthology/W03-0906.pdf⟩ (2003).

[4] Bos, J. and Markert, K.: Recognising Textual Entailment with Logical Inference, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, Association for Computational Linguistics, pp. 628–635 (online), available from ⟨http://www.aclweb.org/anthology/H/H05/H05-1079⟩ (2005).

[5] MacCartney, B. and Manning, C. D.: An extended model of natural logic, *Proceedings of the Eight International Conference on Computational Semantics*, Tilburg, The Netherlands, Association for Computational Linguistics, pp. 140–156 (online), available from ⟨http://www.aclweb.org/anthology/W09-3714⟩ (2009).

[6] Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S. and Zamparelli, R.: SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, Association for Computational Linguistics and Dublin City University, pp. 1–8 (online), available from

ⓒ 2017 Information Processing Society of Japan

⟨http://www.aclweb.org/anthology/S14-2001⟩ (2014).

[7] Bowman, S. R., Angeli, G., Potts, C. and Manning, C. D.: A large annotated corpus for learning natural language inference, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 632–642 (online), available from ⟨http://aclweb.org/anthology/D15-1075⟩ (2015).

[8] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R. and Zamparelli, R.: A SICK Cure for the Evaluation of Compositional Distributional Semantic Models, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S., eds.), Reykjavik, Iceland, European Language Resources Association (ELRA), pp. 216–223 (online), available from ⟨http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf⟩ (2014).

[9] Wang, S. and Manning, C.: Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island, Korea, Association for Computational Linguistics, pp. 90–94 (online), available from ⟨http://www.aclweb.org/anthology/P12-2018⟩ (2012).

[10] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 67–78 (online), available from ⟨http://www.aclweb.org/anthology/Q/Q14/Q14-1006.pdf⟩ (2014).

[11] Rashtchian, C., Young, P., Hodosh, M. and Hockenmaier, J.: Collecting Image Annotations Using Amazon's Mechanical Turk, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, Association for Computational Linguistics, pp. 139–147 (online), available from ⟨http://www.aclweb.org/anthology/W10-0721⟩ (2010).

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).

[13] Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L. and Jin, Z.: How Transferable are Neural Networks in NLP Applications?, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics, pp. 479–489 (online), available from ⟨https://aclweb.org/anthology/D16-1046⟩ (2016).

[14] Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kociský, T. and Blunsom, P.: Reasoning about Entailment with Neural Attention, *Proceedings of The International Conference on Learning Representations (ICLR2016)*, (online), available from ⟨http://arxiv.org/abs/1509.06664⟩ (2015).

[15] Yin, W., Schtze, H., Xiang, B. and Zhou, B.: ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs, *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 259–272 (online), available from ⟨http://www.aclweb.org/anthology/Q16-1019⟩ (2016).

[16] Mou, L., Men, R., Li, G., Xu, Y., Zhang, L., Yan, R. and Jin, Z.: Natural Language Inference by Tree-Based Convolution and Heuristic Matching, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 130–136 (online), available from ⟨http://anthology.aclweb.org/P16-2022⟩ (2016).

[17] Wang, S. and Jiang, J.: Learning Natural Language Inference with LSTM, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, Association for Computational Linguistics, pp. 1442–1451 (online), available from ⟨http://www.aclweb.org/anthology/N16-1170⟩ (2016).

[18] Liu, P., Qiu, X., Chen, J. and Huang, X.: Deep Fusion LSTMs for Text Semantic Matching, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 1034–1043 (online), available from ⟨http://www.aclweb.org/anthology/P16-1098⟩ (2016).

[19] Liu, P., Qiu, X., Zhou, Y., Chen, J. and Huang, X.: Modelling Interaction of Sentence Pair with Coupled-LSTMs, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics, pp. 1703–1712 (online), available from ⟨https://aclweb.org/anthology/D16-1176⟩ (2016).

[20] Cheng, J., Dong, L. and Lapata, M.: Long Short-Term Memory-Networks for Machine Reading, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics, pp. 551–561 (online), available from ⟨https://aclweb.org/anthology/D16-1053⟩ (2016).

[21] Parikh, A., Täckström, O., Das, D. and Uszkoreit, J.: A Decomposable Attention Model for Natural Language Inference, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics, pp. 2249–2255 (online), available from ⟨https://aclweb.org/anthology/D16-1244⟩ (2016).

[22] Sha, L., Chang, B., Sui, Z. and Li, S.: Reading and Thinking: Re-read LSTM Unit for Textual Entailment Recognition, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, The COLING 2016 Organizing Committee, pp. 2870–2879 (online), available from ⟨http://aclweb.org/anthology/C16-1270⟩ (2016).

ⓒ 2017 Information Processing Society of Japan