知識の整理のための根拠付き自然文間含意関係コーパスの構築

林部 祐太
株式会社リクルート Megagon Labs, Tokyo, Japan
hayashibe@megagon.ai

1 はじめに

旅行情報サイト「じゃらん net」[1]には日々レビューが投稿され，さまざまな感想・意見・要望が集積されている．実際に宿泊したカスタマーの目線からなる情報であり，有益な情報源である．これを用いて，どのような宿の特徴がカスタマーに好評であるかという知識を得て整理をすることは，宿を提案する上で有益である．本研究では，1 文を知識の単位として扱い，ある 1 文と含意関係にある複数の文をその 1 文に集約することで知識を整理することを目的とする．そして，その目的のために文間含意関係認識器の学習用コーパスの構築に取り組む．関係の表現形式は前提文が仮説文を含意するか否かの 2 値[2]とする．

これまでの含意関係コーパス構築ではアノテータに 1 文を提示し，事例[3]が与えた関係ラベルをもつようにもう 1 文を作文[4]してもらう方式が一般的である．しかしこのような作成方法は高コストであり，また偏った事例を収集してしまう可能性がある．例えば，Tsuchiya は SNLI コーパス [2] にはもラベルの予測が前提文なしに可能な事例が多数コーパス内に存在することを示した．そこで，本研究ではアノテータによる作文は行わず，**自然文[5]のみで含意関係コーパスを構築**する．また，応用と分析のしやすいコーパスを目指し，原子的な事例のみを含めるため，**複雑な文構造をもつ文は除外**した．加えて，アノテーション作業を確実にし，事例の分析もしやすくするために，非含意の場合はその根拠となる箇所を選択してもらった．著者が知る限りこのような**根拠付きのアノテーション**は含意関係アノテーションにおいて初めての試みである．

以上の特徴をもつ本研究で作成したコーパスは公開を予定しており，アノテーション候補選定のために作成した宿の特徴判定文コーパスや感情極性コーパスも公開する予定である．

[1] https://www.jalan.net/
[2] 含意，矛盾，中立といった 3 値とする先行研究もある．
[3] 本論文では仮説文，前提文，関係ラベルの組を事例とよぶ．
[4] 参考となる文を探索し，所定の関係ラベルになるように改変する形での作文もある．
[5] コーパス構築のために作られたわけではない文を本研究では自然文とよぶ．

2 含意関係アノテーション候補選定

この節では，仮説文と前提文の候補の選定方法について述べる．過去に著者が肯定的な観点を抽出した際は構文解析結果に対して人手で定義したルールを用いた [7] が，得られる文数は少なく，ルールの記述の労力も大きいという問題があった．そこで，機械学習によるフィルタリングを用いることで，大量に肯定的な宿の特徴の文を得る．

まず，旅行情報サイト「じゃらん net」に投稿された宿のレビューの一部を用いて，宿レビューコーパスとした．このコーパスは 2,000 万文以上を含む．そして，ノイズとみなせる文や，複雑な文構造をもつ文を次のフィルタを用いてアノテーション対象外とした．

- 頻度フィルタ：コーパス内頻度が 1 の文を除外
- 文長フィルタ：5 文字未満 41 文字以上の文を除外
- 述語項構造フィルタ：
  JUMAN++ (v2.0.0-rc3)[5][6]の形態素解析，KNP++ (0.9-21cc58c)[3][7]の構文・格解析の結果に基づき，複数の述語項構造を含む文を除外

また，好評な宿の特徴を述べた文のみを対象とするため，BERT を使い（2.1 節で詳述），「宿の特徴フィルタ」（2.2 節で詳述）と「感情極性フィルタ」（2.3 節で詳述）も用いた．こうして 21,868 種類の文をアノテーション候補の文として用意した．

2.1 BERT

BERT[8]は Transformer [6] に基づくモデルで，テキストをベクトルに変換するモデルである．アノテーションされていない大規模テキストから pre-training し，タスクごとに fine-tuning して用いる．すでにウェブテキストや Wikipedia などに用いた pre-training 済みのモデルがいくつか公開されているが，予備実験により宿レビューコーパスを用いて pre-training したモデルを fine-tuning した方が良い性能となった．そのため，pre-training から BERT モデルを作成した．

[6] https://github.com/ku-nlp/jumanpp
[7] 著者より直接入手．
[8] https://github.com/google-research/bert

表 1: 宿の特徴の有無と感情極性のアノテーションの例

文	特徴	感情
部屋も清潔でとても過ごしやすかったです。	有	肯定
全く美味しくない。	有	否定
また利用したいですありがとうございました。	無	肯定
結婚記念日旅行に利用させていただきました。	無	中立

トークナイズには語彙サイズを 32,000 となるように宿レビューコーパスを用いて学習した SentencePiece[9] [4] モデルを用いた．BERT のパラメータは BERT のウェブサイトで公開されているモデル BERT-Base と同じように，バッチサイズは 512，Attention heads の数は 12，レイヤ数は 12，隠れレイヤ数は 12 とした．TPU を用いて 150 万ステップ学習し，Masked token prediction の精度は 56.8，Next sentence prediction の精度は 95.0 となった．

2.2 宿の特徴フィルタ

宿の特徴とは関係ない記述を除外するフィルタの作成のため，クラウドソーシングを用いて関係の有無の 2 値のアノテーションを行った．アノテーションの質を確保するため，アノテーション対象の 10 問にあらかじめ正解を人手でアノテーションしたチェック問題を 2 問加えて，12 問を 1 タスクとした．そして，チェック問題に正解した場合のみタスクの回答を受理し，アノテータに報酬を与えた．5,970 文に対して，各文 5 人がアノテーションした．表 1 にアノテーション例を示す．

5,370 文を用いて BERT を fine-tuning することで分類器を作成した．600 文で評価したところ，分類性能は F 値で 94.0 となった．この分類器で，宿の特徴に関係無いと判断された文を除外した．

2.3 感情極性フィルタ

感情極性が肯定的な文のみを取りだすフィルタの作成のため，文の感情極性が肯定的であるか，否定的であるか，中立的であるかの 3 値で 2.2 節同様にクラウドソーシングを用いて 5,970 文にアノテーションした．表 1 にアノテーションの例を示す．

5,370 文を用いて BERT を fine-tuning することで分類器を作成した．600 文で評価したところ，分類性能は F 値で 90.2 となった．この分類器で，感情極性が否定的もしくは中立的と判断された文を除外した．

3 含意関係アノテーション

3.1 アノテーションの作業方法

含意関係コーパスでは一般に，関係ラベルのみアノテーションされている．そのため，アノテーションの検証や，システム出力の分析が困難である．また，単にラベルを選択するアノテーションであると，アノテータによっては真剣にアノテーションを行わないことがある．これは，アノテータが得る報酬はアノテーションした事例の数によって決まるからである[10]．さらに，予備実験では，文を細かく読まずに，仮説文と前提文が似ていれば含意と安易にアノテーションするアノテータも一定数存在した．

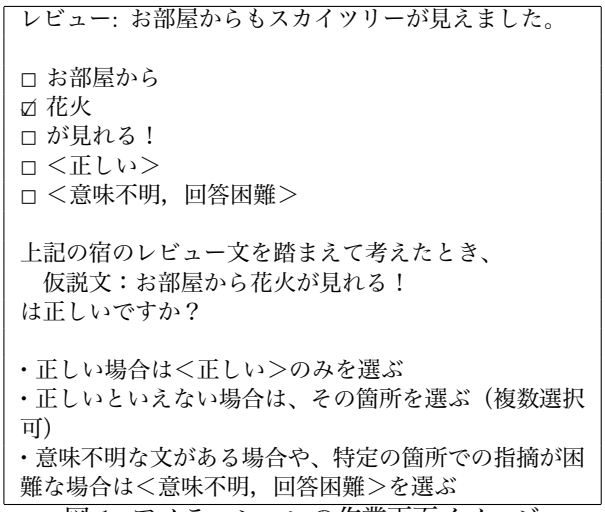

図 1: アノテーションの作業画面イメージ

そこで，非含意の場合はその根拠となる箇所の選択を必須とし，より正確な作業をしてもらえるようにした．作業画面のイメージを図 1 に示す．非含意の場合は根拠となるトークンを選択してもらい，判定困難[11]または含意の場合は該当する選択肢を選択してもらった．トークナイズには，語彙サイズを 8,000 となるように宿レビューコーパスを用いて学習した SentencePiece を用いた．

3.2 類似文を用いた 2 段階含意関係アノテーション

仮説文から事例を作る際に，前提文は意味的に類似している文の中から選んだ．これは，類似している文は含意関係にある可能性が高いという仮定に基づいている．類似文度の計算には文のベクトル表現の内積を用いた．BERT のモデルは各トークンをベクトルにエンコードし，文全体はエンコードしない[12]ため，[1] で提案されている手法を用いて文ベクトルを得た．

より認識が難しい事例にアノテーションするため，アノテーションを 2 回に分け，2 回目のアノテーションは，

[9] https://github.com/google/sentencepiece

[10] 複数の事例とチェック問題をまとめて 1 セットとして行い，チェック問題すべてに正解した場合のみそのセットに対する報酬を与える方式でも，チェック問題のチャンスレートが高い場合は，すべての問題を適当に答え量をこなすことで報酬を得ようとするアノテータが現れやすい．

[11] 評価実験では非含意として扱う．

[12] 文全体を表す特殊なトークン [CLS] のエンコードは可能であるが，fine-tuning なしではエンコードの性能が低く実用的ではない．

表 2: アノテーションした事例数

名称	目的	含意	非含意 (うち回答困難)	合計
BASE	学習	2,826	4,432 (833)	7,258
	評価	278	529 (93)	807
APPEND	学習	2,401	5,699 (692)	8,100
	評価	269	631 (81)	900

1 回目のアノテーション結果を利用した．

第 1 回含意関係アノテーション

仮説文は 2 節で候補とした 21,868 種の文から頻度順に 2,017 種の文を選んだ．ただし，同一の述語をもつ文は最大 5 文までとし，また，含まれる述語項構造と名詞構造が全く同じとなる文が複数含まれないようにすることで，仮説文のバリエーションを増やすようにした．次に，2,017 種の各文に対する類似度が 10 位と 100 から 103 位の 5 文を選んだ[13]．これを前提文とすることで，計 2017×5 件の事例を得た．ここからランダムに 8,065 件抽出し，クラウドソーシングを用いて 1 件につき 3 人がアノテーションした．そのアノテーション結果を元に，著者が最終的なラベルを決定した．このサブコーパスを以下，BASE とよぶ．

第 2 回含意関係アノテーション

21,868 種類の文から頻度順に第 1 回アノテーションと同様の制約で 3,843 文を選んだ．ただし，同一の述語をもつ文は最大 20 文までに緩和した．次に，選んだ各文に対する類似度が 11 位から 20 位までの 10 文を取り出し，2 文の組を 10 組作った．1 つの組からは仮説文と前提文を入れ替えて，2 事例作られる．こうして $3840 \times 10 \times 2 = 76860$ 事例を作成した．これを BASE の学習用データを用いて学習した含意関係分類器[14]を用いて含意関係を 2 値分類し，スコア 0.8 以上の含意とされた事例を 1,000 件，スコア 0.2 以下の非含意とされた事例を 1,000 件，スコア 0.2 超過 0.8 未満の事例を 3,000 件，計 5,000 件抽出した．また，類似度が 951 位から 955 位までの 5 文に対しても同様に事例作成と含意関係の自動分類し，スコア 0.8 以上の含意とされた事例を 2,000 件，スコア 0.2 以下の非含意とされた事例を 500 件，スコア 0.2 超過 0.8 未満の事例を 1,500 件，計 4,000 件抽出した．これら合わせて計 9,000 件に対してクラウドソーシングを用いて 1 件につき 3 人がアノテーションした．そのアノテーション結果を元に，著者が最終的なラベルを決定した．このサブコーパスを以下，APPEND とよぶ．

[13]含意・非含意を均等に取れそうな類似度を経験的に選んだ．

[14]BASE の最終的なラベルは第 2 回アノテーション後も修正したため，4 節の M_{BASE} とは異なる．

表 3: 含意関係認識の性能

評価データ	BASE			APPEND		
モデル	P	R	F1	P	R	F1
M_{BASE}	88.5	96.8	**92.4**	72.3	79.6	75.8
M_{APPEND}	80.4	91.4	85.5	80.0	83.3	81.6
M_{ALL}	87.1	97.1	91.8	82.1	83.3	**82.7**

4 構築した含意関係コーパスの分析

4.1 アノテーションが難しい事例

著者が確認作業を行う中で，アノテータによって判断が分かれやすいと感じた事例や，判断が難しいと感じた事例を記す．

語句に曖昧性がある事例

語義の判断が困難な事例はアノテーションが分かれやすい．例えば，前提文「ご飯が美味しい」が仮説文「米が美味しい」を含意するかどうかは「ご飯」を「食事」として広義にとらえるか,「米」として狭義にとらえるかによって変わる．他の例として「刺身[15]」や「休む[16]」などがある．このような事例はクラウドソーシング前に除外し，別途用意した基準に基いた，専門家によるアノテーションの方が良いと考える．

反例の想定に個人差がある事例

露天風呂は温泉とは限らず，海にビーチがあるとも限らない．しかし，仮説文「目の前がビーチです。」と前提文「目の前が海です。」の事例と仮説文「温泉、良かったです！」と前提文「ゆっくり露天風呂に入れました。」の事例のそれぞれに 1 人が「正しい」とアノテーションした．

また，大きくても柱が多い駐車場は停めにくかったり，熟成させた刺身も美味しかったりするが，前提文「駐車場が大きい。」と仮説文「駐車場が停めやすい。」の事例には 2 人が，前提文「お刺身が新鮮。」と仮説文「刺身が美味しい！」の事例には 1 人が,「正しい」とアノテーションした．このように，どの程度反例を想定できるかは個人差がある．

項の省略がある事例

前提文「のんびり入る事が出来ました。」が仮説文「ゆったりと入浴できました。」を含意するかは前提文の「入る」の省略された項に依存する．3 人中 2 人は（「風呂に」が省略されていると考え）含意とアノテータした．しかしながら，これは提示した仮説文の文言に影響されていると考えられる．

[15]狭義には魚介類の素材の小片だが，広義には馬肉や湯葉なども刺身として提供されうる．

[16]睡眠を含める場合と含めない場合がある．

表 4: 含意関係認識の評価結果の例．E は含意，NE は非含意を示す．

#	仮説文	前提文	正解	M_{BASE}	M_{APPEND}	M_{ALL}
1	wifi 完備。	乾燥機は有料でした。	NE	NE	NE	NE
2	お刺身最高でした！	カニ食べまくりました。	NE	NE	NE	NE
3	立地バツグンです。	駅からのアクセスが抜群。	E	E	E	E
4	部屋から海が見える。	部屋はオーシャンビューで景色がよかったです。	E	E	E	E
5	駐車場も敷地内で便利。	駐車場平地で停めやすい。	NE	E	NE	NE
6	目の前が中華街です。	中華街も徒歩圏。	NE	NE	E	NE

評価対象・評価値が曖昧・不明な事例

前提文「全てにおいて文句なしでした。」と仮説文「接客サービスも素晴らしかったです。」の事例には，2 人が「正しい」，1 人が非含意とアノテーションした．「全て」が指す範囲が曖昧なため，判断が難しいと考えられる．また，前提文「スタッフの方の対応。」と仮説文「お手頃価格。」の事例には，1 人がすべてのトークンを非含意根拠とし，2 人は判定困難を選択した．この前提文の評価値が曖昧[17]なため，判定困難が選択されたと思われる[18]．含意関係アノテーションの前には，このような事例は除外すべきと考える．

評価を直接的に記述していない事例

前提文「金目の煮付けは、最高でした」と仮説文「あの味が忘れられません。」の事例には，2 人が「正しい」，1 人が「意味不明，回答困難」とアノテーションした．「〜が忘れられない」,「〜がおすすめ」,「〜に感謝」といったような評価を直接的に記述していない仮説文をもとに文を集約する必要はないので，アノテーション対象からそのような事例は除外すべきと考える．

強調表現が仮説文にある事例

前提文「朝食は軽食でした。」と仮説文「しかも軽食付き。」の事例には，1 人が「しかも」を非含意根拠とし，1 人が「意味不明，回答困難」とし，1 人が「正しい」とし，アノテーションが分かれた.「しかも」「特に」といったような強調表現を含む仮説文でも文を集約する必要はないのでこのような事例もアノテーション対象から除外すべきと考える．

4.2 含意関係認識の性能評価実験

表 2 に示すように学習と評価に分割し，学習用データを用いて BERT モデルを fine-tuning[19]して含意・非含意の 2 値分類器を M_{BASE}, M_{APPEND}, M_{ALL} の 3 種類作成した．M_{BASE} は BASE の 7,258 事例，M_{APPEND} は APPEND の 8,100 事例，M_{ALL} は BASE と APPEND の 15,358 事例を用いて学習した．

表 3 に評価結果，表 4 に認識結果の例を示す．M_{BASE} と M_{APPEND} を比較すると，BASE の評価セットには M_{BASE} の方が，APPEND の評価セットには M_{APPEND} の方が良い性能を出している．M_{ALL} は双方とほとんど同じか，それを上回る性能を得られている．また，APPEND に対する性能は，BASE に対する性能を下回る傾向にあり，APPEND は BASE より難しい問題が多く含まれていることが分かる．

5 おわりに

本研究ではアノテータによる作文は行わず，自然文からのみなる含意関係コーパスを構築した．また非含意の場合はその根拠となる箇所を選択してもらい，判断根拠もアノテーションした．今後は，アノテーションの対象をより限定してアノテーションの質の向上を図ることと，作成した含意関係認識器を用いてコーパス全体で知識を整理することを考えている．

謝辞：荒瀬由紀准教授からの有益な助言に感謝します．

参考文献

[1] Sanjeev Arora, Yingyu Liang, et al. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR*, 2017.

[2] Samuel R Bowman, Gabor Angeli, et al. A Large Annotated Corpus for Learning Natural Language Inference. In *EMNLP*, pp. 632–642, 2015.

[3] Daisuke Kawahara, Yuta Hayashibe, et al. Automatically Acquired Lexical Knowledge Improves Japanese Joint Morphological and Dependency Analysis. In *IWPT*, pp. 1–10, 2017.

[4] Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *ACL*, pp. 66–75, 2018.

[5] Arseny Tolmachev, Daisuke Kawahara, et al. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *EMNLP*, pp. 54–59, 2018.

[6] Ashish Vaswani, Noam Shazeer, et al. Attention Is All You Need. In *NeurIPS*, pp. 5998–6008, 2017.

[7] 林部祐太. 宿レビューからの肯定的事実と推薦対象の抽出. 言語処理学会年次大会, pp. 554–557, 2019.

[17]感情極性は中立的だが，誤って肯定的と判断されていた．

[18]評価値が曖昧ではなくても，すべてのトークンを選択する代わりに判定困難が選択された事例もあった．

[19]バッチサイズは 32，最大入力トークン数は 25，エポック数は 3 とした．