含意認識タスクに関するかき混ぜ文対データの構築

名取 芙美香 † 松吉 俊 ‡ 福本 文代 ‡
† 山梨大学工学部 ‡ 山梨大学大学院医学工学総合研究部
{t10kg026, sugurum, fukumoto}@yamanashi.ac.jp

1 はじめに

本研究では、含意認識タスクにおいて、**かき混ぜ**という言語現象に着目する。

含意認識とは、2 つのテキスト T1 と T2 が与えられた時に、T1 が真である場合、T2 は真であると推測できるかどうか判断する技術である。これが推測できる場合、T1 は T2 を含意すると言う。以下に、文対の例を示す。事例 (1) は、T1 が T2 を含意する例であり、事例 (2) はそうでない例である。

(1) T1: 太郎と次郎が花子に本を渡す。
T2: 次郎と太郎が花子に本を渡す。
(2) T1: 太郎と次郎が花子に本を渡す。
T2: 太郎が次郎および花子に本を渡す。

日本語も対象とする含意認識に関する評価型ワークショップ NTCIR-10 RITE-2[1]では、文対データに現れる言語現象を分析し、T1 と T2 の関係が単一の言語現象のみにより定まるような文対集合を作成し、ユニットテストデータとして配布している。RITE-2 が整理した、含意関係が成り立つ場合の言語現象カテゴリの 1 つに、**scrambling** (かき混ぜ) がある。この scrambling は、「同じ文節に係る複数の文節の順番を入れ替える」操作と定義される。容易に対応できるように思えるが、scrambling に関するテストデータに、安定して解答できたシステムは存在しなかった [6]。

事例 (1) や (2) のような、構成要素をかき混ぜたような関係にある文対は、含意するかどうかにかかわらず、2 文間の語集合の類似度や文節集合の類似度が非常に高くなる。それゆえ、このような文対の含意認識を行うにあたっては、各文の構造、および、構造における語の意味・特徴を考慮する必要がある。

本論文では、かき混ぜの関係にある文対に関する含意認識システムを作成するための基盤として、現在我々が構築を進めているかき混ぜ文対データについて報告する。本論文は、以下のように構成される。まず、2 章で、本研究におけるかき混ぜを定義する。次に、3 章において、かき混ぜ文対を作成する方法を説明する。4 章では、現状の文対データについて報告する。5 章はまとめである。

2 かき混ぜの定義

本研究では、含意認識タスクにおいて有用であるよう、独自にかき混ぜを定義する。

2.1 関連研究におけるかき混ぜ

文献 [3] によると、scrambling は、Ross [5] が、意味解釈に影響を与えない自由な語順変化を説明するための移動規則として提案したようである。これ以降、言語学の分野では、「語順」として、主題や格成分の長さなどとともに議論されている [4]。

自然言語処理の分野において、藤田は、言い換え事例を分類し、双方向の含意が成立するかき混ぜの事例を、「分裂文の言い換え」や「数量詞の遊離」などに整理している [1]。Hoshino らは、日本語としては非文となるが、入力となる日本語文を英語に直接対応する語順にあらかじめ並べ替えることにより、日英機械翻訳の性能が向上することを報告している [2]。

2.2 本研究におけるかき混ぜ

本研究では、2 つの文 (T1 と T2)[2]の間に次のような関係がある場合、これらはかき混ぜの関係にあると呼ぶ。

T1 に含まれる内容語の bag of words と、T2 に含まれる内容語の bag of words が等しい

本研究では、日本語の品詞体系として、日本語形態素解析システム JUMAN version 7.0[3]におけるものを使用する。主に、品詞に基づき、内容語の集合を以下のように定めた。

動詞、形容詞、副詞、連体詞 内容語とする。形容詞には、イ形容詞とナ形容詞を含む
名詞 形式名詞でない名詞を内容語とする

[1] http://www.cl.ecei.tohoku.ac.jp/rite2/doku.php?id=wiki:メインページ

[2] 前提として、いずれの文も日本語として非文でないとする。

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

接頭辞 名詞接頭辞とナ形容詞接頭辞のみを内容語とする。前者は「初」や「副」などを、後者は「無」や「最」などを含み、これらは、機能語というよりもむしろ、意味を表す重要な要素であると考えるからである

接尾辞 複合名詞の末尾の構成要素になりうるので、名詞性接尾辞は、内容語とする。「的だ」や「っぽい」などを含む形容詞性名詞接尾辞は、内容語とする

「ない」や「ぬ」 形容詞性述語接尾辞「ない」と助動詞「ぬ」は、特別に内容語とする。これらは、内容語であるイ形容詞「無い」と同様の機能を持つと考えるからである。上のかき混ぜの定義においても、特例としてこの 3 語を適宜同一のものとみなす

特殊 (記号類) アルファベットや「+」などを含む「特殊-記号」を内容語とする。ただし、句読点相当である「！」、「？」、「・」は除く

未定義語 実テキストの解析において、品詞を判断することができなかった形態素に「未定義語」というラベルが付く。本研究では、これを名詞相当と考え、内容語とする

1 章の事例 (1) と (2) の文対は、上のかき混ぜと内容語の定義により、いずれもかき混ぜの関係にあると判断される。我々が定義するかき混ぜは、言語学におけるかき混ぜの概念より広い。言語学におけるかき混ぜは、大まかな意味解釈が変わらないことを前提としている。それゆえ、事例 (1) は言語学の定義においてもかき混ぜであるが、T1 の意味解釈と T2 の意味解釈が等しくない事例 (2) は、言語学においては、通常、かき混ぜの関係にあると呼ばない。我々は、内容語の bag of words が等しい 2 つの文において、どのような場合に含意が成立し、どのような場合に含意が成立しないかを知り、そのような文対に対する含意認識システムを作成することに関心がある。このような理由により、言語学が対象とするよりも広いかき混ぜを定義した。

3 かき混ぜ文対の作成法

RITE-2 におけるユニットテストの考え方を取り入れ、かき混ぜの関係にある 2 つの文を集め、各文対に対して、T1 が T2 を含意するか否かのラベルを付けることを目指す。この章では、実際に我々が行っているデータ作成法について説明する。

3.1 予備調査

かき混ぜ文対データを作成するにあたり、どのような種類のかき混ぜ文対が存在するかを調査するため、ブレインストーミングにより、T1 と T2 を自由に作成し、これらの文対を、含意認識課題において必要となる言語知識に基づいて分類した。分類結果と文対の例を図 1 に示す。汎用的な規則のみが必要になるクラスに対して小さい数字を振り、それらの規則に加え、語彙的な知識も必要とするクラスに対して、順に大きい数字を振った。

クラス 1 文節の並列に関する一般的な規則や、名詞句による修飾に関する一般的な規則を必要とする。

クラス 2 述語とその項に関する一般的な規則も必要とする。特に、受身構文や強調構文に関するかき混ぜ文対はこのクラスに属する。

クラス 3 ある構文が別の構文に置き換え可能かに関する一般的な知識も必要とする。例えば、図 1 の (3-Y) では、「V1 ながら V2」と「V2 ながら V1」が多くの場合に置き換え可能であるという知識を必要とする。(3-N) では、「VP1 なら VP2」から「VP2 なら VP1」は演繹的に導けないという知識を必要とする。

クラス 4 上位下位関係や部分全体関係など、2 項関係知識も必要とする。例えば、図 1 の (4-Y) において、「山梨県」と「富士五湖」が部分全体関係にあるという知識があれば、(1-N) のように “含意:NO” と判定することなく、“含意:YES” と判定できる。

クラス 5 述語の項に対する制限に関する語彙的知識も必要とする。例えば、図 1 の (5-N) の述語項構造「NP1 に NP2 が加わる」において、NP1 と NP2 が等価なものでない場合、交換不可能である。一方、述語が「結合する」や「つながる」の場合、この制限はない。

クラス 6 ある事象と別の事象との間の時間関係を推論するための知識も必要とする。

クラス 7 その他。

3.2 作成手順

与えられた文 T1 に対して文 T2 を作成し、かき混ぜ文対を作成する手順を以下に示す。

1. T1 を JUMAN で解析する
2. T1 からすべての内容語を抽出し、リストを作成する。このとき、3.3 節で述べる結合規則により、条件を満たす内容語列を 1 つの内容語とみなす
3. このリストから、ランダムに内容語を 1 つずつ取り出し、並べる。このとき、3.4 節で述べる制約をかける
4. 2. で結合させた内容語があれば、すべて分解する
5. 非文とならないように注意しながら、並べ替えられた後の内容語リストに人手で機能語を補完し、T2 を作成する

クラス 1 係り受け情報が必要

(1-Y) 含意:YES
T1: 彼は、赤色 や 青色 のペンを買った。
T2: 彼は、青色 や 赤色 のペンを買った。

(1-N) 含意:NO
T1: 赤色の 鉛筆はペンの右にある。
T2: 鉛筆は 赤色の ペンの右にある。

クラス 2 述語項構造情報が必要

(2-Y) 含意:YES
T1: 太郎は 昨日 花子に会った。
T2: 昨日 太郎は花子に会った。

(2-N) 含意:NO
T1: 山田 が 鈴木 に 本を渡した。
T2: 山田 に 鈴木 が 本を渡した。

クラス 3 構文の対応に関する知識が必要

(3-Y) 含意:YES
T1: 太郎は踊りながら歌った。
T2: 太郎は歌いながら踊った。

(3-N) 含意:NO
T1: 桜が咲いているなら、太郎は外で酒を飲んでいるよ。
T2: 太郎が外で酒を飲んでいるなら、桜は咲いているよ。

クラス 4 2 項関係知識が必要

(4-Y) 含意:YES
T1: 山梨県の富士五湖 と甲斐市に訪れたことがある。
T2: 富士五湖 と山梨県の甲斐市に訪れたことがある。

クラス 5 項に対する制限に関する知識が必要

(5-N) 含意:NO
T1: EXILE に二代目 J Soul Brothers が加わって EXILE 第 3 章が始まった。
T2: 二代目 J Soul Brothers に EXILE が加わって EXILE 第 3 章が始まった。

クラス 6 時間関係に関する推論知識が必要

(6-N) 含意:NO
T1: 電車に乗り込みながら、手をつないだ。
T2: 手をつなぎながら、電車に乗り込んだ。

クラス 7 上記より複雑な知識が必要

図 1: 必要な言語知識に基づく、かき混ぜの関係にある文対の分類

6. T1 を確認し、T2 のテンスやモダリティ、副助詞を調整する
7. T1 から T2 が含意するかどうか判断する
8. どのクラスのかき混ぜであるか判断する

できるだけ多様なかき混ぜ文対を作成するために、上の 5. は、まずは T1 を見ずに実施する。内容語のリストから、全体の解釈がつかめない場合、T1 を確認する。与えられた順番では、どうしても T2 の作成が困難である場合、人手で内容語の順番を変更する。この場合、文対に特別なマークを付ける。

3. において、完全にランダムに並べる場合、T2 の作成が困難である状況に陥ることが多くなることが予想される。それゆえに、2. において、特定の内容語列が分解しないように、いくつかの結合規則を適用する。加えて、3. において、完全なランダムとならないように、いくつかの制約規則を適用する。利用した規則集合は、文対とともに記録する。

3.3 内容語結合規則

以下の 13 種類ある。

- 連続する名詞をすべて結合させる
- 同じ文節内にあるか、係り受け関係にある場合のみ、連続する名詞をすべて結合させる
- 接尾辞を前の語と結合させる
- 接頭辞を後ろの語と結合させる
- 連体修飾の形容詞と直後の名詞を結合させる
- 連体詞と直後の名詞を結合させる
- サ変名詞と直後の「する」を結合させる
- 「お待ちする」のような場合に、動詞と直後の「する」を結合させる
- 複合動詞を結合させる
- 形容詞と直後の動詞を結合させる
- 並列関係にあり、かつ、連続する 2 つの文節を結合させる
- 開き括弧から閉じ括弧までに含まれる内容語をすべて結合させる
- 「によって」や「を通じて」などの複合辞の核となる動詞を、直前の名詞と結合させる

3.4 かき混ぜ制約規則

以下の 4 種類ある。

- 文頭に動詞が 2 つ並ぶ場合、2 つとも並べ直す
- 文頭に「する」は NG
- 文末に連体詞も副詞も NG
- 「ぬ」が 2 つ並ぶのは NG

表 1: かき混ぜ文対データ内のクラス分布

クラス	事例数	含意の事例数
1	0	0
2	77	12
3	7	5
4	0	0
5	0	0
6	0	0
7	0	0
合計	84	17

4 かき混ぜ文対データ

この章では、我々が構築を進めているかき混ぜ文対データとその現状を報告する。

本研究では、T1 の集合として、『現代日本語書き言葉均衡コーパス』(BCCWJ)[4]のコアデータ内の文を利用した。BCCWJ のコアデータには、次の 6 種類のレジスターのテキストデータが存在する。

Yahoo!知恵袋 (OC)、Yahoo!ブログ (OY)、書籍 (PB)、雑誌 (PM)、新聞 (PN)、白書 (OW)

まず、各レジスターのテキストデータを半分に分割し、その半分[5]から、少なくとも 10 文字以上からなる文をランダムに 40 文抽出した。すべてのレジスターのデータを 1 つにまとめ、240 文のリストとした。含意認識タスクのデータとして相応しくないと考えられるので、このリストから、人手で次のような文を除去した。

- 挨拶
- 見出し文
- 格助詞「と」から始まる文
- 疑問文
- 伏せ字を含む文
- 固有名詞でない、一般語の英単語を含む文
- 表や図を参照している文
- 箇条書きの複数の項目が、まとめて 1 文であると、誤って抽出されてしまったもの

これらを除去した後の 82 文を T1 の集合とした。T1 の文には、どのレジスターのファイルから抽出したかの情報が付与されている。

各 T1 に対して、3.2 節の手順を実施した。3.3 節の結合規則は、少なくとも 9 個以上を同時に適用し、適用する数および種類はランダムに決定した。3.4 節の制約規則も、適用する数および種類をランダムに決定した。

[4] http://www.ninjal.ac.jp/corpus_center/bccwj/

[5] この半分は、開発データ作成用とした。残りの半分は、評価データ作成用とした。

上記のようにして構築したかき混ぜ文対データの現状を図 1 に示す。1 つの T1 から複数の T2 を作成することを許しているので、合計数は、T1 と T2 の対の異なり数で 84 となった。まだ作業した数が少ないので、クラス 2 と 3 以外のかき混ぜ文対は得られていない。約 20%(17/84) の文対が、T1 が T2 を含意するものであった。BCCWJ の PN3d_00002.xml から作成した、含意関係にあるクラス 3 の事例を以下に示す。

(3) T1: ライス大統領補佐官（国家安全保障担当）は同日、ＮＢＣテレビで、隠匿の疑いが濃厚なため、捜索活動が必要だと主張した。
T2: ＮＢＣテレビで、必要なこととしては、隠匿の疑いが濃厚なため、ライス大統領補佐官（国家安全保障担当）は、捜索活動だと、同日に主張した。

クラス 1 や 4 以上のかき混ぜ文対を作成するために、品詞レベルではなく、構文レベルのかき混ぜ制約規則を利用する等、工夫が必要である。現在は、機能語を補完する時に特別な規則は設けていないが、今後はこれも検討する必要がある。

5 まとめ

本論文では、かき混ぜ文対に関する含意認識システムを作成するための基盤として、かき混ぜ文対データを作成する方法を提案し、現在我々が構築を進めているデータについて報告した。

予備調査において作成したかき混ぜ文対を参照しながら、クラス 1 と 2 の文対に対する含意認識システムを実装した。今後は、クラス 3 以上のデータを収集するとともに、これに対する含意認識システムの構築にも取り組みたい。

参考文献

[1] 藤田篤. 言い換えのあれこれ. http://paraphrasing.org/paraphrase.html. (2013/12/23 にアクセス).

[2] Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. Two-Stage Pre-ordering for Japanese-to-English Statistical Machine Translation. In *Proc. of the 6th IJCNLP*, pp. 1062–1066, 2013.

[3] 片岡喜代子. 日本語否定文の構造 かき混ぜ文と否定呼応表現. くろしお出版, 2006.

[4] 日本語記述文法研究会（編）. 現代日本語文法 7. くろしお出版, 2009.

[5] John Robert Ross. *Constraints on Variables in Syntax*. Doctoral dissertation, Massachusetts Institute of Technology, 1967.

[6] Yotaro Watanabe, et al. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proc. of the 10th NTCIR Conference*, pp. 385–404, 2013.