議論文生成における文抽象化のための固有表現抽象化

佐藤 美沙　柳井 孝介　三好 利昇　柳瀬 利彦　丹羽 芳樹
日立製作所 中央研究所
{misa.sato.mw, kohsuke.yanai.cs, toshinori.miyoshi.pd,
toshihiko.yanase.gm, yoshiki.niwa.tx}@hitachi.com

1 はじめに

筆者らの研究グループでは、テキスト内の知識を活用する手段として、人と議論をする人工知能を開発している [6–8]。新聞記事などの既存のテキストから議論に必要な知識を含む文を抽出し、その文を組み上げることによって議論文を生成する。このとき、 既存のテキストの文は、議論には関わりの薄い、具体性の高い情報を含んでいる場合がある。そこで、文を抽象化して、議論に使えるような一般的な主張を表す文を生成する必要がある。

わかりやすい発言のためには、まず冒頭で主張したい内容を簡潔に述べ、次に具体的な事例を用いたサポートを述べるやり方が一般的である。そのため、主張内容と具体的な事例の関係は、前者は後者を一般化したものになっているべきであり、こうした生身の人間のサポートツールとしても具体的な事例からの抽象化は必要とされる問題である。

本稿では、具体的な事例について書かれた文を抽象化する試みを報告する。

本研究では、 文中の固有表現を抽象化することによって、文の抽象化を行う。たとえば、「経済問題」を議題として議論するとき、

(1) Malaria epidemic situation remains normal in **Myanmar**. [1]

という文を材料として、文中の “Myanmar” を “developing countries” に置換することで、

(2) Malaria epidemic situation remains normal in **developing countries**.

という、より抽象的な内容の文を生成することができる。このようにして抽象化した文と、元の具体的な文を複数組み合わせることによって、筆者らの目的とする人との議論に使える文章を作成することが可能になると考えられる。

[1] ©1995–2010 Xinhua News Agency

これまで文の抽象化の研究では、WordNet [1] などの概念辞書から抽象化先を獲得する手法 [3] が提案されている。しかし、こうして獲得された抽象化先には、複数の候補が存在するという問題がある。これらの候補には、その表現そのものの抽象化先としては正しくとも、文の抽象化のためには不適切であるような候補が含まれる。そのため、数ある抽象化先候補の中から、なんらかの評価基準を用いて適切な候補を選択する必要がある。

さらに、適切な抽象化は、文の内容および議論の議題によって異なるという問題点も存在している。つまり、抽象化した後の文の内容が正確な文になるかどうか、また、抽象化の方向が議題と一貫しているかどうか、によって、候補を評価する必要がある。

本研究では、上記の二点の問題を解決し、文の抽象化を行うための手法を提案する。具体的には、材料文中の固有表現に対して、(1) 生成文の内容が事実に照らして正確であること、 (2) 抽象化の方向が議題と関連が強いこと、 という二点を満たす抽象化先を選択する。これらの手法により選択した表現と置き換えることで、文の抽象化を実現する。

2 提案手法

本研究では、議論のテーマを表す議題文と、主張の元となる材料文と、各固有表現に対する抽象化先候補を表す辞書が与えられたときに、材料文中の固有表現と、それを置換する最も適切な抽象化先候補を選択し出力することを目的とする。

具体的には以下の三操作によって文を抽象化する枠組みを提案する。まず、1. 材料文中の抽象化対象の表現を抽出する。本研究では、エンティティを表す固有表現を抽象化の対象とする。抽出した固有表現に対して、抽象化先候補の表現は複数存在する。そこで、次節以降で述べる 2. 正確性の評価と 3. 議題との関連性

の評価を行う。最も評価値の高い候補一つを置換先として選択し、出力する。

2.1 抽象化対象の固有表現の抽出

与えられた材料文に対して固有表現抽出を行う。抽出された固有表現から、議題文中の重要語と関連の深いものを除き、 残ったものを抽象化の対象となる固有表現とする。

2.2 正確性の評価

事象によって抽象化先としての適切さが異なる。

たとえば、“Myanmar” は、(A)“developing countries” の一つであり、(B)“Southeast Asian countries” の一つである [2] ことから、これら 2 つは “Myanmar” の抽象化先候補である。しかし、

(3) Malaria is a major health problem in **Myanmar**.

という文を抽象化するときには、生成される文の内容から、候補 (A) は抽象化先として適切だが、候補 (B) は不適切であると考えられる。

この違いは 、抽象化された文の内容が事実に照らして正確であるかどうかに起因する。文の内容が正確であることを、事象の正確性と呼ぶ。

本研究では、抽象化された文の事象が正確であるときには、その抽象化先候補が、類似の事象が生じているエンティティ同士で共通して候補になっていると考える。上の例では、

(4) Malaria is a major health problem in **Tanzania**.

という、材料文と類似の事象について表す文が存在することが、候補 (A) の適切さを担保している。“Tanzania” は、(A)“developing countries” だが、(B)“Southeast Asian countries” ではない [2]。

本手法では、材料文で述べられている事象と類似の事象が生じている複数のエンティティを比較し、各エンティティに対して複数存在する抽象化先候補の共通度から正確性を計算し、対象の事象が生じる特徴の強い候補を選択する。 そのために、候補である抽象概念に属するエンティティの集合が、材料文で述べられている事象と類似の事象の生じているエンティティをどれだけ含むかによって正確性を評価する。多く含むほどその抽象化先は評価が高く、また、その割合が大きいほど評価が高い。

[2]Wikipedia より

2.2.1 類似事象が生じているエンティティの抽出

類似事象が生じている他のエンティティを獲得するために、類似事象を表す文をコーパスから獲得し、そこから対応するエンティティを抽出する。まず、材料文から置換対象の固有表現を除いた文字列を作成する。そして、作成した文字列をクエリとした大規模コーパスからの連想検索によって、エンティティ以外の情報が類似している文を獲得する。獲得された文に対して固有表現抽出を行い、置換対象の固有表現と異なる文字列かつ固有表現分類が同じであるものを類似事象エンティティとして抽出する。

2.2.2 正確性の計算

より多くの類似事象エンティティにとって抽象表現であるものに高い評価値を与える、かつ類似事象エンティティではないものにとっては抽象表現ではないものに高い評価値を与える、の二観点を反映するように評価値を付与する。

具体的には以下の式 (1) で、ある抽象表現 a の正確性評価値が計算できる。

$$f_{\mathrm{event}}(a) = h(\alpha \times P_{\mathrm{event}}(a), R_{\mathrm{event}}(a)), \qquad (1)$$
$$h(x,y) = \frac{1}{\frac{1}{2}(\frac{1}{x} + \frac{1}{y})}$$

ただし、

$$P_{\mathrm{event}}(a) = \frac{(a \text{ を抽象表現に持つ類似事象エンティティの数})}{(a \text{ を抽象表現に持つ全エンティティの数})},$$
$$R_{\mathrm{event}}(a) = \frac{(a \text{ を抽象表現に持つ類似事象エンティティの数})}{(\text{全類似事象エンティティの数})}$$

を表す。α は重み付け定数である。α を 1 よりも大きくするほど、$R_{\mathrm{event}}(a)$ を重視するようになる。すなわち、より多くの類似事象エンティティが含まれる抽象概念に対してより高い評価が与えられるようになる。

表 1 に、文 (3) における “Myanmar” の抽象化に対する、各抽象化先候補の正確性評価計算結果を示す。セル内の○印は、該当列のエンティティが該当行の表現を抽象化候補として取りうることを示している。表の最下行は、該当列のエンティティが類似事象エンティティであるかどうかを表している。表の最右列は評価計算結果を表す。ここでは、$\alpha = 1$ とし、“developing countries” に対して最高スコアの 1.0 が、“humid countries” に対して次点の 0.8 がそれぞれ与えられている。

表 1: 正確性評価の概要。$h(a,b)$ は、a と b の調和平均を表す。

抽象化先候補 / エンティティ	Myanmar	Tanzania	Japan	Italy	Finland	スコア
developed countries			○	○	○	$h(1*0/3, 0/2) = 0.0$
developing countries	○	○				$h(1*2/2, 2/2) = 1.0$
humid countries	○	○	○			$h(1*2/3, 2/2) = 0.8$
Southeast Asian countries	○					$h(1*1/1, 1/2) = 0.67$
several countries	○	○	○	○	○	$h(1*2/5, 2/2) = 0.57$
類似事象が抽出されたかどうか	●	●				

表 2: 実験に用いた各エンティティの抽象化候補。各列にエンティティとその抽象化候補を示す。

Myanmar	Singapore	penguins
Myanmar	Singapore	penguins
developing countries	developed countries	south polar animals
humid countries	island countries	seabirds
Asian countries	Asian countries	polar animals
several countries	several countries	several animals

2.3 論題との関係性の評価

論題の内容によっても、適切な抽象化先が異なる。節 2.2 では、事象の正確性から、文

(3) Malaria is a major health problem in **Myanmar**.

の “Myanmar” に対して “developing countries” と “humid countries” の両方の候補に高い評価値が与えられている。

ここで、論題が「経済援助を進めるべきである」という内容であったとき、論題に関わりの深い方向へ抽象化するべきと考えられる。この例では、キーワード「economy」と関連度の高い “developing countries” の方が “humid countries” よりも適切である。

本手法では、論題中のキーワードと抽象化先候補間の WordNet [1] のグラフ上の距離を関連度とする。この関連度を論題との関係性の評価値として用いる。

3 実験

3.1 実験設定

英文新聞記事のコーパス Gigaword [5] から選択した数文に対して、提案手法による抽象化を行った。なお、固有表現抽出には Stanford Core NLP [4] を用いたが、一部の固有表現に対しては手動で指定した。

エンティティとそれらの抽象化先候補は、本稿では人手で作成した。実験で用いる材料文のうち、置換対象の固有表現に関する抽象化候補を表 2 に示す。表のとおり、各エンティティに対して、各 5 通りの抽象化候補を準備した。その候補を抽象化先として持つ他のエンティティの情報は、Wikipedia を参照して整備した。たとえば、“developing countries” には、“Myanmar”, “India” 等が含まれる。

また、具体的な事象に付随する情報であるものとして、日付・時間表現を表す句や、数量を表す句、発話文における発話者と発言動詞、等は削除した。なお、類似検索は Gigaword コーパスを対象に上位 500 件とし、$\alpha = 20.0$ とした。

本実験では、正確性評価の上位 2 候補から論題関連性評価の高い一方を最終的な抽象化先として選択した。

3.2 結果

表 3 に提案手法による文の抽象化の結果を示す。各文において、“Myanmar” には “developing countries” が、“Singapore” には “several countries” が、“penguin” には “polar animals” が、置換先に選択された。

4 考察

文の抽象化という目的のためには、本研究で対象としたエンティティの抽象化先候補選択の問題の外にも、いくつか越えなければいけないハードルがある。

まず、抽象化の対象を選択するところに問題がある。本研究では議題のキーワードと同義の語は抽象化の対象から外しているが、不十分である。

(5) ... the fight against smoking “is extremely important because diseases and deaths caused by smoking by far outnumber those induced by AIDS, tuberculosis and **malaria** combined.”. [6]

議題が「smoking」の場合でも、固有表現 “malaria” は抽象化の対象とするべきではない。

[3] ©1995–2010 Xinhua News Agency
[4] ©1994–2010 The Associated Press
[5] ©1994–2010 Agence France Presse
[6] ©1995–2010 Xinhua News Agency

表 3: 実験結果例。材料文は [5] より引用。

論題	材料文	生成文
economy	Malaria epidemic situation remains normal in **Myanmar**. [3]	Malaria epidemic situation remains normal in **developing countries**.
smoking	By Oct. 1 **Singapore** will ban smoking in bus shelters and depots public toilets swimming complexes stadiums and community clubs. [4]	**Several countries** ban smoking in bus shelters and depots public toilets swimming complexes stadiums and community clubs.
environmental problems	Global warming threatens **penguins**. [5]	Global warming threatens **polar animals**.

また、数値表現の抽象化も必要である。

(6) Malaria Kill **90 people** in Central Kenya. [6]

ここでは、文脈において 90 people が多いのか少ないのかを判断し、適切な形容詞に置換する必要がある。

複数のエンティティの対比関係を捉えた抽象化が必要なケースもある。

(7) **Thailand** has handed drugs and pesticides to **Myanmar** in an attempt to eradicate malaria from border areas. [6]

は、以下のように抽象化したい。

(8) **Countries** has handed drugs and pesticides to **the neiboring countries** in an attempt to eradicate malaria from border areas .

議題に特に関わりの深い特徴を残して抽象化することが議論文生成のために有用であるような例もある。

(9) **Brown & Williamson** executives were “most certainly aware” that smoking caused disease. [7]

は、次のように元のエンティティの特徴を明示したい。

(10) Even **tobacco companies'** executives were “most certainly aware” that smoking caused disease.

5 関連研究

自動要約の分野において抽象化の取り組みがある。[3] では、WordNet を用いて複数の概念からの一つの概念への統合を行っている。同一文中における具体的な概念の組合せに応じて抽象化先を決めている点で本手法とは異なる。また、プライバシー保護の分野においても、具体的な情報を伏せるために名詞を抽象化する研究がされている [2]。

WordNet を用いる手法では、概念階層を上に登り、共通する概念を置換先として抽出している。つまり、概念辞書の階層の情報を用いている。本研究で用いる辞書は、各抽象化先候補間の階層関係の情報を持たない。代わりに正確性評価式において各概念を抽象化先候補にもつエンティティの総数を用いることで、エンティティを過剰に含む抽象化先が選ばれることを防ぐ、すなわち、概念階層を上に上がりすぎることを防いでいる。

[7] ©1994–2010 New York Times

謝辞

本研究を進めるにあたり、共同研究先である東北大学の乾健太郎教授には技術的な議論を通して、貴重なご意見を頂きました。感謝いたします。

参考文献

[1] George A. Miller. WordNet: A Lexical Database for English. In *Communications of the ACM*, Vol. 38, pp. 39–41, 195.

[2] Yeye He and Jeffrey F. Naughton. Anonymization of Set-valued Data via Top-down, Local Generalization. *Proceedings of the VLDB Endowment*, Vol. 2, No. 1, pp. 934–945, 2009.

[3] Eduard Hovy and Chin-Yew Lin. Automated Text Summarization in SUMMARIST. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 18–24, 1997.

[4] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the ACL 2014: System Demonstrations*, pp. 55–60, 2014.

[5] Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme. *Annotated English Gigaword LDC2012T21*. Philadelphia: Linguistic Data Consortium, 2012.

[6] 柳井孝介, 柳瀬利彦, 三好利昇, 丹羽芳樹, 佐藤美沙. ディベート人工知能のためのアーキテクチャ. 第 7 回 データ指向構成マイニングとシミュレーション研究会, 2014.

[7] 三好利昇, 佐藤美沙, 柳井孝介, 柳瀬利彦, 丹羽芳樹. ディベートでの立論材料文検索における論点選択方法. 第 7 回 データ指向構成マイニングとシミュレーション研究会, 2014.

[8] 柳瀬利彦, 三好利昇, 柳井孝介, 岩山真, 丹羽芳樹. ディベートの意見文章生成のための分散表現を用いた文の並び替え. 第 7 回 データ指向構成マイニングとシミュレーション研究会, 2014.