発話に対する引用表現の自動生成

上垣外 英剛 †　笹野 遼平 ‡　高村 大也 ‡

† 東京工業大学 総合理工学研究科　‡ 東京工業大学 精密工学研究科

† kamigaito@lr.pi.titech.ac.jp ‡ {sasano,takamura}@pi.titech.ac.jp

1　はじめに

新聞記事中には例 1 のような表現が存在する.

例 1　「公約を一つ一つ着実に実現したい」と抱負を語る.

このような表現は「」内の発話を文末の表現で簡潔に説明している. 以降このような文末の表現を引用表現と呼ぶ. 引用表現に対し “抱負” というクエリで検索を行うことで抱負を表す発話を効率的に収集することができる. しかし, 新聞記事中には例 2 のように引用表現に含まれる情報が少ないものもある.

例 2　「公約を一つ一つ着実に実現したい」と語る.

また対話形式の記事などはそもそも引用表現が存在しないことが多い. そのような発言は引用表現を手がかりにして検索をすることはできない. もし発話に対する引用表現を自動推定して付与することができれば, より包括的な検索が可能になる. このような動機の下, 本稿では入力された発話から引用表現を自動生成し出力する手法を提案する.

新聞記事には例 3 のように限定的な対象を伴う表現も多い.

例 3　「元請けも下請け業者も, 労災発生を隠すケースが増えている」と関係者はみている.

この例 3 で『関係者は』は対象を表しており, 我々が扱う引用表現は『みている』であり, 本稿ではこの部分を自動生成することを目的とする.

2　前処理

2.1　テキスト断片の抽出

訓練データを生成するために, まず新聞記事中の発話とテキスト断片を抽出する必要がある. ここでは述部が句点で終了する単文のみを対象とする. 単文の例を例 4 に, 複文の例を例 5 に示す.

例 4　「大変申し訳ありません. 心からおわびします」と謝罪した.

例 5　「北部同盟はカブールを占拠すべきではなかった」と批判し, 国連主導の和平実現を求めた.

単文を抽出するための条件は以下の通りである.

- 『「」』がひとつ存在し, 『「」』+格助詞『と』の形を取る.
- 話者と発話の係り先が同じである. 話者については『名詞+は』または『名詞+が』の形で述部に係るものとする.
- 格助詞『と』の直後に読点を含んでいない.
- 受動態ではない. 具体的には述部に接尾動詞『れる・られる』を含んでいない.

2.2　引用表現の判別

「」で囲まれた全てのものが発話とは限らないため, 引用表現が発話に関連するものかを判別する必要がある. 引用の種類は引用表現との関係が深い事が知られている [2]. 本節では次節で抽出する引用表現が発話と関係しているかを判断する. まず発話と関係が深いと考えられる次のようなルールを定める.

- 話者が人名か固有名である
- 引用が句点か読点を含む

引用表現が対象データ中で上記のルールを満たした回数を C_y, 満たさなかった回数を C_n とする. 閾値 T_2 に対し式 (1)

$$T_2 \leq \frac{C_y}{C_y + C_n} \tag{1}$$

を満たす引用表現を発話に関する表現として扱う.

3 提案手法

本研究では以下の手順で引用表現を自動生成する.

1. 獲得されたテキスト断片から引用表現を分離する
2. 選択された引用表現をクラスとして扱い, テキスト分類問題として解く.

この手順を図 1 に示す.

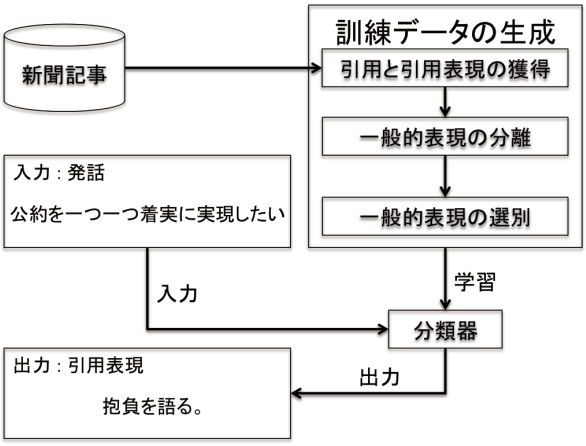

図 1: 提案するシステム

3.1 引用表現の抽出

獲得されたテキスト断片は例 6 のように対象を含んでいる場合がある.

例 6　　憲法改正に慎重な姿勢を見せる

例 6 では『憲法改正に』が対象である. 本稿では引用表現のみの生成を目的とするので, 対象部分を削除し, 引用表現を抽出する. 対象はテキスト断片ごとに異なることが多く, 文節が接続される箇所のエントロピーに着目することで引用表現から切り離すことができると考えられる. 具体的には次のような処理を行う. テキスト断片を文末から数えて i 番目の文節境界から文末までの文字列を G_i と表す. $F(G_i)$ を, テキスト断片集合中での G_i の頻度とする (G_i を真の部分文字列として含むようなテキスト断片はここでは無視する). 与えられたテキスト断片の引用表現候補集合 A を,

$$A = \{G_i \mid F(G_i) > 0, F(G_{i+1}) \leq F(G_i)\} \qquad (2)$$

と定義する. $A = \phi$ となるようなテキスト断片は訓練データには含めない.

さて, 引用表現候補集合 A から, このテキスト断片における引用表現を選択する. 各候補に対してスコアを計算し, そのスコアが最大となる候補を引用表現として選択する. スコアとしては, branching entropy[1] に基づいた値 $s(i)$ を用いる：

$$s(i) \quad = \quad e(i) - e(i-1). \qquad (3)$$

ただし,

$$e(i) \quad = \quad \sum_{b \in B} -P(b \mid G_i) \log P(b \mid G_i) \qquad (4)$$

である. ここで, B は文節の集合である. G_i の直前に分節 b が接続する確率 $P(b \mid G_i)$ はテキスト断片集合に対する最尤推定を用いて,

$$P(b \mid G_i) \quad = \quad \frac{f(bG_i)}{f(G_i)} \qquad (5)$$

と推定する. ここで, $f(G_i)$ は G_i を含むテキスト断片の頻度の総和とする.

さらに, 引用表現の抽出精度を向上させるために, 候補集合 A を改良して新たな候補集合 A' を作成する. 引用表現には例 7 と例 8 のような類似した表現が多い.

例 7　　本音を漏らす.

例 8　　本音を語る.

ここで, 『語る』は様々な対象をとるため, 上で説明した手法では後者の『本音を語る』は適切に抽出されず, 『語る』が引用表現として抽出されてしまうことが予想される. しかし, 例 7 が引用表現として抽出されていれば, 『本音を』という文節が引用表現に含まれていることは分かる. この知識を利用し, 『本音を語る』を抽出することができる. 上で抽出された引用表現のうち頻度が閾値 T_2 以上の表現に含まれている文節の集合を B_{T_2} とする. ただし述語は含めない. また G_i の直前に接続する文節を b_{i+1} と表記する. 新たな候補集合 A' を次のように定義する：

$$A' = \{b \mid F(G_i) > 0, F(G_{i+1}) \leq F(G_i), b_{i+1} \notin B_{T_2}\} \qquad (6)$$

この候補集合の下で再度抽出を行う.

3.2 引用表現の自動生成

抽出された引用表現をクラスとして扱い, 引用表現の自動生成を多クラス分類問題として扱う. 具体的には自動獲得された発話と引用表現を訓練データとする.

それらのデータを用いて分類器を学習し, 入力された発話に対する引用表現を出力する. 本研究では分類器として多項モデルに基づくナイーブベイズ分類器 [3] を用いた. 引用表現は引用の文末表現との関係が深いと考えられるので, 文末がわかるよう引用の文末に文末記号を追加した. ナイーブベイズ分類器は単語が独立して生成されることを前提にしているため, そのままでは文末に記号を追加しても意味は無い. そのため本研究では単語だけではなく n-gram も用いている. また訓練データが少ないクラスは学習がうまくいかないため, データ中の出現回数が閾値 T_3 以上の一般表現のみをクラスとして扱っている.

4 評価実験

4.1 実験データ

発話と引用表現の抽出には以下に列挙する新聞記事データを用いた.

- 日経新聞本紙 1990–2003
- 日経新聞三紙 1994–2003
- 読売新聞 1991–2004
- 毎日新聞 1991–2005

データ抽出の際の閾値はそれぞれ $T_1 = 100$, $T_2 = 0.5$, $T_3 = 100$ とした. 抽出したデータを訓練データと開発用データ, テストデータに分けて用いた. 訓練データは全データ中の 90%となる 310,321 件, 開発用データとテストデータはそれぞれ 5%となる 17,241 件である. またクラス数は 504 クラスとなっている. ナイーブベイズ分類器のスムージングパラメータは 1 とした. また獲得されたテキスト断片のうち一度も文節が接続しない文節境界は分離の候補から除いた.

4.2 評価方法

提案手法を自動評価と人手評価を用いて評価した. 評価は 1-gram から 3-gram までの n-gram に対して行った.

4.2.1 自動評価

発話をクラスに分類し, 分類されたクラスが一致した場合を正解として扱う. そしてその正解率によって提案手法を評価する. ベースラインとしては常に最も高い出現頻度のクラスを出力するものを用いる. またクラスのラベルは引用表現として文の要素も持ちあわせているので機械翻訳で評価に用いられている BLEU[4] を使用する. しかし BLEU は n-gram の完全一致を用いて評価しているためそのまま用いるのでは類似表現の多い引用表現の評価には適さない. そこで n-gram の一致の基準を緩和する. 具体的には単語の原型が同じか日本語 WordNet 上で同一 synset にあれば一致とした. 全ての単語を対象に評価を行うと機能語や非自立動詞等によって関係ないクラス同士のスコアが高くなるので, 以下の条件を満たす単語は取り除いた.

- 日本語 WordNet において動詞『言う』と同一の synset 上に存在する単語
- 動詞『する』
- 『ない』以外の助動詞
- 非自立の名詞・動詞, 句読点, 記号, 助詞

4.2.2 人手評価

自動評価では WordNet を用いても以下の 2 例のような同義の表現に対してフレーズ単位の比較はできない.

例 9　謝罪した.

例 10　頭を下げた.

そのため人手によって出力結果の評価を行う. 分類器はナイーブベイズ分類器を 3-gram と組み合わせたものを使用した. 実際に引用とそれに対する引用表現の出力を読み 5 段階の評価を行う. 最終的にその結果の平均値によって評価する. 今回は発話と引用表現の組み合わせ 100 個に対し二名で評価を行った.

4.3 結果・考察

クラスの完全一致を用いた際の評価結果を表 1 に示す. また同義語を考慮して評価した際の結果を表 2 に示す. 人手評価の結果は評価者 A が 3.29, 評価者 B が 4.09 となった. 人手評価の実例の一部を表 3 に示す.

表 1: クラスの完全一致に対する自動評価の結果

1-gram	2-gram	3-gram	baseline
17.04%	15.87%	17.17%	7.23%

結果についての考察を行う. まず n-gram のうちどれが最も適しているかについて考える. 表 1, 表 2 共に

表 3: 人手評価の実例

発話	引用表現	評価者 A	評価者 B
今までくせで球種を見破られていたのを逆手にとったんですよ	振り返る.	5	5
ミスではない. 整備不良が原因だ	指摘する.	3	5
中学校では, 文法やスペリングにとらわれて『英語は難しい』と思い込んでいる生徒がいる. 『簡単なんだよ』と児童に伝えたい	指摘する.	1	2
衆院解散・総選挙に向けて県民にアピールすると同時に, 自由党県連からも参加してもらうことで準合流大会と位置付けている	強調した.	1	3
元気な限り, 見た人に喜んでもらえるような花を咲かせたい	意気込んでいる.	5	5
大徳寺の他の小寺院は狩野永徳, 探幽など大家の襖絵ばかりで, これに見劣りしない作品を目指して精進してきたつもり. ぜひ多くの人に見てもらいたい	あいさつした.	4	4
だれにでも登れるやさしい山だが, 落石などには注意してほしい	訴えた.	5	5
楽しそうに仕事をしている姿を見て, 任せてみようと思った	振り返る.	5	5
結構なことだ. ソ連, 中国, アラブ諸国の和平のための努力を高く評価する	強調した.	4	3
医師として復帰反対の立場は変わらず, ルール改正ではない. 国内に診断を引き受けてくれる医師がいるかどうかも疑問	指摘する.	1	5
当時の団長にポストを紹介されて就任した. 批判もあるかなと思うが, まだ元気だし, 仕事ができるならとお受けした	説明した.	5	5
うがいをしっかりして予防したい	強調した.	1	5
あまり悠長なことも言っていられない	振り返る.	1	1

表 2: 同義語を考慮した場合の自動評価の結果

1-gram	2-gram	3-gram	baseline
0.23	0.22	0.24	0.11

傾向は同じである. 文末記号がなければ 1-gram が最も高く, 文末記号があれば 3-gram が最も高い値を示している. このことから文末表現が重要な役割を果たしている事が分かる. 文末表現以外の特徴は 1-gram が最も良く考慮できていると考えられる. 基本的には 1-gram を用いて, 文末表現のみを 2-gram 以上の n-gram で表現するなどの方法を用いれば正解率が向上する可能性が考えられる.

5 まとめ

本研究では発話と引用表現を新聞記事中から自動獲得し, 引用表現を抽出することで発話に対する引用表現の自動推定を行った. 現在得られる出力は発話の対象を表す表現が存在しない引用表現なので表現力には限りがある. そこで発話の対象を表す表現も発話から推定し, その出力を引用表現に連結する. 対象を表す表現は現在扱っている引用表現に比べ異なり数が多く, 新聞記事のみを対象とした場合, データ量が足りないと考えられる. よって Web コーパス等新聞記事以外のデータの利用も視野に入れている.

参考文献

[1] Zhihui Jin *et al.,* Unsupervised Segmentation of Chinese Text by Use of Branching Entropy, *COLING–ACL2006*, pp.428–435, (2006).

[2] 鎌田修, 日本語の引用 p.24, ひつじ書房 (2000) .

[3] Andrew McCallum and Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification, In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48, (1998).

[4] Kishore Papineni *et al.* BLEU: a Method for Automatic Evaluation of Machine Translation, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, (2002).