時間表現と固有表現を標識とするウィキペディアからの言い換え知識獲得

市川浩丈　　松崎拓也　　宮尾祐介
東京大学　　東京大学　　国立情報学研究所
{ichikawa1014,matuzaki}@is.s.u-tokyo.ac.jp　　yusuke@nii.ac.jp

1. はじめに

歴史の試験問題を解くというタスクへの応用を念頭に、時間表現と固有表現を標識とすることで、ウィキペディアから言い換え表現を得る手法を提案する。具体的には、一つの時間表現と二つ以上の固有表現を共有する文集合を抜き出し、固有表現を含む係り受け部分木を言い換え表現として抽出する。時間表現と固有表現の一致という強い制約を用いることで、本来、並列性の低い文集合から、同一の歴史事実に対する複数の記述が得られ、高精度な言い換え表現の抽出が期待できる。

獲得した言い換え表現に対し、人手での評価を行い、さらに、NTCIR-9 RITE 大学入試サブタスク[1]のデータを用いて含意関係認識タスクに獲得した言い換え表現を用いた際の寄与を評価した。

2. ウィキペディアデータに対する前処理

日本語ウィキペディア（2011 年 8 月 25 日ダウンロード、約 100 万項目）を元データとして、以下の前処理を施した後、同一歴史事実を記述する複数の文を集めた：

1. 日時記法テンプレートの展開
2. 句点等のパターンを用いた文区切り
3. パターンマッチによる時間表現認識
4. 辞書マッチによる固有表現認識
5. Juman [2] による形態素解析
6. KNP [3] による係り受け解析

以下、前処理の詳細について簡単に述べる。

日本語ウィキペディアには、西暦和暦の併記など、日時の記述を簡単にするためのテンプレートが複数用意されている。まず、これらのテンプレートの内、頻度の大きいものを全て「YYYY 年 MM 月 DD 日」の形に展開した。その後、パターン「(一元号) N+年(和暦年)?(N+月(N+日)?)?」にマッチする表現を全てタグ付けし（N+は数値、「?」は省略可能を表す）、7,249,793 の時間表現を得た。

固有表現の認識は、岩波日本史辞典[4]の見出し語（約 18,000 語）および地名集日本 2007 [5] に収録された地名（約 3,900 語）を併せたリストを作成し、その中の表現と一致するものをタグ付けすることで行った。表記のゆれを吸収するため、ウィキペディアでリダイレクトの関係にあるページ名（例：日比谷焼き打ち事件→日比谷焼打事件）を用いてリストを拡張した。この結果、全部で 80,738,386 箇所がタグ付けされた。ただし、この中には“所”など一般語（の一部）として用いられているものも多数含まれており、全てがいわゆる固有表現という訳ではない。

表 1　固有表現クラス

クラス	固有表現の例
PERSON	卑弥呼、織田信長
ORGANIZATION	愛国婦人会、国際連盟
LOCATION	エトロフ島、薩摩藩
EVENT	戊辰戦争、本能寺の変
RULE	日米和親条約
IDEA	近代主義、啓蒙思想
TEXT	古事記、紫式部日記
BUILDING	江戸城、清水寺

表 2 固有表現クラス識別パターンの例

タイプ	パターンの例
固有表現パターン	*の変 → EVENT *事件 → EVENT *藩 → LOCATION
項目本文パターン	法律 → RULE *生まれ。→ PERSON *団体。→ ORGANIZATION

タグ付けした固有表現の内、言い換え表現抽出に特に有用であるものを選ぶため、固有表現集合の一部を表１に示す８つのクラスに分類し、いずれかのクラスに分類できたものだけを言い換えを含む文を抽出する際の標識として用いた。また、この固有表現クラスは言い換えパターンにおける名詞句の型を表すラベルとしても用いる。

固有表現をクラスに分類する際には、固有表現自身および日本史辞典の項目本文の最初の 2 文の

表３　文ペアの例

・ 宝亀元年（770 年）称徳天皇の崩御に際して、参議として藤原永手らとともに光仁天皇を擁立する。 ・ 宝亀元年（770 年）8 月に称徳天皇が死に、天智天皇の孫の白壁王が践祚した（光仁天皇）。
・ 古郡氏は建暦 3 年（1213 年）の和田合戦において和田義盛の挙兵に参加し、鎌倉で義盛勢が敗退すると古郡経忠・保忠兄弟は和田常盛や横山時兼らと波加利荘へ敗走し自害している。 ・ 1213 年（建暦 3）、和田合戦で和田義盛に加勢した横山時兼とその一党は鎌倉で全滅した。
・ 天正 11 年（1583 年）、清洲会議がきっかけで羽柴秀吉と対立した勝家は賤ヶ岳の戦いで争うも敗北。 ・ 1583 年、清洲会議で羽柴秀吉と柴田勝家の対立が深まり、賤ヶ岳の戦いにおいて秀吉が勝利、勝家の居城である北ノ庄城を攻め落とす。

文末に対するパターンマッチを用いた。固有表現自身に対しては 10 パターン、本文に対しては約 500 のパターンを定義した。その一部を表２に示す。この結果、計 8,701 個の用語が 8 つのクラスのいずれかに分類できた。

3. 同一歴史事実に対する複数記述の抽出

前処理を施したウィキペディアから、同一年を表す時間表現と２つ以上の固有表現を共有する文集合を取り出した。固有表現としては前述の 8 クラスに分類できたもののみを考慮した。

文集合を取り出す際、時間表現認識のエラーが多い西暦 50 年以前と、同一の年について多数の異なる事実が記述される 1961 年以降を表す時間表現は用いなかった。さらに、観察の結果、文数が少ない、あるいは極端に多い文集合は信頼性が低いことが分かったため、30～800 文からなる文集合のみを用いることにした。以上の処理の結果、1,591 個のグループが得られ、各グループ内で総当たり式に文のペアを作ることで、98,281 個の文ペアが得られた。得られたペアの例を表３に示す。

4. 言い換え表現の抽出

得られた文ペアから抽出する言い換えパターンとしては、様々な形式のものが考えられるが、本稿では、最も単純なパターンである名詞句と動詞の依存関係を単位とするパターンを抽出した。例えば、以下の文ペア：

“1895 年に下関条約が結ばれた。”
“1895 年に下関条約を締結した。”

からは、

(下関条約、が、結ばれる)
(下関条約、を、締結する)

という言い換え関係が得られ、さらに、固有名「下関条約」を RULE クラスに抽象化することで

(RULE、が、結ばれる)
⇔ (RULE、を、締結する)

という言い換えパターンが得られる。この例にも見られるように、受動-能動などの文法的に決まる言い換え関係については、今回は特別な考慮をせず、語彙的な言い換えと同列に扱った。また、動詞句による連体修飾を含む文ペア、例えば

“1895 年に下関条約が結ばれた。”
“1895 年に締結された下関条約は、…”

からは、

(RULE、が、結ばれる)
⇔ (RULE、＜連体修飾＞、締結される)

という形の言い換えパターンを抽出した。

より正確には、文ペア（s1, s2）に対し、

- s1 と s2 が共有する固有表現 e を含む文節を N1∈s1 および N2∈s2 とするとき、
- s1 と s2 が共有する時間表現が、動詞を含む文節 V1∈s1 および V2∈s2 と、それぞれ直接あるいは間接的に係り受け関係にあり、
- N1 と V1、N2 と V2 の間にそれぞれ直接の係り受け関係があるとき、

(Class(e)、Rel(N1, V1)、Base(V1))
⇔(Class(e)、Rel(N2, V2)、Base(V2))

という言い換えパターンを抽出した。ここで Class(e) は e の固有表現クラス、Rel は 2 文節間の関係を表す関係ラベル（格助詞または＜連体修飾＞）、Base(V) は V から「た」などの助動詞を除いたものを表す。

結果として、22,328 個の言い換えパターンが得られた。同一のパターンが複数の異なる文ペアから抽出される場合がある。この抽出元の文ペア数を以下抽出回数とよぶ。抽出回数が最も多かったパターンをいくつか表 4 に示す。

表 4　獲得された言い換えパターンの例

（P、が、挙兵する）⇔　（P、が、挙げる）	（P、が、上洛する）⇔（P、が、上洛してくる）
（P、を、て上洛する）⇔（P、を、擁する）	（E、が、起きる）⇔（E、が、起こる）
（L、には、描かれる）⇔（L、が、提出する）	（B、に、籠もる）⇔（B、に、籠城する）
（L、が、提出する）⇔（L、＜連体修飾＞、提出する）	（P、を、暗殺する）⇔（P、が、暗殺される）
（P、が、滅亡する）⇔（P、は、滅亡する）	（P、に、割譲する）⇔（P、に、割譲される）

注：P、L、E 等は表 2 のクラス名の頭文字

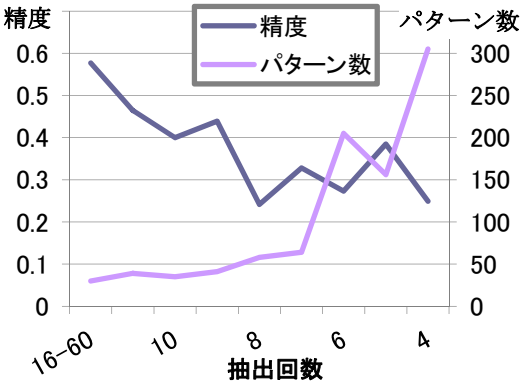

図 1　抽出回数ごとの獲得パターンの精度

5.　獲得された言い換え表現の評価

本節では、獲得した言い換えパターンを評価した結果を報告する。まず、抽出回数が最も大きかった 1000 個のパターンに対し、言い換えとしての妥当性を筆者の 1 人が判定した。判定は、パターン中の名詞クラス（PERSON など）の制約も考慮して行い、文脈なしで同義又は含意関係にあると判断できるとき「正しい」とした。この結果、328 パターンが正しいと判定された。判定結果をパターンの抽出回数ごとに整理したものを図 1 に示す。

図より、抽出回数の大きいパターンは精度が高い傾向が読みとれるが、最も抽出回数の大きいものでも精度は 60%を下回っている。より高精度の言い換えパターンを得るには、抽出元の文ペアから不適切なものを排除し、文ペアからの言い換え抽出手法についてもさらに工夫する必要がある。

次に、上記の評価で正しい言い換えと判定されたパターン 328 個に対し、パターンの動詞部分について、日本語 WordNet [6] の動詞 Synset との重なりを測定した。受け身・連体修飾などの文法的な変形により、同じ動詞の異なる形が現れたパターンを除くと、273 パターンが残り、そこから動詞のみを取り出して 222 個の動詞ペアを得た。これらの内、日本語 WordNet で同じ Synset に含まれるものは 17 ペア、hypernym-hyponym の関係にあるものは 9 ペアであった。

このように、日本語 WordNet から直接には得られない言い換えが多数獲得できた理由として、提案法では歴史分野に特有の「築城する⇔築く」、「討ち死にする⇔討死する」といった言い換え表現が獲得できたこと、また、下記の例：

（PERSON、が、討たれる）
⇔（PERSON、が、討死する）

のように、WordNet からは実質的に得られない、受け身の形と能動形の形が言い換えとなっているようなパターンが獲得できたことが挙げられる。

最後に、獲得した言い換えパターンを用いて含意関係認識器の精度を改善することを試みた。評価には、NTCIR-9 RITE 大学入試サブタスク[1]のデータを用いた。このデータは、大学入試センター試験の社会科系科目の問題と日本語ウィキペディアから作成されたもので、開発/テスト用データとして 499 文ペア/442 文ペアを含む。タスクは、与えられた文ペア（t1, t2）に対し、t1 が t2 を含意するか（Y）否か（N）を識別するものである。

ベースライン手法として、Shima ら[7]の Voting Score 法とほぼ同様のものを用いた。この手法は、まず、t1 と t2 に異なる日時を表す時間表現が含まれる場合は無条件にNを出力し、それ以外の場合、t1/t2 の中の内容語の集合 W1/W2、内容語間の依存関係の集合 D1/D2、および文字の集合 C1/C2 に対し、t2 の各要素の t1 による被覆率 α、β、γ を

$$\alpha = |W1 \cap W2| / |W2|$$
$$\beta = |D1 \cap D2| / |D2|$$
$$\gamma = |C1 \cap C2| / |C2|$$

と算出し、閾値 θ との比較により

$$(\alpha + \beta + \gamma) / 3 > \theta$$

のときY、そうでないときNを出力する。内容語間の依存関係の集合 D1/D2 の要素は言い換えパターンと同様＜修飾語、関係、被修飾語＞の三つ組で表されるが、動詞－名詞の関係には限らない。

このベースラインに対し、獲得した言い換えパターンを利用した改善を以下のように試みた。まず、t2 中の依存関係集合 D2 の内、関係 r にある動詞 v と名詞 n の対を探す。名詞 n が文 t1 中の内

表５　含意性認識による評価（開発データ）

	A(%)	P(%)	R(%)	F1(%)
Baseline	68.16	67.18	43.14	52.54
+言い換え	68.55	67.67	44.12	53.41

表６　含意性認識による評価（テストデータ）

	A(%)	P(%)	R(%)	F1(%)
Baseline	65.16	62.39	37.57	46.90
+言い換え	64.93	64.44	32.04	42.80

容語 W1 に含まれ、動詞 v が W1 に含まれないとき、関係 r および動詞 v と適合する言い換え

（_、r、v）⇔（_、r’、v’）

を検索し、パターンを適用した場合の動詞の言い換えv’ がW1 に含まれるかどうかを調べる[1]。言い換えた動詞v’がW1 に含まれる場合はパターンを適用してt2 を書き換える。このような書き換えをt2 中の全ての名詞―動詞の依存関係について行い、書き換え後のt2 を用いて上述の被覆率を計算する。

ベースラインおよび言い換えを用いた場合の各々について、閾値 θ の調整と識別精度の評価を開発用データ上の 5 分割交差テストで行った結果（平均値）を表 5 に示す。表中の数値は、2 値分類の精度（A）、出力ラベル「Y」に対する適合率（P）、再現率（R）、および F1 スコアである。結果から、言い換えの利用によってわずかに含意関係認識の精度が向上したことがわかる。また、開発用データの 499 文ペアに対し、言い換えが適用された回数は 29 回、その内、目視で妥当な言い換えと判断できたのは 23 回、誤った言い換えは 6 回であった。

最後に、閾値 θ を開発データ全体で調整したときのテストデータに対する識別結果を表 6 に示す。開発データ上での結果と異なり、テストデータでは言い換えの利用により識別精度が低下した。この違いは、データ数が小さい、開発/テストデータの性質が異なる等の理由によると思われる。

6. 関連研究

言い換え獲得に関する研究は既に多数あるが、提案手法に近いものとして、同日のニュース記事（[8]など）や同一の対象に対する複数の定義文[9]からの獲得手法が挙げられる。これらの手法では、機械学習を用いて認識した固有表現[8]や、定義されるものの名前、および定義文に特有な部分文字列“とは”[9]が言い換え発見のための標識として用いられている。これに対し、提案手法は、時間表現および特定クラスの固有名という強い手掛かりを用いることで、同一事実に関する複数記述を文単位で集められること、また、歴史辞典などのリソースを併用することで、文中の固有名のクラスが高い精度で同定可能であり、これを言い換えを含む文ペアの選別、および言い換えパターン抽出のヒントとして利用可能である等の利点がある。

7. おわりに

本稿では、時間表現と固有表現の組を標識として同一の歴史事実に対する複数の表現をウィキペディアから抽出し、そこから言い換えパターンを獲得する手法を提案した。獲得されたパターンに関する評価実験により、高精度の言い換えパターン獲得にはさらに手法の洗練が必要であるものの、WordNet からは得られない言い換えも獲得できたこと、また、含意関係認識のようなタスクに有効に利用できる可能性を示した。今後の課題として、固有表現辞書の拡張によってさらに多数の言い換えパターンを獲得することや、不適切な言い換えの排除、単一の名詞-動詞ペアのパターンを超える複雑な言い換えパターンの抽出などが挙げられる。

[1] この実験では、PERSON など名詞句のラベルを無視してパターンマッチを行った。

参考文献

[1] 宮尾祐介，嶋英樹，金山博，三田村照子．大学入試センター試験を題材とした含意関係認識技術の評価．言語処理学会第 18 回年次大会予稿集，2012.

[2] 黒橋禎夫、河原大輔．日本語形態素解析システム JUMAN VERSION 6.0.

[3] D. Kawahara and S. Kurohashi. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proc. HLT-NAACL2006*, 2006.

[4] 永原慶二ほか．岩波日本史辞典．岩波書店，1999.

[5] 日本国政府．地名集 日本．2007. http://www.gsi.go.jp/kihonjohochousa/gazetteer.html

[6] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi and K. Kanzaki. Enhancing the Japanese WordNet. In *Proc. ACL-IJCNLP2009*, 2009.

[7] H. Shima, Y. Li, N. Orii and T. Mitamura. LTI's Textual Entailment Recognizer System at NTCIR-9 RITE. In *Proc. NTCIR-9*. 2011.

[8] Y. Shinyama, S. Sekine and K. Sudo. Automatic Paraphrase Acquisition from News Articles. In *Proc. HLT-02*. 2002.

[9] C. Hashimoto, K. Torisawa, S.D. Saeger, J. Kazama and S. Kurohashi. Extracting paraphrases from definition sentences on the Web. In *Proc. HLT-ACL2011*. 2011.