教師なし学習による関係抽出に基づくパラフレーズの獲得

長谷川 隆明
日本電信電話株式会社
NTT サイバースペース研究所
hasegawa.takaaki@lab.ntt.co.jp

関根 聡　　Ralph Grishman
Dept. of Computer Science
New York University
{sekine, grishman}@cs.nyu.edu

１．はじめに

自然言語処理の様々なタスクにおいて、タスクに関する情報が付与されたタグ付きコーパスを用いた機械学習に基づく教師付き学習のアプローチは、これまでに大きな成功を収めてきた。しかしながら、このアプローチは、機械学習に適用するためのタグ付きコーパスを大量に用意しなければならず、高度なタスクになる程、コーパス作成のための大量の時間とコストがかかるという問題を内在していた。近年、この問題を回避するために、大規模に収集されたタグなしコーパスから、意味的な知識を獲得しようとする試みが見られるようになった。意味的な知識の中でも、言語表現を言い換えるパラフレーズは、情報抽出や質問応答の研究において特に重要な役割を果たすと考えられる。例えば質問応答においては、多くのパラフレーズが事前に獲得できれば、これまで表現のミスマッチにより答えることができない質問にも答えることが可能となり、回答の精度を向上させることができる。本稿では、特に固有表現に関する情報抽出や質問応答で効果の高いと考えられる固有表現の対の間に存在する特定の関係を表すパラフレーズに着目する。例えば、“Company A’s acquisition of Company B”と“Company B agreed to be bought by Company A”はニュアンスこそ違うが、イベントのレベルから捉えれば両者は共に Company A と Company B が M&A という関係にあることを表現している。従って、これらは M&A の関係についてのパラフレーズであると考えることができる。本稿では、大規模なタグなしコーパスからこのような関係を表すパラフレーズを網羅的に獲得する方法を提案する。

本稿では、まず第 2 節で従来研究とその課題を述べ、第 3 節でパラフレーズを獲得する方法を提案し、第 4 節で評価実験の結果を報告し、第 5 節で結論と今後の課題を述べる。

Unsupervised Paraphrase Acquisition via Relation Discovery by Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman. NTT Cyberspace Laboratories, NTT Corporation and Dept. of Computer Science, New York University.

２．従来研究と課題

パラフレーズ獲得における従来の研究のアプローチのひとつに、出典は異なるが内容がほぼ同一であるコンパラブルなコーパスを用いる方法がある。例えば Shinyama ら[6]は同じ日付の異なる新聞社の新聞記事を対象とし、Barzilay ら[1]は同じ物語に対する複数の翻訳を対象とすることでコンパラブルなコーパスを得ている。しかしながら、コンパラブルなコーパスの文のアライメントの精度の問題に加え、利用可能なコンパラブルなコーパスの種類や量は極めて限られており、得られるパラフレーズの量も多くは期待できない。

パラフレーズ獲得のためのもうひとつのアプローチは、大規模なコーパスから統計量を用いて入力とするフレーズと類似したフレーズを発見する方法である。Lin らは類似した動詞句を収集するために、大規模コーパスにおいて出現する各動詞句の主語と目的語に現れる単語の分布の相互情報量を用いた[3]。また Ravichandran らは、質問応答のタスクにおいて、ある対象とその対象についての属性を質問としたときの答えとなる属性値の対から、その属性に関するパラフレーズを獲得する方法を提案した[4]。このアプローチでは大規模なコーパスがあれば様々なパラフレーズが得られる可能性がある。しかしながら、最初にシードとして、どのような動詞句に関するパラフレーズかあるいはどのような属性を表す対象と属性値の対かを入力しなければならないため、獲得可能なパラフレーズが制限される。

本稿では、入手困難なパラレルコーパスではなく大規模なコーパスを使い、人手によるシードの選択を必要としない教師なし学習によるパラフレーズの獲得の方法を提案する。

３．パラフレーズの獲得方法

３．１　アプローチ

本稿では大規模なタグなしコーパスから関係を表現するパラフレーズを獲得する方法を提案する。はじめに本稿において対象とするパラフレ

ーズについて説明する。一口にパラフレーズといっても、単純な単語や名詞句、動詞句の言い換えからより複雑な文の構造が変わるような言い換えまで幅が広いし、パラフレーズの数量についても膨大であることが想像される。そこで我々は、パラフレーズを限定するために、コーパス内でほぼ常に成立するような、固有表現同士の間にある特定の静的な関係に着目する。なぜならば、コーパスからある特定の静的な関係が得られれば、その関係を表現する複数のフレーズがパラフレーズであることが期待されるからである。この仮定に基づいて、我々は、まずコーパスから固有表現同士の関係を抽出し、次に得られた関係に基づいてパラフレーズを獲得するという新しいアプローチを提案する。我々のアプローチは教師なし学習による方法であり、必要とするものは大規模なコーパスと固有表現抽出器のみである。以下に基本的なアイデアを示す。

1. 固有表現抽出器を用いてコーパスに固有表現を付与する
2. 固有表現の対をそれらが出現する文脈でクラスタリングすることにより、類似した関係にある固有表現の対のクラスタを得る
3. 同一のクラスタを対象とし、固有表現の対に存在する主要な関係を共通に表しているフレーズを選択する

以下では、個別の処理について説明する。

３．２　固有表現抽出

本稿で提案する教師なし学習によるアプローチは、コンパラブルなコーパスや人手により与えられるパラフレーズのシードを必要としない代わりに、固有表現抽出器を用いる。昨今の固有表現抽出器は実用レベルに達していることに加え、識別できるタイプも細分化されている。関係の抽出の精度を高めるためには、細分化された固有表現のタイプは有利に働くと考えられる。そこで我々は、150 種類のタイプの固有表現を識別できる固有表現抽出器[5]を用いる。

３．３　関係の抽出

コーパスから関係を抽出するために、まず同一文内に前後５単語以内で共起して現れる固有表現の対を対象とし、固有表現の対の間のフレーズを文脈としてコーパスから収集する。これらはステミングにより基本形に変換する。固有表現の対の出現順序を考慮して、順序が異なる文脈は別の文脈として扱う。

次に、各々の固有表現の対が共起出現した文脈をベクトル空間モデルで表現する。ベクトルの各要素の値には、文脈中にある単語頻度と文書頻度の逆数の積である tf*idf を用いる。文書頻度はコーパス全体においてその単語を含む文書数とする。また、ストップワードは除く。

得られた各ベクトル間のコサイン距離を調べることにより、固有表現の対同士の類似度を得る。類似度に基づいてクラスタリングを行い、得られたクラスタに対応した関係を得る。関係は、クラスタ内の固有表現の対の持つ文脈に多く共通する単語により規定される。クラスタリングには、あらかじめクラスタの数を決めることができないので、階層型クラスタリングを行うが、本稿では最長距離法(complete linkage)を採用した。

３．４　フレーズの選択

固有表現の対の文脈によるクラスタリングによって得られるクラスタには、特定の静的な関係にある固有表現の対とその関係を表現する多くのフレーズが存在すると考えられる。しかしながら、フレーズは前後５単語以内に共起するという条件で機械的に収集するため、すべてのフレーズがその関係を表現するとは限らない。このため、収集されたフレーズをそのままパラフレーズとすることができない。そこで、我々は特定の静的な関係を表すフレーズだけをフィルタリングするため、次の２つの制約を提案する。

共通フレーズ制約：クラスタ内のフレーズが多くの固有表現の対に共有されるならば、そのフレーズは固有表現の対が属するクラスタの関係を表現することを支持している信頼性が高いと考えられるので、そのようなフレーズはパラフレーズとみなす

共通単語制約：クラスタ内の多くの固有表現の対に共通して出現する単語は、そのクラスタの関係を表現するラベルと考えられるので、そのような単語を含むフレーズはパラフレーズとみなす

共通フレーズ制約の方がフレーズを構成する全単語が一致しなければならないため、共通単語制約よりも厳格な制約である。共通単語制約において、関係を表現するラベルを特定するために、クラスタで共有される単語の割合を調べる。ある単語が２つの固有表現の対のフレーズに共通して出現する場合に、その単語について２つの固有表現の対の間にリンクを張る。クラスタに N 個の固有表現の対が存在する場合には、N 個の組み合わせである $N(N-1)/2$ のうち、いくつのリンクが張られているのかという割合を求める。例えば，ある単語の割合が１の場合は、クラスタ内のすべ

表 1　クラスタの主要な関係とフレーズに共通な単語

クラスタの主要な関係	割合	クラスタ内の共通の単語（比率）
President	17/23	President (1.0), president (0.415), …
Senator	19/21	Sen. (1.0), Republican (0.214), …
Prime Minister	15/16	Minister (1.0), minister (0.875), Prime (0.875), …
Governor	15/16	Gov. (1.0), governor (0.458), Governor (0.3), …
Secretary	6/7	Secretary (1.0), secretary (0.143), …
Republican	5/6	Rep. (1.0), Republican (0.667), …
Coach	5/5	coach (1.0), …
M&A	10/11	buy (1.0), bid (0.382), …
M&A	9/9	acquire (1.0), acquisition (0.583), buy (0.583), …
Parent	7/7	parent (1.0), unit (0.476), own (0.143), …

表 2　人手によるフレーズのクラスの分類

分類	1	2	3	4	計
PERSON-GPE	59	31	17	222	329
COMPANY-COMPANY	102	10	29	177	318

ての固有表現の対がその単語を文脈に持つことを意味する。この割合が高い単語ほど関係のラベルにふさわしいので、そのような単語を使って関係を表現するフレーズを選択できると考えている。

４．評価実験

我々は、関係の抽出とフレーズの選択という連続する２つのステージに分けて実験を行った。実験では、対象言語を英語とし、コーパスとして New York Times 1995 年版を用いた。

最初のステージとなる関係の抽出では、PERSON と GPE[1]の間にある関係(PERSON-GPE)及び COMPANY と COMPANY の間にある関係（COMPANY-COMPANY）の二つのドメインについて実験を行った[2]。PERSON-GPE のドメインでは 177 個の固有表現の対から 38 個のクラスタが、COMPANY-COMPANY のドメインでは 65 個の固有表現の対から 10 個のクラスタが得られた。クラスタの精度は PERSON-GPE が F 値 80、COMPANY-COMPANY が F 値 75 であった。得られたクラスタのうち 5 個以上の固有表現の対を含むクラスタを表 1 に示す。表 1 には、クラスタのすべての固有表現の対の中で主要な関係を持つ固有表現の対が占める割合と、クラスタ内の各固有表現の対が現れるフレーズに多く共通して現れる単語が示されている。

2 番目のステージとなるフレーズの選択では、関係の抽出で得られたクラスタから最も共通の関係を表現するフレーズを選択する実験を行った。まずクラスタ内の各固有表現の対が共起出現するフレーズをコーパスからすべて収集した。ここでいうフレーズは、同一文内の固有表現の対に挟まれる単語列である。ただし、コーパス中に一度しか出現しないフレーズや、記号やストップワードのみからなるフレーズは対象外とした。クラスタごとに得られたフレーズの集合を実験のデータとする。提案した２つの制約を用いてフレーズの集合からパラフレーズとみなすことができるフレーズを選択する前に、フレーズの集合には実際にどの程度パラフレーズとみなすことができるフレーズが含まれているのかを人手で調査した。調査の結果、フレーズを（１）関係を表すフレーズ、（２）余分な単語が入ったフレーズ、（３）より広い意味のフレーズ、（４）関係を表していないフレーズの 4 段階に分類した。分類の結果を表 2 に示す。分類１のフレーズ同士はいずれもパラフレーズとみなすことができる。

評価は、（１）のみ、（１）と（２）をそれぞれ正解として、２つの制約およびその選言（論理和）に対して再現率、適合率および F 値を求めた。実験の結果を表 3 に示す。実験では、共通フレーズ制約においてフレーズを共有する固有表現の対は２つ以上とし、共通単語制約における共通単語の比率を 0.4 以上に設定した。ベースラインとなる制約を用いない場合と比較するため、対象とするフレーズの集合すべての評価も掲載した。ベースラインではすべてのフレーズを選択したことになるので、正解とするフレーズもすべて含むために再現率を 100%としているが、主要な関係を表現しないフレーズも多く含むので適合率は低い。２つの制約によるフレーズのフィルタリング

---

1 米国のプロジェクト ACE (Automatic Content Extraction)により導入された Geo-Political Entity の略語で、統治機能の存在する場所を指す。

2 関係抽出に関する実験の詳細については、文献[2]を参照されたい。

表 3　パラフレーズ獲得の精度

	クラス	ベースライン			共通フレーズ制約			共通単語制約			両制約の選言		
		再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値	再現率	適合率	F 値
PERSON	1 のみ	100	18	31	62	44	51	82	38	52	92	35	50
-GPE	1＋2	100	28	44	53	57	55	80	57	66	89	51	65
COMPANY-	1 のみ	100	33	49	33	77	47	58	82	68	70	76	72
COMPANY	1＋2	100	36	53	30	77	44	57	89	70	68	81	74

表 4　獲得されたパラフレーズの一部

関係	フレーズ
M&A	A bought B
	A has agreed to buy B
	A, which is buying B
	A’s proposed acquisition of B
	A’s acquisition of B
	A’s agreement to buy B
	A’s purchase of B
	A bid for B
	A’s takeover of B
	A merger with B
	A succeeded in buying B
	B, which was acquired by A
	B would become a subsidiary of A
	B agreed to be bought by A

を行うことで、再現率は多少犠牲になるが適合率が大幅に向上することがわかった。PERSON-GPE では共通単語制約が最も精度が高く、COMPANY-COMPANY では 2 つの制約の選言が最も精度が高かった。また、COMPANY-COMPANY の方が PERSON-GPE よりも F 値が高い結果が得られた。さらに、得られたパラフレーズを検証した結果、とりわけ COMPANY-COMPANY ではバラエティに富んだパラフレーズが得られたことがわかった。COMPANY-COMPANY において２つの制約の選言により正しく選択された主要な関係を表現するフレーズの例を表 4 に示す。表 4 のフレーズは単に単語レベルの言い換えだけに留まらず構文的にもバラエティに富み、これらのフレーズはまさにパラフレーズであると言える。

５．おわりに

本稿では、教師なし学習によりコーパスから任意の同一な関係を表現するパラフレーズを獲得する方法を提案した。新聞記事を用いた２つのドメインにおける実験の結果、固有表現のイベントに関する表現が多く現れた COMPANY-COMPANY では特に再現率適合率とも精度良くパラフレーズが獲得できた。一方、PERSON-GPE での適合率が低かった原因は、主要な関係を表現しないフレーズ，とりわけ関係詞を含むフレーズや複文における節にまたがるフレーズを除外できなかったことにあると考えられる。今後は、このような課題を解決することによりさらに精度を高めていきたい。また、否定詞を含むフレーズや特定の文脈でのみパラフレーズとみなされるフレーズの誤った選択も精度を下げる原因となっていた。これらのフレーズを排除する方法についても今後検討していきたい。

参考文献

[1] Regina Barzilay and Kathleen McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-EACL2001)*, pages 50-57.

[2] Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman. 2004. Discovering Relations among Named Entities from Large Corpora. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pages 416-423.

[3] Dekang Lin and Patrick Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pages 323-328.

[4] Deepak Ravichandran and Eduard Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 41-47.

[5] Satoshi Sekine. 2001. OAK System (English Sentence Analyzer). http://nlp.cs.nyu.edu/oak/.

[6] Yusuke Shinyama and Satoshi Sekine and Kiyoshi Sudo and Ralph Grishman. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proc. of Human Language Technology Conference (HLT-2002)*, pages 313-318.