

修士学位論文要約（平成29年3月）

文脈依存選択選好モデルの共参照解析への適用

中山 周

指導教員：乾 健太郎

Coreference Resolution with Context-dependend Selectional Preference

Supervisor: Kentaro Inui

Coreference resolution is one of the important research tasks in natural language processing that analyzes whether or not the phrase in a sentence refers to the same object. It was known that a semantic constraint between words called “selective preferences” is a major clue for performing coreference resolution. In this research, we propose a selectional preference model considering not only semantic classes of words themselves but also how words are mentioned in sentences using with expressing words and relationships between words on a vector space. We confirmed whether the proposed model acquired context-dependent selection preference.

1. はじめに

共参照解析は、文章内の複数の単語や句が同一の指示対象を指しているか否かを推定するタスクであり、人間が使う日本語や英語などの自然言語をコンピュータで処理する自然言語処理において、重要な研究タスクの一つである。

例えば、例文 1) が与えられた時、 $it_{(?)}$ に対して、 $banana_{(j)}$ という名詞が、ある世界において同一の物体を指示しているかどうかということを解析する。

1) A monkey_(i) gets a banana_(j), and eats it_(?).

このように、同一の物体を指示する単語句間の関係を共参照関係と呼び、共参照関係を構成する単語句をメンションと呼ぶ。

例文 1) で、 $it_{(?)}$ と $banana_{(j)}$ は共参照関係にあるが、 $it_{(?)}$ と $monkey_{(i)}$ は共参照関係にないことは、「eat の目的語としてふさわしいのは食べ物である」という知識からわかる。このような、主語や目的語となりうる単語が述語によって制限される選択選好という言語現象があり、共参照解析にも利用されてきた。

しかし、単なる選択選好だけでは共参照を解析するにあたって十分でない場合がある。例えば、例文 2)において、 $him_{(?)}$ と $John_{(i)}$ は共参照関係にあるが、單に動詞だけの選択選好だけを考えると、 $arrest$ の目的語は、人物である $John_{(i)}$ にも $Bob_{(j)}$ にもなりうる。

2) John_(i) killed Bob_(j). Police arrested him_(?).

この場合、動詞だけの選択選好ではなく、選択選好対象の単語がどのように言及してきたのかも考慮するほうが自然である。例文 2)に当てはめて考えると、「誰かに kill された人」よりも「誰かを kill した人」のほうが $arrest$ されやすいため、 $arrest$ の目的語は $John_{(i)}$ であると判断できる。

我々は、ある単語が文章内でどのように言及してきたかという文脈情報により選好性が変容するような選

択選好(文脈依存選択選好)が共参照解析に有効であると考えた。

2. 文脈依存選択選好モデル

選択選好は、単語の意味と、主語や目的語などの単語間の関係性(意味役割)によって選好性が変化する場合があるため、構築するモデルもこれらの情報を取り入れられるように設計する。先行研究として Inoue らの研究¹⁾があり、本研究では Inoue らのモデルで扱えない、任意の構造の文やフレーズの選択選好性を計算可能なモデルを提案する。

提案モデルには、依存構造に基づき、単語と意味役割をベクトル空間上に表現することで構成的な計算を可能にする DCS Vector²⁾を導入する。依存構造とは、文内の単語間の係り受け関係を表したものである。DCS Vector は、単語をベクトル、意味役割を行列として表現することで、ある単語がどのように言及されたのかを依存構造から計算することを可能にする。

提案モデルを、例文 2)を例に挙げて説明する。通常の DCS Vector では、”police arrest”的目的語として him という単語ひとつを予測するように学習する。ここで、him が「Bob を殺した John」であることを表現するために、Inoue らのモデルと同様に、him の単語部分に”John killed Bob”を挿入した依存構造木の DCS Vector を考える(図 1)。この木を Tian らと同様に DCS Vector として学習することで、文脈を考慮した選択選好性を獲得する。

また、Noise-contrastive estimation³⁾に基づき、学習データから生成した負例を学習に利用することで、効率的に学習する。”Police arrested him”的な解析対象の代名詞を含む文をターゲットパス、”John killed Bob”的な先行詞を含む文をコンテキストパスとして、後者をコーパス全体の出現頻度分布に基づきランダムに入れ替えることで、負例を生成する。

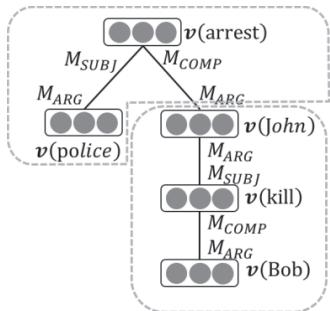


図1 文脈依存関係を組み込んだ DCS Vector

3. 評価実験

モデルの学習に用いるデータには、ClueWeb 12 に Stanford CoreNLP を用いて依存構造解析及び共参照解析を行ったけっかを用いる。Stanford CoreNLP による共参照解析結果には誤りが含まれているため、共参照関係が文内で閉じており、かつ、メンションの主辞の文字列が完全に一致する解析結果のみ抽出することで解析誤りを取り除く。本研究では、得られたデータの 10%を抽出し、約 1 億件を学習データとして使用した。

本実験では、未知の正例データと、学習時の負例の生成方法と同様の方法で生成した負例を判別する二値分類問題を解くことで、提案モデルが文脈に依存する選択選好性を捉えられているか確かめる。以下の三つの評価タスクを解くことで、提案モデルが文脈に依存する選択選好性を正しく捉えることができるのかを検証する。

Pseudo-disambiguation (PD) 単語のみの意味表現が選択選好へ与える影響を評価

Context-sensitive PD (CSPD) 文脈に依存した単語の意味表現が選択選好へ与える影響を評価
選択選好性を決定する上で、どの程度の係り先まで必要なのかを調べるため、ターゲットメンションを根にした DCS Tree と、コンテキストメンションを根にした DCS Tree を、それぞれ深さ 0, 1, 2 またはすべてを使う条件で実験した。

提案モデルの PD による実験結果を表 1 に示す。正例と負例をランダムに選択した場合、精度は 0.5 である。また、表 1 の no matrix は、意味役割行列を使用せず、単語ベクトルのみを用いて DCS Vector を計算した場合の実験結果である。

各値は、正例と負例を弁別する二値分類の精度を表す。表 1 の結果では、ランダムや意味役割行列を使用しない場合より、単語ベクトルと意味役割行列と共に使用した場合の精度が高い。特に、target depth が 1 の時、つまり二つの単語間の選択選好が、高い精度で弁別ができる。

次に、CSPD による実験結果を表 2 に示す。target depth 及び context depth が 1 の時、意味役割行列を使った場合と使わなかった場合の精度を比較すると、意味役割行列を精度が高いという結果となった。一方で、no matrix と比較して、depth を深くした場合、つまり係り受け関係をたくさん考慮した場合、精度が

表1 提案モデルの実験結果 (PD)

target depth	context depth	random	PD	
		accuracy	accuracy	accuracy (no matrix)
1	1	0.5000	0.7602	0.6904
2	1	0.5000	0.7022	0.6683
all	1	0.5000	0.6866	0.6671

下がっていく傾向がある。これは、深い係り受け関係を構成的に計算できるほど行列が学習できていないためだと考えられる。

また、表 2 の target depth, context depth がどちらも all の場合に特に精度が高い。これは、評価データの作成方法に由来する。評価データは、共参照関係にあるメンションが同一文内に出現していることを条件にフィルタリングしているため、ターゲットとコンテキストの DCS Tree は、同一の文から生成される。そのため、ターゲットとコンテキストの DCS Tree を深く辿るほど、それぞれを構成する単語の共通部分が大きくなるに従い、精度が高くなる。

表2 提案モデルの実験結果 (CSPD, CSPD-X)

target depth	context depth	CSPD	
		accuracy	accuracy (no matrix)
1	1	0.6441	0.6360
2	1	0.6279	0.6247
all	1	0.6465	0.6348
1	2	0.6213	0.6370
2	2	0.6250	0.6391
all	2	0.6492	0.6600
1	all	0.5926	0.6464
2	all	0.5954	0.6580
all	all	0.6971	0.6946

4. まとめ

本論文では、文脈依存選択選好モデルを提案した。既存の文脈依存選択選好モデルで問題となっていた、モデルに適用できる単語の出現パターンが制限される問題を、単語と意味役割の意味表現の構成的な計算を行う DCS Vector を導入することで解決した。また、提案したモデルの有効性の検証した結果、二単語間の選択選好性は獲得していることを確認した。

文献

- 1) N. Inoue, Y. Matsubayashi, M. Ono, N. Okazaki, and K. Inui. Modeling Context-sensitive Selectional Preference with Distributed Representations. 2016. COLING.
- 2) R. Tian, N. Okazaki, and K. Inui. Learning Semantically and Additively Compositional Distributional Representations. 2016. ACL.
- 3) M. Gutmann and Aapo Hyvarinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. 2012. Journal of Machine Learning Research.