E-034

概念辞書を利用した日本語テキストのエンティティ間の意味役割分類

Semantic Role Classification between Entities in a Japanese Text based on Semantic Dictionary

井口宜久† 石塚満†
Nobuhisa Inokuchi Ishizuka Mitsuru

1 はじめに

現在、電子データ化されたテキストデータが急激に増加しており、それら膨大なデータから情報を取り出し構造化することで、ウェブ検索や情報抽出に役立てようとする研究が広く行われている。特に最近では、自然言語テキストの中から述語項構造を特定しそれを分類する意味役割分類（Semantic Role Labeling）というタスクがよく研究されており、これは自然言語処理分野の様々なアプリケーションにおいて重要な役割を担うものである。現在、FrameNet や PropBank といった大きなコーパスが存在し、これらを用いて高い精度で意味役割を分類するシステムが既に作られている。

一方横井らによる研究[1]は、述語項構造のみでなく、文全体を一つの意味構造に翻訳することを目的とした。横井らはそれを概念記述言語(Concept Description Language)と呼び、述語とその項だけでなく、文中のすべてのエンティティに対して、それがどのエンティティとどの意味で関係するのかを特定し、文の意味を表現している。

本論文では自身で作成したコーパスを基に自然言語テキストにおける文節間の関係を分類する分類器を機械学習により作成しその精度の評価を行った。その際、データのスパースネス問題に対応するため、概念辞書を利用し、どの程度精度の向上が見られるかを分析した。

2 背景

2.1 Semantic Role Labeling

Semantic Role Labeling とは、文中で表された意味を表現する述語とそれに対してなにか関係を持つ語句（項）との間にどのような意味があるかを同定するタスクであり、述語項構造解析とも呼ばれるものである。

英語における Semantic Role Labeling(以下では SRL と呼ぶ)の例を以下に示す。

The girl on the swing whispered to the boy beside her.

→	Agent	: The girl on the swing
	Predicate	: whispered
	Recipient	: the boy beside her

SRL は文の意味解析における重要な要素技術の一つであり、統計的機械翻訳、質問応答、含意関係認識などの自然言語処理の高度なアプリケーションにおいて、Semantic Role を利用することの有効性が示されている。

SRL は一般に次の 3 つのステップに分けて考えられている。第一に述語－項関係の特定（Identification）、第二にその関係の分類(Classification)、最後に大域的最適化（Global Scoring）で、これは上の二つの段階の結果を文全体として自然な文になるように調整するものである。

†東京大学情報理工学系研究科電子情報学専攻

2.2 CDL.nl

CDL(Concept Description Language)[1]は，自然言語テキストだけでなく他も含む広いメディアー般が表す概念を表現するために設計された汎用で基本的な枠組みである．単純な３つの組表現（<実体１，関係，実体２>，<主語，述語，目的語>あるいは<実体，属性，属性値>などを表す）を基礎とし，グラフ表現だと実体を表すノードと関係を表すアークから成る。

CDL.nl(CDL natural language version)は、自然言語テキストの意味概念を汎用的に表現する概念記述言語である．構造を表現するうえでの二つの基本的な要素は、実体(Entity)と関係(Relation)であり、実体とは文の意味の構成要素の一つを表すものである。

自然言語テキストから CDL.nl への変換は、英語について Y.Yan らによる研究[3]が既に行われており、一定の成果をあげている。その研究では、Identification をルールベースで行ない、Classification は機械学習でやる、という手法をとっている。

2.3 WordNet

WordNet は英語の概念辞書であり、各単語が synset と呼ばれる同義語の集合（＝概念）に分類されているものである。また、各 Synset は上位・下位の関係（IS A 関係）によって定義されるような階層構造にまとめられている。

WordNet データベースは現在 115,000 の synset に分類された 150,000 語・200,000 語義（語と意味の組み合わせ）を収録している。また、多言語への翻訳も行われており、特に日本語 WordNet は既に公開済みで、56,741 の synset に分類された 92,241 語・157,398 語義が収録されている。

3 手法

3.1 問題設定

SRL での考えと同じく日本語の CDL への変換も、1.Identification、2.Classification、3.Global Scoring の３段階に分けて考え、今回の研究では 2.Classification だけを行うこととした。また、分類する意味関係としては SRL に含まれる、述語項構造の関係のみに限定して、分類を行うこととした。このため、分類に使用するラベルは CDL のタグ全 44 種類の内の 27 種類である。

つまり、本稿で解く問題は、係り受け解析器の結果と、どの二つの文節（エンティティ）が関係を持つかという 1.Identification の結果を入力とし、その関係の分類を出力することである。

3.2 学習手法

学習アルゴリズムとしては最大エントロピー法を採用した。最大エントロピー法とは、モデルのパラメータを決定する際に、コーパスにおいてのエントロピーが最大になるようにパラメータを定めるという手法である。実際にどのようにパラメータを計算するかについては様々な数学的手法が存在するが、本研究では

L-BFGS 法を用いた。

3.3 特徴量

3.3.1 文節の特徴量

・自立語について

自立語そのもの、およびその品詞。また、その自立語が属する概念辞書上での区分も特徴量として利用した。

・付属語について

自立語に付属した助詞をすべて特徴量とした。また、句読点の有無も特徴量として利用した。

3.3.2 その他の特徴量

まず文節間の距離を特徴量として利用した。距離の定義は文節間にいくつ他の文節があるかとした。

また、述語項構造の述語側の文節が、他にどのような係り受けを受けているかを、各文節の末尾の助詞を元に特徴量として利用した。

4 実験

4.1 実験概要

京都大学コーパスから取り出した３００文に対して人手で CDL のラベルをつけたもの（９６５事例）をコーパスとして学習を行った。また、概念辞書としては WordNet 日本語版を利用し、特徴量としての使い方をいくつか変えて精度を比較した。

また述語に係る文節が末尾にどんな助詞が付いているかについても、特徴量の定義を変えて制度の比較を行った。

精度の計算については、１０分割したコーパスのうち９個を利用して学習を行い、残りの一つを利用して精度を測定し、これを１０回繰り返して平均をとることで分類器の精度を測定した。

結果は表１、表２にある通りである。

表 1 WordNet の使用による精度の比較

素性の定義	F 値
WordNet なし	68.3%
単語の WordNet-synset	71.5%
単語の WordNet-synset の上位 synset	73.2%
単語の WordNet-synset の二つ上の synset	74.9%
単語の WordNet-synset の三つ上の synset	74.0%

表 2 述語に係る文節の助詞による素性の定義による精度の比較

素性の定義	F 値
各助詞が１素性に対応	74.9%
助詞の組み合わせが１素性に対応	74.0%
助詞の並びが１素性に対応	71.4%

4.2 評価

実験結果は表 1 のとおりである。まず分類器の精度は F 値で最大 74.9%であった。WordNet の synset の素性としての使用については、これを用いない場合で F 値が 68.3%であり、単語の synset そのものを用いた場合で 71.5%で、その一つ上位の synset を用いた場合だと 73.2%であり、さらにその上位の synset では 74.9%となった。さらにその上位の synset を用いると精度は 74.0%と下がってしまっていた。

これは、上位のものを用いれば用いるほど、特徴量の次元が減り、データのスパースネスが起こりにくくなる一方で、各単語の意味が抽象化されすぎ、分類がうまくいかなくなるという直感的な考えに合致する結果であった。

また、そもそも今回の実験では synset の特定に、単に synset の頻度情報を使っていた。これでは単語に対して間違った synset を特徴量として利用してしまう可能性が大いにある。これはいわば語義曖昧性解消の問題であるが、語義と意味役割の間には依存関係があり、どちらか一方を先に解決するというのは非常に困難である。そこで、Global Scoring において synset の特定も行うことで語義と意味役割を同時に同定することができないかと考えている。

また、助詞の属性値の取り方を変え比較した実験では、助詞の並びを属性値とした場合が最も悪いという結果になった。この原因は、まず助詞の並びが一番属性値の次元が大きくなってしまうことが上げられる。そのため、少ないコーパスではデータスパースネスにより精度が下がってしまったのであろう。また、もうひとつの理由として、日本語という言語の語順の曖昧性があると考える。日本語は語順に対する制約が緩く、そのため助詞がどの順番に係っていたかまで考慮することによる寄与が大きくなかったと推測される。

結果として F 値で 74.9%という結果であったが、マクロでみたときの F 値は 58%と低く、これは各ラベルごとにおいて精度の差が大きいということを意味している。この原因の一つは、コーパスにあると思われる。つまり、コーパス中に出現する回数にラベルによってかなり偏りがあるため、必然的にラベルによって制度が大きく異なってしまっていると思われる。この問題の解消のためには、さらにコーパスを拡張することが必要であると考える。また、コーパスの拡張において、頻度の低いラベルが現れるような事例を積極的に集める必要もあるかもしれない。

5 まとめ

本研究では、日本語テキストを CDL.nl へと変換するタスクのためのサブタスクとして、日本語テキストの中のエンティティ間の意味分類を CDL のラベルを基に行った。分類器の学習には独自に作成した小規模なコーパスを利用し、そのために起こるデータのスパースネスの解決のために概念辞書（日本語版 WordNet）を利用した。

結果として分類精度は F 値で 74.9%であった。また、特徴量を変更して行った比較から、確かに概念辞書を利用することが有効であることが示された。

今後の研究課題としては、述語項構造に限らないエンティティ間の関係の分類、Identification、Grobal Scoring についての問題の解決などがあると考えている。

参考文献

[1]T.Yokoi, H.Yasuhara, H.Uchida, et al. CDL (Concept Description Language): A Common Language for Semantic Computing. In WWW2005 Workshop on the Semantic Computing Initiative(SeC2005)

[2] GA Miller, WordNet: a lexical database for English, Communications of the ACM, 1995

[3] Y.Yan, Y.Matsuo, M.Ishizuka, et al Annotating an Extension Layer of Semantic Structure for Natural Language Text, the IEEE International Conference on Semantic Computing, 2008