3S-8

日本語テキストCDL化のためのエンティティ間関係の判別*

井口宜久† 石塚満† 内田裕士‡
東京大学大学院情報理工学系研究科† UNDL財団‡

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日本語テキストにおける文節間の関係を分類しCDL表現へと変換するシステムを作成しその性能の評価を行った．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．

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一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．

CDL への変換の問題の一つに、それが照応関係や省略表現の補完といった高次の意味解析も含むことが挙げられる.そこで本研究では各文節が基本的に 1Entity で、文節の係り受けに沿ってのみ Relation があると仮定して問題を簡略化している．

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 STEP1:関係の特定

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段階に分けて考えた.本研究では基本的に文節が Entity で係り受けが Relation と仮定しているため、関係の特定のためにはテキストを係り受け解析しさえすれば良い.ただし、一部の文節は Entity とならないため、それについては例外ルールを設け対処した．

3.2 STEP2:関係の分類

関係分類は、後述する特徴量を用いた分類器により分類を行った.分類器のパラメータ推定には最大エントロピー法を用いた．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．

分類器で用いる特徴量は次の通りである．

3.2.1 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の synset.用いる synset のレベルを変えて精度の比較を行った．

*Classification of Relationship between Entities for CDL labeling of Japanese text
Nobuhisa Inokuchi† Mituru Ishizuka† Hiroshi Uchida‡
†Gratuate school of Information Science and Technology, The University of Tokyo
‡UNDL Foundation

・付属語について

自立語に付属する助詞・句読点をすべて特徴量とした.

3.2.2 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た.

3.3 STEP3:リランキング

文全体として良い推定を行うためにはラベル間の依存関係を踏まえた全体としてのラベル列の推定が必要である.この推定には、HMM を用いた bi-gram モデルを採用した.つまり、ある述語に対して r_1～r_n までの関係があるとき、以下の式を最大化するラベル列 L_1～L_n を最終的な推定結果とした.

$$\prod_{i=1}^{n} P(Li|Li-1) \cdot \frac{P(Li|ri)}{P(Li)} \quad (1)$$

4 実験

4.1 実験概要

京都大学テキストコーパスから取得した 300 文に対して人手で CDL のラベルをつけたもの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, 特徴量として利用する synset のレベルを変えて比較した. また、Wikipedia から取得した 50 文をラベル付けし別ドメインのテストデータとして利用した.精度の測定は 5 : 1 の交差検定で行った.

4.2 評価

まず関係の特定・関係の分類・及びシステム全体の性能の評価を示したものが表 1 である.ここで KC・WC はそれぞれ京都大学コーパス・Wikipedia を基にしたテストデータである.関係の特定における精度はそれぞれ 91.3%と 90.5%で、一方関係分類・リランキングにおいては 77.3%と 71.8%であった.関係の特定における差が少ないことから、特定に用いた例外ルールは別ドメインでも適用可能と考えられるが、関係の分類では 5.5%の精度の差が出てしまっていた.

概念辞書の効果の検証結果を表 2 に示す.分類器の精度は概念辞書を用いない場合 68.3%ともっとも低く、対象単語の二つ上位の synset を用いた場合最大で 76.4%であった.また、さらに上位の synset を用いた場合は 75.1%と逆に精度が低くなっていた.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.

リランキングの効果の検証結果は表 3 の通りで、0.9%精度が向上していた.これは今回用いた bi-gram モデルの HMM が、日本語テキストの解析に不向きだったためではないかと考えられる. 日本語においては文節間の順序に対する制約が緩く、ご順の制約の厳しい英語などとは異なるモデルで解析を行う必要があるかもしれない.

4.3 考察

実験結果を細かく分析すると、解析エラーの原因は大きく二つあると考えられた.

一つは、出現頻度の低いラベルの存在である.コーパス中に出現する回数にはラベルによってかなり偏りがあるため, 十分に学習出来ていないラベルが存在し、それが制度を低くしていると考えられる.この対策としては, 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.

もうひとつは、WordNet の synset の同定ミスである.WordNet に登録されていない単語ではもちろん、登録されている単語でもその synset が複数ある単語の場合、正しい synset を選ぶことが難しく、結果として解析ミスにつながっていた.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できないかと考えている.

表 1 各テストデータでのシステムの評価

	KC	WC
Step1	91.3%	90.5%
Step2+3	77.3%	71.8%
Step1+2+3	70.6%	65.0%

表 2 WordNet の使用による関係分類精度の比較

素性の定義	F 値
WordNet なし	69.3%
単語の WordNet-synset	72.4%
単語の WordNet-synset の上位 synset	74.5%
単語の WordNet-synset の二つ上の synset	76.4%
単語の WordNet-synset の三つ上の synset	75.1%

表 3 リランキングの効果の検証

条件	F 値
リランキング無し	76.4%
リランキングあり	77.3%

5 まとめ

本研究では, 日本語テキストを文節の係り受けを基にした CDL 表現へと変換するシステムの開発を行った. エンティティ間関係の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. また文全体として良い構造となるように HMM を用いた bi-gram モデルでリランキングを行った.システム全体の精度は F 値で 70.6%であり、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, リランキングに用いるモデルの改善, 語義曖昧性問題の解決、コーパスの拡充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