1 はじめに

高度な自然言語処理を行うためには、形態素や構文などの表層的、情報だけではなく、深層格に代表される意味的な情報の利用が必要であると考えられる。しかしながら、現在のところ、そのような意味的、情報が誰にでも利用できる形で整備されているわけではない。意味の情報は人手で作成することは莫大な労力がかかり、さらに、作成者の主観による影響を大きく受ける。以上のような理由により、人手を介することなくコーパスから自動的に学習することが望ましい。

本稿では、帰納的学習を用いることで、単語の、内容語と機能語のタグ、依存関係という表層的、情報だけが付与されたコーパスから、意味の情報を自動的に獲得する手法を提案する。意味的、情報は、深層格、依存格を決定するための規則、下位分類要素、選択制限に関わる規則、上位下位概念に基づいた階層構造に相当するものである。

意味の情報を獲得する研究として、文献[1, 2, 3]などがある。彼らの手法は、事前に与えられた表現形式や共通語文によるパターンを用いて獲得する。このような手法は、用意されたパターンだけで全てを網羅できるのかという問題がある。また、彼らの手法は、上位下位概念に依存する名詞間の関係、動詞下位分類化された名詞間の関係を対象とするものである。本手法では、パターンを用いずヒューリストイックスを指標に、動詞、名詞、形容詞といった対象を限定せずに学習できる。

我々の目的として、特定の言語に依存した知識を使う学習するということがある。

表層的な情報である上にあげた知識は、以下に挙げる手法をタグなし文に用いて明示することが可能であるとと考えられる。我々は、文献[4]においてタグなし文から依存関係を学習する手法をすでに提案している。内容語と機能語のタグ付けにおいては、EDRコーパスを用いた我々の実験で、出現頻度を基に90%近い精度で分類できることを確認している。単語境界においても、中国語であるが文献[5]の91%という値から高い精度で示すことが可能であるとsofarされる。誤りの部分が及ぼすであろう影響については、今回実験では無視することにした。

2 基本的な考え方

2.1 意味表現

本手法では、文や語の意味をフレーム構造の一種で表現する。意味表現をフレーム構造で表現したときにスロット値となりうる語を内容語と呼び、そうでない語を機能語と定義する。また、1つの内容語と2つ以上の機能語で構成されるものをhamlet、複数のhamletでリスト化されたものをboroughと定義し、それぞれ文節と句の意味表現に相当する。特に、
入力文全体に対応するborough は country と呼ぶ。文節や句の依存関係は、hamlet の組 (pair of hamlet, PH) で表現する。全ての borough には、中心的な役割を果たす hamlet が存在し、これを head hamlet (HH) と呼ぶ。HH 以外の hamlet は complement hamlet (CH) となり、HH に対してある深層的な役割をもって存在する。この深層的な役割を duty と呼び、格文法で用いられる深層格も duty の一種とみなす。duty には、動詞と名詞間の関係だけでなく、名詞と名詞、名詞と形容詞などの関係も含まれる。

2.2 派生

規則を用いて解析するときに問題となるのは、規則を適用することのできない事例をどう処理するかということである。大規模なコーパスから単純に大量の規則を獲得したとしても、そのコーパスが現実世界に起こり得る全ての事例を含んでいるという保証がない限り、規則の完全性も保証できない。また、大量の規則を獲得することは、規則間の整合性の欠如と適用時の競合を引き起こす。これらの問題は、規則を枠内規則に収め適用できないことに原因があると考えられる。

我々は、整合性のとれた少数の基本的な規則を保持し、直接適用できない事例に対して、適用時に若干規則を変形させることで解決する方法をとした。事例に合わせて規則を変形させることを派生と呼ぶ。本論文での派生は、規則を構成する要素の挿入、削除、倒置の三つである。要素の置換については、削除と挿入を行うことで可能である。

派生の効果は適用時にのみ及ぶものではない。学習時において、派生により生成されたと思われる規則を原形に戻すことにより、規則の一般化を行い学習を促進させることができる。これは、データベーススベース問題の解決にも繋がる。それゆえ、事例をそのまま保持し適用時にのみ類推して解決しようとする用例ベースのアプローチとは、一線を画して

いる。

3 処理過程

3.1 全体の流れ

本システムは、下に示す 5 つの処理によって構成されている。

(1) PH の作成。
(2) HH の決定。
(3) borough の作成。
(4) 階層構造の作成。
(5) 規則の一般化。

本手法では、各処理がインタラクティブに学習していく。各処理順序は一定ではなく、条件を満たしたものを処理を行う。

3.2 PH の作成

依存関係が明示された文から、PH を抽出する。各 PH は一つの duty を持つ。PH は、以下の三種類に分類される。

| PH1 | PH の両方あるいは一方の hamlet が不明である。
| PH2 | PH の hamlet が両方とも明確であるが、どちらが HH になるか示されていない。
| PH3 | PH の hamlet が両方とも明確であり、HH になる hamlet が示されている。

例えば、依存関係を括弧を用いて表現した (((太郎は)((本を)(読む))) という入力から、[[本を][読む]という PH2 と [[太郎は][@1]] という PH1 が作成される。変数@1 には、PH2 の HH が対応する。PH2 の HH が決定次第、PH2 が PH3 に、PH1 は PH2 に再分類されることになる。このように作成された PH3 を用いることで、文中の依存関係を意味表現に変換することが可能になる。

— 473 —
3.3 HHの決定

HHは、ヒューリスティクスと出現頻度を基に決定される。使用するヒューリスティクスは、以下のように導き出した。例えば、以下に示すA、B、Cのhamletに関するPHが存在したとする。

(1) [[A][B]]
(2) [[@2][C]]
(3) [[A][C]]

(2) の@2には(1) のHHが入るとする。この場合、(1) においてAをHHと仮定すれば、(2)は(3) と同一になる。しかしながら、BをHHと仮定した場合、(3) とは別に[[B][C]]を作成しなくてはならない。規則性を増加させるという観点から、AをHHとした方が良いことが導き出される。逆のhamletをHHとするヒューリスティクスも同様に導き出される。

ヒューリスティクスだけで決定できない場合、階層構造を用いてPH中の要素を抽象化していくことでHHを決定する。

3.4 boroughの作成

PH3だけでdutyを決定することは出来ない。例えば、「太郎は次郎が運んだ」という文脈では、「太郎は」は動作主ではなく対象とみなすべきである。これは「運んだ」に対する「太郎は」と「次郎が」のdutyが競合していることに由来すると考えられる。従って、各HHごとに取り得るdutyの組をboroughを基に学習する。boroughはcountyを分割することにより求める。boroughの条件として、1つのHHを持ち、複数のCHが同一のdutyに対応することは無い。従って、1つのboroughの中に複数のHHが存在する場合は、boroughの分割を行う。このときCHをどちらのHHに割り振るかという問題が起きる。依存関係から直接下位範疇化されているCHが分かる場合は、一意に決定する。そうでない場合、HHと共起する回数により決定する。

3.5 階層構造の作成

内容語、機能語、dutyのそれぞれについて以下のクラスタリングを行い、階層構造を作成する。本手法の階層構造は木構造を仮定している。

3.5.1 クラスタリング

本手法では、文献[4]と同様のアプローチでクラスタリングを行う。まず、学習する事例に対して独自にシンボルを割り当て、その後、「周囲の環境が類似したシンボルは類似している」という指標に基づいて階層化していく。このとき問題となるのは、分類するクラスタの個数についてである。我々は、教師なし学習であることと、理論的な裏付けが無いということから、事前にクラスタの個数を決定することを避ける。そこで、シンボルを一つの木構造になるまで階層的にクラスタリングすることで、この問題を解決する。階層的な木構造が求められていれば、各要素の密接度(close)を計算するのは難しいことではなく、必要に応じてクラスタの個数を変更することができる。

本手法の階層構造は、3.5.4に後述するシンポル間の類似度(similar)から求められる。最初にシンポル一つだけからなるクラスタの初期集合を求める。初期集合のクラスタ間の類似度はシンポル間の類似度と同じである。それから、最も類似度の高い組から、その二つ目のクラスタを包含するクラスタを作成し、ボトムアップに組み上げていく。本手法は二つ木であることを仮定していないので、作成されたクラスタに三つ以上の要素が含まれる可能性がある。そこで、クラスタを作成したときの類似度を基準に、クラスタ内の要素との類似度の差が閾値以上のクラスタも、そのクラスタに包含させる。新しく作成されたクラスタとの類似度は、そのクラスタに含まれる各要素との類似度を平均して求める。以上の作業を一つのクラスタになるまで繰返す。

次に、作成された階層構造からクラスタの
3.6 規則の一般化

作成されたPH3は、依存関係を意味表現に変換する最も具体的な規則である。しかしながら、具体的であるために他の事例への適用という点で問題が残る。2.2節で述べた派生の逆手順を行い、個々の規則の一般化を行う。

4 今後の予定

本手法を実装したシステムを作成し、実験を行うことで本手法の有効性を確かめる予定である。

参考文献


