

5R-1

Web テキスト文からのルール知識の抽出

小林 大祐[†] 谷口 智哉[‡] 石塚 満[‡]

東京大学工学部電子情報工学科[†]

東京大学情報理工学系研究科[‡]

1. はじめに

近年、インターネットの普及によって、Web 上において莫大な量の知識を得られるようになってきた。その知識は自然言語で書かれているものがほとんどである。しかし現在、コンピュータにおいて自然言語文をそのまま扱うのは、自然言語の意味理解の難しさもあり、難しい。

また、その知識は、個人が書いたものから組織が記したもので、雑多な自然言語文章が存在している。これらをひとまとめに処理しようとすると大変である。

そこで、コンピュータが扱いやすいように統一された知識表現を導入することが重要であると考えた。そのために考案されたのが KRNL (Knowledge Representation for Natural Language) [1][2]である。KRNL は 1 階述語論理をベースとしていて、表現に一定の緩い制約を課すことで述語論理の表現を一意に定めることができ、また自然言語文で表現された知識の利用を効率的に行うことができる。

このような特徴を持つ KRNL であるが、自然言語文を用いて効率的な推論を行うためには、IF-THEN のルール知識を抽出することが重要である。

そのため、この研究では、自然言語からルールの知識の抽出を行うことを目的とした。今回はそのための実験と、その実験の結果を報告する。

2. KRNL

従来、知識を組み合わせるための表現として、述語論理や、命題論理などの論理表現が研究されてきた。しかし、自然言語文を論理に変換する場合、表現が一定に定まらないという問題がある。

例えば、「この山は高い」を述語で表現しようとする場合、下の例のように、述語や引数を

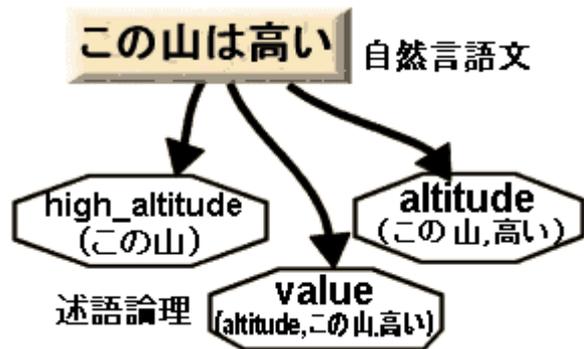


図 1

どう設定するかによって、表現に任意性が生じる。

high_altitude(この山)

altitude(この山, 高い)

value(altitude, この山, 高い)

同じ知識でも複数の表現がされていると、知識の組み合わせによる推論が困難になる。知識が複数ユーザによって入力されることを考えると、この表現の任意性の問題は大きい。

このように異なる表現をとることが起こりうるのは、ひとつの現象についても、それを捉えるさまざまな視点があるからである。しかし、このような複数の表現がある場合、例えば、あるユーザから 1 番目の知識が入力されたときに、他の人が 2 番目の知識を得ようと思っても、それにマッチしないという事になってしまう。

そこで、このようにならないように表現を制限した知識表現を介せば、この問題をある程度解消することができると思われる。

そこで、KRNL では、表現にゆるい制約を加え、さらに自然言語に内在する規則をあらかじめ与えておくことで、自然言語文で表現された知識の利用を簡単かつ効率的に行えるようにした。

また、KRNL には world という概念が考えられている。これは、各人の持っている知識を、とりあえずその各人の world の知識だとして、各 world 内は矛盾がないようにする。

こうすることで、Web 上のあらゆるところから知識をとってこようとしても、立場の異なる人による知識で矛盾が生じてしまう、ということ

Extraction of Rule knowledge from Document on Web

[†] Dept. of Information and Communication Engineering,
The University of Tokyo

[‡] Graduate School of Information Science and Technology,
The University of Tokyo

を防ぐことができる。

3. 提案手法

Web テキストからのルール知識の抽出ということで、Web 上のあらゆる場所から知識を取得できることが望ましいが、いろいろな場所から取ってくると知識の矛盾が起こることも考えられる。よって、今回は、原子力の百科辞典と言うべき、原子力図書館 ATOMICA (<http://mext-atm.jst.go.jp/atomica/>) に対象を絞って解析を行うこととした。

そして、実際のルールの抽出にはサポートベクタマシン(Support Vector Machine, 以下 SVM)で学習する、という手法をとることとした。

3.1 サポートベクタマシン

今回、IF-THEN のルールを抽出する方法として SVM を用いてモデルの学習を行った。

SVM は高い汎化性能を持ち、また少ない学習数でも効率よくモデルを学習することができる。

3.2 モデル学習に利用する統計量

SVM にモデルを学習させるためには、入力となる文章から素性ベクトルを生成する必要がある。

本研究では、文章 D を Chasen[3]で形態素解析し、その結果を N-gram (N = 10) で抽出して、その tfidf 値を素性ベクトルの各要素の値として採用した。

$$\text{tfidf}(t, D) = \text{tf}(t, D) \times \text{idf}(t)$$

$\text{tf}(t, D)$ = t が文章 D に出現した回数

$\text{idf}(t) = \log(\text{総文章数} / t \text{ が出現した文章数})$

ただし、 tf (出現回数)が 3 以上となる N-gram のみ採用した。さらに、各素性ベクトルは大きさが 1 となるように正規化されている。

3.3 正例と反例の判別

IF-THEN のルール構造をもつ文であるかの判断基準は、個々の判断に任せると、あいまいになりやすい。そこで、今回は RuleML による定義を用いて判別することとした。

RuleML とは、Rule Markup Language の略であり、メタ言語である。Semantic Web と非常に近い関係にある。

RuleML で記述される「ルール構造を持つ文」とは、

1. 必ずしなければならない条件が書いてある
ex) たばこは 20 歳になったら吸うことができる
2. 強制ほどではないが、前提条件である
ex) 傘があまっているならば、断りなしに持って行ってよい
3. 何かの動作が起こることが、次の動作を引

き起こす

ex) スイッチを入れると、電気がつく

4. 前提条件(IF)を変化させて得られるもの

ex) 一冊の本があれば、それを作者と本のタイトルにまとめられる

5. プロダクションルール

主にこのような文章のことである[4]。今回はこの判断基準を用いることとする。

4. 実験

原子力図書館 ATOMICA に掲載されている文章から 2000 文を対象とした。

この中から、正例と反例からそれぞれ 50 文、150 文用意し、それを学習データとした。

そして、残りの 1800 文をテストデータとして、適合率、再現率、F 値を求めた。

まず、学習データの数をさまざまに変えて、学習データの個数が妥当であるかどうかを調査した。

また、N-gram として取得する単語を、助詞・助動詞・動詞のみ、それに副詞・接続詞を加えたもの、さらにそれに名詞を加えたもの、で比較を行った。

なお、この比較手法として、パターンマッチングで IF-THEN のルール構造を含むと思われるものを分類するスクリプトをつくり、比較した。

5. おわりに

本研究では、Web 上からの文章を取得してきてそれを解析し、その中から IF-THEN ルール構造を持つ文を抽出することを試みた。今後はそれを実際に利用して Web 上の知識の取得、解析を行う。

参考文献

- [1] 友部 博教, 谷口 智哉, 松尾 豊, 石塚 満: “知識表現 KRNL による知識共有システムの構築”, 第 18 回人工知能学会全国大会 (2004)
- [2] 谷口 智哉, 友部 博教, 松尾 豊, 石塚 満: “自然言語と論理をつなぐ知識表現 KRNL の UNL による拡張”, 第 18 回人工知能学会全国大会 (2004)
- [3] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原 正幸: 日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書 (2000)
- [4] Gerd, W., Said, T., and Harold, B.: MOF-RuleML: The Abstract Syntax of RuleML as a MOF Model (2003)