人工知能学会研究会資料
SIG-SW&ONT-A302-01
2003-11

自然言語処理と蓋然的なセマンティックウェブ

Natural Language Processing and the Probabilistic Semantic Web

荒川直哉 (Arakawa, Naoya) [1]

In this paper, I shall argue that “knowledge” generated by NLP should be accompanied with meta-information including that of probability if it is to be published on the Web. Moreover, I shall discuss kinds of “probabilistic knowledge” generated by NLP to serve for intelligent information processing on the Web.

1. はじめに

この論文では、自然言語処理により生成され、ウェブ上の知的な情報処理に役立つような知識について考察する。以下での主な論点は、「ウェブなどの情報源から自然言語処理により生成される『知識』は蓋然的なものであって、公開利用される場合には確実度などのメタ情報を伴うべきである」ということである。また、生成される蓋然的な「知識」について、獲得方法、表現方法および利用に関する個別的な議論も行う。自動生成された「知識」は多くの場合、情報処理システムが主な利用者になる意味論的な情報だと考えられ、Semantic Web (the) [1]が提供しようとしている情報と重なる部分が多い。このため、Semantic Web についても参照を行いながら議論を進める。

近来、自然言語処理の世界ではコーパスに基づく統計的な手法が開発されてきたが、こうした手法が生み出す情報はもっぱら蓋然的なものである（100%の確実度を持たない）[2][3]。一方、人手で作成された Semantic Web の情報は（作成者が信頼されるならば）通常蓋然的なものとして扱われない。

Semantic Web と自然言語処理の境界では、「知識」をウェブページ上のテキストや表などから自動生成する研究が行われている。ここで問題になるのは自動的に生成された「知識表現」の信頼性である。生成された「知識」が正しい確率は、抽出元となったテキストなどの性質や抽出方法によって大きく異なる。

自動生成した情報が（Semantic Web 上で）公開されるならば、利用者は情報の信頼性を知りたいと思うであろう。とすると、自動生成された情報を公開する際には、メタ情報として信頼性やデータの抽出方法、データソースの情報を付け加えてはじめて情報提供サービスとして完結するといえる。

こうしたメタ情報は、利用者が比較的確実と思われる情報源を選ぶ、というよう

[1] フリー：naoya. arakawa@nifty. com

な目的での利用もできる。しかし、確率（統計）情報のより重要な利用方法は、蓋然的かつ合理的な意志決定理論（例えばゲーム理論）を用いるような情報処理への応用である。そこでは複数の情報源（命題）に関する統計情報からベストと思われる決定を行うことが問題になる。

ここで議論している信頼性（確率）は Semantic Web で扱う Web of Trust の Trust とは異なる。Trust は情報発信者への信頼であり（レベルはあるかもしれないが）確率を表すわけではない。自然言語処理でも情報発信者の信頼度は役に立つが、それは確率的なものとしてとらえられる場合である。例えばウェブ上のテキストから知識を抽出するような自然言語処理プログラムは、情報の抽出精度に情報源の信頼度（正しいことを言う確率）を掛け合わせて全体の信頼度を計算することができる。

情報源の（確率的な）信頼度をどうやって獲得するかは今後の研究を待つことになるであろうが、ウェブ上の「メディアリテラシー」の問題としても興味深い[4]。例えば、他の多くのメディアの情報と一致する場合は、より信頼度が高いとしてよいかもしれないし、今までの情報の正確性や、テキスト内での一貫性も信頼度の判断基準になるだろう。[2]

2. 自然言語処理で生成・利用する蓋然的な知識

次に、自然言語処理システムが生成・利用する情報で、ウェブ上で共有できるものにどんなものがあるかをより具体的に見ていくことにしよう。こうした情報には純粋に言語学的（例えば統語論的）なものから、概念（オントロジ）的なもの、世界で起きている事柄に関する知識、事柄に関する一般規則などがある。

一般的に、ウェブ上（など）のテキストから抽出された情報には、情報源に関する情報や、取得日時、抽出された情報の信頼度といったメタ知識を付与することができる。抽出された情報の信頼度は、情報源の信頼度や抽出技術の信頼度などから計算される。また、検索のために、知識表現の部分構造にはインデックスを作成しておくことができる。

2.1. 言語学的な知識

自然言語処理システムがウェブ上（など）のテキストから収集できる言語学的な知識には、例えば、単語の出現頻度や共起頻度、係り受け（あるいは文法規則）の統計などがある。これらは言語の部分的な構造の統計であるといえる。

統計ではないが、こうした部分構造をインデックスとして情報源を検索することが考えられる。通常の全文検索は、部分構造が単語あるいは文字連鎖である場合に対応する。係り受け解析結果をインデックスとするような研究も行われている[6]。

---

[2] このトピックはより一般的には知識の哲学（認識論）で扱われる [5]。

2.2. 事例に関する知識

個物に関する知識

自然言語処理システムは、ウェブ上のテキストやユーザーの教示などから個物（個人）に関する情報を収集することができる。ここでポイントとなるのは、複数の個物（個人）が同一のものかどうかを判断し、同一と判断した場合は単一化する技術である（名前が同一だからといって常に単一化できるわけではない)。

エピソードに関する知識

エピソードも事例の一種（状況の事例）である。自然言語処理システムは、ウェブ上などの記事からエピソードを表すテキストを抽出し、意味解釈を行って保持する。個物の場合と同様、複数のエピソードが同一のものかどうかを判断し、同一と判断した場合は単一化する。

エピソードの各構成要素（個物）について、所属するクラスについて抽象化エピソードも知識として生成することができる（例えば、ポチが犬の場合、「ポチが走った」という事例から「犬が走った」や「動物が走った」という抽象化事例を作ることができる)。抽象化エピソードには使用したオントロジを注記する。複数のオントロジから得られた抽象化エピソードはよりよい（信頼性の高い）抽象化エピソードであるといえる。

一方、自然言語処理システムは、数多くのテキストからエピソードを抽出するのではなく、個々のテキスト中にあらわれるエピソードの解析結果を別々に公開することもできる。その場合、解析結果は機械が理解できる個別コンテンツのサマリーのようなものととらえることができる。

2.3. オントロジ

自然言語処理システムは（著作権をクリアできれば）シソーラスや辞書からクラスや属性に関する知識を抽出し、一種のオントロジを生成することができる。この際、最も注意すべき点は、単語が多義性を持ちうる（polysemy）ということである。すなわち、単語（の見出し）とクラスの関係は多対多の関係になる。また、複数のシソーラスの分類は通常一致しないなどの事情にも留意する必要がある。

クラスに関しては、クラス階層やインスタンスが持つ属性に関する知識が重要であり、属性や関係については引数の値の範囲や、他の属性、関係との関係が記述される。クラスについても、事例と同様、複数の情報源から得られた複数のクラス表現が同一のクラスを表すと推定される場合は、マージするか同一視を宣言する。逆に、あるクラス表現に関して複数の情報源から得られた情報が矛盾している場合、複数のクラスを表現していると推定して、分割することを考える。後者は１つの単語が複数の語義を持つことを発見することと類似のプロセスである。

自然言語処理システムは、オントロジを作成するための情報源としてウェブ上のテキストやユーザーからの教示も用いることができる。例えば、語の定義や、あるクラスによくあらわれる属性（車と色など）をテキストから抽出する。通常、クラスは一般名詞に対応し、属性や関係は形容詞や動詞あるいはそれらの名詞形に対応する。

2.4. 確率データベース

事例はさまざまな要素間の関係として表現される。ここで、事例中でなりたつ要素や関係の間の条件付き確率（例：「ライオン」と「シマウマ」が言及された時に「襲う」という関係が言及される条件付き確率）を集計したデータベースを考える。このようなデータベースを使うと、ある要素の組み合わせが特定の関係を持つ確率を求めたり、ある要素が特定の関係を持つような別の要素を推測（連想）したりすることができる。（実際のデータベースは、母集団の大きさなどの統計情報を利用するために、確率値を保持するより事例の生起回数を保持するほうがよい。）

自然言語処理システムがテキストから抽出する関係には、例えば、1) テキスト中の動詞などにより表される関係、2) 接続詞により表される文の間の関係、あるいは 3) エピソードが生起する時間や場所の関係、といったものが挙げられる。また、シソーラスやオントロジを参照すると、字面レベルより抽象的な概念（単語）間の関係を抽出することも可能になる。これらの関係は、もちろん上記の事例に関する知識の構成要素としても利用できる。

確率データベースは、自然言語処理システムを含むさまざまな知的システムが、判断や推定を行う際に参照し、より尤もらしい解を得るために使用することができる。確率データベースの統計情報は、コーパスとしてのウェブの情報を集約して提供したものといえる。

3. 知識ベース表現について

この節では、自然言語処理システムが生成する蓋然的な知識をどのように表現するべきかについて考察する。人間が知識表現を作成する場合に比べ、自然言語処理システムは、テキスト中の用語の指示対象について不確かであることが多い。Semantic Web 的な言い方では、自然言語処理システムはテキスト中の表現を（universal に同定される）リソースに一意に結び付けることが困難である、ということになる。自然言語処理システムはテキストを分析して、何らかのものごとの間の関係を示す「意味表現」を出力することができるが、その何らかのものごとが何を指すかについては確かなことはいえないのである。

人工言語（知識表現）では、何を指すか特定できないものを表現するのにしばしば変数を用いる。自然言語処理システムが生成する知識表現も、変数を使って指示

対象を表すことができる。変数の使用は、テキスト上の表現や使用するオントロジ上の表現と実体に関する議論を分離するのに役立つ。例えば、自然言語文を（変数 x を含む）述語論理の式に変換すれば、x の指示対象の多義性と文自体の曖昧さを分離して議論できる。変数を使わない場合、指示対象と構文のそれぞれの多義性について、複数の可能性を列挙しなければならないようなことになる。これは経済的でない上、理解もしづらい。また、変数を使った表現は、述語論理や SDRT [7]などの理論とも相性が良い。

自然言語システムの意味表現中では、各変数にはテキスト中の表現と同じものを指す、という関係を持たせることができる。例えば変数 x は「犬」という単語で指示されるものを指す、という関係を expr(x, “犬”) で表すことができる。

例文を使ってさらに具体的に説明すると次のようになる。「太郎が饅頭を食べる」という文を、expr(o1, “太郎”), expr(o2, “饅頭”), expr(r1, “食べる”), expr(a1, “が”), expr(a2, “を”), arg(r1, a1), argobj(a1, o1), arg(r1, a2), argobj(a2, o2)と表すことにする。ここで、o1, o2 はオブジェクト、a1, a2 は引数、r1 は関係を表す変数とする。expr(x,y) は x が言語表現 y で表されることを意味し、arg(x,y) は y が関係 x の引数であることを示し、argobj(x,y)は引数 x のオブジェクトを表す変数が y であることを表している。（これらの関係をグラフとして図 1 に示す。）ここでは、動詞が表す関係や述語の引数関係も変数として表現されている。引数関係を変数で表しているのは、例えば助詞「は」が表現する引数の役割（role：Agent, Object など）を自然言語処理システムが常に正確に判断できないことに対処するためである。

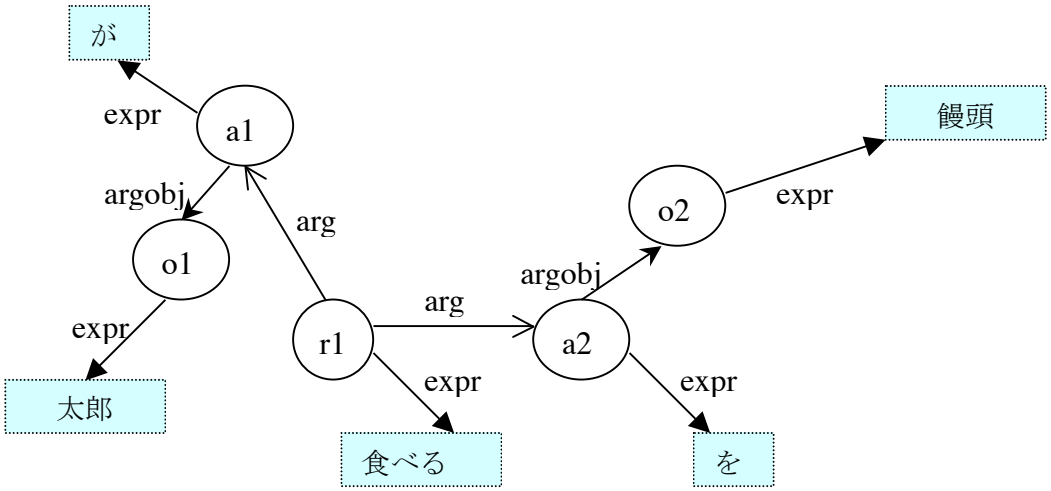

図 1

意味解析結果として、図 1 にいくつかの要素を追加することができる。

1) 引数の役割（Agent, Object など）を明示することができる。述語の引数の役割を明示しているような他のオントロジや意味表現と相互運用を行うためには、引数の

役割を明示しておいたほうがよい。ただし、役割は蓋然的にしか決定できないことも多い。引数の役割を表現するためには例えば argrole(x, y, p) のような関係を用いる。ここで、x は引数（上の図で例えば a1）、y は役割名（Agent など）、p は確率値である。１つの引数は複数の役割と蓋然的な argrole 関係を持つかもしれない。

2) 大域的なリソースへの参照。固有名詞やクラスを表す変数に関しては、例えば同じものを指示すると思われる Semantic Web 上のリソースを１つ以上参照することができる。

3) 構文的多義性。上の例文は多義性を持たないが、例えば動詞が複数ある場合など、選考する文節がどの動詞に係るかなどの多義性が生じる。係り受けの多義性は上の例では例えば argobj(x,y)に確率の引数 p を足して argobj(x,y,p)とすることで表現できる。多義性により x（引数）と y（オブジェクト）の関係は１対１ではなくなる。[3]

4) 時間の表記。実際の文章には過去や未来などの時制や進行形などの相（aspect）が存在している。時制や相は文や節が表す事態と状況との時間的な関係を表しているので、状況の表現を明示的にして参照したほうがよい。時間の他にも、条件文や様相的（modal）な表現についても状況を明示したほうがよいが、この点については次節で述べる。

なお、RDF を使って変数を表現する場合、変数はローカルに定義したリソース（id）ととらえることができる。変数をリソースと考えると、上の例のようなグラフは容易に RDF で表現できる（確率の表現方法はいくつか考えられる）。

4. 状況とエピソードについて

自然言語システムはテキストを分析して情報を抽出するが、テキスト中の文は状況を表すといえる。前節で述べたように、ある文は時制によってテキストが書かれた状況と表現する状況との関係を表したり、接続詞や引用、様相を表す表現などによってテキスト中の別の文が表現する状況と関係を持ったりする。[4]

ウェブ上のテキストから情報（命題）を抽出する場合、その命題がいかなる状況で生じたのかを明らかにすることが重要である。例えば「田中さんは鈴木さんがパソコンを持っていることを信じなかった。」という文から「鈴木さんはパソコンを持っている」という命題を取り出して主張することはできない。テキスト中には、一般的に状況が入れ子になって表現されており、抽出された命題はそうした入れ子の命題に埋め込まれた形で提示されなければならない。（埋め込みが複雑になると自動処理の正確さも信頼できなくなる。入れ子はそれぞれの処理の信頼度を持つべきであるし、あまり信頼度のない情報は公開しないほうがよいだろう。）

---

[3] 日本語に類似した言語でなくても係り受け（依存）という概念はなりたつ。[9][10]

[4] こうした考察では「メンタルスペース理論」も参考になる。[11]

状況を形式的に表現する１つの方法は、状況を実体としてやはり変数として表すことである（状況理論 [8]、SDRT [7]）。ある命題がある状況に含まれていることを示すには、前節で述べたような述語形式では contains(s, o)（s は状況、o は状況や関係を含む任意のオブジェクトを表現する変数）などと表せばよい。XML のような言語を使うのであれば、要素間の包含関係を使って暗黙的に包含関係を表すこともできる。

5. 応用

5.1. 会話システム

今まで説明してきたような知識ベースを使って会話を行うような会話システム（Chat Engine）を作ることができる。会話システムは、次のような目的で知識ベースを使用することができる。

1) 入力文の解析：構文上の統計的情報や統計（確率）情報付きのオントロジは、入力の構造を解析し、多義性を解決するのに役立つ。

2) 応答内容としての利用：会話システムは、インデックスと統計を使って、入力文に最も関連する話題（コンテンツ）を検索し、見つかった話題に関する情報をその信頼度情報に応じた表現（「～のようです」など）で提示する。

3) 学習（知識ベースへの貢献）：会話システムは、ユーザーから知識を得て知識ベースに追加したり、知識ベースから得た情報の間違いを直してもらったりすることができる。（もちろんユーザーが常に正しいわけではない。この点、ユーザーはウェブ上のテキストと同様、信頼度の付与対象となる。）なお、ユーザーから得た知識に関しては、プライバシーに留意して利用する必要がある。

5.2. 情報検索

知識ベースには情報ソースの情報が埋め込んであるので、必要に応じてソースを参照することができる。検索にあたって、複雑な論理式などを使うことができるかどうかは、クエリインタフェイスによる。上記のような会話システムもクエリインタフェイスとして使用できる。

5.3. 多義性解決

自然言語処理で常に問題になるのが多義性（語彙レベル、統語レベル、論理式のスコープなど）である。上記の知識ベースは多義性解決に役に立つ意味上の制約あるいは選好を与える。知識ベース自体の構築にも自然言語処理を用いているので、知識ベースの増大および高精度化は、知識ベース自体の高精度化をもたらす（これは一種のブートストラッピングである）。

5.4. The Probabilistic Semantic Web

Semantic Web ではデータ作成がボトルネックになる。ここで説明してきた知識ベースが Web 上のテキストを入力として自動的に生成され、公開されれば一種の Semantic Web として機能する。従来の Semantic Web と異なる点は、確率情報と信頼性というメタ情報を付与していることである。このことは、会話システムのような自然言語応用にとって重要な意味を持つ。まず、自然言語処理システムは確率情報なしでは（意思決定理論的な意味で）合理的な決定を行うことができない。また、会話システムは信頼性情報なしでは適切な情報の提示を行うことができない（例えば不確実な情報を真であるとして主張するのは適切ではない）。最後の点は次のように換言することができる。知識をテキストから自動生成するのであれば 100%の精度は期待できないのであるから、生成されたものには確率と信頼度の情報を付加するのが誠実というものである。

6. References

[1] http://www.w3.org/

[2] 自然言語処理 (岩波講座ソフトウェア科 15). 長尾真 (編), 岩波書店 (1996).

[3] 確率的言語モデル (言語と計算 4). 北研二, 東京大学出版会 (1999).

[4] Web の信頼度に関しては、例えば次の Stanford 大学のサイトを参照されたい：
http://www.webcredibility.org/

[5] An Introduction to Contemporary Epistemology. Dancy, J., Blackwell (1985).

[6] “大規模テキストベースに基づく自動質問応答－ダイアログナビ－” 黒橋 et al., 自然言語処理 Vol. 10, No. 4 pp. 145-175 (2003).

[7] *The Logic of Conversation*. Asher, N. et al., Cambridge University Press (2003).

[8] *Logic and Information*. Devlin, K., Cambridge University Press (1991).

[9] Word Grammar: http://www.phon.ucl.ac.uk/home/dick/wg.htm

[10] The GDA Tag Set: http://www.i-content.org/GDA/tagset.html

[11] メンタルスペース. Fauconnier, G., 白水社 (1996).