オントロジー学習

Ontology Learning

人手によるオントロジーの構築には多大なコストがかかるため、様々なリソースからオントロジーを半自動的に構築するための技術が研究されている。学習のためのリソースとしては、自然言語で記述されたテキスト、機械可読辞書、XML や RDF などで記述された半構造化データなどがあげられる。なかでも自然言語テキストは、WWW の普及を背景にして、現在のところ最も大きな知識源といえる。

自然言語テキストからのオントロジー学習のために広く用いられている手法としては、Hearst による語彙構文パターンを使った手法がある [1]。これは、ある特定の言い回しが概念間の関係を表現していることを利用する。例えば、

Bruises, wounds, broken bones or other injuries ...

というような文があれば、bruise が何かを知らなくてもそれが injury の一種、すなわち bruise が injury に対して is-a 関係にあることを推測することができる。したがって、特定の概念関係を表現するような語彙構文パターンを用意すれば、テキスト中からそのパターンにそれにマッチする部分を抜き出すことで概念関係を抽出することができる。

Is-a 関係を抽出するためのパターンとしては、次のようなものがある（NP は名詞句を表す）。

1. NP {,NP}* {,} or other NP
2. such NP as {,NP}* {*or*|*and*} NP

Part-of 関係についても、NP's NP のようなパターンと統計情報を組み合わせることにより、ある程度の精度で抽出できるとの報告がある。

オントロジーを学習する上では、概念を数値化して表現できると便利である。テキストを利用すると、概念の意味を、その概念がテキスト中に出現した文脈を利用して表現することができる。表現方法としては例えば、その用語と係り受け関係にある動詞との共起情報や、同一名詞句中に共起する名詞との共起情報を、ベクトルの形で表現する。このように概念の意味をベクトルとして表現すると、概念間の類似度をベクトル間の類似度として計算することが可能になり、オントロジーを自動構築する際に必要な様々な処理が行える。例えば、既存のオントロジー中に、新しい概念を追加しようとする場合、その概念をどこに追加するべきなのかを、既存のオントロジー中に存在する概念との類似度を比較することによって決めることができる。また、ベクトルをクラスタリングすることにより、類似した概念をまとめることも可能になる [4]。

テキストからオントロジーを学習するための要素技術としては、他にも、テキストからの自動用語抽出、語義曖昧性解消などの処理などが必要とされる。特定分野に関するオントロジーの場合は、その概念を表現する言葉に多義性があることは比較的少ないが、WordNet のように一般的な単語によって表される概念を対象とする場合、語義の曖昧性は重要な問題になる。

これまで述べてきた手法では、基本的に、用語がテキスト中に現れる文脈を利用して、その概念や概念関係を抽出しようとする手法である。それに対して、用語（特に複数の一般的な語からなる複合語）の概念を、それを構成している個々の単語の意味を用いて構成的に規定しようとする試みもある [3]。例えば、transport service という用語の概念を規定したいとする。その場合、構成要素の単語、すなわち transport と service にはそれぞれ多義性が存在する。そこでまず、それぞれの候補の語義を WordNet と意味タグ付きコーパスでの共起情報などを利用して、意味ネットワークとして表現する。そうすると、意味ネットワーク間の整合性を評価することによって、それぞれの単語に関してどの語義が適切なのかを決定することができる。さらに、transport と company が purpose 関係にあることを C4.5(機械学習の一手法）を利用して決定することで、この複合語の概念を規定することができる。このようにして、いったん複合語の概念を WordNet 中の概念の組み合わせで表現することができれば、複合語どうしの Is-a 関係などもルールベースで比較的簡単に決定することができる。

自動学習されたオントロジーにはノイズが多く含まれている。そのため、実際に利用可能なオントロジーを構築するためには、人間の介入が不可欠である。すなわち、学習のためのリソースを選択、

自動学習によってオントロジーを拡張、得られた結果を人間が修正、というサイクルを繰り返す必要がある。したがって、オントロジー構築に必要なのは、概念や概念関係を抽出するための個々の要素技術だけではなく、作業者とのインターフェースも含めたシステムを考える必要がある。近年では、そのようなサイクルを実現するための環境を提供するオントロジー構築ツールがいくつか提案されている[2]。最近は特に Semantic Web における必要性からオントロジー構築に対する需要が高まっており、OntoWeb ポータル (www.ontoweb.org) では、オントロジー構築手法や構築ツールに関する豊富な情報が公開されている。

参考文献

[1] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 539–545, 1992.

[2] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[3] Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31, 2003.

[4] Fernando C. N. Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

執筆者：鶴岡慶雅