橋本　力　（黒橋教授）

「Knowledge Acquisition from the Web for Text Understanding」（テキスト理解のための Web からの知識獲得）

平成 23 年 9 月 26 日授与

人間と言葉で通じ合うことのできる機械を作ることは人類の夢の一つであると言える。計算機科学の黎明期から取り組まれてきた、計算機によるテキスト理解の研究は、まさにその夢の実現を目指している。我々は当面のゴールとして、あるテキストの対が与えられた時、一方のテキストがもう一方のテキストを意味的に含意するかどうかを判定するシステムの実現を目指している。例えば図 1 では、1 行目の文が 2 行目と 4 行目の文を含意するが、3 行目の文は含意しない。

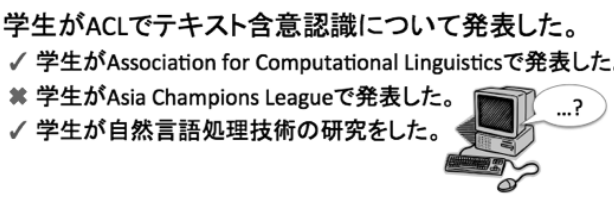

図 1　テキスト理解の例

計算機によるテキスト理解を実現する上での難問の一つに知識獲得ボトルネックと呼ばれる問題がある。これは、人間がテキストを理解する上で無意識のうちに使いこなしている膨大な量の知識を、現実的なコストで、計算機に教え込むにはどうすればよいか、という問題である。テキスト理解研究の初期の段階では、テキスト理解に必要な知識を人間の手で一つ一つ計算機に入力していたが、これではコストがかかりすぎることが分かり、近年では、大量の文書集合から（半）自動で知識を獲得する研究が盛んになった。

本研究では、テキスト理解にとって重要でありながら獲得技術が発展途上にある 3 種類の知識を対象に知識獲得技術を開発した。1 つは我々がドメイン知識と呼ぶ、単語とそれが属するドメインに関する知識である。例えば「テキスト理解研究」と「Association for Computational Linguistics」、「自然言語処理」は＜科学・技術＞ドメインだが、「AFC Champions League」は＜スポーツ＞ドメインである。2 つ目は動詞含意知識で、例えば「学会発表した、ということは、研究した、ということを含意する」というような知識を獲得した。3 つ目は言い換え知識で、「ACL で発表する」と「Association for Computational Linguistics で発表する」のような言い換え関係にあるフレーズ対を獲得した。本研究の知識獲得手法はいずれも、Web という世界最大の文書集合を獲得源とし、Web の特性を活かしたものとなっている。ドメイン知識獲得では Web 検索エンジンを徹底活用した。動詞含意知識獲得では、低頻度かもしれないが多くのドメインの動詞をカバーしていると考えられる Web の「ロングテール」の部分からも高精度に知識を獲得する工夫をした。言い換え知識獲得では Web の冗長性を利用した。具体的には、Web 上ではある 1 つの概念について、複数の人々が異なる表現で定義を与えているが、それらの定義文から言い換えを獲得した。本手法で獲得した知識は、人手によるチェックを経て、世の中に広く配布されている（言い換え知識は今年度末配布予定）。つまり本研究は社会的貢献も果たしている。

文書集合から知識を獲得するパラダイムは過去 20 年にわたって研究コミュニティを支配し、一定の成果を上げてきた。しかし、必要な知識の全てが文書に書かれているとは限らない。常識的な知識ほどテキスト理解に重要だが、常識的であればあるほど文書に書かれることは少ない。今後の課題は、文書に明示的に書かれていない、あるいは全く書かれていない知識を如何にして獲得するか、である。