3E−2

情報量から見た自然言語処理システムに対する考察*

野美山 浩†
日本アイ・ビー・エム株式会社 東京基礎研究所‡

1 はじめに

1.1 自然言語処理における曖昧性

自然言語処理では、1つの入力に対して複数の解釈が存在し、その結果が一意には定まらない場合がある。そのため、最も確からしい候補を選択するための何らかの手段が必要となる。

この最尤候補選択には、様々な手法が提案されている。しかし、これらの手法の多くは、実験的に有効性が検証されているものの、根本となる明白な理論を持たない経験的な手法であり、尤度、コスト、類似度といった最尤候補選択の基準となる尺度を決めるための客観的な指針が得られない。そのため、"どのようにすれば、より良い尺度が得られるか" という点は、開発者の経験的直観に依存しており、工学的には健全でないと言える。

情報量の観点から、このような問題を扱ったいくつかの研究が報告されている [1, 2]。本稿でも、同様な観点に基づき、従来提案されている最尤候補選択手法を考察する。

1.2 自然言語処理システム

自然言語処理システムは、一般に、図1のように表現できる。

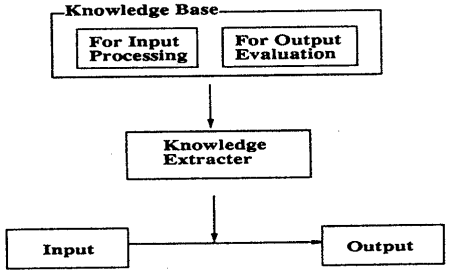

図 1: 自然言語処理システムの概略

処理のために必要なすべての知識は、知識源に記述されている。入力処理用の知識は、"規則" に相当する。知識抽出系は、知識源中の知識をうまく組み合わせて、入力から出力を生成する。

情報量の観点から知識抽出系がとるべき戦略は、2つ考えられる。1つは、出来るだけ確からしい規則を用いるというものである (最適規則選択)。これは、実際は、より特定的な条件で成り立つ規則は、より確からしいとするヒューリスティクスとして実現される。したがって、処理に用いたすべての規則に対し、それが適用されうる条件の範囲を最も狭くすればよい。情報量として見ると、用いたすべての規則に対する条件部の情報量を最大にすることを意味する。もう1つは、結果として得られるものの一般性をより高くするということである (最尤結果選択)。これは、結果として得られるものの情報量を最小にすることを意味する。

これらの2つの観点から、形態素解析と機械翻訳の幾つかの手法について考察する。

*Discussions on Natural Language Processing Systems from the Viewpoint of the Amount of Information
†Hiroshi Nomiyama
‡IBM Research, Tokyo Research Laboratory

2 個々の手法に対する考察

2.1 形態素解析

形態素解析は、一般に、正規文法でうまく記述できる。その規則の一般形は、$S \rightarrow aA$ と表せるが、これは、状態 Sのとき a が来たならば、状態 A に遷移するととらえることができる。よって、条件部の情報量は、$-\log_2 P(a/S)$ で表すことができる。

最長一致法 ある1つの状態において、最も長い単語を優先する手法である。2つの文字列 w1,w2 において、w2 が w1 の最左部分列とする。この時、文字列 w の出現確率を $P(w)$ とすると、明らかに、$P(w1) \leq P(w2)$ が成り立つ。よって、条件部の情報量 $(-\log_2 P(a/S))$ を $-\log_2 P(a)$ で大雑把に近似すれば、より長い文字列を選択することで、より条件の厳しい規則を選択することになる。よって、最長一致法は、最適規則選択の手法である。

形態素数最小法 これは、形態素の数が最小のものを候補として選択する手法である [3]。これも、基本的な考え方は、最長一致法と同様であり、最適規則選択のヒューリスティクスであるといえる。ただ、最長一致法が、局所的に最適な解を選択していたのに対し、文全体から大局的に最適化するものである。

簡単な例で考える。ある文に対し、$[w1]$ と $[w21, w22]$ の2つの候補が残った場合、最初の候補の条件部の情報量は $-\log_2 P(w1)$ で近似される。2番目の候補に対しては、2つの知識が用いられている。使用した知識全体の適用範囲を考えると、互いの知識は排反であると仮定すれば、$-\log_2(P(w21)+P(w22))$ で近似できる。文字列として比較すれば、明らかに $P(w1) \leq P(w21), P(w1) \leq P(w22)$ であり、$-\log_2 P(w1)$ の方が大きな値を取る。

文節数最小法 これは、文節数最小の候補を選択する手法である [3]。これも、基本的な考え方は、前2つの手法と同様である。ただ、一般に、P(付属語) $>$ P(自立語) であるので、より付属語を少なくする (文節数を少なくする) 方向を導入している。

規則に対するコスト最小法 個々の規則に対してコストを記述し、そのコストを最小にする方法がある [6]。コストは、その規則の一般性に応じて付与される。すなわち、最尤結果選択の手法である。

確率的形態素解析 これは、得られた結果の生起確率を N-gram を用いて推定し、最も確率の高い (一般性の高い) ものを選択する手法である [5]。これは、最尤結果選択の手法である。

2.2 機械翻訳変換過程

目的言語の知識を用いた訳語選択 この手法は、目的言語側の知識から、目的言語側の一般性を判断することによって、結果の情報量を最小にする基準に基づく [7, 8]。

Example-Based Machine Translation(EBMT) これは、規則集合の中から入力と最も類似した事例を組み合わせることによって翻訳を行なう手法である [9, 10, 11]。

この手法は、類似性によって規則の適用の制御が決定される。類似性は、語の類似性と構造の類似性の２つがある。構造の類似性は、一般に、一致しているノードの数が多いほど類似しているという計算式となっている。これは、明らかに一致するノードの数が多ければ、当然その生起確率が低くなり、その情報量は多くなる。語の類似性は、シソーラスから計算される。例えば、語 $w1$ と、語 $w2, w3$ の類似性を比較することを考えた場合（図2参照)、そのシソーラス上の最も特定的な共通の概念ノード (Most Specific Common Abstraction,MSCA) を想定し、そのノードが、より $w1$ に近ければ、より類似していると判断する。図2中、$w1$ と $w2$ の MSCA は Y、$w1$ と $w3$ に対しては Xであり、Yは Xより特定的であるため、$w2$ の方が類似していると判断される。これは、より特定的な概念を共有するものを類似しているとするものである。よって、語および構造のどちらの類似性も特定的な規則を優先するものであり、最適規則選択の手法であるといえる。

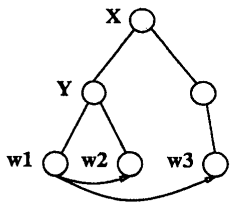

図 2: シソーラスを用いた語の類似度の計算

3 議論

3.1 知識の構成

実際に、システムを構築する上では、２つの立場がある。１つは、知識を処理する側としての立場である。この立場から言えば、記述する能力をうまく制限することによってより少ない計算量で効率的に実行できるものが良いシステムである。もう１つは、規則そのものに重点を置く立場である。この立場から見ると、規則の記述のしやすさ、獲得・蓄積のしやすさ、全体としての冗長性のなさ・無矛盾性などが評価の基準となる。

従来の自然言語処理では、処理する側の立場に重点が置かれてきた。これは、(1) 実行時の効率に重点が置かれたために、その方式によって知識の記述方法が制限を受けた[1] (2) どの規則が重要であるかを客観的に評価することが難しい、などの理由による。

EBMT などの新しいアプローチや、文法規則を宣言的に記述しようという方向は、規則そのものに重点を置いた方向である。しかし、いかに効率的に最適な知識を抽出するという観点から、知識源を構成するための研究を進める必要があるであろう [12]。

[1]例えば、形態素解析では正規文法で記述すれば、効率的に処理できる (O(n))。

3.2 最適規則選択と最尤結果選択との協調

最適規則選択と最尤結果選択は、尺度を計算する基準が異なっており、単純には比較できない。しかし、両方の観点を加味して判断する必要がある場合が存在する。

例えば、慣用句の翻訳の場合を考える。“頭が切れる” は “smart” と “hurt one's head” の２つの可能性がある。これを知識として実現するには、“頭が切れる” という規則に対し、“smart” という訳を指定すれば良いが、“頭が切れる” は、常に “smart” と翻訳すれば良いのではない。医師の怪我人を前にした会話であれば、“hurt one's head” の方が一般性が高い。このように、２つの手法を相補的に用いただけでは、解決が困難な場合がある。

4 おわりに

自然言語処理システムを情報量の流れから考察した。曖昧性を解消するための戦略が、(1) 知識源から如何に適した規則を選択するか (2) 如何に一般的な結果を選択するかの２つあることを示し、従来の手法を考察した。知識源から最適な規則を効率的に取り出すための知識の構成方法については、あまり研究されていないが、１つの解決法として、EBL(Explanation-Based Learning)[13] 的手法の導入が考えられる。しかし、自然言語のように、予め完全に定義することが困難な問題には、帰納学習との統合が必要であろう。また、MDL(Minimal Description Length)[14] の考え方は、処理系自身の効率化から見ても有用な指針を与える。

参考文献

[1] 大須賀, “知識の表現に関する一考察,” 情報処理学会論文誌, Vol.25, No.4, 1984.

[2] Sumita, K., et al., “Disambiguation in Natural Language Interpretation Based on Amount of Information,” IEICE Trans., Vol.E 74,No.6,pp.1735-1746,1991.

[3] 吉村, 日高, 吉田, “文節数最小法を用いたべた書き日本語文の形態素解析,” 情報処理学会論文誌, Vol.23, No.6, 1982.

[4] 中村他, “接続コスト最小法による日本語形態素解析の評価実験,” 電子情報通信学会技術研究報告,NLC91-1, 1991.

[5] 丸山他, “確率的形態素解析,” ソフトウェア科学会, 第８回全国大会, 1991.

[6] 西野, 藤崎, “漢字複合語の確率的構造解析,” TRL Research Report TR87-0026, 1987.

[7] 野美山, “目的言語の知識を用いた訳語選択とその学習性,” 情報処理学会自然言語処理研究会 86-8, 1991.

[8] 村木他, “機械翻訳における意味解釈、訳語選択の適切さ尤度の抽出,” 人工知能学会第５回全国大会,pp.479-482,1991

[9] Sumita, E. et al., “Translating with Examples: A New Approach to Machine Translation,” Proc. of the 3rd Int. Conf. on Theoretical and Methodological Issues in Machine Translation, 1990.

[10] 佐藤, “MBT2:実例に基づく翻訳における複数翻訳例の組合せ利用” 人工知能学会誌,Vol.7,No.1,1992.

[11] Watanabe, H., “A Similarity-Driven Transfer System,” Proc. of Coling '92, Vol. II, pp.770-776,1992.

[12] Nomiyama, H., “Machine Translation by Case Generalization,” Proc. of Coling '92, Vol. II, pp.714-720,1992.

[13] Mitchell T.M. et al., “Explanation-Based Generalization: A Unifying View,” Machine Learning,Vol1,No.1,pp47-80,1986.

[14] Rissanen, J., “Stochastic Complexity in Stastical Inquiry,” World Scientific,Series in Comp. Sci.,Vol.15,1989.