言語獲得システム（自然言語認識について）　2014/11/27（木）

画像認識を情報源として、言語を獲得する為の、自然言語認識機構について考えていって見たいと思います。

１．はじめに

人間の赤ちゃんのように、真っ新な状態から言語を獲得していく人工知能について、どんな機構を備えているべきかを考えていきます。赤ちゃんは、音素列を基本にして、音節パターンとか、強勢のあり方とか、色々な手がかりをもって、単語の切れ目を推論して、単語を獲得し、言葉が語られた状況との対比で意味と文法を得ていきます。その詳細はまだまだ研究の途上ということで、本論では、文の比較から単語切り出しもできるということを示したいと思います。ただ、本考察の方法で、単語と文法を得ることができれば、そのほかの特徴的な手がかりも収集することができます。同じ特徴パターンとか、それらの間の関係とかの重みをカウントしていけば良いからです。

学習の初期においては、あまり複雑な事ができません。手がかりが少なく、また、推論に必要な知識が使えないからです。ある程度、文法や単語を獲得したら、知識が豊富になりますから、推論の手がかりが多くなり、複雑なことにも対応できるようになります。学習システムは２段階のフェーズを踏むと考えられる所以です。本論では、簡単な文の比較のみで、単語と文法を得ていく機構を提案することになります。

２．文の比較による単語と文法の獲得

２つ以上の単純な文を比較することで、単語と文法を推定していきます。

（例文１）これは本です。

（例文２）これはノートです。

この２つの文を比較すると、違う部分は「本」と「ノート」です。したがって、単語の候補は、「これは」、「です」と「本」、「ノート」です。

（例文３）それは本です。

この文と例文１を比較すると、「それ」、「は」、「本」、「です」が単語候補として推論できます。

文法ですが、人間の生来の能力として、格の知識と「be 動詞」が有るはずなのです。格の基本は、主格と目的格と from 格と to 格と with 格ですね。from 格と to 格と with 格がきっちりあれば、原因格・結果格、とか、発生格・終了格とかは、その亜種として学習していくことができます。バスで行くのも with 格でいいし、ナイフで切るのも with 格でいいです。また、「道を歩く」の「を」も目的格と捉えることもできます。つまり、動詞は頭（主格）と手が４つあるオブジェクトとして解釈していくことが文法獲得の基本だということです。

また「be 動詞」ですが、例文１のように何かを提示しているのが基本中の基本の言語パターンです。赤ちゃんが生まれて初めて獲得する文でしょう。それはまた、次の文と等価のものでもあります。

（例文４）そこに本があります。

この文は、「本」というものが何であるかを提示する文であるわけで、「be 動詞」です。

（例文５）私はボールを投げる。

この文から格を獲得するには、ボールを投げる画像が必要です。主語は誰それと明確なことから、文頭に主語が来て、「が」が主格であることを推量できます。手が動いている図から、それが with 格で手を結びつけます。ボールが手から離れていくことから、フォーカスはボールに当たっていて、ボールが目的格であることが推量されます。しかも、それは「を」で表されることも知ります。これが文法の獲得です。

３．単語獲得のアルゴリズム

切り出した単語候補を総当たりで比較していきます。切り出した単語候補だけでなく、単語候補列に分割する前のものも総当たりの対象とします。そうしないと、間違った分割も訂正できません。したがって、単語分割のシーケンスはネットワークになります。ネットワークのパスをなぞりながら更に正確な単語分割を推定していくことになります。だから、大規模なデータを用いた学習には向かないのです。

この単語ネットワークはパスの通った回数を覚えておいて、単語の共起の解析に利用できます。実際、人間は単語列の共起に敏感です。

（例文６）プランを立てる。

であって、「プランを作る」ではないですね。

また、単語切り出しに有利な手がかりとなる項目（音節とか強勢とか拍とか）も、同時にカウントしていって、高いパターンを抽出していくことも大切です。それにより、更に正確、高速に単語と文法の学習ができるようになるでしょう。その中で、「r」と「l」の違いが無くなっていくのが日本語です。

４．おわりに

文法と単語が十分多くなったら、助詞や助動詞は知識にあるでしょう。あとは、動詞とか名詞とか、形容詞・副詞の未知単語を獲得していくことです。動詞は「う」で終わる単語とか、形容詞は「い」で終わるとか、副詞は「く」で終わるとかの規則性がありますし、未知単語の推定は重み付き投票で他の既知単語から推定できます。ビックデータ解析に有利なアルゴリズムは別にあるでしょう。

おわり