

自然言語認識システムはどうあるべきか、人工知能の中でどんな位置を占めるべきか、それを考えていきたい。文章の意味理解システムとしてどんな仕様が必要十分であるのか考察しておくことは、これから知識ベースとして自然言語を利用していく上で大変重要なことである。また、マンマシンインターフェースとして将来自然言語は重要な位置を占めるであろう。それは人間にとって記号システムの中でもっとも自然なインターフェースとなるからである。人工知能は自然言語を基盤として構築されていかなばならないのである。だから、この意味理解を深く考察しておく事は、新しい人工知能技術への扉を開くターニングポイントになると感じる次第である。

1. 日常業務に必要なデータ

機械による文章の理解とは、結局の所、文章を解析した結果からいかに有効な情報を得て利用していくことが出来るかということで計られる、計算機の行動である。だから、意味理解システムを考察する最初にするのは、どんなデータが有れば人間を満足させる事ができるかを抽出することである。よく、5W1Hという事象のまとめ方があるが、そういうデータのまとめ方はどれくらいあるのかを考えていくことから考察を始めたい。

しかる後、そのデータの形式を論じていきたい。どういう形に文章の解析結果を持てば、必要なデータを取り出しやすくなるか、これを考えたい。必要なデータが決まれば、自ずと見えてくる仕様である。

1.1 要素データの種類

事象の説明をするときのポイントは5W1Hを把握しておく事であるとよく言われることである。この5W1H（これに、「感情」、「思惟」を加えて5W1H1Fとしよう）を更に分解していってみたい。

これは、いかの情報である。

(1) ステージ（時間系列、空間系列）

何時、何処で何が行われていたかを把握する事は全ての情報の基本である。

(2) アクター

誰が、何に、何を用いてというふうに、事象の記述はアクターの間関係という形式で記述される。だからステージへの登場人物は誰かということは基本的な情報になる。

(3) アクター間関係

アクターの包含関係とか、隣接関係などは重要な情報である。特に比喻などはこの情報によって実現される。

(4) アクターの行動（格、属性、アクター間関係、モダリティ）

行動、操作は言葉で伝える情報のアンカーである。

(5) 原因結果関係

人は物事を原因と結果という概念で把握しようとする。文章があるのはこの原因と結果を表現したいためであるとも言える。

(6) 文の持つ感性(感情、思惟、話者の立場)

感情は重要である。実際に起きた事か、推測かというものや、好悪といったことは文章表現には付属的であるが、重要な情報である。

これらは、世界を記述する根本パラメータで、これらのパラメータは全てを尽くしていると言える。

だが、自然言語処理の立場からはもっと複雑になる。イメージを基盤にすれば、(1)から(6)の概念を「点」として把握していけばいいのであるが、言語表現では、様々な表現が「点」に対して存在する。そうした一つの点に対応する表現をまとめて一つの記号で代表させるという処理が必要になるのである。しかも、この対応は状況に依存する。例えば、「歩く」も「走る」も「泳ぐ」も「移動」という状況では一つの記号で表現されるが、陸上という状況では「泳ぐ」と「歩く」、「走る」は別の記号で表現されるだろう。このような、自然言語処理で浮かぶ記号を(7)としよう。

全部で7個のパラメータが見つかった。利用されるのはこの組み合わせの数だけあることが分かる。その内、完備なパラメータセットを持つものが5W1H1Fであることも分かる。

1.2 データの形式

データの形式はなんであるか。ステージとかアクターは逐一変わっていくが、それらが組み合わさって特別な意味を持つという事はない。あくまでも個別なデータである。しかし、アクターの行動は時系列を経ると、特別な意味を持つてくるのが普通である。個々の行動だけでない、纏まったときの意味というものが現れる。しかも、意味は視点によって変わってくる。意味は元々あるとしても、ある状況では重要な意味を持ち、ある状況では価値のない意味となる。

文の時系列の様子を次に示す。

(例文) 私は牟礼駅で列車に乗った。列車は直ぐに出発して、豊野を過ぎて長野駅に20分で到着した。

この文章と次の文とは等価である。

(例文) 私は列車で長野へ行った。

後者の立場では「豊野」は無価値な情報として捨てられているのである。必要なのは何処にいったか、とかどこを経由したのかといった外的条件で変わってくるのである。

従って、状況を与えてからアクターの行動の時系列を解析して意味を抽出して、利用していくという手続きになる。データの形式は、このアクターの行動をどう表すのが、解析しやすいかという視点で設計しなければならない所以である。

だから必要なデータ構造は、状況によって意味が異なってくることと、データが固まり

として意味を表現してくることから、フレーム技術を採用し、さらにフレームが格や属性などの構造を持ったデータを表現できるものということになる。基本的には XML で表現したフレームということになるだろう。木構造の記号、値セットを表現できる定番の技術であるからである。

またフレームは句や文節、文、文章が名詞化されて表現されていることに対応できなくてはならない。

(例文) 私が昨日行ったのは本屋だ。

この文では「私が昨日行った」が名詞化されている。この名詞のプロフィールとして「本屋」がある。フレーム内の意味データの検索で、「私は昨日本屋に行った」とマッチングできなくてはならない。こういった、構造上の問題をフレームは抱えている。とにかく、名詞化(オブジェクト化)は言語の重要な機能なので、十分対処していかななくてはならない。

また、他のシステムへ働きかけるアクションを定義できなくてはならない。それも、XML などのメタデータを付けてフレームの中に埋め込めるものであることが必要である。

筆者はこのフレームを Prolog 文を拡張した文法で記述している。左辺にパターンを記述し、右辺に実行アクションを記述する。また、様相も表現できるようにして、この文の成立する文脈を指定出来るようにしている。ただ、項に記述するのはシステムで定めた値だけで、その項の組み合わせで全ての処理、文の表現などを行えるようにしている。もちろん、変数の指定は可能である。例えば、

- (1) 格は Case
- (2) 属性は Profile
- (3) 係り受けは Attribute
- (4) 連想は Relation
- (5) 文脈記号は Information
- (6) 場面は Stage
- (7) アクターは Actor()

である。

2. 自然言語解析

自然言語処理では、多重並行処理により曖昧性を低減させながら次の処理を行っていく。

(1) 形態素解析

日本語処理では単語が接着しているため、単語の切り出しとして重要な意味を持つ。

(2) 係り受け解析

修飾関係は単語間の関係のスケルトン（関係基盤）である。

(3) 格解析

単語間の関係で特に意味的に重要な情報が格である。動詞、名詞は格によって他の名詞とどんな関係が有るかを明確に表現される。

格は通常助詞とか前置詞で表現される。日本語の「の前に」も助詞相当語として格を表している。また動詞は助詞のままで係るが、名詞の格は助詞に「の」がつけられる。

(例) 校舎の前の庭

(4) 文脈解析

指示代名詞の解決、欠落語補填処理

(5) ステージ解析

文が成立している時間、場所の推移を解析する。

これらの処理は曖昧性が多数あり、格ステップの結果を文として組み合わせると、無数の候補文ができてしまう。組み合わせ爆発である。それを防ぐには、単語に係り受けなどの関係情報を全部ポインタで保持させて、ポインタの有効性を確信度で評価できるようにする。コーパス情報や共起情報を用いて、確信度を操作し、最後にもっとも確信度の高い関係のものを選択して、文として出力するようにする。解析しながら曖昧性を刈り込んでいく事は重要な処理になる。

また処理は手戻りが避けられない。全体が決まらなないと部分の評価も決まらない事が沢山あるからである。弛緩法を用いて順次曖昧性を低減していく処理にすべきである。

この処理の結果は基本的に原文の骨格を色濃く残したものが望ましい。できるだけ加工しない方が事実情報が多いからである。

そうして、語られている場面をステージ情報から推定して、あらかじめ知識としてあるフレームに埋め込んでいく。

次のステップはこれをもっと意味的な解析が進んだフレームに落とし込んで行く処理が必要である。

(1) アクターのプロフィール

- ・名前、色、堅さなどの属性である。(is_a 関係)
- ・頭がある、手があるの属性である。(has_a 関係)

(2) アクター間の関係を表現するプロフィール

- ・親愛、好悪、兄弟、親子などの関係である。位置関係、時間関係もある。

(3) アクター毎の行動をまとめたトレース

- ・格とか行動の変遷である。

動詞完備型の標準構文パターンを文の意味とする。自然言語の持つ多様さは無くして表現していく。パターンマッチングを簡単に行えるようにするためである。

例えば、次の2つの文は相同で、パターンマッチングするのであればどちらかの形式に統一しておく方がよいだろう。

【例文 a】データのまとめ方の考察を始めたい。

【例文 b】データをまとめる方法を考察し始めたい。

(4) 文の意味をパターンにまとめた文プロフィール

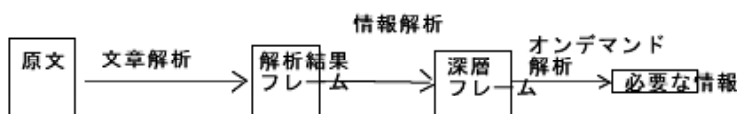
- ・疑問/ファジーさ、予想、好悪などである。

(5) 原因結果の把握プロフィール

- ・「～したのは～だから」のスロット表現である。

これらが必要十分な情報源である。

フレームデータの利用に当たっては、視点によって文章から取り出す意味がことなるから、オンデマンドでパラメータを与えて処理を進めなくてはならないことに注意がひつようである。



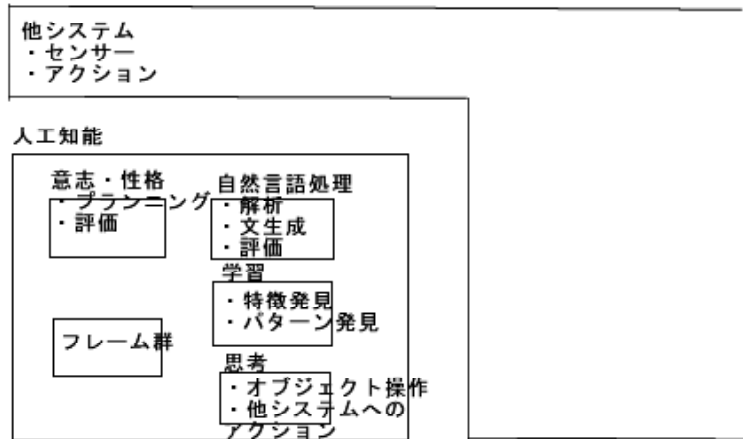
解析結果フレームから深層フレームへの情報解析部分は結構複雑な処理を必要とする。例えば、うなぎ文である。

(例文) 私はうなぎだ。

「私はウナギを食べる」のか「私の名前はウナギだ」なのか曖昧である。前後の文を解析しないと決定出来ない事柄である。

3. 人工知能システムの中での位置づけ

自然言語処理は計算機のマンマシンインターフェースを担うことになる。人工知能は計算機システムのなかの知能となる部分、知識ベースを担う事になるから、自然言語処理は人工知能の基盤になっていることが分かる。下図にこの様子を示す。



・人工知能は他システムに埋め込まれて存在する。
・自然言語システムはマンマシンインターフェースを担う。

対話処理では、計算機と人間との会話を成り立たせるために、人間との話題を選択し、話題を維持していくこと、話題を変える事、戻す事、こういうことを実時間で制御して行かなくてはならない。何が重要な事か、何が関連事項かを把握し、足りない情報を人間から聞き出していかねばならない。人間からの情報も組織だってフレームに組み込んでいかねばならない。これは、意志とかプランニングの問題である。なにが重要かを判断するのは、計算機の持っている価値観であり、性格であるといえる。そうしたものが人工知能には必要なのである。

思考は、意志システムと直結して、オブジェクトを操作し、問題を解決していく過程である。そのベースとなる自然言語処理システムは意志に依らない無意識過程として思考システムを支えることになる。思考過程で試行錯誤の結果得られたプロセスは、無意識システムで利用出来なくてはいけない。これが思考をベースとする学習機構になる。

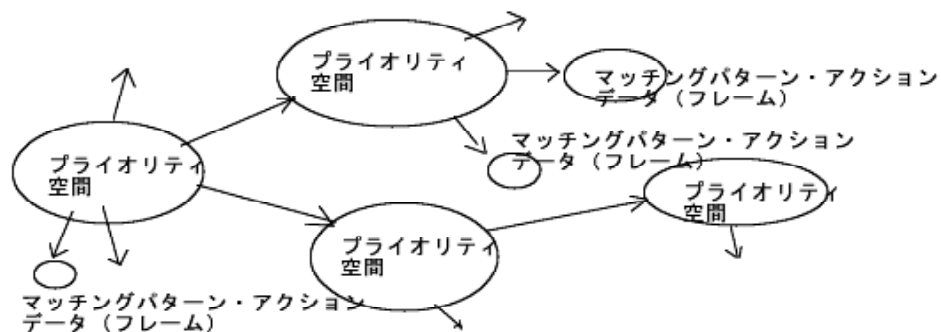
4. 意志・プランニング

自然言語で扱うデータは変化に富んでいる。だから、データを取捨選択していく機構が必要である。また、ある目的の元に関連データを集中して獲得していく行動力が必要になる。意志とか、プランニングが重要な機構となる所以である。プランニングとは話題をとくか発話内容を計画する機能で、意志なるメタデータ、メタプロセスで知識ベースの内容をズーム、パン、チルトそしてフォーカスしていく過程である。

意志・プランニングもフレームで表現出来る。それは、文章を解析して得られたフレームと同じ形式であることが望ましい。それは、文章として意志やプランニングを設計していきたいという事と、通常の文章の何が未定であったり、文章をさらに詳細に知る必要があるかを判断し処理を計画していくのが、意志でありプランニングであるからで、文章と密接に関係しているからである。未定であれば処理をするというコマンドが文章に埋め込まれていたら、そのプランニングは単に優先順位を割り振っていくという処理だけになるのである。

人間の頭も、意識的に行動するときでも、意識に登る前に前もって神経活動が起こっているという。決して、超越した自由が有るわけではない、データ駆動型のシステムなのだという。

意志、プランニングでは無数に知識ベースにある事柄の中から特定のデータを選んで発話するのであるから、発話の優先度決定は大がかりのものであるはずである。各データはローカルにパターンマッチングして特定のアクションを選択しようとする。しかし、システムとしては1つに統一して行動しなくてはならないわけで、ここに最適制御システムというものが必要になってくる。それはn次元のプライオリティ空間で、プライオリティ決定の視点によってプライオリティ空間値が異なってくるような、ダイナミックなシステムとして設計されると考えられる。n次元と変量にしたのは、評価軸となる量は学習によって獲得していくものだからである。



5. 学習

自然言語処理はデータが基本のシステムである。小さな意味の単位が膨大に組み合わせられて、使えるデータになるのである。とても人手によってのみで構築出来る量ではない。必然的に学習機構が要請されるのである。

(1) パターン発見能力

- ・思考能力からのパターン発見
- ・無意識処理としてのパターン発見

プリミティブな特徴量を重み付きで組み合わせていくのがパターン発見である。

(2) 抽象化能力

コーパスなど具体的な要素で生データは構築されていくが、そこから共通な要素を抽出してより汎用的なパターンを生成できなくてはならない。

曖昧性を消してデジタルなカテゴリーにしていくこととか、カテゴリーの合成カテゴリーを作るなどである。

(3) 推論能力

抽象化と関連するが、要素の共通項から自然言語としての規則を予測して試していく機能は欲しいものである。

カテゴリーを試して、人間からの評価を得ていく過程でもある。

学習に当たっての自然言語データとしては次のものが考えられる。

(1) コーパス

単語と単語の組み合わせとか単語間の共起関係である。

(2) コーパスを抽象化してカテゴリー記号を付加したもの

品詞とか、作業的なカテゴリー記号群である。

学習は、ある意味ローカルな特徴からマッチング用のパターンとアクションとのコネクションを作る事であり、そのコネクションは何ら確固とした基準があって行われるわけではない。基準そのものを学習していく必要が有るからである。従って、無数のゴミが生まれやすい。

だから、詳細に検討すると無意味に見えるデータであるが、大局的判断には貢献するものが知識ベースの各データである。つまり、学習されるデータは曖昧性のあるシステムとなっているという事で、意識・プランニング機構によって評価され、正しく使われて始めて意味を成していく。このことに注意していく必要がある。

6. 思考

思考とはどんな作業であるか。物事を考えるとき、従来の計算機ならば、モデルを使って有限要素法のように計算するか、記号の連想によって解に行き着くか (Prolog) であった。

これをもう少し拡大して考えてみれば、他の既存の時系列操作をモデルとして当てはめ、今の問題を解決していくという方法論にいきつく。モデルはイメージデータでもいいし、記号列でもよい。オントロジーみたいなシステムで、プリミティブな時系列操作群として接地されていればよい。この方法を「**比喩**」的問題解決と呼ぶ事にする。

筆者としては圏論とオントロジーで明晰にこの辺の理論は構築出来ると感じるのだが、どうだろうか。

比喩の能力によって、意識下 (意志・プランニングの下での) でオブジェクトを手繰る能力は学習システムには必須になるだろう。学習を大げさに実現すれば、この思考機能を必然的に構築していく事になるからである。また、イメージベースのオントロジーを基盤にしなくては人間の持っている知識を完全には記述出来ないの、従って人間と会話はできないので、この思考機構はいつかは実現していかなばならないものである。



筆者はこの比喩能力がオントロジーを基盤にした思考の理論化に重要と考える次第である。

7. おわりに

自然言語システムはどうあるべきかを考えてきた。自然言語処理は人間の日常の行動である、世界をどう見ているかを問う問題である。従って、データとして何を捉えていけばいいかは、日常の生活がどうゆうパターンで記述をしていけば十分かを考えて行く事になる。そうすると、時間、場所、アクター、アクターの属性、アクターの行動、アクター間の関係というもので表現できることが分かる。しかもそれが全てである。後は、物事には原因と結果があるとか、言葉を発信した人の心の動きとかの問題となる。それらをパラメータとしてデータを表現していくことで、完備なシステムとすることができる事が分かる。

自然言語は確かに捉えがたいが、人間の脳がやっていることである。どう捉えるかは不可能な事ではない。言葉が発せられた時に表出されるパラメータを調べ、まとめていけば体系だった理論を構築出来るのではないか。この論文はそれへの一歩となるものである。自然言語処理システムを作り、言葉を手繰る試行をしていく中で、自然言語を分析していけば、最終的に確固たる言語理論ができると信ずるところである。

おわり