

# 自動処理できる言語とはどのようなものか

山田隆弘

宇宙航空研究開発機構 宇宙科学研究所

tyamada@isas.jaxa.jp

## 1 はじめに

自然言語処理技術は、すでに存在している自然言語の文章に対して計算機を用いて何かしらの自動的な処理を行い、質問応答や機械翻訳などの特定の応用目的を達するためのものされている[1]。しかし、自然言語の文章には曖昧性などの問題があり、自動処理を完全に行うことは困難である場合が多い。そのために、統計的な処理等も活用し、ほぼ正しい答えを確率的に導き出すという場合も多い[1]。

ところで、「もし自動処理できるような言語があったとしたら、それはどのようなものであるか」という問いに対する回答を見出すことは、自然言語処理の限界を見定めるために有効ではないだろうか。また、自動処理が行える言語がどのようなものであるかが分かれば、自動処理を行いたい場合には、そのような言語を使用する（あるいは、自然言語をそのような言語に近づけたものを使用する）ことによって自動処理が行えるようになる（あるいは、自動処理が行いやすくなる）はずである。

本稿では、上記のような動機に基づいて、自動処理できる言語が持つべき条件を考察する。ただし、本稿では、議論を分かりやすくするために、対象とする言語を限定して考えることにする。まず、本稿で対象とする言語を規定するために本稿で前提とするコミュニケーション・モデルを第2節で定義する。第3節では、言語の自動処理が行われるための条件について考察し、第4節において、自動処理が行える言語が持つべき条件について検討を加える。

なお、本稿は、筆者が以前発表した[2]の内容をさらに厳密化させるために、検討の対象を機械による自動処理に限定し、議論の内容も精密化させたものである。

## 2 コミュニケーション・モデル

この節では、本稿で前提とするとする言語を用いたコミュニケーション・モデルを定義する。

まず、本稿では、言語を「ある文法規則と語彙によって生成される文の集合」とであると定義する。

また、ここでは、言語は「ある世界における事実を伝えるために用いる」ものとする。すなわち、感情を伝えたり、修辞技法を示したりするための言語使用は除外することにする。また、上記の「世界」とは、現に存在している世界のみでなく、存在することが予想される具体的な仮想世界も含むことにする。例えば、来年実施すべき計画を議論する時に使用する言語は、来年になれば実現する具体的な世界についての事実を伝えるためのものであるから、本稿の対象となる。このように対象となる言語を限定しても、実用的に使用されている言語のかなりの部分を包含していることになるであろう。

言語による表現は、ある世界における事実を話者から聴者に伝える（あるいは、著者から読者に伝える）ために使用されるものとするが、どの世界も非常に多くの情報を含んでいるのが普通である。しかし、話者から聴者への情報伝達は、特定の目的のために行われるわけであり、話者から聴者へ伝えるべき情報は、世界の中の特定の部分に関するものである。また、世界の特定の部分における事実を話者から聴者へ伝える場合、その特定部分に関する知識を話者と聴者で既に共有していることがほとんどである。従って、話者から聴者への情報伝達は、ある世界の特定部分に関する共有知識に対する差分を伝えるものであると言って良い。

さらに、話者から聴者への言語による情報伝達は、情報伝達の経済性を高める（すなわち、最小の言語表現で必要な情報を伝える）ために、聴者が既に共有知識を保有していることを前提とし、共有知識に含まれている情報は、必要最小限のものだけを言語表現に含めるべきである。この考えは、Grice [3]の協調の原理(cooperative principle)に従ったものであるし、原島[4]の知的符号化の考え方も一致する。

聴者は、話者から受信した言語表現を自分が保有している共有知識と照らし合わせた上で言語表現に含まれている知識の差分を復元し、それを自分の共有知識に組み込み、新たな言語表現の受信に備える。

以上のように定義したコミュニケーション・モデルを図で示すと図1のようになる。

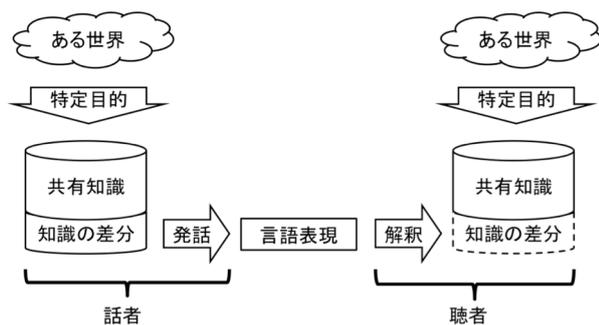


図1 コミュニケーション・モデル

### 3 言語の自動処理とは

本稿のテーマは、言語の自動処理であるが、これは図1における聴者が機械（あるいは計算機）に置き換わり、聴者の処理が機械によって自動的になされることに相当する。この場合、自動処理が行われる（すなわち、自動処理に成功する）とは、どのようなことを言うのであろうか。それは、受信した言語表現を聴者（機械）が保有している共有知識を参照して処理を行い、話者が発信した知識の差分が復元でき、聴者の共有知識に組み込まれることであると言いつけることができる。すなわち、言語の自動処理は、受信した言語表現が話者の意図通りに聴者の共有知識に組み込まれた時に成功する。

この議論をさらに進めるためには、共有知識の内容とそれがどのように表現されているかについての何らかの想定が必要になる。共有知識の内部構造とその表現方法については、詳細な検討が必要であるが、ここでは共有知識はオブジェクト指向モデル[5]によって表現されていると想定する。あるいは、オブジェクト指向モデルからオントロジーを構築することも容易であるので[6]、共有知識はオントロジーとして表現されていると想定しても実質的には同じである。また、共有知識の内容は、特定の文化に依存するような知識を除けば、使用される言語には依存しないものとする。

オブジェクト指向モデルあるいはオントロジーの内容は、概念に関する知識と個々の物体や事象に関する知識の二つに分けて考えることができる。概念に関する知識は、オブジェクト指向モデルではクラス・モデルとして表現することができ[5]、オントロジーとしては TBox (Terminology Box)として表現される[7]。個々の物体に関する知識は、オブジェクト指向モデルではオブジェクト・モデルとして表現することができ[5]、オントロジーとしては ABox

(Assertion Box)として表現される[7]。個々の事象(event)を表現する方法としては、いくつかの重要な研究が存在し[8][9][10]、それらを適用することができる。

ここでクラスとは、共通の性質を有する個々の物体（オブジェクト）の集合であり、そのクラスに属する個々の物体は全て共通の属性を持つ。属性は、オブジェクトの性質を表現する手段であり、属性が有する値によってオブジェクトの性質が表される。また、複数の同一あるいは異なるクラスに属するオブジェクトの間に関係を定義することができ、この関係をリンクと呼ぶ。特定の（同一あるいは異なる）クラスのオブジェクト間の同種のリンクの集合のことをアソシエーションと呼ぶ。また、オブジェクトはオペレーションと呼ばれる動作を実行することができ、どのようなオペレーションが実行できるかは、そのオブジェクトが属しているクラスによって決定される。

筆者は、言語表現の意味もオブジェクト指向モデルで表現できることを指摘したが[11]、ここでも言語表現の意味は[11]で示したオブジェクト指向モデルによって表現されると考える。

さて、上では「受信した言語表現が話者の意図通りに聴者の共有知識に組み込まれた時に言語の自動処理が成功する」と書いたが、これは以下のように二つの問題に分解して考えることができる。一つ目は、受信した特定言語における言語表現を話者の意図通りのオブジェクト指向モデル（あるいはオントロジー）として復元できるかどうかという問題である。これは、受信した言語表現に対する構文解析や意味解析の結果として言語表現の意味を表すオブジェクト指向モデルが構築でき、それが話者の意図したものと同じであるかどうかという問題として言い換えることができる。二つ目は、復元されたオブジェクト指向モデル（あるいはオントロジー）を聴者の保有している共有知識に正しく組み込むことができるかどうかという問題である。これは、受信した言語表現より構築したオブジェクト指向モデルが聴者の共有知識と整合して、聴者が自分の知識として今後活用できるかどうかという問題であると言いつけることができる。

### 4 自動処理可能な言語とは

ここでは、前節で述べた二つの問題が解決できるような言語とはどのようなものであるかを考えることにする。

まず、一つ目の問題は、(1) 受信した特定言語に

おける言語表現をオブジェクト指向モデルとして表現できるかどうかという問題と、(2) 表現されたオブジェクト指向モデルが話者の意図通りのものであるかという二つの問題に分解できる。

(1)の問題は、言語表現が表していることからの種類 ([11]の表1参照、例えば、アトリビュート値の変更) が同定でき、さらに、必要な項の情報が揃っているか (例えば、アトリビュート値の変更の場合は、どのオブジェクトがどのオブジェクトのどのアトリビュートの値をどのように変更するか) という問題に還元できる。

(2)の問題は、(1)を行う過程で曖昧性が存在し、可能性のある解釈が二つ以上存在し、話者が意図した解釈とは異なる解釈が選ばれてしまうという問題である。これには、語彙や構文の曖昧性 (例えば、動詞の意味が目的語の種類によって異なる) の問題と項の同定の問題 (代名詞の参照先の同定や省略されている項の同定の問題) が含まれる。

上の(1)も(2)も言語依存の問題であるが、どちらの問題も、特定言語における言語表現のパターン (例えば、特定の動詞と名詞の組み合わせ) と意味構造 (例えば、あるクラスのオブジェクトの特定アトリビュートの値の変更) との間の1対1の対応表を作成し、この対応表に合致した言語表現を使用している限り、自動処理可能と見なせることになる。この対応表は、特定の言語毎に作成することになるが、言語に依存しない一般的な対応表の構成方法を検討することは、重要な将来の課題である。言語に依存しない一般的な対応表の構成方法を開発するためには、一方では、言語表現のパターンを一般的に扱うための枠組みが必要となり、他方では、言語表現の意味構造を特定言語には依存せず、かつ共有知識との対応も取れるように表現する方法が必要になる。

さらに、文と文の間の関係 ([11]の表3参照、例えば、ある文が別の文の理由を表す) についても、言語表現のパターンと意味構造との間の対応表を言語毎に作成する必要がある。これについても、言語に依存しない一般的な対応表の構成方法を検討すべきである。

第3節で述べた二つ目の問題は、受信した言語表現の意味をオブジェクト指向モデルとして表現できたとして、それを聴者が自分の知識に正しく組み込むことができるかどうかという問題である。上に論じた一つ目の問題は、特定の言語に依存する問題であったが、二つ目の問題は、特定言語には依存せず、知識のレベルにおける問題である。

単文に相当する言語表現の意味は[11]の表1のよ

うに分類される。この分類の各々に対して、その意味を聴者が自分の知識に正しく組み込むことができるための条件 (すなわち、その言語表現が自動処理されるための条件) を以下の表1に示す。ただし、[11]の表1は、クラスに関する言語表現に対する意味構造の記述が欠けていたので、本稿の表1はそれを補ってある (表1の最初の4つの行がそれに相当する)。

表1 オブジェクト指向モデルとして表現された言語表現の意味が聴者の知識に組み込まれる条件

言語表現の意味	言語表現の意味を聴者の知識に組み込める条件
クラスの存在	そのクラスと何かしらのアソシエーションを有するクラスが聴者の知識に既に存在する
アトリビュートの存在	そのアトリビュートを有すべきクラスが聴者の知識に既に存在する
アソシエーションの存在	そのアソシエーションを有すべきクラスが聴者の知識に既に存在する
オペレーションの存在	そのオペレーションを有すべきクラスが聴者の知識に既に存在する
オブジェクトの存在	そのオブジェクトが属するクラスが聴者の知識に既に存在する
アトリビュート値	そのアトリビュートと話題となっているオブジェクトが聴者の知識に既に存在する
リンクの存在	そのリンクが属するアソシエーションと話題となっているオブジェクトが聴者の知識に既に存在する
オブジェクトの生成・消滅	そのオブジェクトが属するクラスが聴者の知識に既に存在する
アトリビュート値の変更	そのアトリビュートと話題となっているオブジェクトが聴者の知識に既に存在する
リンクの生成・消滅・変更	そのリンクが属するアソシエーションと話題となっているオブジェクトが聴者の知識に既に存在する
オペレーションの実行	そのオペレーションと話題となっているオブジェクトが聴者の知識に既に存在する

簡単な例を用いて表1の意味を説明する。例えば、聴者が受信した言語表現をオブジェクト指向モデルとして表現した結果が「キリンは動物である」(クラスの存在) であったとする。「キリンは動物である」

という知識を既に聴者が保有している場合は、何の情報伝達も起こらない。キリンという知識を聴者が保有していないが、動物がクラスであるという知識は保有している場合は、「キリンは動物のサブクラスである」という知識が聴者の知識に組み込まれる。このような場合は、受信した言語表現の自動処理が可能であると言える。聴者がキリンという知識も動物という知識も保有していない場合は、「キリンは動物である」という知識を聴者が今後活用することはできないので、聴者の知識に組み込まれたとみなすことはできず、自動処理も失敗するとみなすべきである。

上記の議論は、個々の言語表現ごとに自動処理が可能であるかどうかを論じたが、実際には、談話（複数の言語表現の集合）によって情報伝達を行う場合もあり、そのような場合も加味して上記の理論は拡張する必要がある。これは今後の研究課題である。

さらに、ここでは問題を言語依存の問題と言語非依存の知識の問題の二つに分けたが、これらの二つが干渉する場合も考えられる。すなわち、言語非依存の知識を用いなければ特定言語の解釈を行えない場合もあり得る。これも今後の研究課題である。

さらに、本稿のコミュニケーション・モデルは、聴者に事実を伝えることを情報伝達の目的としたが、別の目的のコミュニケーション（例えば、聴者の知識を問い合わせる）もあり得る。このような場合にも適用できるように本理論を拡張する必要もある。

## 5 将来の課題

本項で検討した方式に対する将来の課題は、以下の通りである。

1) 共有知識の内容の定義とその表現方法を開発する。

2) 言語表現のパターンと意味構造との対応関係を規定するための（言語に依存しない）一般的な方法を開発する。

3) 談話によって情報伝達を行う場合にも適用できるように理論を拡張する。

4) 言語非依存の知識を用いて特定言語の解釈を行う場合にも適用できるように理論を拡張する。

5) 聴者に事実を伝えること以外の目的のコミュニケーションにも適用できるように理論を拡張する。

## 6 おわりに

本稿では、自動処理できる言語（文の集合）とはどのようなものであるかについて検討を行ってきた。

まず、ここで前提とする言語とそれを用いたコミュニケーション・モデルを提示した。それらの前提に基づいて自動処理が行える言語が持つべき条件を考察した。ただし、第5節で列挙したように、本稿の考察には不十分な点も多く、将来の課題も多い。今後は、将来の課題を解決する方向で研究を進める予定である。また、本稿の内容が言語処理技術に関する研究や開発に何らかの刺激を与えることができれば幸いである。

## 参考文献

- [1] Jurafsky, D., Martin, J.: *Speech and Language Processing, Second Edition*, Pearson Education (2009).
- [2] 山田隆弘: 意味の理解とは何であるのか — 通信理論からのアプローチ, 2016年度人工知能学会全国大会, 303-OS-04a-4 (2016).
- [3] Grice, H.: *Studies in the Way of Words*, Harvard University Press (1989).
- [4] 原島博: 知的画像符号化と知的通信, テレビジョン学会誌, Vol. 42, No. 6, pp. 519-525 (1988).
- [5] Blaha, M., Rumbaugh, J.: *Object-Oriented Modeling and Design with UML, Second Edition*, Prentice Hall, 2005.
- [6] Gasevic, D., Djuric, D., Devedzic, V.: *Model Driven Engineering and Ontology Development*, Springer (2010).
- [7] Baader, F., et al.: *The Description Logic Handbook: Theory, Implementation and Applications, Second Edition*, Cambridge University Press (2010).
- [8] Davidson, D.: *Essays on Actions and Events, Second Edition*, Oxford University Press (2001).
- [9] Parsons, T.: *Events in the Semantics of English*, MIT Press (1990).
- [10] Kaneiwa, K., Iwazume, M., Fukuda, K.: Upper Ontology for Event Classifications and Relations, Proceedings of the 20th Australian Joint Conference on Artificial Intelligence, pp. 394-403 (2007).
- [11] 山田隆弘: オブジェクト指向に基づく意味構造の記述について, 2014年度人工知能学会全国大会, 214-OS-08a-4 (2014).