

機械翻訳システム CONTRAST における 文脈情報の利用

7J-7

石崎 俊 井佐原 均 内田 ユリ子 橋田 浩一 元吉 文男

電子技術総合研究所

1: はじめに

筆者らは、文脈を考慮した機械翻訳システム CONTRAST を提案し、要素技術を開発してきた。(図1) 前回の発表では、英日翻訳システムについて報告したが、今回は、引き続き開発中の日英翻訳システムを中心に説明を行う。ここで対象とする分野は、現在はテロリズムに関する新聞記事である。本稿で説明に用いる新聞記事を図2に示す。

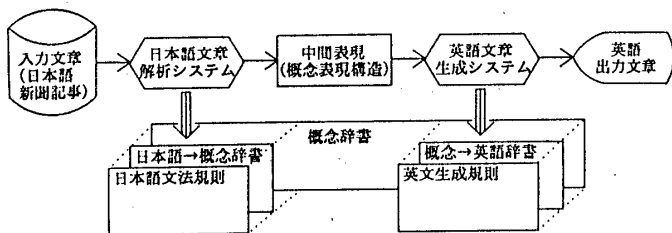


図1 機械翻訳システム CONTRAST の構成

誘拐3時間 無事に解放 ス페인大使 (スペイン) 十日前に共同西... (transcription of the newspaper article text)

図2 取り扱う新聞記事

(昭和59年10月11日朝日新聞夕刊より)

2: アクティブマッチングを用いた文脈理解

CONTRASTの入力文章理解過程は構文解析、意味解析、文脈解析の3つの段階を統合している。意味解析、文脈解析の過程において、概念間の同一性の検出が大きな課題となる。筆者らは意味的にひとまとまりの文章において、表層上でいくつかの表現があって、それらの指示物を同定するとき、異なるスロットに様々な情報が格納されていても、同一スロットに異なる値があるような明確な矛盾がないかぎり、積極的に同一物としてマッチングを取る手法を採用しており、これを「アクティブマッチング」と名付けた。これは我々人間が、文章を読んだり、人の話を聞いて理解するとき、大変効率的に理解が行われることの一つの基準を示すものであるが、詳細は別稿に譲る。

さて、図2に示す新聞記事から、図3に示されるような中間表現構造を得る。事前知識として文脈表現構造は、「誘拐し交渉し解放する」という「topic-kidnap」という概念表現構造と、その「誘拐」シーンを「拘束し、移動し、監禁する」というように細分化する

“topic-kidnap-act”という概念表現構造などから組み立てられた構造である。中間表現構造は、トピックを表わす適切な文脈構造に入力新聞記事から得られるボトムアップの知識を埋め込んで得られたものである。図3のAとOはそれぞれ共通に、図の上部にある構造を指す。このAとO、すなわち、加害者である「犯人」と被害者である「大使」とが表層では明示的には表現されていないような事象(たとえば、“release”)において、それぞれ“actor”と“object”になるということの決定は上に述べたアクティブマッチングによる。また、“convey”の“situation”は、“blindfold(目隠しする)”によって引き起こされる状態のもとで“convey”が行われたことを、入力文から抽出し、表現している。

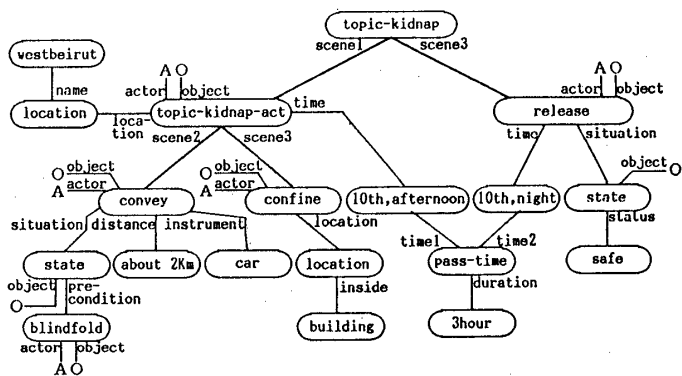
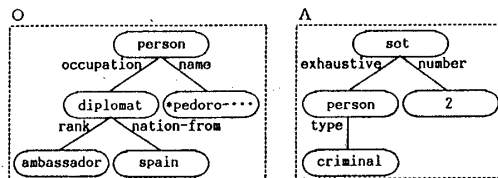


図3 図2の記事の中間表現

3: 英文生成における文脈情報と生成規則の使用

生成部は、生成すべき文のもととなる中間表現構造を入力情報として、文脈表現構造、概念辞書、その概念辞書を英語に変換する部分および英文生成規則からなる英語辞書を用い、英語文章を出力する。

以下に生成規則の概要と図3の構造に対する生成例を示す。ここでは、生成規則の中で、文脈表現構造と密接な関係をもつ主要な生成規則について述べ、一般英文法による細則は省略するものとする。

3-1: 英語文章全体(複数パラグラフ)の構造の決定

(図4参照)

(規則1) 第1パラグラフで、 S_1, S_2, S_3 について生成する。これを $[S_1 S_2 S_3]$ と表わす。このとき各 S_i (scene)が完結していることを確かめ、未完のものについては生成しない。(このパラグラフ自体の内部構造は規則2に従う)

第2パラグラフで $[S_{11} S_{12} S_{13} \dots]$ を生成し、第3パラグラフで $[S_{21} S_{22} \dots]$ を生成する。以下第2レベルについてはこれをくり返す。第2レベルの生成が終了したら第3レベルのnode(これはsceneではなくeventになるのが一般的である)について $[S_{111} S_{112} S_{113} \dots]$ を生成する。以下同様。

3-2: パラグラフ内部構造の決定

(規則2) 同レベルのscene(たとえば $[S_1 S_2 S_3 \dots]$)についてのパラグラフ生成については、“Focus”のあたっているscene(文脈表現構造におけるFocus Listに従う)をそのパラグラフの第1文の主文(主節)とし、そのsceneと、文脈表現構造中に示される因果関係などの関係にあるsceneを埋込文(従属節)として生成する。(これは規則3に従う) focus指定がなければ、sceneは第2文以下に關係の強さに従って並べる。(例えば時間順)図3の場合には $[S_1 S_3]$ において S_3 のほうがFocus Listの上位にあるので S_3 に対する生成(すなわち、“release”による生成)を主文とし、図5のような結果となる。

3-3: 埋込文、従属節の生成とそれに伴う接続詞の決定

S_n に対する生成文 C_n と S_m に対する生成文 C_m を、埋込や主従関係で接続するときは次のようにする。(埋込文、複文、重文などの文体の選択は、現在のところルールを検討中である。)

(規則3) S_n と S_m のslotを調べ、同一のvalueをもつものがあれば関係代名詞節又は関係副詞節 C_n をつくり C_m に埋め込む。

同一のvalueを持つslotの組が複数あるときは関心度の大きい方をとる。関心度は、文脈表現構造中の必要なnodeに与えるものとしている。関心度の指定がない場合はactor, object, time, locationなどの順とする。図2の記事の場合、“topic-kidnap-act”と“release”は共通のactorとobjectを持つが、誘拐事件の場合は被誘拐者は有名人、富豪などで、誘拐者よりも読者の関心度が高いとみて、被誘拐者のところに埋め込まれている。(図5)

3-4: 動詞を中心とした文、節、句の生成と選択

(規則4) 各nodeのconcept slotの動詞概念から動詞を決める。動詞の型と、中間表現におけるactor slot, object slot, situation slotなどの組合せにより基本文型を決める。たとえば、“Bomb exploded.”では、中間表現でslotは“object = bomb”、動詞“explode”は V_i と V_t があるので、文型 $S + V_i$ をとり生成した。situation slotは一般に S または O の状況を表わしており

動詞が V_i であれば $S + V + C$

動詞が V_t であれば $S + V + O + C$ (C は補語)

の文型になる。situation slot中のstateで表現される概念が C に対応する。状況が動作結果であるときは、図2の例の出力図5のように過去分詞型(“blindfolded”)となり、status slotの場合は、概念をそのまま英語化する。

基本文の動詞は副詞句(または節)で修飾されるがこれらを生成する規則は、slot名ごとに決められるので詳細は省略する。

3-5: 主語選択と態の決定

能動態で生成するか受動態で生成するかはactor, objectの両方のslotがあるときどちらを主語とするかという問題である。

(規則5) 主語は同定度の高い方を選択する。

たとえば名前が指定されていれば同定度は高い。図2の例に対してはactor(“2 criminals”)は名前がないのでobjectより同定度が低い。)

3-6: 名詞(句)の生成

名詞句は一般的に次のような構造を持つ。

(前位修飾語) (主要語) (後位修飾語)

一般の名詞句の内部構造の決め方については、一般の英文法の規則に従う[1, 2]。以下では特に人間に関する名詞句の生成の特徴について述べる。

(規則6) 主要語には、有名人はnameを用いる。(人名の綴りと有名人か否かは問題ごとに指示する。) それ以外の場合の主要語は、第一パラグラフにおいては名前以外のslot情報(職業、性別など)を用いる。第一パラグラフで再び生成するときは原則として代名詞化する。第二パラグラフの第一回目に生成するときは代名詞化せずにnameがあればそれを主要語とし、一般に一回目の主要語を修飾語として組合せる。

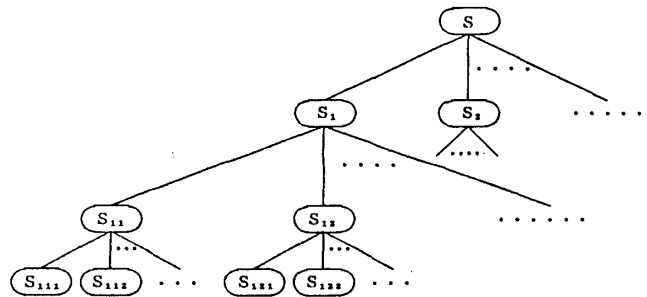


図4 中間表現構造

(各nodeはsceneまたはeventを表わす。下位nodeは上位nodeの構成要素を表わす。)

A Spanish ambassador who had been kidnapped by 2 criminals in the afternoon on the 10th west Beirute was released safe by them at night 3 hours after kidnapping.

The ambassador *pedro was conveyed by car blindfolded about 2 kms and he was confined inside a building by them.

図5 図2の記事から得られる出力文章

参考文献

[1] Winograd, T., Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, MIT, 1971
 [2] Longman Dictionary of Contemporary English, Longman, 1978
 [3] 石崎俊・井佐原均・横山晶一・内田ユリ子, 文脈情報を用いた機械翻訳システムCONTRASTの特徴, 第32回情報処理学会全国大会, 1986
 [4] 井佐原均・石崎俊・半田剣一・橋田浩一, 文脈理解のための拡張ソース知識表現法, 第33回情報処理学会全国大会, 1986