パターン処理に基づく情報抽出システムの概要 – MUC7, MET2 参加システム –

福本 淳一, 下畑 光夫, 桝井文人, 佐々木 美樹, 杉尾 俊之
沖電気工業 (株) 研究開発本部 関西総合研究所
{fukumoto,simohata,masui,sasaki,sugio}@kansai.oki.co.jp

1 はじめに

本稿では、Tipster プロジェクト [1] の一環である情報抽出に関する国際会議 [2], Message Understanding Conference (MUC) 7, Multi Lingual Entity Task (MET) 2 に向けた抽出システムの概要について述べる。今回の会議では、抽出のタスクとして NE, CO, TE, TR, ST の5種類のタスクが設定されており、それぞれのタスクについて各参加団体による抽出結果の評価が行われる。

本稿では、各タスクの概要及びそれに向けた抽出システムの情報抽出手法について述べる。抽出は表層パターンによる認識と機械翻訳システムの解析モジュールから得られた構文解析木に対するパターン処理を基にしている。

2 MUC7, MET2 の概要

MUC7, MET2 は、設定されたタスクについて抽出精度を各参加システムについて評価を行う会議である。MUC7 において設定されたタスクは、Named Entity (NE), Coreference (CO), Template Element (TE), Template Relation (TE), Scenario Template (ST) の5つがあり、TR は今回初めて設定されたタスクである。また、抽出対象テキストは、英語の新聞記事が設定されている。MET2 は英語以外の言語について NE のタスクのみが設定されている。今回の会議では、日本語、中国語、タイ語が設定されている。

2.1 評価スケジュール

まず、会議参加の申し込みの後、会議参加者に対して抽出結果評価用のタスク定義、トレーニング用データがリリースされる。ここでのタスク定義は、最終評価用シナリオのコンテキストには直接依存しない一般的なものであり、シナリオについてもいくつかのサンプルという形式でリリースされる。その後、それぞれのタスクについて、よりしぼられた形式でのシナリオ、それに基づくタスク定義、トレーニングデータ、スコアラーがリリースされ、各参加者で評価用のシナリオに合わせるようなチューニングが行われる。最後に、一週間の評価期間 (Dry Run, Formal Run がある) の初日に実際の評価対象のテキストがリリースされ、最終日までに抽出結果を送り返すことで評価が行われる。評価結果としては、precision, recall, F-measure の各評価値がある。なお、評価結果については、どの団体がどのようなスコアであったのかは原則として公開されないとのことである。

今回は、MUC7 の NE task と MET2 の Dry Run, Formal Run についてはそれぞれ 97/9 月末、98/3 月始めに設定され、MUC の NE 以外のタスクについては、Dry Run, Formal Run ともに NE の約一ヵ月後に設定された。

2.2 参加団体

MUC7 への参加を予定している団体[1] および MET2 への参加を予定している団体 (日本語タスクへの参加) は以下の通りである (*は日本からの参加団体)。

[1] MUC7 のホームページ http://www.muc.saic.com/ より抜粋

MUC7

American University in Cairo
BBN
Cymfony Research
Institute of Systems Science
Isoquest
MITRE
National Taiwan University
New Mexico State University/CRL
New York University
*Oki Electric Electric Industry Co., Ltd.
Quinary
SRI
Southeast Asian Software Research Center
TASC
Teragram Corporation
University of Durham
University of Edinburgh
UMIST
University of Manitoba
University of Massachusetts
University of Southern California
University of Pennsylvania
University of Sheffield
University of Surrey
Wayne State University

MET2 (日本語)

BBN
Lockheed-Martin
*NTT Data Corporation
NYU
*Oki Electric Industry Co., Ltd.
SRI
University of North Texas

3 タスク定義

MUC7 は英語の文書を対象に NE, CO, TE, TR, ST の 5 つのタスクについて設定された定義に基づき、それぞれについて抽出結果の評価が行われる。MET2 では、MUC7 の NE と同様のタスク定義が設定されているが、言語による表現上の違いによる部分については別に MET2 のタスク定義で示されている。評価用のテキストは SGML タグ付きの新聞記事であり、テキストの本文、タイトル、日付などの部分が抽出対象となっている。

Named Entity task (NE)

NE タスクでは、テキスト中の固有名及びそれらの省略表現などの要素を抽出する。抽出する要素は、Entity として人名、組織名、場所名を、Times として日付、時間を、Quantity として金額、割合表現を抽出する。抽出された要素は SGML タグとしてテキスト中に埋め込まれる。

Coreference task (CO)

CO タスクでは、テキスト中の代名詞, "the"+名詞, 固有名詞などの名詞のみについての参照関係を抽出する。動詞及び動詞を含むものについては対象外である。参照関係をもつ要素としては、テキスト中で意味的に同一のもので互いに参照しているものがある。抽出された参照関係の情報は NE と同様、SGML タグとしてテキスト中に埋め込まれる。

Template Element task (TE)

TE タスクでは、Entity と Location についての属性的情報を抽出する。Entity としては、組織 (政府関連、一般企業関連、その他の細分類)、人 (軍関連、一般市民の細分類)、乗り物 (空、地上、海の細分類) といった要素があり、これらの要素の属性的情報として、それを記述または参照する名詞句を抽出する。Location としては、City, Province, Country, Region, Water, Airport, その他といったタイプの情報を認識する。各要素に対して Descriptor (要素に関するその性質等の表現を表したもの) 情報も抽出する。抽出の際、意味的に同一の要素はひとつの要素として抽出し、それにともない Descriptor 情報もまとめて抽出する。抽出結果は、BNF 記法で表現される。

Template Relation task (TR)

TR タスクでは、TE で認識された要素 Location (L), Entity (E) 間について location_of(L, E), employee_of(E, E), product_of(E, E) の関係を持つものを抽出する。抽出結果は、TE と同様に BNF 記法で表現される。

Scenario Template task (ST)

ST タスクでは、シナリオとして設定されたイベントの抽出を行う。MUC7 で設定されたシナリオは、航空機の衝突や事故及びそれに関連するものが設定された。抽出する情報としては、事故に関連した航空機、事故の場所、時間、事故の犠牲者、事故調査の状況、事故の原因などがある。それに関連する情報として、航空機の出発、到着時刻などを含むフライト情報などもある。

4 システム概要

今回の会議に我々が参加したタスクは、MUC7 では ST を除くすべてのタスクと MET2 の日本語 NE のタスクである。MUC7, MET2 用の抽出システムとして、機械翻訳システム [3] [4] の形態素·構文解析の処理モジュールをそのままの形式で用い、各解析モジュールにパターン処理を行うモジュールを組み合わせることで抽出システムを実現した。MUC7 用には、英日機械翻訳システムの解析モジュールを、MET2 用には日英機械翻訳システムの解析モジュールを利用した。

図 1 に抽出システムの構成を示す。まず、NE 処理用として、表層レベルのパターンを認識する表層パターン処理部 [5] と形態素·構文解析処理の解析途中の木構造データに対するパターンを認識する構文パターン処理部を設定した。表層パターン処理は Perl で記述し、構文パターン処理は翻訳システム用の文法記述言語を用い、形態素·構文解析処理部にそのまま組み込む形式で実現している。また、一文を超える処理である CO, TE, TR についても翻訳システム用の文法記述言語を用いて抽出規則の記述を行った。CO, TE, TR の各処理部は、一文単位の形態素·構文解析処理の全文の解析結果を入力にして抽出処理を記述している。CO 処理では、NE 処理で認識された要素の情報を用い、代名詞、名詞句の参照関係の解析を行っている。TE 処理では、NE 処理で認識された要素の情報と CO 処理の参照関係の解析情報から要素の属性的情報の抽出を行っている。TR 処理では、TE 処理で認識された要素の情報を基に要素間の関係の認識を行っている。

4.1 NE 処理

NE 処理では、入力テキストの SGML タグから抽出対象の部分をとりだし、表層パターン処理を行い、認識された要素に対し NE タスク定義で設定されたタグの付与を行う。ここでの要素の認識は、接辞 (Mr., Dr. など)、機能語 (Bank, Association など)、固有名リスト (Boeing, NASA など) を利用している。また、固有名については認識された要素からの省略形の自動認識も行っている。

次に、表層パターン処理で付与された NE タグを形態素·構文解析用の構文木に取り込むためのタグ処理を行った後、形態素·構文解析処理部で文の構文構造の認識を行う。構文構造レベルで認識されるパターンとしては、括弧による挿入句や特定の動詞の主語などがある。最後に、認識された要素に対して原文中に NE タグを付与することにより NE 評価用の出力を得る。

4.2 CO 処理

CO 処理では、NE と同様に入力テキストの SGML タグから抽出対象の部分をとりだし、表層レベルでのパターン処理と形態素·構文解析処理部で文の構文構造の認識から NE で認識される要素の認識を行う。表層パターン処理における人名の認識では Mr. Bean の様に認識された要素の一部の要素である Bean が再びテキスト中に現れた時、それも人名として認識を行う処理を利用し、表層パターンレベルでも CO の処理を一部行っている。

形態素·構文解析処理の後、解析結果の全文を CO 処理の入力とする変換を行った後、CO 処理を行う。CO 処理としては、代名詞、“the”+名詞などの参照関係の解析を順次行った後、解析結果の構造から CO タグ情報を原文にタグとして付与することにより CO 評価用の出力を得る。参照関係の解析は、照応語を含む文内では文を右から左へと先行詞を探索し、その前文では文頭から探索している。

4.3 TE 処理

TE 処理では、認識する要素の属性情報として NE 処理、CO 処理の出力結果を用いている。要素の認識は、NE の解析結果を用い、それに関する文内の属性は属性記述のパターンを構文解析結

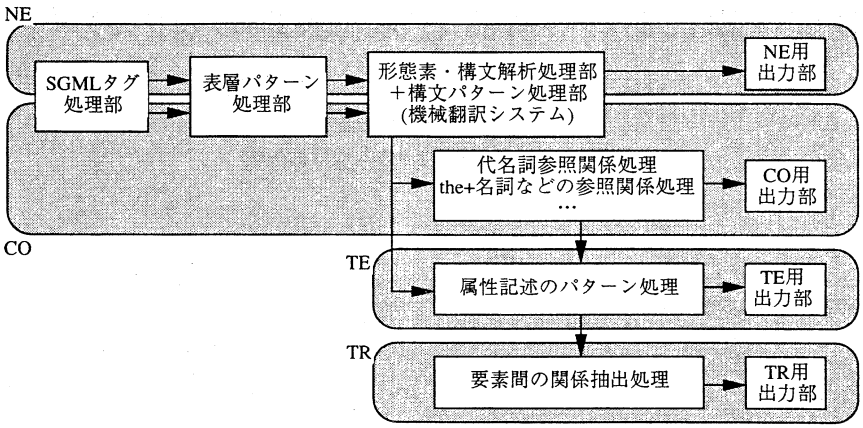

図 1: 情報抽出システムの構成

果に適用することで抽出を行う。また、文書内の同一要素の認識は、CO 処理の解析結果から参照関係を利用し、文書内に現れる同一要素の属性表現を認識する。最後に、抽出結果は出力部によって BNF 記法に変換され、評価用データとなる。

4.4 TR 処理

TR 処理では、TE 処理で認識された要素間の関係として location_of, employee_of, product_of の各関係を認識する。TE で認識された要素について、これらの関係を表す表現パターンとの照合および要素に関する辞書的知識 (例えば、組織の場所や業種などの知識) も利用することにより要素間の関係を抽出する。抽出結果は、TE と同様に BNF 記法に変換され、評価用データとなる。

5 おわりに

本稿では、MUC7, MET2 の会議の概要とそこで設定された各抽出タスク及び抽出システムの概要について述べた。また、各会議に向けた我々の抽出システムの概要についても述べた。情報の抽出は、表層パターンによる認識と、機械翻訳システムの解析モジュールに抽出パターン処理を埋め込むことで実現した。表層パターン処理以外の抽出規則は、機械翻訳システム用の文法記述言語で記述した。

MUC7, MET2 の会議は、1998.4.29 - 5.1 に開催予定であり、会議終了後はそこで行われた評価についても報告の予定である。

参考文献

[1] TIPSTER TEXT PROGRAM Phrase II, DARPA, (1996).

[2] Proceedings of 6th Message Understanding Conference (MUC-6), DARPA, (1995).

[3] Masui, F., Tsunashima, T., Sugio, T., Tazoe, T. and Shiino, T.: "Analysis of Lengthy Sentences Using an English Comparative Structure Model", System and Computers in Japan, pp.40–48, SCRIPTA TECHNICA Inc., (1996).

[4] 北村, 甲斐, 岡田, 永田："拡張性を重視した日英機械翻訳システム", 電子情報通信学会技術報告 NLC, No.91-24 (1991).

[5] 下畑, 福本, 杉尾："パターンと構文情報による固有名の情報抽出 – MUC7, MET2 参加システム –", 「テキスト要約の現状と将来」言語処理学会第 4 会年次大会併設ワークショップ予稿集, (1998).