自然言語処理 126－15
(1998. 7. 24)

固有名詞抽出における日本語と英語の比較

福本淳一, 下畑光夫, 桝井文人
沖電気工業(株) 研究開発本部 関西総合研究所
〒540-6025 大阪市中央区城見 1-2-27 クリスタルタワー
e-mail: {fukumoto, simohata, masui}@kansai.oki.co.jp

あらまし

本稿では、7th Message Understanding Conference (MUC-7) において報告した日本語と英語の固有名詞抽出システムの抽出精度の比較結果について報告する。システムの比較は、抽出を行った抽出項目と抽出のための各モジュールの観点から行った。

まず、各システムの構成と抽出方式について述べた後、MUC-7 の formal run テストで用いた新聞記事 100 テキストを用いた場合のシステムの精度について述べる。最後に抽出結果から両システムの比較について述べる。

キーワード

情報抽出, パターン認識, 形態素解析, 構文解析

Comparison of Japanese and English Named Entity Recognition

Jun'ichi Fukumoto, Mitsuo Shimohata and Fumito Masui
Kansai Laboratory, R&D Group, Oki Electric Ind. Co., Ltd.
Crystal Tower 1-2-27 Shiromi, Chuo-ku, Osaka 540-6025 JAPAN

Abstract

In this paper, we discuss comparison between Japanese and English Named Entity (NE) Recognition System which have been submitted to the 7th Message Understanding Conference (MUC-7). The comparison is done on the point of the kinds of NE types and recognition modules.

Firstly, we will show system architecture and recognition mechanism of both NE systems. Then we will report system performance of the NE systems using 100 texts which are utilized in the formal run test of MUC-7 evaluation. Finally, we will discuss the comparison of the systems.

Keywords

Information Extraction, Pattern Recognition, Morphological Analysis, Syntax Analysis

1 はじめに

我々は、1998年4月に開催された情報抽出に関する国際会議 Seventh Message Understanding Conference, MUC-7, (Tipster プロジェクト [1][2] の一環として開催) に参加し、日本語と英語の抽出システムについて報告を行った [3][4]。参加したシステムは、日本語の Named Entity (NE) タスク (Multilingual Entity Task, MET-2, として日本語の他に中国語が設定) および英語については Named Entity (NE), Coreference (CO), Template Element (TE), Template Relation (TE) 各タスクである [5][6][7]。固有名詞の認識タスク (NE) の処理を行うシステムとしては、今回、日本語と英語の両システムを開発し、それぞれ評価を行った。MUC-7 において設定されたタスクでの固有名詞の認識とは、指定された新聞記事より人名、組織名、場所名、日付、割合の各要素に相当する固有名詞の抽出を行うものである。

日本語、英語の各抽出システムは、言語による違いはあるがほぼ同様の処理構成をとっている。まず、表層レベルの情報に対するパターン処理によって固有名詞要素の認識を行い、その結果を SGML タグとしてテキストに付与する。次にこの SGML タグ付きテキストの形態素解析処理、構文解析処理を行い、その解析木に対するパターン処理によって固有名詞要素の認識を行う。認識された要素は、解析木に値を付与することにより表現される。最後に、原文中の認識された要素を SGML タグ付けすることにより評価用のテキストを得る。

本稿では、固有名詞抽出において、日本語と英語に対して同様の抽出方式をとった場合の抽出精度の違いについて述べる。まず、各言語における固有名詞の抽出システムについて述べた後、同様の構成をとった場合における日本語と英語の抽出精度と抽出に用いる辞書や規則等がそれぞれどのような貢献をしているのかについて各システムを比較検討する。最後に、日本語と英語の固有名詞抽出システムの違いについて議論する。なお、抽出結果の評価として用いたデータとしては、MUC-7, MET-2 の formal run で用いた新聞記事 100 文書を利用した。

2 日本語 NE システム

図1に示すように、日本語 NE システムは表層パターン認識処理部、SGML タグ処理部、構造パターン認識処理部、抽出フィルタの各部から構成されている。

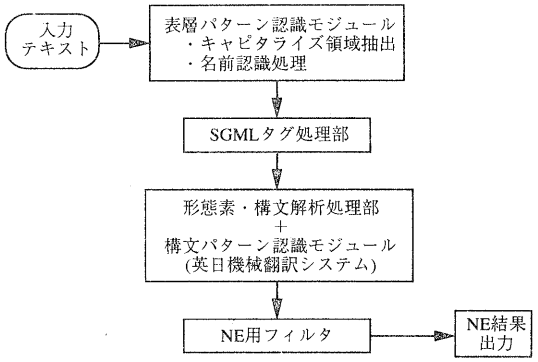

図 1: 日本語 NE システムの構成

入力されたテキストは、まず、表層レベルの処理として、平仮名、漢字、数字などの文字種の情報によって要素に分割される。分割された要素は単語リストや接辞リストなどの情報を用いることにより名前の認識が行われ、その要素に SGML タグが付与される。次に、テキストの各文の形態素、構文解析が行われ、得られた構文木に対して構文レベルのパターンを適用することで名前の認識を行う。表層レベルの認識処理およびテキストの情報として予め付与されている SGML タグの情報は、構文木の内部状態であるノード属性として取り込まれており、ここで新たに認識された名前についても同様にノード属性が付与される。最後に、構文木のノード属性を用いて、認識された名前に対するタグ付けを行う。

2.1 文字種による要素分割

表層レベルの処理として文字種によって要素への分割を行う。分割は、まず、平仮名や句読点でない文字の2つ以上連続するものを取りだし、さらに、あらかじめ登録されたいくつかの平仮名機能語、固有名詞、いくつかの記号の連続するものについては要素としてひとまとまりにする。以下に例を示す。例では、記号 "≪", "≫" で囲まれたものが切り出された要素である。

≪日本人観光客17人≫の乗った≪マイクロバス≫が≪トラック≫と≪衝突≫、≪静岡県沼津市≫の≪原山行雄さん≫(≪59≫)が≪死亡≫、≪5人≫が≪重傷≫、≪3人≫が≪軽傷≫を負った。

≪原山行雄さん≫は、漢字列≪原山行雄≫と平仮名機能語≪さん≫によって得られている。また、一文字漢字については、ここでは抽出されず、以下で述べるヒューリスティックによって抽出される。

固有名詞リストとしては、≪東京·霞が関≫が、固有名詞リストとして登録されている≪霞が関≫によって抽出される。≪和歌山県伊都郡かつらぎ町中飯降≫についても≪かつらぎ町≫という固有名詞リストによって抽出される。

2.2 分割要素からの名前の認識

分割された要素からNE要素を認識するため、表1に示された情報を用いることで要素の再分割·統合が行われる。

表1: 日本語、英語NE要素の認識用情報

リスト名	要素数	例
人名リスト (*pn*)	58	高橋, 斉藤
組織名リスト (*on*)	255	住友, 三菱
地名リスト (*ln*)	10083	大阪, 山口
人名接辞リスト (*ps*)	90	さん, 被告, 社長
組織名接辞リスト (*os*)	155	社, 株式会社
地名接辞リスト (*ls*)	66	市, 町, 村
組織修飾語リスト (*onm*)	124	銀行, 電気, 産業
人名ストップワード (*swp*)	48	農業, 大学生
組織名ストップワード (*swo*)	37	大手, 主要

図2に分割·統合の規則を示す。図中の規則は(1)から順に適用される。

(1) $pn + ps \rightarrow$ 人名 (p) + ps
(2) $on + os \rightarrow$ 組織名 (o)
(3) $ln + ls \rightarrow$ 地名 (l)
(4) $ln + ps \rightarrow$ 人名 (p) + ps
(5) $pn + onm^* \rightarrow on$
(6) $on + onm^* \rightarrow on$
(7) $ln + onm^* \rightarrow on$
(8) $on \rightarrow$ 組織名 (o)
(9) $ln \rightarrow$ 地名 (l)

図2: 分割·統合規則

規則(1)(2)(3)では、各NE要素のタイプはその名前と接辞によって認識されることを示している。規則(4)においては、地名リスト要素であっても人名接辞を取る場合には、人名と認識されることを示している。これは、地名も人名も同様の名前が多く用いられるためである。規則(5)(6)(7)では、組織名修飾語が名前と同時に現れる場合には、全体として組織名の候補となることを示している。規則(8)(9)は、組織名、地名として認識されないための要素が現れた場合の規則を示している。

2.3 名前認識のためのヒューリスティックス

前節の基本的な認識規則に加えて以下のようなヒューリスティックスがNE要素の認識に用いられる。

1. 地名の前部にある“東”, “西”, “南”, “北”は地名と結合される。
2. “前”, “新”, “元”, “現”, “副”と限定された人名接辞(大統領、首相、監督、大臣、社長など)が連続する場合、全体として人名接辞となる。
3. “月”, “日”が分割処理の後、独立した要素である場合、それらは場所として認識される(定義から衛星名は場所名である)。
4. “日”, “米”, “英”のような一文字国名が連続して現れる場合、それぞれ場所名として認識される。また、「と」「、」を挟んで連続する場合もそれぞれ場所名として認識される。
5. 人名接辞を伴わない人名リスト要素は、人名として認識されない。
6. 「(同|両|各)+接辞」の場合、接辞情報を無効にし名前と認識しない。
7. 人名接辞「さん」や「ちゃん」の前に名前以外の要素が存在する場合、その人名接辞「さん」「ちゃん」を処理対象の接辞から除去する。
   - 例:「(父|母|赤)ちゃん」など
8. 「地名+未知語+数字+“−”+数字+“−”+数字」は全体として地名と認識する。
   - 例:「< l大阪市>< l中央区>< l城見1-2-27>」(ここで、「城見」は地名として登録されておらず、また、数字とハイフンの組み合わせは住所の一部として解釈される。)

9. 「カタカナ未知語+一部の接辞」はカタカナ未知語が接辞と連続する場合、全体として固有名として認識する。

   - 人名： 容疑者、被告、大佐など（社長や監督は除外）
   - 地名： 区、湾、県、市、町、など（村は例外が多いので除外）
   - 組織名：社、銀行、空港など

表2に、分割要素からの名前認識の例を示す。表中の各タグの名前は、表1にイタリックで示された要素名である。例えば、"< *pn* xxx>" では、xxx が人名リストの要素であることを示す。

2.4 構文解析結果による日本語 NE 要素の認識

表層レベルで認識された各 NE 要素は、テキスト中のその部分にSGML タグを付与することで示されている。この SGML タグ付きテキストを形態素、構文解析するため、これらのタグ情報は構文解析の構文木のノード属性として表現される。同様に構文解析結果に対して認識された要素もノード属性として表現される。

構文解析結果を用いた NE 要素の認識処理は構文木の構造に対するパターンマッチの規則を文法記述言語で記述し、それらを構文解析と同様の枠組で処理することによって認識される。構文レベルの抽出規則では、構文解析で用いられる意味情報も用いることで規則が記述される。例えば、人間の意味を持つ単語間に未知語が存在した場合、それらはまとめて人名として認識することができる。以下の文では、"中村" は人名であり "副操縦士" は人間の意味を持つ単語であることから、全体として "中村貴洋副操縦士" は人名であると認識される。ここで、"貴洋" は人名として登録されていないため未知語となっている。

> 中村貴洋副操縦士(30)を起訴猶予処分にした。

2.5 SGML タグ処理

構文レベルのパターン処理の後、構文解析木において名前であると認識された要素のノード属性情報から、入力テキストに対してSGML タグを付与することによって評価の対象となるタグ付きテキストを得る。この処理モジュールは Perl によって実現した。

3 英語 NE システム

英語 NE システムは、日本語 NE システムと同様の構成をとっている(図3)。

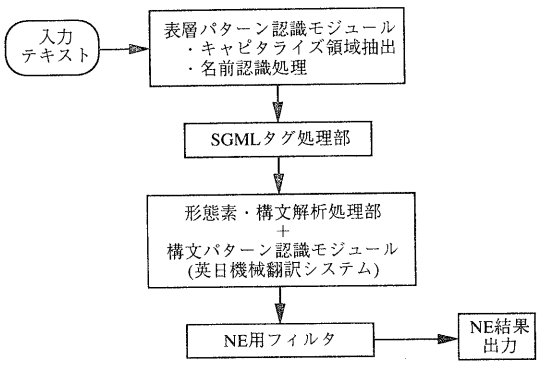

図3: 英語 NE システムの構成

まず、入力されたテキストからキャピタライズ領域を抽出し、それらからいくつかの単語リストと機能語リストを利用することで NE 要素の認識を行う。また、一旦認識された要素のいくつかについては照応処理を行うことで機能語を含まない部分についての認識も行っている。以上の表層レベルで認識された要素には SGML タグが付与される。次に、テキストの各文の形態素、構文解析が行われ、得られた構文木に対して構文レベルのパターンを適用することで NE 要素の認識を行う。ここでの SGML タグ処理については、日本語 NE システムと同様の処理が行われ、最後に、構文木のノード属性を用いて、認識された名前に対するタグ付けを行う。

3.1 キャピタライズ領域の抽出

表層レベルの処理として固有名詞を表すキャピタライズ化された領域の抽出を行う。ここで、キャピタライズ化された連続した単語が一つの領域として抽出される。但し、省略形を示すピリオド以外のピリオドを文末であると判断する文区切り処理により、文頭語が領域から削除される。また、文頭の接続詞や冠詞なども領域から削除される。ここで抽出された領域が NE 要素の認識のための元データとなる。その他、以下のようなキャピタライズ領域の修正も

表 2: 日本語 NE 要素の認識例

原形	分割形	認識結果
静岡県沼津市	< *ln* 静岡>< *ls* 県>< *ln* 沼津>< *ls* 市>	< *l*静岡県>< *l*沼津市>
大阪府出身	< *ln* 大阪>< *ls* 府>出身	< *l*大阪府>出身
北京	< *ln* 北京>	< *l*北京>
広瀬さん	< *pn* 広瀬>< *ps* さん>	< *p* 広瀬>< *ps* さん>
越田典子さん	越田典子< *ps* さん>	< *p* 越田典子>< *ps* さん>
原山行雄さん	< *ln* 原山>行雄< *ps* さん>	< *p* 原山行雄>< *ps* さん>
日本大使館	< *ln* 日本>< *os* 大使館>	< *o* 日本大使館>
セイコー電子工業	< *on* セイコー>< *onm* 電子>< *onm* 工業>	< *o* セイコー電子工業>
欧州連合	< *ln* 欧州>< *onm* 連合>	< *o* 欧州連合>
花巻空港	< *ln* 花巻>< *os* 空港>	< *o* 花巻空港>
交通事故	< *onm* 交通>< *onm* 事故>	交通事故
交通安全委員会	< *onm* 交通>< *onm* 安全>< *os* 委員会>	< *o* 交通安全委員会>
米道路交通安全局	米< *ls* 道路>< *onm* 交通>< *onm* 安全>< *os* 局>	< *o* 米道路交通安全局>

行われる。

1. 小文字でも固有名に入ることがある語はキャピタライズ領域として扱う。
   - 例：“von”, “de”
2. キャピタライズ領域末尾に付く「’s」「;」「:」は領域から除く。
   - 例：<John Dow’s> → <John Dow>’s

3.2 NE 要素の認識

領域の統合

いくつかの機能語や “for”, “of” などの前置詞情報を用いることで領域の統合を行う。例えば、機能語 “Bank” と前置詞 “of” は、NE 要素 “Bank of Tokyo” の認識に用いられれ、機能語 “University” と前置詞 “of” NE 要素 “University of Tokyo” の認識に用いられる。

また、キャピタライズ領域直後に “airport”, “university”, “sea” などのように属性を決定する語がある時は、その語を領域と統合する。但し、最終出力では、その語は領域外へ出される。

NE タイプの認識

NE 要素のタイプの認識は、領域に存在する各タイプの接辞情報によって判断する。例えば、“Mr.” は人名の認識に、“Bank” は組織名の認識に、“City” は地名の認識に用いられる。

また、以上のような基本的な接辞によるタイプの認識処理に加えて、次のようなヒューリスティックスが NE 要素の認識に用いられる。

1. 「未知語’s (mother | father | sister | brother)」の場合、未知語を人名とする。ここで未知語とは固有名の単語リストとして登録されていないものを示す。
2. 人名特有の表記方法による認識として、「John F. Kennedy」のようにキャピタライズ領域が 3 語で構成され、2 語目が英大文字＋ピリオドである場合には、この領域を人名と認定する。
3. 人名特有の表記による認識として、キャピタライズ領域の最後の語が「Jr.」であれば、この領域を人名とする。
4. 人名特有の語による認識として、キャピタライズ領域中に「Von」か「von」があれば、この領域を人名とする。

辞書引き

以上の認識処理で残った領域の部分に対して各タイプの単語リストを利用することで NE 要素の認識を行う。認識のための各タイプの単語リストは新聞記事、地図インデックス、会社情報などから人手によって作成した。また、キャピタライズ領域とそうでない部分の連続する以下のような場合、それらは個別に登録することで、NE タイプの認識を行う。

「United States of America」

「National Aeronautics and Space Administration」

表3に候補の領域に対して NE 要素を認識するために用いる各情報をまとめたものを示す。

表 3: 英語 NE 要素の認識用情報

リスト名	要素数	例
人名リスト (姓, 名) (*pn*)	340	John, Ford
組織名リスト (*on*)	64	Microsoft, IBM
地名リスト (*ln*)	795	Washington
人名接辞リスト (*ps*)	54	Mr., Dr.
組織名接辞リスト (*os*)	49	Ltd., Bank
地名接辞リスト (*ls*)	82	City

照応表現の認識による NE 要素認識

以上の表層レベルの NE 要素の認識処理が終了した後、それまでに認識された NE 要素となっている単語情報が繰り返し現れていたものを同じタイプの NE 要素であると認識する。例えば、"Mr. John Doe" が一旦認識されると、それ以降のテキスト中で "John" や "Doe" が現れた場合、それらも "Mr. John Doe" に対して付与された NE 要素のタイプ名が付与される。また、"FAA" のように認識された組織名から先頭の文字を組み合わせることによって自動的に生成される省略表現についても同様に認識される。

また、認識固有名の再利用において、「非地名＋組織接辞」で認識した組織は、非地名部分についても繰り返し要素の認識に用いられる。例えば、「Valujet Airlines」の "Valujet"、「Ford Motors」の "Ford" がこれに該当する。しかし、「American Bank」の場合、"American" は地名として登録されているため "American" がテキスト中で繰り返し現われていても NE 要素とは認識しない。

3.3 構文解析結果による英語 NE 要素の認識

表層レベルで認識された各 NE 要素のタグ情報は、日本語の場合と同様に構文解析の構文木のノード属性として表現される。構文解析結果に対して認識された要素も同じくノード属性として表現される。

構文木に対して適用される NE 要素認識のための抽出パターンとしては以下のものがある。

- "say", "die", "play" などのような動詞句の主語の名詞句は人名である。
- 関係代名詞 "who" の前にある名詞句は人名である。
- 埋め込み句として人の性質をとる名詞句は人名である。また、その逆も同じ。
- "employee", "spokesman" などの語句の前に続く名詞句は組織名である。
- "in", "at", "near", "over" などの前置詞を伴う名詞句で埋め込み語句が組織名であるものは組織名である。

3.4 SGML タグ処理

日本語の場合と同じく、構文解析木において名前であると認識された要素のノード属性情報から、入力テキストに対して SGML タグを付与することによって評価の対象となるタグ付きテキストを得る。解析対象となるテキストのヘッダ (全文字がキャピタライズ) の部分については、テキスト本文でタグ付けされた要素と同じものであれば、それに対して同じタイプの NE 要素であると認識する。

4 日本語、英語 NE システムの比較

MUC-7, MET-2 の formal run で用いた新聞記事 100 文書についての日本語 NE システム、英語 NE システムの評価結果 (再現率、適合率) を抽出項目別にまとめたものを表 4に示す。再現率、適合率はそれぞれ以下の式で与えられる。

$$\text{再現率} = \frac{\text{システムによる正解数}}{\text{全正解数}}$$

$$\text{適合率} = \frac{\text{システムによる正解数}}{\text{システムの全出力数}}$$

4.1 抽出項目による比較

各システムの抽出項目別の f-measure の値をまとめたものを表 5に示す。f-measure は以下の式で与えられる。

$$\text{f-measure} = \frac{2 * \text{再現率} * \text{適合率}}{\text{再現率} + \text{適合率}}$$

表 5 から分かるように、日本語 NE システムと英語 NE システムを比較すると日本語システムの方が良い結果を得ている。また、各項目について比較す

表 4: 日本語、英語 NE 評価結果 (再現率、適合率)

抽出項目名	日本語 NE システム				英語 NE システム			
	再現率		適合率		再現率		適合率	
組織名	70	(444/637)	97	(444/456)	58	(1061/1839)	93	(1061/1147)
人名	50	(70/139)	99	(70/71)	82	(725/884)	94	(725/775)
場所名	93	(847/910)	98	(847/863)	90	(1183/1310)	90	(1183/1311)
日付	97	(540/558)	97	(540/556)	89	(1075/1210)	96	(1075/1124)
時間	92	(110/120)	100	(110/110)	81	(154/190)	97	(154/158)
金額	100	(71/71)	97	(71/73)	93	(200/215)	97	(200/206)
割合	100	(42/42)	95	(42/44)	100	(100/100)	95	(100/105)
計	85	(4210/4954)	97	(4210/4346)	77	(8887/11496)	92	(8887/9652)

表 5: 日本語、英語 NE 評価結果 (f-measure)

抽出項目名	日本語 f-measure	英語 f-measure
組織名	81.2	71.1
人名	66.7	87.4
場所名	95.5	90.3
日付	96.9	92.1
時間	95.7	88.5
金額	98.6	95.0
割合	97.7	97.6
計	90.5	84.1

ると、場所名、日付、金額割合についてはほぼ同等の結果を得ているが、組織名、人名、時間については言語による差が出ている。人名については英語が良い結果を得ており、組織名と時間については日本語がよい結果を得ている。これは、今回のシステムでは日本語の人名の登録はあまり行っておらず主に接辞による認識に依っており、実験したテキストの打ち上げのイベントについては「～宇宙飛行士」などのように一般にはあまり現われない接辞情報が存在したためである。しかし、英語ではこのような表現は修飾語句によって表現され、接辞レベルの情報では一般的な形態で表わされているものが多く存在した。英語の組織名は人の名前の組み合わせなどバリエーションが多く、それらのパターンを捉える規則が不足していたためである。これは時間表現についても同様で、英語では日本語よりも時間表現のバリエーションが多かったためである。

4.2 抽出モジュールの比較

各 NE システムは図 1、図 3 に示す通り、表層パターン認識モジュール、構文パターン認識モジュール、NE 用フィルタの各部によってタグ付けが行われている。英語 NE タスクの場合、テキストのヘッダ部分はすべての文字がキャピタライズされており、このため現在の認識方法ではタギングができない。そこで、英語 NE システムではテキスト中のタグ付け情報を基に、NE フィルタでヘッダへのタギングを行っている。一方、日本語 NE タスクではこのような処理はなく、認識モジュールでヘッダに対するタギングも行っている。

以上のように各言語による認識方法の違いは存在するが、NE 要素のタグ付けにおいて各モジュールがそれぞれどの程度の貢献を行っているのかをまとめたものを表 6に示す。表 6では、表層パターンでは表層パターン認識モジュール終了後までの結果を、構文パターンでは構文パターン認識モジュール終了後までの結果を、NE 用フィルタではフィルタ処理によるタグ付け終了後までの結果を示している。

日本語 NE システムにおいては、構文パターンによって NE 要素を認識することはできなかった。この理由の一つとして、システムで利用した構文解析モジュールがトランスファー方式の機械翻訳システムの解析モジュールを用いており、特に日本語の構文解析において解析の途中段階で現言語の構造がターゲット言語の構造に変換されるため、構造パターンをうまく捕えることができなかったことがある。

英語 NE システムにおいては、構文パターン認識モジュール終了後までの正解数 8024 件のうち 7840 件が表層パターンで認識されたもので、残りの 184 件が構文パターンによって認識されたものである。

このように構文パターンによる貢献は少いが、こ

表 6: モジュール毎の評価結果

モジュール名	日本語 NE				英語 NE			
	再現率		適合率		再現率		適合率	
表層パターン	85	(4210/4954)	97	(4210/4346)	68	(7840/11478)	96	(7840/8186)
構文パターン	85	(4210/4954)	97	(4210/4346)	70	(8024/11478)	95	(8024/8414)
NE 用フィルタ	85	(4210/4954)	97	(4210/4346)	77	(8887/11496)	92	(8887/9652)

れも日本語と同様にトランスファー方式の機械翻訳システムの解析モジュールを用いたことも原因として考えられる。しかし、3.3 節でも述べたように、ある特定の用言の主語であるといった情報や関係代名詞節などの埋め込み句の情報は構文解析結果がなければ得られない情報であり、接辞などの表層的な手掛かりが存在しない場合に時に有効である。英語 NE システムにおいて、表層パターン、構文パターンの各抽出項目ごとの評価値 (f-measure) を表 7に示す。表からわかるとおり構文解析の情報は実際に人名の認識において大きく貢献していることがわかる。金額の項目については、pound の重さと金額の違いを認識するための規則が当てはまった結果である。

表 7: 英語 NE 項目別評価結果 (f-measure)

抽出項目名	表層パターン		構文パターン	
組織名	62.0	(839)	62.4	(854)
人名	81.9	(621)	85.0	(666)
場所名	84.1	(971)	84.8	(988)
日付	92.1	(1075)	92.1	(1075)
時間	88.5	(154)	88.5	(154)
金額	92.7	(191)	95.3	(201)
割合	97.6	(100)	97.6	(100)

( ) 内は正解数を示す。

5 おわりに

本稿では、MUC-7, MET-2 で報告を行った日本語、英語の NE システムの構成と MUC-7, MET-2 の formal run で用いた新聞記事 100 文書を利用してそれぞれのシステムの抽出精度の違いについて述べた。評価では、それぞれのシステムについての抽出項目による精度の違い、および、言語による抽出モジュールの違いについて比較を行った。

今後は、さらに大規模な評価を行いながらシステムの改良を行い、テキストの分野による違いについても考察する予定である。また、これらのシステムを対訳辞書の作成にも利用することも考えられる。

参考文献

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

[2] TIPSTER TEXT Phrase III 18-Month Workshop, DARPA, 1998.

[3] Fukumoto, J., Masui, F., Shimohata, M. and Sasaki, M., "Oki Electric Industry : Description of the Oki System as Used for MUC-7", Seventh Message Understanding Conference (MUC-7), DARPA, 1998.

[4] Fukumoto, J., Shimohata, M., Masui, F. and Sasaki, M., "Oki Electric Industry : Description of the Oki System as Used for MET-2", Seventh Message Understanding Conference (MUC-7), DARPA, 1998.

[5] 福本, 下畑, 桝井, 佐々木, 杉尾 : "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 –", 言語処理学会第 4 回年次大会発表論文集, pp.230–233, 1998.

[6] 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 回年次大会併設ワークショップ予稿集, pp.44–49, 1998.

[7] 桝井, 福本 : "新聞記事からの要素間関連情報の抽出", 信学技報 NLC98, 1998.