自然言語処理　114－12
（1996. 7. 18）

英語テキストからの情報抽出

ＭＵＣ第６回大会の参加報告

若尾 孝博

シェフィールド大学コンピューター・サイエンス学部

［１９９６年滞在先］日本電気（株）情報メディア研究所　，音声言語研究部
〒２１６神奈川県川崎市宮前区宮崎４丁目１－１
Tel：044-856-2152　Email：*wakao@hum.cl.nec.co.jp*

あらまし　電子化されたテキストからの重要な情報を抽出する技術は近年米国を中心に盛んに研究されて来ている。この報告では米国のＡＲＰＡが支援するＭＵＣ(Message Understanding Conference)で行われて来た情報抽出研究の過去９年間の移り変わりと最新のＭＵＣ(第６回大会、９５年１１月)の研究成果について詳しく紹介する。これまでの大会では、予め定められたテンプレートを埋めることが情報抽出作業の中心であったが、第６回大会では作業が４種類に分割され、各作業別にシステムの評価が行われた。

キーワード　情報抽出、固有名詞抽出、照応関係認定、ＡＲＰＡ、ＭＵＣ

Information Extraction from English Text

Report on the sixth Message Understanding Conference

Takahiro Wakao

University of Sheffield, Computer Science Department

[visiting] NEC Corporation, Information Technology Research Laboratories,
Human Language Research Laboratory
1-1, Miyazaki 4-Chome Miyamae-ku, Kawasaki, Kanagawa, 216, Japan
Tel：044-856-2152　Email：*wakao@hum.cl.nec.co.jp*

Abstract

Information extraction (IE) has been actively researched in recent years in the United States. In this report, an ARPA- supported US project, the Message Understanding Conference (MUC) is introduced including its 9-year history, and the results of the latest MUC convention (the 6th convention, November 1995) are reported in detail.

key words　Information extraction, Proper name extraction, Coreference resolution, ARPA, MUC

1 はじめに

近年新聞記事など大量のテキストが電子化されてきているのに伴い、電子化テキストから重要な情報だけを抽出したいという要求が高まって来ている。従来の情報検索(Information Retreival)は、大量のドキュメントからユーザーのクエリー(問い合わせ)に適合するドキュメントを取って来るものであるが、それに対して情報抽出(Information Extraction)では、取るべき情報を予め定めて、インプットされたテキストからその情報を抽出して来ることになる。取るべき情報としては、テキスト中の

- 事実(facts)
- 事実間の関係 (relations among the facts)

が中心となる。

例えば、新聞での人事異動に関する記事を対象とする情報抽出では、認定すべき事実を、会社名、人名、地名、役職名、年月日とし、認定すべき事実関係としては、会社と人名の関係、特に、誰が特定の役職を辞め、誰が新しくその役職に就いたか、又その日時などと定めることが出来る。情報抽出システムのアウトプットとしては、これらの事実と事実関係を示すテンプレートと呼ばれる関係データーベース (relational database)でのレコードのような形にすることが一般的である。(具体例は後出のテンプレートを参照のこと)

2 MUCとは

米国においては Tipster([1]), Murasaki, MUC ([2] [3]) と言った情報抽出のプロジェクトがあるが、その中でも最も長期にわたり情報抽出に関する研究を押し進めて来ているのがMUC (Message Understanding Conference) と呼ばれる大会である。この大会の特徴は以下の通りである：

- 米国国防省の一機関であるARPA (Advanced Research Projects Agency) により運営、支援されている
- 従来の学会ではなく、情報抽出システムの評価とフレンドリーな競争を行う
- 情報抽出システムの開発の促進とテスト、評価のための共有出来る資源を増やすことを目的としている
- 9年前の87年から始まり、長期間にわたり情報抽出の研究を進めている

3 MUCの歴史

MUCは過去6回開催された。対象となるテキストは第1・2回大会では海軍関係の短いメッセージ、第3・4回は中南米でのテロリストの活動に関する記事、第5回は企業の合弁・提携に関するビジネス分野とコンピューターのチップ製造に関するマイクロ・エレクトロニクス分野、2分野での新聞記事が対象になった。第5回大会では対象となる言語がそれまでの英語だけから日本語にも拡張され、日本語新聞記事からの情報抽出研究が進められた([2])。最新の第6回では、言語は英語のみとなり、対象は企業での人事異動の記事であった([3])。第1回～6回大会までをまとめると表1のようになる。

大会	開催時期	対象テキスト	言語
第1回	1987年	海軍メッセージ	英語
第2回	1989年	海軍メッセージ	英語
第3回	1991年	テロリスト記事	英語
第4回	1992年	テロリスト記事	英語
第5回	1993年	合弁・提携記事	英・日
		チップ製造記事	英・日
第6回	1995年	人事異動記事	英語

表 1: MUC過去6回の大会のまとめ

4 MUC6

第1回～5回大会までは、作業内容は対象となるテキストから予め定められたテンプレートを埋めるという内容であったが、前回の第6回大会(MUC6)から作業内容が4つに細かく分かれた。その4つの作業は以下の通りである：

- Named Entity
  固有名詞(組織名、人名、地名)、
  時間表現(曜日、年月日等)、
  数表現(金額、パーセント)の認定
- Coreference
  照応関係の認定
- Template Element
  MUC5のテンプレートより規模を縮小
  組織名、人名、製品名の抽出
- Scenario Template
  シナリオが渡され、それによって必要とされる情報だけを抽出する

ＭＵＣ６参加者は上の４つの作業のどれか一つ以上を選び大会に参加した。実際に参加し、作業するにあたっては、作業毎に何をすべきかを説明した詳しい説明書が手わたされた。対象となった新聞記事は Wall Street Journal 紙の記事であった。

次に、４つの作業の内容をより詳しく説明する。尚、例として用いられている記事は、実際にＭＵＣ６で使われた記事の一部を編集してあり、各作業の結果も示してあるが、これもあくまで参考例であることを御承知おき頂きたい。

4.1 Named Entity

新聞記事から固有名詞(組織名、人名、地名)及び時間表現 (曜日、年月日等)、数表現(金額、パーセント)を認定し、テキストに直接タグを付けるものである。タグはＳＧＭＬ (Standard Generalized Markup Language) タグである。

組織名は会社名、政府組織機関、国際的機関名などを含む。人名は記事中の人名を認定するものであるが、人名に付いたタイトル (President, Mr., Dr. など) は含まない。地名は基本的に大陸、地域、国、州、市の名前である。

時間表現としては、日付、曜日、季節、会計年度などであり、数表現としては通貨単位を含む金額、%のついた表現が抽出の対象となった。

記事の一例[1]

```
New York Times Co. named Russell T. Lewis, 45,
president and general manager of its flagship
New York Times newspaper, responsible for all
business-side activities. Mr. Lewis succeeds
Lance R. Primis, who in September was named
president and chief operating officer of the
parent.
```

Named Entity のタグが付けられた結果

```
<ENAMEX TYPE="ORGANIZATION">New York Times Co.
</ENAMEX> named <ENAMEX TYPE="PERSON">Russell
T. Lewis</ENAMEX>, 45, president and  general
manager of its flagship <ENAMEX TYPE=
"ORGANIZATION">New York Times</ENAMEX>
newspaper,responsible for all business-side
activities. Mr. Lewis succeeds <ENAMEX TYPE=
"PERSON">Lance R. Primis</ENAMEX>, who in
<TIMEX TYPE="DATE">September</TIMEX> was named
president and chief operating officer of the
parent.
```

[1]これは例として分かりやすくするために編集してある。元の記事は Copyright Dow Jones, 1994

4.2 Coreference

この作業では、記事中の照応関係 (Coreference) を認定し、Named Entityの時のようにテキストに直接その照応関係を示すＳＧＭＬのタグを書き込む。照応関係を認定すべき対象は次の４つである。

- 代名詞
  it, they, he, she, ...
- 固有名詞
  NEC Corporation, NEC, ...
- 限定名詞句
  the company, the president, ...
- 同格の表現
  Mr. Yamada, a senior manager of NEC's research lab, ...

照応関係の認められる語句は各々に固有の番号が付けられ、照応関係は照応する語句のタグ中にその照応先の番号を明示することによって示される。

上記の記事に照応関係を付けて見ると

```
<COREF ID="1">New York Times Co.</COREF> named
<COREF ID="2">Russell T. Lewis</COREF>, 45,
president and general manager of <COREF ID="3"
TYPE="IDENT" REF="1">its</COREF> flagship
New York Times newspaper, responsible for all
business-side activities.  <COREF ID="4" TYPE=
"IDENT" REF="2">Mr. Lewis</COREF> succeeds Lance
R. Primis, who in September was named president
and chief operating officer of <COREF ID="5"
TYPE="IDENT" REF="1">the parent</COREF>.
```

4.3 Template Element

従来のＭＵＣではテンプレートを埋めることがすなわち情報抽出であった。特に、第５回のＭＵＣでは埋めるべきテンプレートがかなり複雑になり、システムの評価の結果もあまり良くなかった。第６回の大会ではテンプレートが複雑になり過ぎたとの反省にたち、テンプレートを埋める作業を２つに分けた。Template Element と Scenario Template の作業である。

まず、Template Element の作業ではテキスト中から基本的な事実だけを認定し、テンプレートを埋めることとした。対象となる記事の分野に関わらず基本のテンプレートを作成するのが狙いである。Scenario Template の作業ではTemplate Element で認定されたものの間の関係を抽出して来ることが狙いである。具体的にはTemplate Element での認

定すべき対象は、記事中に現れる全ての組織名、人名、製品名(artifact) である。

例えば、上記の記事からは、製品名はないので組織名、人名に注目し、次のようなテンプレートが生成される。

```
<ORGANIZATION-1>
    NAME          : "New York Times Co."
    TYPE          : Company
<ORGANIZATION-2>
    NAME          : "New York Times"
    TYPE          : Company

<PERSON-1>
    NAME          : "Russell T. Lewis"
    ALIAS         : "Lewis"
    TITLE         : Mr.
<PERSON-2>
    NAME          : "Lance R. Primis"
```

4.4 Scenario Template

Template Element の結果に基づき、より複雑な関係を定義したシナリオに沿ってテンプレートを埋めていく。今回の大会で対象となったのは人事異動に関する新聞記事で、シナリオで要求されたのは、誰がどの会社のどのポストを辞め、新しく誰がそのポストを埋めるのか、そして、その時期（年月日）などの関係を認定して、テンプレートを埋めていくことであった。

Template Element の時とは違い、この作業ではシナリオに該当する情報だけを抽出してくるので、もし記事中に抽出すべき情報がないと判断された場合は空のテンプレートととして結果を出す。関係する情報が記事中にある場合でも、記事全ての組織名、人名が対象となる訳ではなく、あくまで、該当するもののみを取り出してくる。シナリオ自身は評価の１ヶ月前に公表された。つまり、Scenario Template 用のシステムの開発は１ヶ月間だけであった。

例の記事で役職とその引き継ぎ(SUCCESSION)、つまり、誰がそのポストを辞め(WHO_IS_OUT)、誰がそのポストに新しく就任(WHO_IS_IN) したのかに注目してテンプレートを埋めてみると以下のようになる。<ORGANIZATION-2>、<PERSON-1>、<PERSON-2> はポインターで、Template Elementの作業で生成されたものを指している。役職にあたる POST は記事中の文字列をそのまま抜き出して来ている。

```
<SUCCESSION-1>
    ORGANIZATION : <ORGANIZATION-2>
    POST         : "president"
    WHO_IS_IN    : <PERSON-1>
    WHO_IS_OUT   : <PERSON-2>

<SUCCESSION-2>
    ORGANIZATION : <ORGANIZATION-2>
    POST         : "general manager"
    WHO_IS_IN    : <PERSON-1>
    WHO_IS_OUT   : <PERSON-2>
```

5 評価

5.1 評価の方法

評価は「答え」とシステムのアウトプットを比較して算出された。この「答え」は全て人間の手作業で作られた。評価（スコアーづけ）自身は専用のプログラムを動かし自動的になされる。主要なスコアーは情報検索で良く用いられている二種類、Recall （再現率）と Precision （適合率）である。計算の方法は以下のようである。

$$\text{再現率} = \frac{\text{システムのアウトプット中の正解数}}{\text{テキスト中の正解の総数}}$$

$$\text{適合率} = \frac{\text{システムのアウトプット中の正解数}}{\text{システムのアウトプットの総数}}$$

この他にも、再現率や適合率とは異なる方法での評価、参加グループのシステムのスコアー間の統計的差異の分析、システムの結果と人間が同作業をした結果の比較などがなされているが詳しくは参考文献 [4] [5] などを参照されたい。

5.2 評価テキスト

評価に使われたテキスト数は、Template Element, Scenario Template 用に１００テキストで、その内の３０テキストがNamed Entity, Coreference 用にも用いられた。これらのテキストは全て評価時にシステムにとって今までに見ていない新テキストのセット(blind test set) であった。

１００テキスト中５４テキストだけが Scenario Template で埋めるべき情報のある(relevant) テキストであった。

6 参加者

ＭＵＣ６への参加者は全部で１６グループである。作業別に見た参加者は表２の通りとなる。

内訳は企業が８グループ、大学が８グループであり、企業からの参加者は全て米国の企業であるが、大学からの参加者には、米国以外の大学もあった。英国の大学が２校(Sheffield, Durham)とカナダの大学１校(Manitoba)が入っている。

参加者	NE	CO	TE	ST
BBN Inc.	X		X	X
Knight-Rider	X			
Lockheed/Martin	X		X	X
Mitre Corp.	X		X	
SAIC(McLean)	X			
SRI	X	X	X	X
SRA	X		X	X
Stering Software	X		X	
New Mex. State	X			
New York U.	X	X	X	X
U. Duhram	X	X	X	X
U. Mass	X	X	X	X
U. Manitoba	X	X	X	X
U. Penn		X		
U. Sheffield	X	X	X	X
Wayne State	X			

表 2: 作業別に見たＭＵＣ６の参加者

(注： NE = Named Entity, CO = Coreference, TE = Template Element, ST = Scenario Template)

7 結果

次にシステムの評価の結果を作業別に示す。

7.1 Named Entity のスコア

Named Entity には４つの作業中最も参加者が多く、１５グループが参加した。結果は表３の通りである。

この結果は予想よりも良く、ベストのシステムのスコアーは人手で「答え」を作成した時の人間の出来と同じぐらいの出来である。それに加え、上位半分のグループのシステムが再現率、適合率ともが９０％以上であり、英語新聞記事中の固有名詞を確実に認定、区分出来ている。これより Named Entity の作業については英語では技術がほぼ確立したと言うのが大会での大方の見解であった。

参加者	再現率	適合率
SRA	96	97
SRI	92	96
BBN Inc.	94	93
U. Manitoba	92	95
Stering Software	92	93
Mitre Corp.	91	91
Lockheed/Martin	91	91
U. Sheffield	84	94
New York U.	86	90
New Mex. State	85	87
Knight-Rider	80	92
U. Mass	82	89
U. Duhram	62	74
SAIC(McLean)	27	52
Wayne State	1	68

表 3: ＭＵＣ６ Named Entity のスコア

尚、今回の英語新聞記事からの固有名詞抽出の具体的な方法や技術を示した１例として [6] などがある。

7.2 Coreference のスコア

照応関係 (Coreference) の作業は今回の４つの作業中最も難しいものであった。各グループとも再現率を上げるのが困難であったようである。人手でこの作業をすると再現率、適合率とも８０％以上にはなるので、まだまだシステムの改善、照応関係抽出の研究の余地がある。

スコアーを表４に示す。

参加者	再現率	適合率
SRI	59	72
U. Manitoba	63	63
U. Sheffield	51	71
U. Penn	55	63
New York U.	53	62
U. Mass	44	51
U. Duhram	36	44

表 4: ＭＵＣ６ Coreference のスコア

7.3 Template Element のスコア

まずまずの出来であったが、まだ改善の余地はある。上位半分のグループのシステムのスコアーにはほとんど差が見られない。人手でした場合のスコアーは再現率、適合率ともに９０％を越えているのと比べるとまだ多少低くなっている。

参加者	再現率	適合率
SRA	74	87
Mitre Corp.	71	85
Stering Software	72	83
Lockheed/Martin	76	77
SRI	74	76
U. Manitoba	71	78
BBN Inc.	66	79
New York U.	62	83
U. Sheffield	66	74
U. Mass	53	72
U. Duhram	49	60

表 5: ＭＵＣ６ Template Element のスコア

7.4 Scenario Template のスコア

シナリオが１ヶ月前に公表されたので、システムのこの作業用の開発の期間は１ヶ月しかなかったことになるが、その割には出来が良いとの評であった。前回の大会のＭＵＣ５ではテンプレートがもっと複雑でベストのシステムが再現率、適合率とも５０％台であった。それと比べると少なくとも適合率が改善されている。

表６には示されていないが、Scenario Template で埋めるべき情報のある(relevant) テキストが１００テキスト中５４テキストであり、各システムとも９０％前後の正確さで関係あるテキスト(relevant text) の選り分けが出来ていた。

8 結果のまとめと今後の課題

英語新聞記事からの固有名詞の抽出については技術が一応確立したと言える。実際にこの大会後固有名詞抽出システムの商品化が行われている。照応関係の認定はまだまだ改善の余地があることが判明した。テンプレートを埋める作業については技術が向上しつつあり、基本的な情報の抽出が出来るシステムを短期間で特定の用途に適したシステムに変更することが可能であることが分かった。

参加者	再現率	適合率
New York U.	47	70
BBN Inc.	50	59
SRA	47	62
Lockheed/Martin	43	64
SRI	44	61
U. Sheffield	37	73
U. Manitoba	39	62
U. Mass	36	45
U. Duhram	33	33

表 6: ＭＵＣ６ Scenario Template のスコア

現行の作業に加えて、今後のやるべき事として考えられているのは以下のような作業である。

- 複数のドキュメントを対象とした情報抽出
  現在までは、情報抽出の対象は１つの新聞記事毎であったが２つ以上の関連した新聞記事からの情報抽出も必要か。
- 要約の生成
  抽出された情報を基にしてテンプレートのみならず、要約文を生成してはどうか。
- 英語以外の言語での情報抽出
  ＭＵＣ６の延長として、９６年４月に日本語、中国語、スペイン語での固有名詞抽出システムの評価が行われた。結果はまずまずで今後もこれらの言語での情報抽出研究を進める必要がある。

ＭＵＣ７が９７年に開催されるのは間違いないようである。どの言語でどれだけの作業をすることになるのかは、現在のところまだ決定されていない。今後、米国政府関係機関と参加予定者の議論・調整がなされる予定である。尚、ＭＵＣ６についての説明は [7] においても述べられているので参照して頂きたい。

参考文献

[1] *Proceedings of the 24th Month Workshop of Tipster Text Program (Phase I)*, Virginia, September 1993, Morgan Kaufmann Publishers Inc. 1994.

[2] *Proceedings of the fifth Message Understanding Conference (MUC-5)*, Baltimore, Mary-

land, U.S.A. 1993, Morgan Kaufman Publishers Inc. 1993.

[3] *Proceedings of the sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, U.S.A. 1995, Morgan Kaufman Publishers Inc. 1996.

[4] Sundheim, Beth M., "Tipster/MUC-5 Information Extraction System Evaluation" In *Proceedings of the fifth Message Understanding Conference (MUC-5)*, Baltimore, Maryland, U.S.A. 1993, Morgan Kaufman Publishers Inc. 1993.

[5] Chinchor, N., L. Hirschman, and D. Lewis, "Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)" *Computational Linguisitcs* 19(3), 1993.

[6] Wakao, T, R. Gaizauskas, and Y. Wilks "Evaluation of An Algorithm for the Recognition and Classification of Proper Names" In *Proceedings of the 16th International Conference on Computational Linguistics (Coling 96)*, 1996.

[7] Ralph Grishman, and Beth Sundheim "Message Understanding Conference - 6: A Brief History" In *Proceedings of the 16th International Conference on Computational Linguistics (Coling 96)*, 1996.