日本語テキストからの固有表現抽出システムの開発と評価

1N－3

竹元 義美　山田 洋志　福島 俊一
NEC ヒューマンメディア研究所

1. はじめに

情報洪水と言われる今日、大量のテキストを利用者の興味に合うもののみに絞り込む「情報検索」技術に加えて、テキスト中から利用者に必要な情報のみを取り出す「情報抽出」技術の重要性が高まっている。情報抽出においては、テキスト中のキー要素を判別することと、それらの間の関係を認定することが必要になる。この前者の機能を、できる限り情報抽出の個別タスクに依存しない形で実現しようというのが「固有表現抽出」である。「固有表現」とは、まだ一般には馴染みのない日本語であるが、キー要素として一般性のある、組織名・人名・地名などの固有名詞類や時間表現・数値表現を、総称したものである。

固有表現抽出の技術は、米国の情報抽出コンテスト MUC（および MET）によって発展してきた[3][4]。筆者らも 1996 年の MET-1 に参加し、日本語を対象とした固有表現抽出に早い段階から取り組んでいる研究チームの 1 つである[1][2]。このような取り組みを通して、固有名詞辞書の充実とパタンマッチルールの整備が、高い抽出精度達成への最も確実なアプローチだと共通認識されるようになってきた[2][3]。筆者らのシステムもこのアプローチにしたがって改良を重ねてきており、本稿では、その抽出方式と最新の評価結果を報告する。

2. 固有表現抽出

米国での MUC や MET の流れを汲みながら、1999 年 5 月に情報抽出コンテスト IREX-NE が日本で独自に開催された。この IREX-NE では、日本語の新聞記事テキストからの固有表現抽出が、コンテスト課題として設定された。その実施に先立っては、固有表現の定義に関しても真剣な議論が交わされた[4]。

本稿における固有表現の定義や評価データは、IREX-NE で使われたものを用いている。固有表現抽出システムは、与えられた日本語テキストに対して、表 1 に示した 8 種類の固有表現を見つけ、SGML 形式のタグで囲む。

表 1 固有表現の種類

種類		タグ	例
固有名詞	組織名	ORGANIZATION	NEC、西武ライオンズ
	人名	PERSON	クリントン、山田太郎
	地名	LOCATION	東京都、新大阪駅
	固有物名	ARTIFACT	カローラ、芥川賞
時間	日付表現	DATE	5 月 14 日、6 月下旬
	時刻表現	TIME	午後 5 時 25 分
数値	金額表現	MONEY	500 億円
	割合表現	PERCENT	120%、5 分の 1

3. 抽出処理ステップ

筆者らの固有表現抽出システムでは、以下のようなステップで抽出処理を実行する。

(1)**形態素解析**：入力文を単語列に分割する。

(2)**複合語分割**：(1)の精度や効率をチューニングする技法として、複合語（長単位語）の単語辞書登録がしばしば行われる。しかし、例えば「来日」の「日」を地名として抽出できるように、この段階で「来日」は「来」+「日」に分割しておきたい。(2)では、このような分割パタンを事前に辞書化しておき、それを参照して(1)の結果をさらに細分割する。

(3)**固有表現情報付与**：表 1 の 8 種類に該当する様々な固有表現に加えて、固有名詞の共起語、時間・数値表現の単位も収集し、約 9.2 万語の固有表現辞書を構築した。(2)の結果に対して、この固有表現辞書と合致した箇所をマークする。

(4)**低信頼語推定**：(1)で未知語と判定された箇所だけでなく、例えば「三（数詞）」+「星（名詞）」+「電子（名詞）」）のように信頼性の低い箇所を 1 つにまとめるようなルールも用意した。

(5)**固有表現判定**：(1)～(4)までで得られた情報に対して、人手で作成したパタンマッチルールを適用することにより、固有名詞や時間・数値表現を判定する。(3)の辞書照合だけで決定できる部分もあるが、多くの場合、固有表現と一般名詞との曖昧性（例えば「森」は人名または一般名詞）や固有表現の種類の曖昧性（「福島」は人名または地名）などが発生する。このような曖昧性に対し、それを解消したり、可能性の高い選択を行うために、各種共起語（例えば人名共起語の「氏」「社長」、地名共起語の「県」「市」、組織名共起語の「委員会」「社」

Development and Evaluation of a Japanese Named Entity Extraction System
Yoshikazu Takemoto, Hiroshi Yamada, Toshikazu Fukushima
Human Media Research Laboratories, NEC Corporation
8916-47, Takayama-Cho, Ikoma, Nara 630-0101, Japan
E-mail: {takemoto, h-yamada, fuku}@hml.cl.nec.co.jp

表 2 各固有表現の抽出精度

	出現	抽出	正解	R 値	P 値	F 値
組織名	389	402	313	80.5	77.9	79.2
人名	355	340	305	85.9	89.7	87.8
地名	416	390	341	82.0	87.4	84.6
固有物名	49	21	16	32.7	76.2	45.8
日付表現	277	292	257	92.8	88.0	90.3
時刻表現	59	66	51	84.7	75.8	80.0
金額表現	15	15	13	86.7	86.7	86.7
割合表現	21	18	16	76.2	88.9	82.1
合計	1581	1544	1312	83.0	85.0	84.0

「出現」は各固有表現の出現数(人手判定での正解)。
「抽出」はシステムが各固有表現として抽出した数。
「正解」は「出現」のうちで「抽出」に該当した数。
「R 値」は再現率＝「正解」／「出現」。
「P 値」は適合率＝「正解」／「抽出」。
「F 値」は R と P の統合指標＝2×R×P／(R＋P)。
IREX-NE の総合ドメインの評価用テキストでの精度である。

など)に着目するなど、パタンマッチルールの充実が必要になる。

(6)**判定結果調整**:(5)の結果を調整・リファインするルールをいくつか用意した。例えば「宮前区宮崎4の1の1」のような住所は、(5)までで「宮前区」「宮崎」が地名とわかり、(6)の段階で「4の1の1」も含めて全体が地名とまとめられる。

(7)**固有名詞省略表現判定**:(6)までで決定された固有表現のうち信頼性の高いものをもとに、省略表現を同定する。例えば、テキストの初出現位置では「横浜市」と表記され、それ以降は「横浜」と略記された場合、「横浜」は「横浜市」の省略表現だとみなす処理である。「横浜」は、単独だと地名と組織名の曖昧性をもつが、「横浜市」(地名)の省略表現と解釈すれば曖昧性が解消できる。

(8)**判定結果出力**:SGML タグを挿入する。

4. 評価

IREX-NE での訓練用テキスト(記事数:46)と評価用テキスト(記事数:70)により、抽出精度を算出した[1]。表 2 には各固有表現に分けた抽出精度を示し、表 3 には処理ステップの組み合わせを変えた場合の抽出精度を示した。

- 平均として 84%の精度が得られた。
- 固有物名の抽出は難度が高い。固有物名に該当する作品や製品の名前は、一般名詞や他の固有表現との解釈の競合が発生しやすく、辞書に網羅することも現実的でない。

[1] IREX－NEでは毎日新聞の記事データを用いており、使用を快諾くださった毎日新聞社に感謝する。

表 3 各処理ステップの貢献度

処理の組み合わせ		訓練	評価
S1	(1)＋(3)＋(8)	51.1	34.9
S2	S1＋(5)時間・数値ルール	69.5	53.9
S3	S2＋(5)固有名詞ルール	85.9	80.0
S4	S3＋(4)	86.2	80.2
S5	S4＋(6)	86.2	81.2
S6	S5＋(2)	89.6	83.3
S7	S6＋(7)	89.8	83.8
S8	訓練テキストの単語登録	93.9	84.0

IREX-NE の総合ドメインの訓練用テキストと評価用テキストの各々に対する抽出精度を F 値で示した。

- 各処理ステップは、効果の大小にばらつきがあるが、いずれも精度向上に貢献している。固有表現辞書の役割は大きいが、パタンマッチルールと組み合わせないと 80%の精度には至らない。
- 抽出洩れの原因としてパタンマッチルールの不足があるが、特に組織名の共起語は種類が多いので拡充が必要に思えた。
- 抽出誤りの原因では、文脈の考慮が必要と思われるものが目立つ。特に「羽田空港」のように組織名か地名かの曖昧性の解釈に文脈を必要とするケースが多い。時間表現も、前後の表現を考慮してパタンを強化すべきケースが目に付いた。

5. おわりに

日本語テキストからの固有表現抽出システムを開発した。新聞記事テキストでの評価を通して、固有名詞辞書の充実とパタンマッチルールの整備というアプローチが、高い抽出精度の達成に有効なことを示した。コーパスの統計分析によるルール整備や辞書拡充など、チューニングやカスタマイズを容易にする仕組みの構築、および、新聞記事以外への対象の拡大などが課題である。

参考文献

[1]Y.Takemoto, et al., Description of NEC/Sheffield System Used for MET Japanese, Proc. of Tipster Text Program (Phase II), 1996.

[2]竹元・他、日本語新聞記事からの固有名詞情報抽出、情処 53 全大:7L-3、1996 年。

[3]江里口・他、パターンマッチング手法による名称特定処理の有効性の検討、情処研報:NL-115-10、1996 年。

[4]関根・他、固有表現の定義の困難さ－IREX における NE 定義の事例から、言語処理学会第 5 回年次大会:B2-1、1999 年。