日本語人名辞書を用いた中国語文書からの人名抽出

Xiao Liying　新納 浩幸　佐々木 稔　古宮 嘉那子
茨城大学大学院　茨城大学工学部
理工学研究科　情報工学科
{14nm721x@vc, shinnou@mx, msasaki@mx, kkomiya@mx}.ibaraki.ac.jp

1　はじめに

我々の研究室では、未知外国語からの情報抽出に取り組んでいる。未知外国語とはシステム作成者にとっての未知の外国語という意味である。ここではシステム作成者を中国語を知らない日本人、情報抽出タスクを中国語文書からの人名抽出に設定し、この課題について考察する。

中国語を全く理解できないという条件下では、中国語文書から人名抽出を行うことは不可能であり、なんらかの手がかりが必要である。我々はこの手がかりとして中国語文書中の日本人名に着目した。中国語文書中の日本人名は日本語の漢字表記のまま表記されることが多い。このため日本人にとっては日本人名であれば、対象が中国語文書であってもそれを取り出すことができる。そこで日本語人名辞書を利用することで、中国語文書内から日本人名を取り出し、その人名の前後の単語組を抽出規則とすることで、任意の中国語文書内から人名抽出が行える。

ただしこうして得た抽出規則だけでは人名抽出を精度良く行うことはできない。ここからなんらかのヒューリスティクスを利用することで抽出精度を高める必要がある。未知外国語という条件下では、このヒューリスティクスに対象言語の知識を使えない。ここでは規則の発火回数と抽出単語の IDF 値を利用する。

実験の結果、抽出精度は低いがいくつかの人名を抽出することができた。対象言語の知識が皆無という条件下では人名抽出は困難である。いかに少量の対象言語の知識により、抽出規則の精度を高めていくかが今後の課題である。

2　中国語文書からの人名抽出

2.1　中国語の単語分割

未知外国語という設定で、単語分割システムを利用するのは不自然かもしれないが、単語分割は大規模なコーパスさえあればある程度可能である。そのためここでは中国語の単語分割システムが利用可能として、対象文書は単語分割された中国語文書とした。

具体的に単語分割には KyTea[1] を利用する。配布元から提供されている中国語単語分割モデルの msr-0.4.0-1.mod を利用して単語分割を行う。

2.2　日本語人名辞書を利用した日本人名の抽出

図 1 はある中国語文書の一部である。

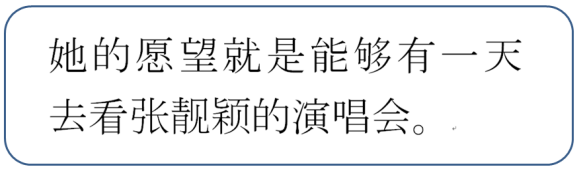

図 1: 中国人名を含む中国語文

中国語を全く知らない人にとっては、この文書から人名を取り出すことは不可能である。

一方、図 2 の中国語文書の場合、日本人であれば、この文書中にある人名を少なくとも 1 つは推測できる。それは日本人名の「横山大観」である。図 2 のように中国語文書であっても日本人名はその漢字表記のまま記載されることが多い。このため日本語人名辞書を利用して、中国語文書中の日本人名を抽出することが可能である。

[1] http://www.phontron.com/kytea/index-ja.html

为了看横山大観的画展，他逃了下午的课。

図 2: 日本人名を含む中国語文

KyTea が正しく単語分割できていれば、その単語が日本語人名辞書に登録されているかどうかを調べれば良いが、KyTea は日本人名を正しく単語分割できない場合も多い。そこで 3 単語列まで組み合わせて、その文字列が人名となっているかどうかを調べることにした。その手順を以下に示す。

```
INPUT: w_1, w_2, … , w_n

i = 1
while( i < n + 1) {
    if (w_i in F-dic or N-dic) {
        OUTPUT w_i; i++
    } else {
        w = w_i + w_{i+1}
        if (w is PersonName) {
            OUTPUT w; i += 2
        } else {
            w = w_i + w_{i+1} + w_{i+2}
            if (w is PersonName) {
                OUTPUT w; i += 3
        } else i++
   }
}
```

上記手順の F-dic が姓辞書、N-dic が名辞書である。また w is PersonName の判定だが、これはまず w が姓辞書あるいは名辞書に登録されていれば真である。また w の文字列を任意の箇所で 2 分割し、前半文字列が姓辞書に登録され、かつ後半文字列が名辞書に登録されていれば真となる。

2.3 前後単語による人名抽出規則

図 2 の「横山大観」が人名とした場合、その前後の単語の組（「看」「的」）を 1 つの人名抽出規則とする。この人名抽出規則を図 1 に適用することで人名を抽出できる（図 3）。

3 言語非依存の知識によるフィルタリング

前章までに説明した人名抽出規則だけでは精度良く人名を抽出できない。ここからは何らかのヒューリスティクスを利用して、抽出精度を上げていく必要がある。ただしここでは未知外国語という設定のため、対象言語に依存したヒューリスティクスを使うことができない。

ここでは言語に依存しないヒューリスティクスとして以下の 2 つのフィルタリングを試すことにする。

(1) 規則の発火回数

抽出規則が何度も発火するのは規則として正しくない可能性が高い。例えば (”的”,”。”) という規則があったが、これは 165 回も発火していた。ここでは規則の発火回数が 4 回以上あった場合、その規則での抽出は削除することにした。

(2) 抽出単語の IDF 値

通常、同じ人名が様々な文書に出現するのはおかしい。逆に様々な文書に出現する単語は、一般の名詞である可能性が高い。ここでは抽出した単語の IDF 値が 3.16 以下であった場合に、その抽出は削除することにした。

なおここでの IDF 値は以下で計算した。

$$\log \frac{|D|}{|\{d : w \in d\}|}$$

ここでの実験では $|D| = 142$ なので、抽出単語が現れた文書数が 6 以上の場合に、その抽出を削除する形になっている。

4 実験

学習用データは、日本人名が含まれやすい中国語文書として、日本に関する中国語のネットニュース記事を対象にした。以下の URL から政治の記事を中心に約 15Mbyte の中国語文書を取り出した。

http://asahichinese.com/

次に JUMAN と一緒に配布される姓辞書（1498 種類）、名辞書（1901 種類）を用いて、前述した日本人名の抽出手順に従って、先の中国語文書から合計 8596 個の日本人名を抽出した。ここから日本人名を囲む直前の単語 W_a と直後の単語 W_b の組 (W_a, W_b) を取り

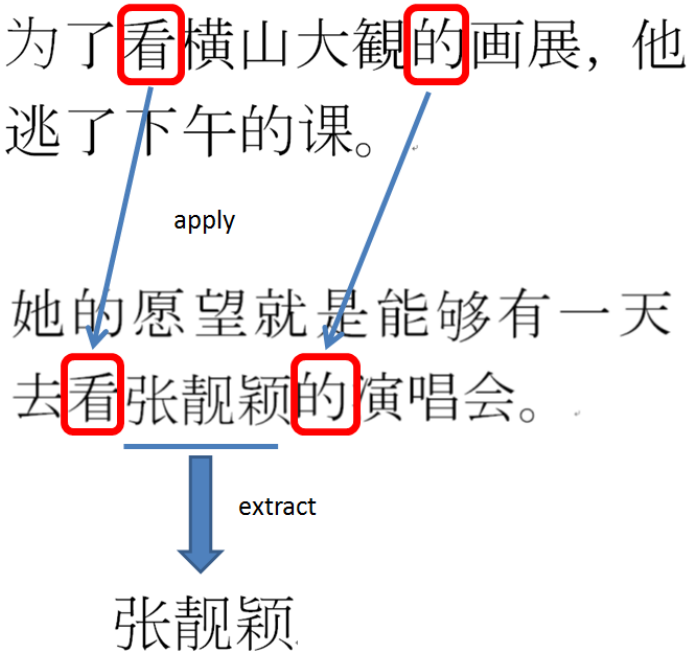

図 3: 抽出規則による人名抽出

出した。この (W_a, W_b) が抽出規則に対応する。取り出された抽出規則は 4519 種類であった。

次に、テスト用文書は以下の URL で示される中国のネットニュース記事を対象にした。ここから政治の記事を中心に約 700Kbyte (142 文書) を取り出した。

http://china.huanqiu.com/

テスト用文書に対して抽出規則を適用することで、3455 個の単語を取り出した。これら取り出された単語が実際に人名となっているかどうかは手作業でチェックした。結果、101 個が実際に人名になっていた。これは正解率 2.9% であり、かなり低い。

次に前述したフィルタリングを行った。規則の発火回数のフィルタリングを行った結果、抽出数は 435、うち正解は 32 個となり、正解率は 7.4% まで向上した。その上で更に IDF 値によるフィルタリングを行った結果、抽出数は 282、うち正解は 32 個となり、正解率は 11.3% まで向上した。

5 考察

ここでは人名抽出規則として日本人名の前後の単語のペアを用いたが、前後 2 単語、あるいは前 2 単語後 1 単語や前 1 単語後 2 単語などの抽出規則を用いることもできる。この場合の実験結果を表 1 に示す。

抽出規則	規則数	抽出数	正解数	正解率 (%)
前後 1 単語	4519	3455	101	2.9
前 1 後 2 単語	7924	206	7	3.4
前 2 後 1 単語	7167	118	9	7.6
前後 2 単語	8126	10	2	20.0

表 1: 各抽出規則による正解率 (%)

前後の単語を増やして抽出規則の条件をより厳しくすれば、正解率は向上する。また「前後 2 単語」の抽出規則による抽出のあとに IDF 値によるフィルターをかけると抽出数は 6、 正解数は 2 となり、正解率は 33.3% まで向上した。このように正解率は向上するが、再現率はかなり下がるはずであり、抽出規則は用途によって使い分けた方がよい。

本研究は核になるアイデア（日本人名は中国語文書でもそのまま表記される）の可能性を試すことに傾注したため、抽出システムとしてはいくつかの不備が存在する。まず抽出規則 (W_a, W_b) があった場合、W_a と W_b の間は 1 単語に限定している。このため未登録単語や KyTea の解析誤りに対応できない。さらに中国語の人名が 1 単語として扱われるというのは誤りの可能性もある。次に日本人名を KyTea では正しく形態

素解析できないことが多く、検出できていない日本人名もある。これは使用したモデルが中国語のモデルのためだが、KyTea は辞書を追加して学習できるので、KyTea の適切な調整で、日本人名を正しく検出できるはずである。

核になるアイデアにも不備があった。日本語のある種の漢字は中国語では別の漢字が当てられていた。気がついたものは日本語の辞書に中国語の漢字表記のものを加えて対処したが、網羅はしていないと思われる。

本研究はブートストラップ型の固有表現抽出 [1][2][3][4] と関連が深い。ブートストラップ型の固有表現抽出では、概略、以下のような処理となる。まず種となる小さな固有表現の集合を準備し、その集合内の固有表現をコーパスから探す。そして見つかった固有表現の周辺文脈を抽出規則として構築する。次にこの抽出規則をコーパスに適用し、新たな固有表現を抽出し、それを先の固有表現の集合に追加し、先の手順を繰り返す。我々のタスクの設定では、対象言語が未知外国語という設定なので、中国語の人名を種にすることができない。我々の手法は日本人名を種にしていると見なせる。我々の手法はコーパスから日本人名を見つけた後の抽出規則の構築が簡易すぎるために精度が悪かった。ただし良質の抽出規則を、対象言語の知識なしで構築するのは非常に困難である。いかに少量の対象言語の知識により、抽出規則の精度を高めていくかが今後の課題である。

6 おわりに

本論文では、システム作成者が全く中国語を理解できないという条件下で、中国語文書からの人名抽出を行った。核となるアイデアは、中国語文書中の日本人名は日本語の漢字表記のまま表記されることが多いために、日本語人名辞書を利用することで、中国語文書内から日本人名を取り出し、その人名の前後の単語組を抽出規則とする、というものである。ただしこうして得た抽出規則だけでは人名抽出を精度良く行うことはできないため、規則の発火回数と抽出単語の IDF 値をフィルターとして利用した。小規模の実験を行った結果、いくつかの人名を抽出することはできた。また設定したフィルターも機能した。しかし正解率は低かった。抽出精度を高めるため、対象言語についての何らかのヒューリスティクスが現実的には必要である。いかに少量の対象言語の知識により、抽出規則の精度を高めていくかが今後の課題である。

参考文献

[1] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, pp. 100–110, 1999.

[2] Alessandro Cucchiarelli and Paola Velardi. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, Vol. 27, No. 1, pp. 123–131, 2001.

[3] David Nadeau, Peter Turney, and Stan Matwin. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Proceedings of the 19th Canadian Conference on Artificial Intelligence*, pp. 266–277, 2006.

[4] Ellen Riloff, Rosie Jones, et al. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pp. 474–479, 1999.