人名構成文字確率を用いた文字ベース CRF による中国語人名検出

新納浩幸
茨城大学工学部情報工学科

全太俊
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1　はじめに

人名検出は固有表現抽出の一部であり，情報抽出の要素技術として重要である．本論文では，中国語の人名検出を行う．

一般に，固有表現抽出は系列ラベリング問題として定式化し，条件付き確率場（CRF: Conditional Random Fields）[1] を用いることで精度よく行える．ただし，通常，CRF の素性として単語の品詞を利用するが [2]，中国語の場合，標準的に利用できる形態素解析システムが存在しないために，単語分割，品詞付与の処理が容易ではない．そこでここでは文字列ベースの CRF を用いる [3]．また中国語のコーパスと中国人名のデータベースを利用して，各文字が人名の構成要素となる確率を推定し，その確率がある閾値よりも高いか低いかのラベルを素性として加えることで，単純な文字列ベースの CRF を改善する．

実験では人名のタグを付与した中国語の 1,000 文書を用いた．10 分割交差検定により本手法の有効性を示す．

2　系列ラベリング問題

m 個のデータの系列 $\{x_1, x_2, \cdots, x_m\}$ が与えられたときに，系列中の各データ x_i に対する適切なラベル $y_i \in Y$[1]を求めて，ラベルの系列 $\{y_1, y_2, \cdots, y_m\}$ を出力する問題を系列ラベリング問題という (図 1)．

人名抽出の問題は，系列ラベリング問題として取り扱うことが可能である[2]．例えば，以下の単語切りされた文は，各単語がデータとなるデータの系列である．

私 / の / 名前 / は / 全 / 太俊 / と / 申し / ます

系列中の各単語に対して BI ， BO ， O のいずれかのラベルを与える．ラベル BI は人名の開始単語で

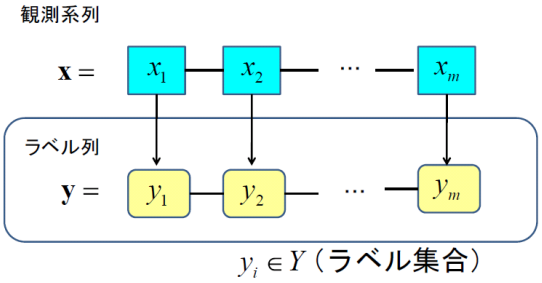

図 1: 系列ラベリング問題

あることを意味し，ラベル BO は人名内の単語であることを意味し，ラベル O は人名ではない単語を意味する．この例の場合，以下のようなラベルが与えられ，ラベルの系列が完成する．

私	/	の	/	名前	/	は	/	全	/	太俊	/	と	/	申し	/	ます
O		O		O		O		BI		BO		O		O		O

このラベルの系列から「全太俊」が人名として抽出できる．

3　文字ベースの CRF

3.1　CRF

系列ラベリング問題は HMM などのモデルを使って解くことも可能であるが，現在は，条件付確率場 (CRF: Conditional Random Fields) を用いて解くのが精度上好ましい．

CRF は対数線形モデルの一種であり，データの系列 $\boldsymbol{x}$ に対応するラベルの系列が $\boldsymbol{y}$ となる確率 $P(\boldsymbol{y}|\boldsymbol{x})$ を以下の式でモデル化する．

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp(\boldsymbol{w}\cdot\phi(\boldsymbol{x},\boldsymbol{y}))$$

[1] Y はラベルの集合
[2] 一般にチャンキングの問題が扱え，固有表現抽出はチャンキングの一種である．

ここで$\boldsymbol{w}$は素性に対する重みベクトル，ϕ は素性ベクトルに変換する関数，Z は正規化のための以下で定義される数である．

$$Z = \sum_{\boldsymbol{y}} \exp(\boldsymbol{w} \cdot \phi(\boldsymbol{x}, \boldsymbol{y}))$$

訓練データ $\boldsymbol{D}$ は以下の形をしている．

$$\boldsymbol{D} = \{(\boldsymbol{x^{(1)}}, \boldsymbol{y^{(1)}}), (\boldsymbol{x^{(2)}}, \boldsymbol{y^{(2)}}), \cdots, (\boldsymbol{x^{(|D|)}}, \boldsymbol{y^{(|D|)}})\}$$

$\boldsymbol{x^{(i)}}$ はデータの系列，$\boldsymbol{y^{(i)}}$ は $\boldsymbol{x^{(i)}}$ に対応するラベルの系列である．CRF はこの $\boldsymbol{D}$ を利用して $\boldsymbol{w}$を学習する[3]．

実際の識別は以下で行える．

$$\boldsymbol{y^*} = \arg\max_{\boldsymbol{y}} \boldsymbol{P}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \exp(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}))$$

ここで

$$\boldsymbol{\phi_k}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{t}} \boldsymbol{\phi_k}(\boldsymbol{x}, \boldsymbol{y_t}, \boldsymbol{y_{t-1}})$$

となるように $\boldsymbol{\phi}$ を設計すれば，以下の形となり，ビタビアルゴリズムなどで解が求まる [5]．

$$\boldsymbol{y^*} = \arg\max_{\boldsymbol{y}} \sum_{\boldsymbol{t}} \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y_t}, \boldsymbol{y_{t-1}})$$

3.2 利用する素性

文字列ベースの CRF は，文字だけを利用して素性を構成する．ここで利用した素性は 6 種類である．注目している文字が $\boldsymbol{c_0}$ として，以下の文字列を例とする．

$$\cdots \boldsymbol{c_{-2}} \boldsymbol{c_{-1}} \boldsymbol{c_0} \boldsymbol{c_1} \boldsymbol{c_2} \cdots$$

1. $\boldsymbol{c_{-2}}$：2 文字前の文字
2. $\boldsymbol{c_{-1}}$:1 文字前の文字
3. $\boldsymbol{c_0}$：注目している文字
4. $\boldsymbol{c_1}$：1 文字後の文字
5. $\boldsymbol{c_2}$：2 文字後の文字
6. $\boldsymbol{c_{-1}c_0}$：直前文字と注目文字の 2 文字列
7. $\boldsymbol{c_0c_1}$：注目文字と直後文字の 2 文字列

[3]学習手順は [5] が参考になる．

4 人名構成文字確率

4.1 確率の推定

日本語や中国語では漢字文字に意味がある場合が多く，人名に使われる漢字には傾向がある．例えば，悪い意味の漢字を人名に使うようなことはない．

ここでは漢字文字 $\boldsymbol{c}$ が人名になる確率 $\boldsymbol{P(c)}$ を推定し，$\boldsymbol{P(c)} \geq \boldsymbol{\theta}$ となる $\boldsymbol{c}$ には H というラベルをつけ，$\boldsymbol{P(c)} < \boldsymbol{\theta}$ となる $\boldsymbol{c}$ には L となるラベルを付けて，このラベルの情報を素性に加えて CRF を学習する．

$\boldsymbol{P(c)}$ の推定手順を示す．まず以下のサイトから中国語の 15,147 文書を得た．

`http://download.csdn.net/detail`
`/finallyliuyu/2123131`

各文書はプレーンな中国語のテキストであり，平均 1196.5 文字からなっている．このコーパス中の文字 $\boldsymbol{c}$ の頻度 $\boldsymbol{f(c)}$ をカウントする．次に，このコーパス中で文字 $\boldsymbol{c}$ が人名の構成文字として使われた回数 $\boldsymbol{g(c)}$ をカウントすることで，$\boldsymbol{P(c)} = \boldsymbol{g(c)}/\boldsymbol{f(c)}$ と推定できる．

$\boldsymbol{f(c)}$ のカウントは容易だが，$\boldsymbol{g(c)}$ の正確なカウントは困難であり，ここでは以下のような処理で $\boldsymbol{g(c)}$ を推定した．

まず，全ての文字 $\boldsymbol{c}$ に対して $\boldsymbol{g(c)} = \boldsymbol{0}$ とおく．次に，以下のサイトから中国語の人名のリストを得る．このリストには 65,537 人の人名が記載されている．

`http://www.gangzi.org/article/463.htm`

この中から文字列の長さが $\boldsymbol{k}$ の人名に注目する．

$$\boldsymbol{c_1 c_2 \cdots c_k}$$

単純な文字列一致を利用して，この人名が先のコーパスに現れる頻度 $\boldsymbol{h}$ をカウントし，$\boldsymbol{g(c_i)}$ += $\boldsymbol{h}$ $(\boldsymbol{i = 1 \sim k})$ とする．これを $\boldsymbol{k = 2}$ から 7 まで動かすことで，$\boldsymbol{g(c)}$ を得る．得られた $\boldsymbol{P(c)}$ の上位 10 文字を図 2 に示す．

人名の文字列長とその種類数を表 1 に示す．

表 1: 人名の文字列長とその種類数

長さ	種類数
2	15,438
3	48,056
4	1,488
5	402
6	70
7	30
8 以上	53
合計	65,537

図 2: 人名構成文字確率の上位 10 文字

4.2 素性の追加

上記で得た $\boldsymbol{P(c)}$ を用い，その値が $\boldsymbol{\theta = 0.01}$ 以上となる $\boldsymbol{c}$ には H というラベルをつけ，$\boldsymbol{\theta}$ 未満となる $\boldsymbol{c}$ には L というラベルを付ける．このラベルをここではラベル P と呼ぶことにする．このラベル P を品詞のように扱い，以下に示す 12 種類の素性を CRF の学習に追加した．

注目している文字が $\boldsymbol{c_0}$ として，以下の文字列を例とする．

$$\cdots \boldsymbol{c_{-2}c_{-1}c_0c_1c_2} \cdots$$

$\boldsymbol{c_i}$ に対するラベル P を $\boldsymbol{p_i} \in \boldsymbol{\{H, L\}}$ とおく．

1. $\boldsymbol{p_{-2}}$：2 文字前の文字のラベル P
2. $\boldsymbol{p_{-1}}$：1 文字前の文字のラベル P
3. $\boldsymbol{p_0}$：注目している文字のラベル P
4. $\boldsymbol{p_1}$：1 文字後の文字のラベル P
5. $\boldsymbol{p_2}$：2 文字後の文字のラベル P
6. $\boldsymbol{p_{-2}p_1}$：2 文字前の文字のラベル P と直前文字のラベル P の列
7. $\boldsymbol{p_{-1}p_0}$：直前文字のラベル P と注目文字のラベル P の列
8. $\boldsymbol{p_0p_1}$：注目文字のラベル P と直後文字のラベル P の列
9. $\boldsymbol{p_1p_2}$：直後文字のラベル P と 2 文字後の文字のラベル P の列
10. $\boldsymbol{p_{-2}p_{-1}p_0}$：2 文字前の文字のラベル P と直前文字のラベル P と注目文字のラベル P の列
11. $\boldsymbol{p_{-1}p_0p_1}$：直前文字のラベル P と注目文字のラベル P と直後文字のラベル P の列
12. $\boldsymbol{p_0p_1p_2}$：注目文字のラベル P と直後文字のラベル P と 2 文字後の文字のラベル P の列

5 実験

先に述べたコーパスから 1,000 文書をランダムに取り出し，それらの文書中の人名にタグを付与した．このデータセットを用いて 10 分割の交差検定を行うことで，文字ベースのみの CRF の検出と人名構成文字確率の情報を追加利用した CRF の人名検出能力の評価を行った．前者 (文字ベースのみ CRF) の結果を表 2 に後者 (本手法の CRF) の結果を表 3 に示す．これらの表により人名構成文字確率の情報を利用する効果が確認できる．なお CRF のツールとしては CRF++ [4]を用いた．

表 2: 文字ベースのみの CRF

data	人名数	抽出数	正解数	F 値
1	914	622	561	0.7305
2	1088	738	685	0.7503
3	895	550	497	0.6879
4	1144	668	596	0.6578
5	911	596	531	0.7047
6	1040	764	706	0.7827
7	979	595	530	0.6734
8	642	465	414	0.7480
9	879	633	565	0.7474
10	722	580	525	0.8065
平均	921.4	621.1	561.0	0.7289

表 3: 人名構成文字確率の情報を追加利用した CRF

data	人名数	抽出数	正解数	F 値
1	914	653	596	0.7607
2	1088	771	715	0.7692
3	895	577	513	0.6970
4	1144	782	713	0.7404
5	911	614	551	0.7226
6	1040	790	730	0.7978
7	979	595	531	0.6747
8	642	488	434	0.7681
9	879	673	598	0.7706
10	722	622	556	0.8274
平均	921.4	656.5	593.7	0.7528

[4] http://crfpp.sourceforge.net/

6 考察

本論文では人名構成文字確率を H と L の二値の素性として導入したが，素性の形式には連続値のまま導入したり，三値以上にすることも考えられる．ただし連続値は CRF の素性としては使えないので，なんらかの方法で離散化する必要がある．何種類の離散値にすればよいかは問題だが，訓練データの量に依存していると考える．つまり種類数が多い場合は，その素性の並びのパターンが多いという長所があるが，そのパターンの頻度が少なくなるという短所がある．本論文では少量の訓練データを想定したので二値としたが，三値以上の精度は未確認である．

また H と L を区別する閾値の設定の問題もある．本論文では訓練データ中の人名構成文字の人名構成文字確率の分布を調べ，訓練データ中の人名構成文字の約半数が 0.01 以上であったので，この値を閾値とした．この閾値を小さくすれば，多くの人名に H のラベルが付くので，抽出が容易になりそうだが，人名以外の文字にも H のラベルが付くケースも増えるために，単純ではない．最適な閾値は交差検定などを行うことで推定できる．

またここでは比較的大規模な人名リストを利用できた．そのような人名リストがあれば，文字列一致だけを利用して文書中の人名候補を取り出せる．候補になった文字列の文字に H，それ以外の文字に L の素性を与えれば，本論文と，厳密には異なるが，ほぼ同様の情報を学習に取り入れることになる．このような素性は冗長でも同時に利用することができるので，その点で改良が可能だと考える．

本実験の未検出と誤検出の原因としては文字ベースゆえの学習能力の低さがあると思われるが，誤検出についてはタグ付けの誤りも多かった．本論文で使用した人名タグつきのコーパスは中国人学生 1 名が手作業で構築したものであり，多くの誤りが含まれている．実際に交差検定で使われた data1 のテストデータについて調べると，文字ベースの CRF では 61 個の誤り中 35 個は，タグ付けの間違いであり，実際には正解となるべきものであった．また人名構成文字確率の情報を付与した本手法では，同じテストデータで，57 個の誤り中 37 個が実際は正解であった．このようにかなりの数のタグ付与されていない人名があるために，本実験での実際の正解率は，ここで出された正解率よりもかなり高いと思われる．ただし本手法の優位性は変化はないことは注記しておく．また未検出については 1 文字で構成される人名も目立った．例えば data1 のテストデータについて調べると，179 個の 1 文字による人名が存在したが，文字ベースの CRF では 105 個，本手法では 103 個が未抽出となっている．1 文字で構成される人名の検出が困難であることは [4] でも論じられている．中国語の姓は 1 文字で構成されるものが多いために，この点は何らかの対処が必要である．

今後の課題としては，人名以外の固有表現（地名や組織名など）への応用が考えられる．他の固有表現であっても構成文字にある傾向が存在すると考えている．本手法は比較的大規模な固有表現のリストがあれば利用可能である．そのようなリストは Web 上から比較的容易に収集できるために，本手法を人名以外の固有表現の抽出に試したい．

7 おわりに

本論文では中国語の人名抽出のタスクを文字列ベースの CRF を用いて行った．中国語のコーパスと中国語の人名リストを利用して，各文字が人名文字列の構成要素になる確率を推定した．この確率がある閾値以上のものに H，その閾値未満のものに L というラベルを与え，この素性を文字列ベースの CRF に追加利用することで，単純な文字列ベースの CRF の精度を改善することができた．

今後の課題としては人名以外の固有表現にも，その表現の構成要素にとなる確率という情報が CRF の有効な素性として利用できるどうかを調べたい．

参考文献

[1] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289, 2001.

[2] L. Li, Z. Li nad Z. Ding, and D. Huang. A Hybrid Model Combining CRF with Boundary Templates for Chinese Person Name Recognition. *Advanced Intelligence*, Vol. 2, No. 1, pp. 73–80, 2010.

[3] CW. Wu, SY. Jan, RTH. Tsai, and WL. Hsu. On Using Ensemble Methods for Chinese Named Entity Recognition. In *the Fifth SIGHAN Workshop on Chinese Language Processing*, pp. 142–145, 2006.

[4] X. Zhu, M. Li, J. Gao, and CN. Huang. Single character Chinese named entity recognition. In *the Second SIGHAN Workshop on Chinese Language Processing*, pp. 125–132, 2003.

[5] 高村大也．言語処理のための機械学習入門．コロナ社, 2010.