中国語専門用語抽出におけるCRF法とブートストラップ法の比較

李 寧*,　小川 泰弘,　大野 誠寛,　中村 誠,　外山 勝彦 (名古屋大学)

lining@kl.i.is.nagoya-u.ac.jp

1　はじめに

日本，中国，台湾，韓国，ベトナムなど漢字文化を共有する国と地域は，漢字文化圏と呼ばれる．漢字文化圏の人々は，漢字で記録した情報をある程度共有できるが，国や地域によって，同じ漢字で構成した用語であっても意味が異なっていたり，同じ意味を持つ用語であってもを構成する漢字が異なっていたりすることがある．

漢字文化圏において円滑な情報共有を行うためには，ターミノロジーの構築が 1 つの解決手段である．ターミノロジーの構築を支援するため，本稿ではコーパスの中から専門用語候補を自動的に抽出する方法に注目する．

中国語で書かれた文書から機械学習によって専門用語を抽出する研究では，特に近年，CRF などの系列ラベリングモデルを構築する手法が使われている．それに対して，本稿では，ブートストラップ法に基づく抽出アルゴリズムである Monaka[1] を中国語に適用する．CRF 法と比較することにより，抽出される専門用語にどのような違いがあるかを調査した．

以下，2 章では Monaka について紹介する．3 章では CRF 法と Monaka を用いた手法との比較実験について述べ，結果について考察する．4 章は本稿のまとめである．

2　Monaka

萩原ら [1] は，日本語コーパスから単語分割を行わないで専門用語を抽出するアルゴリズム Monaka を提案した．中国語は，日本語と同様に，文が分かち書きされない言語であり，Monaka は中国語にも適用できると期待される．

Monaka は，ブートストラップ法に基づく手法であり，与えられたシードインスタンスに対して，それと同じ意味カテゴリに属する表現を抽出する．Monaka の概略を図 1 に示す．

まず，入力として与えられたシードインスタンスからパターンを抽出する．Monaka では，インスタンスに隣接する文字 n グラムをパターンとする．例えば，

> S1：委員及び臨時委員は学識経験のある者のうちから，内閣総理大臣 が任命する．

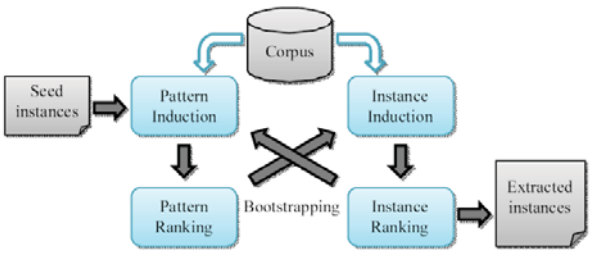

図 1: Monaka アルゴリズムの概略

という文からインスタンス “内閣総理大臣” に対応するパターンを抽出すると，“，#”，“ら，#”，“から，#” などの左側パターン及び “#が”，“#が任”，“#が任命” などの右側パターンが得られる．ここで，#はインスタンスのスロットを表す．

次に，抽出したパターンを評価するため，パターン p の信頼度 $r_\pi(p)$ を以下のように求める．

$$r_\pi(p) = \frac{1}{|I|} \sum_{i \in I} \frac{pmi(i,p)}{max_{pmi}} r_l(i) \ . \tag{1}$$

ここで，I はインスタンスの集合である．$pmi(i,p)$ はインスタンス i とパターン p との自己相互情報量で，

$$pmi(i,p) = \log \frac{|i,p|}{|i,*||*,p|} \tag{2}$$

によって計算する．ただし，$|i,p|$ は i と p の共起頻度を表し，$*$ はワイルドカードを表す．また，max_{pmi} は $pmi(*,p)$ の最大値である．

その後，信頼度の高い上位 n 個のパターンを用いてインスタンスを抽出する．具体的には，パターンのスロットの位置に存在する文字 n グラムをコーパスから抽出する．例えば，文 S1 にパターン “#が任命” を適用すると，“臣”，“大臣”，...，“内閣総理大臣”，“，内閣総理大臣” などのインスタンスが抽出される．

最後に，抽出したインスタンスの評価を行う．インスタンス i の信頼度 $r_\iota(i)$ は，式 (1) と同様に，

$$r_\iota(i) = \frac{1}{|P|} \sum_{p \in P} \frac{pmi(i,p)}{max_{pmi}} r_\pi(p) \tag{3}$$

により求める．ここで，P はパターンの集合である．

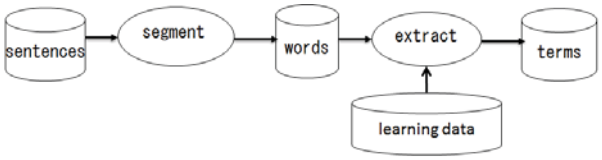

図 2: CRF による専門用語抽出の概略

このようにパターンを定義することにより，分かち書きに頼らない抽出が可能となる．しかし，上述のように，正しく分かち書きされていないインスタンスが大量に抽出されてしまう．そのため，Monaka では，「信頼度の高いインスタンスは，信頼度の高い右側文脈と左側文脈に挟まれなければならない」という**両側隣接制約**を導入している．具体的には，各インスタンスに対して左側信頼度 r_l と右側信頼度 r_r を求める．左側信頼度 r_l は，式 (3) の P として左側パターンのみを用いて計算する信頼度である．右側信頼度 r_r も同様である．インスタンスの信頼度 r_ι は r_l と r_r の一般化平均を用いて，

$$r_\iota(i) = \sqrt[m]{\frac{1}{2}(r_l(i)^m + r_r(i)^m)} \qquad (4)$$

として求める．一般化平均は，算術平均や幾何平均など各種平均の一般化であり，m によって両側隣接制約の強さを調節する．例えば，m を 0 に近い値にすることにより，r_l と r_r の両方の信頼度が高いときにのみ r_ι が高くなるという制約を表現することができる．これにより，Monaka では分かち書きの正しいインスタンスを高い精度で抽出することに成功している．

Monaka では，抽出されたインスタンスをシードインスタンスとして追加し，上述のステップを繰り返すというブートストラップにより，インスタンスを増やしていく．

3 実験と分析

本稿では，CRF 法による実験結果と Monaka を用いた方法による実験結果を比較することにより，中国語専門用語抽出における Monaka の有効性を検証する．

CRF 法では，あらかじめタグ付けされた学習データから CRF によって系列ラベリングモデルを構築し，それを入力データに適用することにより専門用語を抽出する [4][5]．その際，学習データ及び入力データは形態素解析されている．CRF 法の概略を図 2 に示す．一方，比較対象となる Mokana の実験では，既知の専門用語をシードとして与え，専門用語を抽出する．

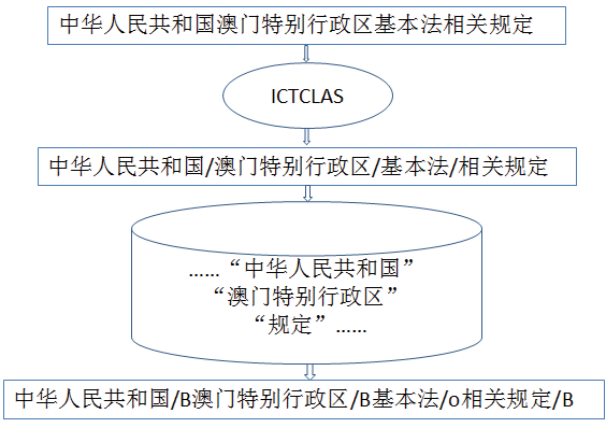

図 3: CRF 法学習データ作成の例

3.1 実験設定

実験に使用したコーパスは，中国大陸法律コーパス [2] 中の法令 223 本，総文数 29,909 文である．既知の専門用語としては，“搜狗”(Sogou)[1] 法律用語集に収録された 5,494 語を使用した．

3.1.1 CRF 法 (形態素解析+CRF)

本実験では，CRF 法の再現として，既知の専門用語をコーパス中でタグ付けして学習データとし，そこから系列ラベリングモデルを構築した．

具体的には，まず，コーパスを形態素解析ツール ICTCLAS[2] によって分割した後，Sogou 法律用語集にある専門用語を最長一致法により検索し，BIO タグを自動的に付加した．例えば，「中华人民共和国澳门特别行政区基本法相关规定」に対し，単語分割の結果は“中华人民共和国/澳门特别行政区/基本法/相关/规定”である．そこで，既知の専門用語として“中华人民共和国”，“澳门特别行政区基本法”，“规定”という 3 つが含まれるとする．この場合，学習データは“中华人民共和国/B 澳门特别行政区/B 基本法/I 相关/O 规定/B”になる（図 3）．

次に，この学習データを CRF++[3] で学習させ，系列ラベリングモデルを構築した．その際に使用した素性はウィンドウサイズ 5 以内にあるトークンの 3-gram，2-gram，1-gram である．

専門用語の抽出のためには，学習データと同じコーパスに対して，上記の系列ラベリングモデルを適用した．これはクローズド・テストであるが，学習データになかった専門用語も抽出される．

[1] http://pinyin.sogou.com/dic/
[2] http://ictclas.nlpir.org/
[3] http://crfpp.googlecode.com

表 1: 分割精度

抽出方法	新規獲得語数	正しい分割語数	分割精度
CRF 法	689	686	99.6%
Monaka	100	92	92.0%

表 2: 用語としての適切性

抽出方法	新規獲得語数	正解獲得語数	精度
CRF 法	689	678	98.4%
Monaka	100	92	92.0%

3.1.2 Monaka を用いた手法

Monaka では，繰り返し回数と繰り返し毎に追加するインスタンスの数がパラメータとなる．今回の実験では，繰り返し回数を 10 回，追加するインスタンス数を 10 とした．よって，100 個の語が最終的に抽出される．シードインスタンスとしては，Sogou 法律用語集に収録された 5,494 語をすべて使用した．なお式 (4) の m は 0.1 とした．

3.2 実験結果

実験の結果を表 1 から表 4 に示す．新規獲得語とは，抽出の結果のうち，既知の専門用語に含まれない用語である．表 1 の分割精度は，新規獲得語のうち，分かち書きが正しい語の割合である．

表 2 は新規獲得語が専門用語であるかどうかの判定をした結果である．判定の際には，インターネット上で検索した結果を参考にした．

CRF 法による抽出結果は 689 語であるが，その一部を表 3 に示す．表 4 に Monaka による抽出結果を示す．表 3 と表 4 では，分割を誤ったものに「*」を付加した．枠で囲んだものは専門用語として不適切なものである．

3.3 考察

3.3.1 分割精度と用語としての適切性

CRF 法は，形態素解析を行っているため，抽出結果の分割精度は 99.6%に達した．一方，Monaka による手法は形態素解析を行わなくても 92%の分割精度であった．このことから Monaka の両側隣接制約が中国語でも有効であったと考えられる．

Monaka では分割に失敗した例を考察すると，多くの誤り結果に出現する最後の文字は“的”である．例えば，“合同的”，“犯罪的”，“企业的”などがある．中国語文では，“的”の隣接文字列に名詞が現れる可能性が高い．Monaka では，用語インスタンスは信頼度の高いパターンによって抽出されるが，信頼度の高いパターンには，名詞の割合が高いことがその原因だと考えられる．実際，抽出された信頼度の高いパターンの一部を表 5 に示す．“#公民”，“#物证”，“#消费”，“#外资”など，名詞であるパターンが多く含まれている．

表 3: CRF 法による抽出結果（一部）

免税 技术转让 税率
奖金 废止 投资总额
建设用地 国有土地使用权出让 土地出让金
集体企业改制 企业改制 债权保全
资产损失 国有资产产权 变更登记
经营范围 职工代表大会 辞职
合同约定 保密义务 书面合同
公共利益 独创性 赔偿数额
国际条约 委托创作 合作开发
委托开发 软件著作权 非法人
软件著作权人 司法判决 转让合同
明知 不履行合同义务 违法所得
经营许可证 限期改正 废弃物
合理使用 对外经济贸易部 非法所得
对第三人* 先进先出法 捐赠
同一海事请求 非专利 利润分配

抽出結果が専門用語である精度について，CRF 法は 98.4%であった．Monaka の抽出精度は 92.0%であったが，これは分かち書きが正しくなかったためである．

3.3.2 新規獲得語の比較

Monaka による手法によって抽出した正解用語 92 語のうち，CRF 法によって抽出できなかった語は 89 語である．すなわち，CRF 法と Monaka による手法では異なる専門用語を抽出している．表 4 にある下線は，CRF 法と Monaka による手法の両方とも抽出できたものを示す．

抽出結果の長さを比較すると，CRF 法の抽出結果の平均長は 4.2 文字であり，Monaka の抽出結果の長さは 2.9 文字であった．また参考として，既知の専門用語の平均長は 4.9 文字である．

この差が生じた原因を調べてみる．CRF 法による結果では複数の単語からなる新規獲得語が多かった．今回の実験では，新規獲得語のうち，複合語は 564 語あり，新規獲得語の 8 割以上を占めた．一方，Monaka による結果では，複合語は 18 語であり，新規獲得語の 2 割未満であった．

Monaka では，新規獲得語が既知の専門用語の部分文字列である結果が多く見られる．例えば，“进口”という用語は，既知の専門用語である“进口药材管理办法”や“进口药品管理办法”などに含まれる．このように既知の専門用語の部分文字列である新規獲得語は 80 語であった．

この現象をさらに考察するために，信頼度の高いパ

表 4: Monaka による抽出結果

繰り返し回数	抽出結果
1 回目	港口 公司债券 外商投资企业 经济合同 进口 经营　企业 销售 中外合作办学机构 消费者
2 回目	食品 进境 广告 文物 军事设施 船舶 枪支 投资 工会　进行
3 回目	价款 裁决 技术 付款 声明 公路 价格 货物的* 境外的* 商业银行
4 回目	投标 承运人 环境 变更 标的 药品 过境 外商 生产　政府
5 回目	有下列情形之一的* 政府批准 进出口商品 境外 进出境 劳动 计算 拍卖 城市规划　成立
6 回目	公民的* 妇女 纳税人 产品 居民身份证 卫生 标底 企业破产 国旗　请求引渡
7 回目	外汇 利益的* 营业　外国人 税收 中国 各级人民政府 出口 民用航空器 职业教育
8 回目	撤销 外资保险公司 高等学校 档案 注册 使用 借款 投标人 海关　生产经营
9 回目	招标人 施工 国籍的* 成立后* 建筑工程 股份 交通 劳务 领海　合作企业
10 回目	股票 发生 拍卖标的　解散 法律 要求 依照本法 法律、法规* 保险公司 货物

ターンを調べると，名詞で始まる右側パターンが多かった．これにより，複合語の先頭部分が抽出されたと考えられる．

ここで，萩原ら [1] による日本語への Monaka の適用と比較する．日本語に適用した場合，信頼度の高い右側パターンには平仮名で始まるものが多かった．すなわち，日本語では専門用語の直後に区切りとなる平仮名が出現しやすく，それがパターンとして採用されやすいと考えられる．一方，中国語では平仮名と漢字といった字種の区別がないため，専門用語の直後に別の名詞が連続して出現し，それがパターンとして採用される．

これをまとめると，Monaka を日本語に適用した場合，信頼度の高いインスタンスと信頼度の高いパターンには，異なる種類の文字列が出現する傾向があるが，中国語に適用した場合は，信頼度の高いインスタンスと信頼度の高いパターンに，同じ種類の文字列が出現する傾向があると言える．また，これは先述の単語分割の誤りの原因にもなっていると考えられる．

すなわち，分かち書きされない言語に適用可能な Monaka であっても，適用する言語の特徴により，性能に差が出ると言える．

表 5: 信頼度の高いパターン

#按规定,#部门对举,#从事有
#另一方投资入,#的无形资产,#企业应
#物证,#抵押人所,#药品管理法
#公民,#中外合,#法官
#合伙企,#消费,#外资
册资本#, 意见#, 收益#, 险费#
义务的#, 广播、电视节目#, 境外#, 职务#
物品#, 商品#, 外汇#, 年的#
市场的#, 息管理#, 定不服#, 视为#,...

4　おわりに

漢字文化圏のターミノロジーの構築を支援するために，本稿は CRF による手法と Monaka による抽出手法を中国語コーパスに適用した．その 2 つの手法を比較した結果，両者には共通する用語が少ないことがわかった．よって，この 2 つの手法を併用することにより，より多くの種類の専門用語の抽出ができる．今後は，多言語のターミノロジー構築のため，同じ意味を持つ専門用語の間での対応付けする方法を検討する．

参考文献

[1] Masato Hagiwara, Yasuhiro Ogawa,and Katsuhiko Toyama, Bootstrapping-based Extraction of Dictionary Terms from Unsegmented Legal Text. New Frontiers, *in Artificial Intelligence: JSAI 2008 Conference and Workshops, Revised Selected Papers, Lecture Notes in Computer Science*, Vol.5447, pp.213-227, Springer (2009).

[2] Sun Hongren, Yang Dingjian., Construction of Parallel Corpus of Chinaś Legal Documents: Philosophy,Processes and Functions. *Journal of Shaoxing University*, Vol.2, pp.48-51 (2010).

[3] Lance A. Ramshaw, Mitchell P. Marcus, Text Chunking using Transformation-Based Learning. In*Proc. of the Third Workshop on Very Large Corpora*, pp.82-94 (1995).

[4] Jia Meiying, Yang Bingru, Zheng Dequan, Yang Jing, Research on automatic military intelligence term extraction using CRF model. *Computer Engineering and Applications*, Vol.45(32), pp.126-129, *in Chinese* (2009).

[5] Li Lishuang, Dang Yangzhong, Zhang Jing, Li Dan, Automotive term extraction based on conditional random fields. *Journal of Dalian University of Technology*, Vol.53(2), pp.267-272, *in Chinese* (2013).