2009－NL－190 (11)
2009／3／26

素性にモーラとシラブルを用いた略語の自動推定

和田 健太* 近山 隆* 横山 大作† 三輪 誠‡

*東京大学大学院工学系研究科 †東京大学 IRT 研究機構 ‡東京大学大学院情報理工学研究科

検索システムや文章要約においては同義語の獲得が必要である。検索システムでは同義語が分かっていれば同義語を用いて検索を行うことが可能であり、文章要約では同義語の一種である略語が分かっていれば、より短い文章を生成することが可能である。

本研究では同義語の一種である略語を推定する手法について提案する。略語は人間が生成するものであるため、人間の感覚が重要であると考えられる。そこで人間の感覚を取り込むため、CRF の素性にモーラとシラブルを用いることにより入力された原語から略語を推定する。

Automatically Figuring Out Abbreviations using Mora and Syllable as Feature

Kenta Wada* Takashi Chikayama* Daisaku Yokoyama† Makoto Miwa‡

*Graduate School of Engineering, The University of Tokyo
†IRT Research Initiative, The University of Tokyo
‡Graduate School of Information Science and Technology, The University of Tokyo

The acquisition of the synonym is necessary to a IR (information retrival) system and summary production. When understanding synonym by IR system, it is possible to search using synonym. And, when understanding the abbreviation which is one kind of synonym by summary production, it is possible to generate a shorter sentence.

I suggest technique to figure out an abbreviation which is kind of the synonym automatically. Since human generates an abbreviation, feeling of human is important for it. There, mora and syllable is used for the feature of CRF to take a human sence in, an abbreviation is presumed from input original language.

1 はじめに

我々人間は、異なる単語を用いて同じ意味を持つ文を生成することができる。例えば「会議に出る」と「会議に出席する」という二つの文章は、表層上の文字と発音は異なるものの、それぞれの文が意味するものは同じである。これは文に限らず単語にも同じ事が言え、表記・発音は異なるが同じ意味を持つ語を同義語と呼ぶ。同義語の例としては「私」に対して「僕」「手前」「我」などがあり、「東京大学」に対しては「東大」のような語が挙げられる。

自然言語処理ではこのような同義語を獲得することが重要な意味を持っている。たとえば、検索システムにおいて検索キーワードとして用いるクエリが他の同義語を持つと分かっていれば、その同義語でも検索を行うことができる。また、一つの文章にて

何度も用いられる比較的長い名詞というのは省略される事が多いが、文章要約システムにおいては略語があらかじめ分かっていればそれを用いて要約を行うことが出来る。

ところで略語というのは人間が生成するものであるため、人間の感覚というものが重要になってくる。[7] では音韻論の立場から略語がどのように生成されているかを考察し、モーラが略語と深い関係にあることを示した。

そこで、本研究ではモーラとシラブルを CRF の素性として用いて、同義語の一種である略語を自動推定する手法を提案する。本研究で対象とする略語は日本語略語である。具体的には Table 1 に挙げるような略語を考える。以下では省略される前の語を「原語」、省略された語を「略語」と呼ぶ。

Table 1: 本研究で扱う略語の例

原語	略語	略語の特徴
文部科学省	文科省	「文部」「科学」「省」の頭文字を取った頭字語と呼ばれる略語
最高裁判所	最高裁	「最高」の文字を全て取った略語
マイクロコンピュータ	マイコン	カタカナ略語
航空自衛隊	空自	頭文字の「航」ではなく「空」を取ったタイプの略語
原動機付自転車	原付	「自転車」が完全に抜け落ちた略語
アメリカ	米	原語には使われていない文字が使われている略語

2 関連研究

2.1 原語・略語対の獲得

酒井らは原語と略語の対をコーパスから抽出する手法を試みた [1]。この手法は以下のような手順による。

1. 名詞の文字情報を用いて名詞 X の略語である可能性のある名詞 (略語可能性名詞)Y をコーパスから探し出す。
2. 次の条件を満たすときに略語可能性名詞 Y が名詞 X の 略語であると判定する。
   (a) 略語可能性名詞 Y に使われている文字が全て原語である名詞 X に含まれている。
   (b) 原語 X と略語可能性名詞 Y の先頭の文字が一致する。
   (c) 原語 X と略語可能性名詞 Y に使われている文字の出現順序が一致する。

この手法により酒井らは 74%の精度で略語を抽出する事に成功したが、「アメリカ」↔「米」のような略語は (a),(b) の条件に当てはまらないので抽出することができていない。また、「大阪大学」↔「阪大」のような略語は (b) の条件に当てはまらずに、やはり抽出することができない。この手法では表層上の文字のみを用いているため、獲得しうる略語に強い制限があるという欠点を持っている。

岡崎らもやはり原語と略語の対を獲得しようとした [3]。対象はコーパスではなく新聞記事で、以下に示す手順により略語を抽出した。

1. 新聞記事から括弧表現「X(Y)」を抽出する。
2. X と Y が相互に言い換え可能であるか SVM により判定する。
3. SVM の素性には次のようなものを用いた。

(a) X と Y の文字種 (漢字、ひらがな、カタカナ、アルファベット)

(b) X と Y の品詞

(c) X 及び Y の係り受け関係 (コンテキストの分布距離)

(d) 文字の包含

(e) χ^2 による共起度

(f) その他

この手法では酒井らが獲得することができなかった「大阪大学 (阪大)」のような略語も獲得することができる。しかし品詞を用いるために、略語が辞書に登録されている必要があり、また、括弧表現にない略語を獲得することはできないという欠点がある。新聞記事のような形式がある程度整っている文章では有効だが、昨今の web 上にあるような文章を対象とした場合、この手法が有効であるかには疑問符をつけざるを得ない。

2.2 略語の推定

村山らは略語をコーパス中から抽出するのではなく、原語から略語を自動推定する手法を提案した [4]。ベイズの定理を用いて、原語 $\boldsymbol{x}$ から尤もらしい略語 $\boldsymbol{y}$ を推定する手法である。尤もらしい略語を $\hat{\boldsymbol{y}}$ とすると

$$\begin{aligned} \hat{\boldsymbol{y}} &= \underset{\boldsymbol{x}}{\operatorname{argmax}} P(\boldsymbol{y}|\boldsymbol{x}) \\ &= \underset{\boldsymbol{x}}{\operatorname{argmax}} \frac{P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})}{P(\boldsymbol{x})} \\ &= \underset{\boldsymbol{x}}{\operatorname{argmax}} P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x}) \end{aligned} \tag{1}$$

となる。

このようにして原語 $\boldsymbol{x}$ が与えられた時の条件付確率 $P(\boldsymbol{y}|\boldsymbol{x})$ を最大化するような $\boldsymbol{y}$ を選ぶことにより、尤もらしい略語を推定することができる。しかしこの手法は生成モデルを用いており、一般に識別モデルと比べて精度の面で劣るという欠点がある。実際、この研究では推定した略語の上位一位では再現率が 0.135、上位 10 位で 0.518 であり、高精度の手法であるとは言い難い。

2.3 略語とモーラ・シラブル

略語の抽出・生成とは別に、略語とモーラ・シラブルの関係を調べた論文もある。まずモーラとシラブルについて簡単に説明する。

モーラは拍とも呼ばれ、日本語の音のリズムを表すものである。俳句や短歌における 5・7・5 や 5・7・5・7・7 といった数は、このモーラ数を数え上げたものである。また、シラブルは一つの母音を中心とした音のかたまりの事であり、音節とも呼ばれている。Table 2 に実際に単語がどのようなモーラとシラブルで構成され、計何モーラ、何シラブルを持っているかを示す。なお、括弧内の数字がモーラ数及びシラブル数である。

Table 2: 単語のモーラ数とシラブル数

語	モーラ	シラブル
はこ	は / こ (2)	は / こ (2)
しゃしん	しゃ / し / ん (3)	しゃ / しん (2)
きって	き / っ / て (3)	きっ / て (2)
おうさま	お / う / さ / ま (4)	おう / さ / ま (3)

モーラでは拗音はその前の音とまとめて数えられるため、「しゃしん」は「しゃ」で 1 モーラとなる。シラブルでは拗音だけでなく撥音、長音、促音も前の音とまとめて数えられるため、「しゃしん」は「しゃ / しん」のように 2 シラブルとなる。

鈴木は外来語略語とモーラ・シラブルの関係について調査した [5]。それにより次に示すような事が明らかになった。

1. 略語のモーラ数は 4 モーラ (43.5%) と 3 モーラ (25.5%) が多い
2. 複合語の頭字語略は 2 モーラ+2 モーラのパターンが最も多い (83.6%)
3. 2 シラブル 2 モーラで始まる略語のパターンが最も多い (52.1%)

この調査により、略語とモーラ・シラブルの間には何らかの関係があることが明確になった。

3 提案手法

上記の関連研究を踏まえ、本輪講では CRF の素性にモーラとシラブルを用い、識別モデルにより略語の自動推定を行う手法を提案する。この手法では「アメリカ」↔「米」のような原語に含まれていない文字が略語に使われるパターンや、「大阪大学」↔「阪大」のような頭文字が一致しない略語でも推定することができる。また、略語を推定するため「X(Y)」のような括弧表現に依存することがない。さらに、モーラとシラブルを考慮することにより、より人間の感覚にマッチした略語を推定する事ができる。

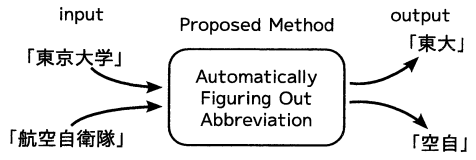

Fig 1: 提案手法

3.1 略語の推定方法

本輪講で提案する手法の具体的な手順を述べる。複合語であればまず原語 $\boldsymbol{x}$ を区切り、それにたいして CRF を用いて略語ラベル $\boldsymbol{y}$ を付与することにより略語の推定を行う。Fig 2 にその例を示す。ここでは「住民基本台帳ネットワークシステム」に対する略語を推定する事を考える。「住民基本台帳ネットワークシステム」は複合語であるので、Fig 2 に示すよう、「住民-基本-台帳-ネットワーク-システム」と原語 $\boldsymbol{x}$ を区切る。これは MeCab によって実現できる。その後、可能性として考えられる略語 $\boldsymbol{y}$ をいくつか生成し、これを CRF によって評価してもっともらしい略語を選択する。

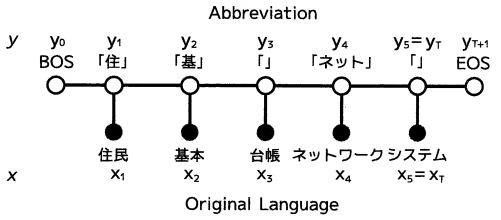

Fig 2: 略語推定

CRF の素性としては原語・略語の特徴を表すものとして次のものを用いる。

観測素性 f_O ：原語・略語間の特徴を表す。ある原語 x_i に対してどのような略語 y_i が付加されるかを示す。

遷移素性 f_T ：略語間のモーラ・シラブルのつながりを表す。略語 y_i の持つモーラ数に続いて、y_{i+1} が何モーラであるかを示す。シラブルについても同様である。

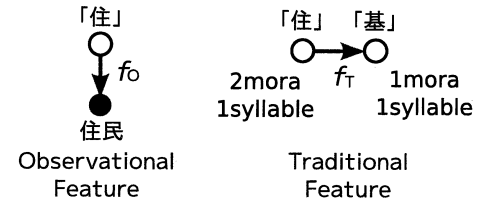

Fig 3: CRF の素性

具体的には Fig 3 に示すように、観測素性 f_O は「住民」に対して「住」を付加する素性であり、遷移素性 f_T は「住」のモーラ数及びシラブル数 (2 モーラ 1 シラブル) に続く「基」が何モーラ, 何シラブルを持っているか (1 モーラ 1 シラブル) を表す素性である。

素性の集合を $\boldsymbol{F}$ とし、これらの素性 $f_i \in \boldsymbol{F}$ が原語と略語のペア $(\boldsymbol{x}, \boldsymbol{y})$ にて成立する回数を ϕ_{f_i} とする。また、素性の重要度を考え、それを θ_{f_i} とする。

さらに、ϕ_{f_i} および θ_{f_i} を並べたベクトルを $\Phi(\boldsymbol{x},\boldsymbol{y})$ ならびに Θ と表記する。

$$\begin{aligned}\Theta &= (\theta_{f_1}, \theta_{f_2}, \cdots)\\ \Phi(\boldsymbol{x},\boldsymbol{y}) &= (\phi_{f_1}(\boldsymbol{x},\boldsymbol{y}), \phi_{f_2}(\boldsymbol{x},\boldsymbol{y}), \cdots)\end{aligned}$$

この $\Phi(\boldsymbol{xy})$ が CRF におけるパラメータである。

CRF を用いた略語の推定を行うにあたり、まず式 (2) に示す確信度を定義する。

$$< \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) >= \sum_{f \in \boldsymbol{F}} \theta_f \phi_f(\boldsymbol{x},\boldsymbol{y}) \tag{2}$$

しかしこの確信度は負値をとることもあれば、1 を越えることもある。そこで指数の肩に乗せることにより条件付確率分布 $P(\boldsymbol{y}|\boldsymbol{x})$ を定義する。

$$P(\boldsymbol{y}|\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{\exp < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) >}{\sum_{y \in Y} \exp < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) >} \tag{3}$$

式 (3) を最大化する際に分母は無関係であるため、式 (2) に代入することにより

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in Y}{\operatorname{argmax}} < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) > \tag{4}$$

となる。式 (4) を満たすような $\hat{\boldsymbol{y}}$ を求めることにより、原語 $\boldsymbol{x}$ から略語 $\boldsymbol{y}$ を推定することができる。

3.2 実験結果

実験には、wikipedia の漢字略語の項目に上げられているものを学習用データとして用いた (349 語)。評価には [8] の漢字略語の項目のうち、学習用データに含まれていない 70 語を用いた。

再現率と適合度を式 (5), (6) とした場合の実験結果を Table 3 に示す。

$$\text{再現率}\ (n) = \frac{\text{上位 } n \text{ 位の出力のうち、正解に含まれる数}}{\text{正解の略語数}} \tag{5}$$

$$\text{適合度}\ (n) = \frac{\text{上位 } n \text{ 位の出力のうち、正解に含まれる数}}{\text{上位 } n \text{ 位の全出力数}} \tag{6}$$

先行研究 [4] では Table 4 に示すような結果が得られている。これと比較すると本研究では F 値が劣っているが、先行研究に対して学習用データを約半分程度しか用いていないという点に着目すると、同程度の結果が得られているのではないかと考える。

Table 3: 実験結果

	再現率	適合度	F 値
上位 1 位	0.129	0.129	0.129
上位 2 位	0.157	0.079	0.105
上位 3 位	0.214	0.071	0.107
上位 5 位	0.257	0.051	0.086
上位 10 位	0.357	0.036	0.065
上位 20 位	0.486	0.024	0.046
上位 30 位	0.514	0.017	0.033

Table 4: 先行研究

	再現率	適合度	F 値
上位 1 位	0.135	0.153	0.143
上位 2 位	0.214	0.121	0.155
上位 3 位	0.253	0.096	0.139
上位 5 位	0.307	0.079	0.126
上位 10 位	0.518	0.060	0.108
上位 20 位	0.614	0.038	0.072
上位 30 位	0.684	0.030	0.057

また、モーラに関する重要度 Θ_m とシラブルに関する重要度 Θ_s の分散 s_m^2, s_s^2 をそれぞれ求めると

$$s_{\mathrm{m}}^2 = 9.12 \times 10^{-2} \tag{7}$$

$$s_{\mathrm{s}}^2 = 1.77 \times 10^{-2} \tag{8}$$

となった。これより、略語を生成するにあたってはシラブルよりもモーラの方がその重要度が高いということが分かった。

4 おわりに

本研究ではモーラとシラブルを用いて略語を自動推定する手法を提案した。また、CRF のパラメータより、シラブルよりもモーラが略語の生成において重要な役割を担っているという知見が得られた。

今後の課題としては学習用データを増やし、精度の向上を図ることが最初に挙げられる。また、本手法ではCRFの素性として隣り合うモーラ・シラブル同士の素性しか考慮していない。しかし、鈴木らの研究[5]によれば略語全体のモーラ数は4モーラ及び3モーラが多いということが示されている。そこで略語全体におけるモーラ数・シラブル数を考慮する事が改善手法として挙げられる。さらに、素性として音訓の情報を用いることが考えられる。湯桶読み、重箱読みが少なくなるような略語の生成を行うことにより、精度を向上させられるのではないかと考えている。

参考文献

[1] Hiroyuki SAKAI, Shigeru MASUYAMA, "Knowledge Acquistion of Relation between Abbreviations and Their Original Words", Institute of Electronics, Information and Communication Engineers. Vol.J85-D-II No.10(2002) pp.1624-1628

[2] Hiroyuki SAKAI, Shigeru MASUYAMA, "Improvement of the Method for Acquiring Knowledge from a Single Corpus on Correspondences between Abbreviations and Their Original words", Journal of natural language processing Vol.12 No.5 pp.207-231, 2005

[3] Naoaki Okazaki, Mitsuru Ishizuka, "Abbreviation Recognition in Japanese Newspaper Articles", The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007

[4] 村山紀文, 奥村学,「Noisy-channel modelを用いた略語自動推定」, 言語処理学会第12回年次大会発表論文集 pp.763-766, 2006

[5] 鈴木俊二,「外来語の略語の構造 –音節・モーラ・フット・語」, 国際短期大学 [編] Vol.11 pp.21-44, 1996

[6] 桑本裕二,「日本語におけるモーラ的鼻音の特徴」, 東北大学言語学論集 第11号, pp.93-104, 2002

[7] 窪薗晴夫,「新語はこうして作られる」, 岩波書店, 2002

[8] K's Bookshelf 辞典, http://ksbookshelf.com/DW/Ryaku/index.html