E-028

確率モデルを用いた略語の自動推定

Automatic Estimation for Abbreviated Words that Uses Probabilistic Model

大工廻 史裕† 三輪 貴大‡ 浦谷 則好†‡
Fumihiro Takue Takahiro Miwa Noriyoshi Uratani

1. まえがき

本論文では Web と確率モデルを用いた略語自動推定法を提案する．

確率モデルを用いることで，尤度による略語候補の順位付けをすることが可能である．

字数制限下の文章作成においてより多くの事柄を伝達する時に，略語の利用が有効である．しかし，略語は日々生成され変動しており，既存の辞書では未知語となる略語が常に増加している．しかし，略語を人手で推定するには膨大な労力と時間が必要である．そこで，略語を原語から自動的に推定できれば，文章解析の手間を大幅に減らすことができる．

また，略語から原語が分かれば，固有表現抽出の精度や，情報検索における検索効率の向上に寄与できると考えられる．

2. 関連研究

略語の推定や略語辞書の作成に関しては，これまでに次のような研究が報告されている．

桝井らは以下のような手法を試みている．カタカナ語の略語に限定し，原語から略語を推定するいくつかのルールを用意して略語候補を生成し，Web での出現数や接辞辞書を用いて略語を推定する．[1]

村山らは以下のような２つの手法を試みている．原語となりうる可能性のある語のリストが与えられるとしたとき，正解データからの学習によって得られる原語―略語候補対の文字列的な正当性，Web 検索エンジンから得られる統計値から得られる意味的・社会的な正当性の双方向を考慮し，略語を推定する．[2]

「略語関係」にある対を，「同義関係にある語の集合の中で，一方がもう一方から表層的に短縮されているような語の対」とし，原語から表層的な情報のみを利用して確率モデルにより，略語候補を生成する．[3]

酒井らは以下のような手法を試みている．[1]と同様にルールを用いて絞り込み，略語候補と原語の候補の名詞間類似度を計算することで，コーパスから略語とその原語との対応関係を自動的に獲得する．[4]

和田らは以下のような手法を試みている．人間の感覚を考慮するため，音韻論の立場からのモーラ・シラブルを CRF の素性に活用し，略語を自動推定する．[5]

土田らは以下のような手法を試みている．Web の更新速度と情報の網羅性に着目し，目的語の説明や定義を Web から抽出し，複数の抽出文における類似文判断により説明文を特定する辞典システムを構築する．[6]

†東京工芸大学大学院工学研究科電子情報工学専攻
‡東京工芸大学工学部コンピュータ応用学科
uratani@cs.t-kougei.ac.jp

3. システムの概要

3.1 前提

以下では省略する前段階の語を「原語」，省略された語を「略語」とする．

また，略語の生成型として，単語の後部の文字を省略するものを「後略」，前部の文字を省略するものを「前略」，一文字も省略しないものを「不略」，全ての文字を省略するものを「全略」と呼ぶ．

さらに，三文字以上で構成される単語の場合，文字列の途中のみを省略する省略形はないものとする．例えば，「工学院」の場合，「工院」は考慮しない．

毎日新聞 2007 年データと Wikipedia から約二千単語の略語を抽出し，原語の単語ごとの略語生成型の組合せ関係をデータベース化した．この結果，文献[4]と同様に略語文字の順序は原語の文字順序とおおよそ等しいことが分かった．このため，略語文字の順序は逆順にならないものと考える．

3.2 システム構成

本節では，本研究で構築する略語推定システムの構成図を図 1 で示す．

1）略語推定をしたい原語を入力し，MeCab による形態素解析を行う．

2）形態素解析された原語に，単語区切りマーカーを単語境界に挿み，単語区切りを明確にする．同時に，単語を一文字ごとに区切る．

例えば，「修士論文」が原語であれば「修士」と「論文」にわけ，「修士」をさらに「修」と「士」とする．

3）単語区切りマーカーを用いて，単語ごとに｛後略・前略・不略・全略｝の順で単語の省略形を記憶する．

例えば，「修士」と言う単語であれば｛「修」・「士」・「修士」・「」｝となる．

4）3)で求めた単語の省略形を用いて，単語と単語での組合せを生成する．

このとき全単語が「不略」のみ又は，「全略」のみになる組合せは生成しないものとする．また一文字となるものを除く．

例えば，「修士論文」が原語であれば生成されるものは，｛「修論」・「修文」・「修論文」｝，｛「士論」・「士文」・「士論文」｝，｛「修士論」・「修士文」・「修士」｝，｛「論文」｝になる．

5）4）の結果を Web 検索し，略語候補を取得する．

6）種々の重み付けの付与をする．

例えば，単語の先頭文字の優先度が高い傾向が先行研究により見られるので，重みを増す．

7）6)で得られた略語候補上位を人手で判定し，原語の正確な略語であるかを判定する．

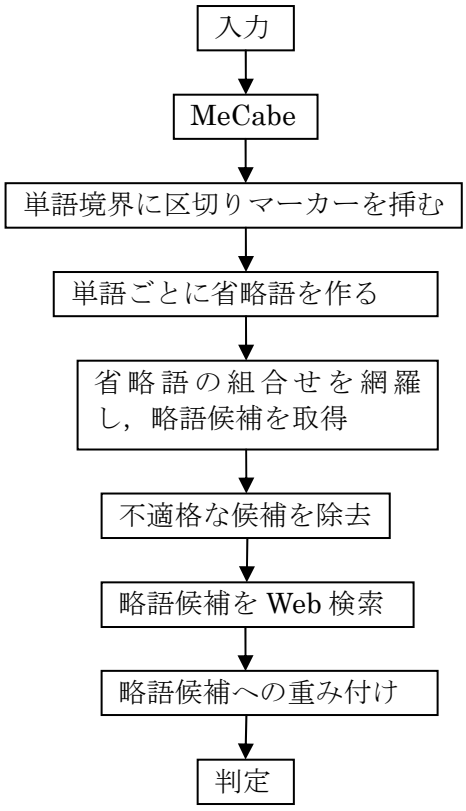

図 1：略語推定システムの構成

4. 結果

提案手法での実験を行った結果を報告する．実験では Yahoo! JAPAN での検索を用いた．

本提案手法での Web 検索する前段階での略語候補は 3.2 の 4）に示す例のようになる．この略語候補に重み付けをせず Web 検索した場合について表 1，表 2 に示す．表での数値は検索該当数である．

表 1，表 2，で上の 2 語は「不略+全略」，「全略+不略」で略語の資格を持たない．

上の 2 語の頻度を f_1，f_2 とし，重みを w とした場合 f_1*f_2*w を以上となるものを略語と推定する．

表 1：「修士論文」を原語とする略語候補

	Yahoo
論文	94,400,000
修士	19,800,000
修論	1,440,000
修文	1,390,000
士文	252,000
士論文	155,000
修士論	15,900
士論	3,800
修論文	1,930
修士文	12

表 2：「結婚活動」を原語とする略語候補

	Yahoo
結婚	491,000,000
活動	143,000,000
婚活	29,600,000
結活	42,200
婚活動	12,800
結動	941
結婚動	913
結婚活	897
結活動	573
婚動	399

5. おわりに

「修士論文」，「結婚活動」のどちらでも正確な略語である「修論」，「婚活」は上位に入っている．

表 1.において，略語候補「修論」と「修文」の検索該当数の差がほかに比べて少なく，二つとも略語と認定される．

重みを適切に決めることで，略語をかなり正確に推定できることが確認できた．

今後文献[1]のように接辞による適否判断を加えて，より精度の向上を目指していく．

参考文献

[1]桝井文人，松田良一，野呂康洋，河合敦夫，井須尚紀:World Wide Web を知識源としたカタカナ語省略形の自動生成．2004 年度電子情報通信学会ソサイエティ大会講演文集，A-13-1，2004.

[2]村山紀文，奥村学:Web 情報を利用した確率モデルによる略語推定．情報処理学会研究報告，2009-NL-183

[3] 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.736-766，2006

[4]酒井浩之，増山繁:略語とその原型語との対応関係のコーパスからの自動獲得手法の改良．自然言語，Vol.12，No.4，pp． 207-231，2005.

[5]和田健太，近山隆，横山大作，三輪誠:素性にモーラとシラブルを用いた略語の自動推定． 情報処理学会研究報告，20009-NL-190.

[6]土田正明，松井藤五郎，大和田勇人:World wide web を用いた辞典システムの構築．第 18 回人工知能学会全国大会， 1A3-04，2004.

[7] MeCab，http://mecab.sourceforge.net/