漢字を中心とした複合語の略語の自動生成

ー音訓を考慮したルールを用いてー

9月14日（月）
豊川　幸秀

目的

- 略語は過去から現在に至り、様々な場面で使われてきた。

↓

その仕組みを理解することは、文書検索や文書要約といった場面において有用である

・原語：元になる語
・略語：その全ての文字が原語中に同じ順で出現する短縮語

過去の関連研究

○「ワールドワイドウェブを利用した用語説明の自動生成（桜井裕、佐藤埋史）」[1]

○「Noisy-Channel model を用いた略語自動推定（村山紀文、奥村学）」[2]

いずれも、略語自身の読みや、漢字の音訓の情報が利用されていない。

全体の流れ

１．略語生成
-1:略語の生成ルール取得
-2:このルールを用いて、略語候補群から最適なものを出力

２．上記のシステムにおける、音訓ルールの有効性について実験にて証明

＊尚、今回扱う略語は漢字のみで構成されたものに限定し、使用するデータは文献[3]から抽出してきた原語と略語のセット678組を用いる。

略語の生成ルール

以下、3つのルールを用いる。

①形態素の省略

②文字数の減少

③音訓の組み合わせと読み

適切な略語の候補を決定するために、これらのルールについて見ていく。

形態素の省略(1)

	原語 →	形態素解析 →	略語
(1)	短期大学	短期、大学	短大
(2)	関西国際空港	関西、国際、空港	関空
(3)	伊豆急行	伊豆、急行	伊豆急
(4)	経営財団	経営、財団	営団

(1)先頭文字を抽出　(2)使用しない形態素が存在
(3)短縮しない形態素が存在　(4)後方文字を抽出

形態素の省略(2)

- 主に(1)が主流だが、他3つのパターンも決して少なくはない。
- 一つの言語に対し、複数の法則が適用されている例もある。
- 同様に、一つの形態素に対し、複数の法則が適用されている例もある。

➡ 形態素の省略法則だけでは、適切な規則の発見は難しい。

文字数の減少

- 直感的に、略語は原語よりも文字数が減少する。
- その中でも特に、短縮語は2文字もしくは3文字になる可能性が高い。

  (訓練データでは、2文字は約41%、3文字は約38%であった)

音訓の組み合わせと読み(1)

モーラ：1モーラ＝仮名1文字もしくは「きゃ」や「しゃ」といった拗音１つ分

フット：１フット＝２モーラ

・元の読みが1モーラの場合を除いて、新しい語は原則最小１フットの長さを持つ

・日本語は、２モーラの長さが使われることが多い。

音訓の組み合わせと読み(2)

Ⅰ．A「名古屋駅（ナ）」→B「名駅（メイ）」

Ａ：訓・１モーラ

Ｂ：音・２モーラ

上記のような場合に、変換時に音訓が変わることがある。（逆も然り）

Ⅱ．A「静岡大学（シズ）」→「静大（セイ）」

Ａ：訓・２モーラ

Ｂ：音・２モーラ

上記のようにどちらも同じ場合は、音を優先して変換することが多い。

音訓の組み合わせと読み(3)

これらから、適切な略語を選択するためには、略語のモーラ数および漢字の音訓の情報にも着目することが必要だと分かる。

略語生成システム

- 先程までの生成ルールから、略語候補それぞれの評価値を算出して適切な略語を割り出していく。大まかな流れは以下のとおり。

１．略語の定義に該当する候補群を取得

２．候補を生成ルールで評価

３．複数の上位候補を略語候補として出力

評価値の合計の大きい順に出力する。

評価(1)

(1)形態素の省略

４通りの法則があるが、どれを適用するかの判断は困難・・・

↓

形態素ごとに、各法則の適用頻度を訓練データから調べ、形態素辞書に付加

評価(2)

略語候補a_kにおける評価値m_kは以下の式により算出する。

$$m_k = P_m(a_k) = \sum_{1}^{\text{形態素数}} \frac{(\text{法則の適用頻度})}{(\text{形態素の出現頻度})}$$

例.「損害保険」→「損害・保険」→「損保」

損害：先頭文字の法則

保険：先頭文字の法則

それぞれの法則の頻度と形態素自身の頻度を上の式に当てはめることにより算出できる。

評価(3)

(2)文字数の減少

２文字または3文字の略語が多いが、それ以上である可能性もある。よって、原語の文字数と略語の文字数の関係から、生成されやすい文字数の略語候補の評価値 n_k が高くなるようにした。式は以下。

$$n_k = P_n(a_k) = \frac{(a_k\text{と同じ文字数の略語の出現頻度})}{(\text{原語と同じ文字数の原語の出現頻度})}$$

評価(4)

(3)音訓の組み合わせと読み

モーラ数と音訓のルール2つについて考える。

〇モーラ数評価

略語において各語は、原則最低１フットの長さを持つため、２＋２モーラ（２＋２＋２モーラ）の長さの読みになることが多い。これを踏まえ、訓練データから文字数とモーラ数の関係を調べ、各略語候補のモーラ数に対する評価値を求める。

評価(5)

○音訓評価

モーラ数の都合や、重箱読み、湯桶読みの回避のために音訓が変化するケースを訓練データから抽出し、それを元に、求める略語候補に音訓の組み合わせの評価値を与える。

＊音訓情報の取得にはSKK辞書(http://openlab.jp/skk/wiki/wiki.cgi)を利用した。

評価値y_kを与える式は以下。

$$y_k = P_y(a_k) = (\text{モーラ数の評価値}) + (\text{音訓の組合せの評価値})$$

実験(1)

- 音訓の組み合わせのルールを適用した場合としない場合について比較する。

対象データ：「フリー百科事典Wikipedia(http://ja.wikipedia.org/wiki/)」の「漢字略語一覧」の原語・略語セットから抜粋した１４８組のデータセット（本論文の略語の定義と一致し、且つ訓練データと重複しないもの）

形態素解析：「Chasen(http://chasen-legacy.sourceforge.jp/)」を用いて行った。しかし、適切に分解できなかったため、最終的に手動で分けた。

実験(2)

- 再現率と精度を以下で定義する。

$$\text{再現率}_n = \frac{(\text{上位}n\text{位のうち、本来の略語と一致する数})}{(\text{データ数})}$$

→上位n位で本来の略語が得られる確率

$$\text{精度}_n = \frac{(\text{上位}n\text{位のうち、本来の略語と一致する数})}{(n \times \text{データ数})}$$

→本来の略語を得るためのコスト

実験(3)

	音訓評価なし		音訓評価あり	
	再現率	精度	再現率	精度
上位1位	0.493	0.493	0.514	0.514
上位2位	0.635	0.327	0.622	0.311
上位3位	0.676	0.225	0.689	0.230
上位4位	0.736	0.184	0.730	0.183
上位5位	0.791	0.158	0.770	0.154
上位10位	0.845	0.085	0.872	0.087
平均順位	28.03		27.15	

実験(4)

- 上位1位や上位10位までの再現率、平均順位などが改善している。
- 一部、正解の略語の順位が下がってしまった例もあるが、全体としては改善が見込めた。
- 各形態素に対し、先頭文字の法則のみを適用した場合の再現率は約24%であり、生成ルールに基づいた評価の有用性が見て取れる。

→ 音訓の組み合わせのルール、およびこのシステムが有効であることが分かった。

まとめ

- 本論分において提案したシステムにより、より正確な略語候補が得られることができる。
- 課題は残るものの、音訓のルールは有効であることが証明できた。

今後の課題としては、音訓のみならず音韻関係にも着目し、読みの省略ルールの強化、新たな生成ルールの発見などが挙げられる。

参考文献

[1]桜井裕、佐藤埋史:ワールドワイドウェブを利用した用語説明の自動生成, 情報処理学会論文誌, Vol.43, No.5, pp1470-1480, 2002.

[2]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, pp.763-766, 2006.

[3]石野博史:マスコミによく出る短縮語・略語読解辞典, 創拓社, 1992.

[4]田窪行則、前川喜久雄、窪薗晴夫、本多清志、白井克彦、中川聖一:岩波講座 言語の科学2 音声, 岩波書店, 1998.

[5]田中章:最適性理論と日本語のいくつかの問題.新潟経営大学紀要 vol.3, pp.191-208, 1997.