2K-03

出現単語の概念関係を用いたテキストの誤り訂正

佐久間丈貴† 井手一郎‡ 坂井修一‡ 田中英彦‡
†東京大学工学部 ‡東京大学大学院工学系研究科
{takeki, ide, sakai, tanaka}@mtl.t.u-tokyo.ac.jp

1. はじめに

テキストの電子化方法として OCR (Optical Character Recognition) は有効な方法であるが、文字認識には必ず誤りが伴う。そのため最終的な認識精度を向上させるためには、後処理として何らかの方法で誤り訂正を行う必要がある。

既存の OCR の文字認識誤り訂正手法として、語の共起関係を利用するものがあるが [1]、大量の学習データを用いたとしても頻出語以外の共起関係は得難く、訂正は困難である。本研究では、この問題の解決法として、語の属する概念の共起関係を利用することにより、より効果的な誤り訂正を目指す。

2. 概念関係を用いたテキストの誤り訂正

2.1 手法の概要

語の属する概念を用いることにより、語の共起関係が得られにくい場合にも、同一概念に属する語のいずれかが共起していれば、概念の共起関係を得ることができる。例えば、学習データで「総理」と「議会」という2語が共起していなければ、これらの語の間の相補的誤り訂正は難しい。しかし、学習データに「首相」と「国会」という2語が共起していれば、属する概念でみると「内閣総理大臣という地位」と「立法権をもつ国家機関」の共起であり、「総理」と「議会」の属する概念が共起していることになる。

このようにして得られた概念の共起関係を用いることにより、より効果的な訂正を行えると考えられる。更に概念を利用することにより、表現の差異を吸収した、より文意に沿った訂正が期待できる。なお本研究では、テキスト中の誤りは OCR で一般的に見られる一字置換誤りのみを対象とした。

"Error correction of misrecognized characters considering conceptual relations", Takeki Sakuma, Ichiro Ide, Shuichi Sakai, Hidehiko Tanaka , Faculty of engineering, The University of Tokyo

2.2 提案手法の処理の流れ

図 1に示す提案手法の処理の流れを説明する。

1. 誤りを含むテキストを形態素解析し、辞書に登録されている語（辞書登録語）と、されていない語（未知語）に分割する。
2. 未知語は語中で置換誤りが生じたために発生したとみなし、一字置換して単語辞書中に存在する語（一字置換語）を求める。
3. 概念辞書を参照し、辞書登録語と一字置換語が属する概念を求める。
4. 概念共起辞書及び概念頻度辞書により、学習データにおける概念の共起頻度と出現頻度を求め、訂正指標を計算する。
5. 訂正指標が大きい順に一字置換語を並べ、訂正候補として提示する。

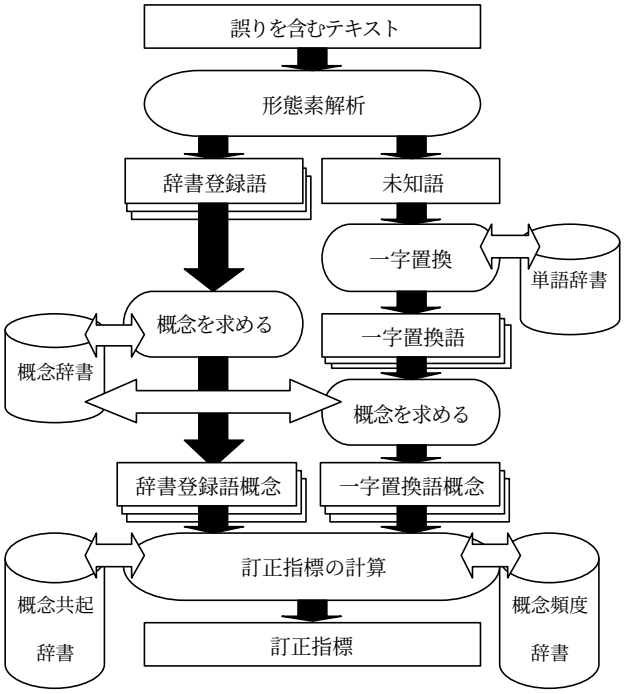

図 1: 提案手法の処理の流れ

3. 誤り訂正実験

3.1 実験の概要

新聞記事中に1ヶ所人為的に置換誤りを発生させ、単語・概念・上位概念各々の共起を用いて、正しい訂正候補が提示できるかを比較した。ここで上位概念とは、語が属する概念の1つ上の親概念のことである。

実験条件は以下の通りである。

◆学習データ：毎日新聞95年度版にRWCテキストコーパス中のRWC-DB-TEXT-97-1[2]（毎日新聞形態素解析差分データ）を適用したもの。

名詞の共起関係を1文毎に収集した。ここで名詞はRWC-DB-TEXT-97-1において、普通名詞、固有名詞（人名・組織名・地域名・国名）、サ変名詞語幹、形容動詞語幹に分類されたものに限った。

◆形態素解析：日本語形態素解析システムJUMAN[3]

◆単語辞書及び概念辞書：EDR電子化辞書[4]

◆訂正指標：各一字置換語毎に

共起回数／辞書登録語の出現頻度

を、全ての辞書登録語との組み合わせで計算したものの総和。

実験データとして、図2のような文を含む新聞記事を使用した。網掛け部の「分豚」は、「分隊」を人為的に誤入力したものであり、図3に示すような「分豚」の一字置換語の各々について訂正指標を計算し、訂正候補の順位を決める。

～ 米英連合軍のノルマンディー作戦の最中に、8人の分豚が敵地で悪戦苦闘しながら、行方不明のライアン二等兵を捜索して歩く～

図2：実験に使用した新聞記事（抜粋）

分○：分離,分割,分配,分前,分業,分隊,分与
　　分裂,分立,分家,分譲,分類,分布,…
○豚：養豚,酢豚,黒豚　　（全66語）

図3：「分豚」の一字置換語

3.2 結果と考察

実験結果を表1に示す。正しい候補は単語共起で7位、概念共起で18位、上位概念共起で2位となり、単語共起に比べ概念共起ではかえって悪化したが、上位概念共起では良い結果を得られた。

単語共起			概念共起			上位概念共起		
順位	候補	訂正指標	順位	候補	訂正指標	順位	候補	訂正指標
1	分野	.060	1	分会	.173	1	分県	2.28
2	分割	.051	1	分署	.173	2	分隊	.796
3	分析	.049	3	分野	.088	3	分団	.792
4	分離	.038	4	分類	.085	3	分会	.792
5	分布	.007	5	分析	.061	3	分署	.792
6	分担	.006	6	分娩	.054	6	分家	.591
7	分隊	.005	7	分散	.036	7	分布	.434

表1：実験結果

概念や上位概念を用いた場合にも、文意から外れた語が候補に選ばれているが、これは比較的浅い階層に属する語の概念・上位概念が非常に一般的なものとなり、他の語との共起が多くなるためである。階層の深さに応じて、概念や上位概念の利用を制限することが必要であると考えられる。

4. おわりに

本稿ではテキストの誤り訂正に概念及び上位概念の共起を用いる手法を提案し、実験によって単語の共起のみでは訂正が困難であった誤りを訂正できる可能性があることを示した。

今回の提案手法では、訂正にOCRから出力されるような画像的特性（文字間の画像的類似度など）、第2候補以下の文字などは利用せず、単なる誤りを含むテキストとして扱った。そのため、OCR以外にもワードプロセッサにおける仮名漢字変換誤りの訂正などに応用することも可能であると考えられる。

参考文献

1) 竹内孔一, 松本裕治：“共起情報と統計的形態素解析による OCR 誤り訂正”, 情報処理学会自然言語研究会報告, NL-121-3, pp.17-24, 1997.

2) 技術研究組合新情報処理開発機構：“RWC テキストデータベース第 2 版”, 1997.

3) 京都大学大学院情報学研究科知能情報学専攻言語メディア研究室：“日本語形態素解析システム JUMAN 第 3.6 版”, 1998.

4) (株)日本電子化辞書研究所：“EDR 電子化辞書第 1.5 版”, 1996.