

機械可読辞書の更新について（その2）

1R-5

—三文字漢字列をもちいて—

田 中 康 仁

愛知淑徳大学

0) はじめに

機械可読辞書は各企業、研究所、ソフトウェア会社等で作られている。しかし、一度、作成してしまえばそれで終わりというものではない。常に更新を行ってゆかねばならない。ここでは辞書更新の問題とその解決の方法について述べる。

同様な標題で1994年3月の情報処理学会全国大会で発表している。

1) 辞書更新の問題点

前回は次の点をあげ述べた。

- i) 辞書は年々古くなっていく。
 - ii) 未知語対策。
 - iii) 更新のための要員の確保と維持、仕事量の見積。
- 今回は次の点について述べる。

機械可読辞書の更新データはどのようにして得られるか、未知語の対策はどのようにするか？

自然言語処理のシステムを開発するにあたっては、次の三つのタイプの辞書に対する保守を行ってゆかねばならない。

(1) 問題対応処理

新しい用語（未知語）が発生した時、その語の表層的特徴から意味、その他をつかむことにより処理する。文処理中の対応。これはあまり正確には行えない。

(2) 利用者の要求による保守

利用者からの要求により、辞書に新しい用語を追加していく。

利用者に提案してもらう。例えば、通信ネットワーク・システムの中に電子掲示板を設けておき、それによって収集する。情報提供者にたいしては無償で行うのか何かの特典を与えるのかで収集の度合いがことなる。

利用者の合意が必要であるが、もし可能なら個々の利用者の追加している単語、複合語、慣用表現、

専門用語、共起関係等を集め、その中から一般的なもの自然言語処理システムの辞書に組み込んでいく。

(3) 予防的保守 (Preventive maintenance)

常に大量のコーパスの中から新しい用語を抽出し、辞書に登録する。出現頻度による数量分析が可能である。新しい分野のコーパスを用いる。積極的な保守点検であり、非常に重要な作業である。

製品の出荷前になるべく多くの欠点を見つけるべきである。製品の出荷後であれば大変な費用を要する。

ここでは(3)の予防的保守について考える。

2) コーパスからの用語の抽出

コーパスからの用語の抽出は文を解析し、用語を抽出する方法がある。しかし、これは処理時間がかかるし、辞書も完成していなければならない。しかし、辞書が完成することはない。

もう一つの方法はコーパスの中の文を機械的に文字種の変わり目に区切り用語の抽出を行う方法である。この方法では誤って用語の抽出もなされるので辞書への追加には慎重に行わなければならない。コーパスからの用語と辞書の用語を照合し、不一致の中の高頻度のものを追加する候補として考えることができる。この方法は機械的処理が多く、しかも、多くの候補がみつかるため良い方法である。

そこで朝日新聞一年分の記事データから三文字漢字列を抽出し分析を行った。

コード	種類	延べ語数	コード
10	35,623	406,827	10…一般的な名詞
80	10,427	96,946	80…固有名詞（人名）
90	3,312	24,536	90…固有名詞（地名）
95	763	3,292	95…数詞
99	27,941	112,711	99…その他
合計	78,066	644,312	

3) 新しいファイルでの照合

読売新聞社一年分のファイルから三文字漢字列を抽出し、朝日新聞のファイルと統合してみた。その結果は次のようになった。

読売新聞社一年分(三文字漢字列)

コード	種類	延べ語数
10	19,960	432,426
80	3,571	91,261
90	1,837	31,221
95	443	6,948
99	6,969	57,219
なし	66,121	234,208
合計	98,901	853,283

コードなしは約9万9千種類の中約6万6千種発生している。これは約67%がアンマッチデータとなったことを示している。延べ語数では27.4%がアンマッチなのである。

アンマッチが多い原因の一つは機械的抽出のため多くなっていることと、数表現によるものがあると考えられる。しかし、もっと多くの語を登録しなければならないことも事実である。

コード10の一般的な用語をみると約2万種類が一致し、その延べ語数は43万2千件にのぼっている。一種類平均21.6件である。これは、重要な用語は朝日新聞の分析結果に含まれていることが判る。コードなしは6万6千種で延べ語数23万4千件である一種類平均3.5件である。今後はこのような語も丁寧に1つ1つ分析、用語として登録しなければならない。約6万6千種は分析中である。

4) 分析の方法について

既知のファイルとの照合は分析の第一歩であるがこれでは約30%強しか一致しない。そこで、アンマッチになったデータについては次のような経験的な分析方法を行うとよい。

しかし、必ず例外があることも考えておかなければならない。

- (1) 次の文字を含む三文字漢字列は地名(90)である。
「県、市、区、郡、町、村、字、省、また最後の文字が(浦、駅、沖、河、海、街、岳、狭、橋、郷、局、園、原、湖、江、港、崎、山、坂、寺、州、沼、川、線、台、沢、谷、地、島、峠、灘、岬、野、路、湾)」であるもの。
- (2) 次の文字列を含む三文字漢字列は数表現(95)である。
「一、二、三、四、五、六、七、八、九、十、拾、百、千、万、億、兆」
- (3) 次の文字列を含む三文字漢字列は氏名、固有名詞(80)である。
～氏、～子(末尾の文字)
- (4) 次のコードが最初にくるものは(99)である。
「案外、以下、以外、以上、以後、以前、以来、依然、一応、一括、一見、一言、一向、一時、一瞬、一生、一切、一層、一体、一度、一回、一般、一晚、一番、一部、一步、一面、一目、一律、何

人、何度、過去、回、極力、月末、現在、現代、午後、今回、今後、今月、今春、今週、今度、今日、今年、再三、再度、最近、最終、最大、最低、最高、昨年、氏(先頭文字)、時折、時代、若干、最終、終日、週間、週末、十分、従来、順次、将来、場合、常時、随分、数日、数年、世紀、世代、絶対、先月、先週、先日、戦後、戦前、前回、前後、前日、前年、前夜、全国、全身、全然、全部、全面、早速、相当、即刻、多少、多数、対応、断然、直接、通常、程度、当時、当初、当然、当分、当日、当面、到底、同年、突然」

これ以外にも種々の特徴ある語がある。これらについては分析中である。

さらにこれら二つのファイルを統合してみると次のようになる。

朝日新聞、読売新聞(三文字漢字列)

コード	種類	延べ語数
10	35,623	839,253
80	10,427	188,207
90	3,312	55,757
95	763	10,240
99	27,940	169,930
なし	66,121	234,208
合計	144,186	1,497,595

コードなしはコードを付加中である。

6万6千件の分析が完了すればさらによい分析ファイルとなるであろう。また、種類で60~70%が自動的にコード付けが可能であろう。

5) 今後の課題

辞書に追加しなければならないデータについての方針の立案について一つの提案を行うことができた。

さらに辞書の見出の増大と共に質の向上(内容の充実)が望まれる。語の使用条件の明確化と、語の関係の把握を行う。上位語、下位語、反対語、関連語、専門語、慣用表現などについての内容の充実である。

三文字漢字列については五文字漢字列の分割によっても得られることに注意しておかなければならない。

接頭語、接尾語についても、定義と語構成について検討しなければならない。

6) 参考文献

田中康仁：機械可読辞書の更新について
情報処理学会第48回(平成6年前期)
全国大会 3Q-1 1994. 3. 23

7) 最後に

平成7年4月1日より次へ移ります。

兵庫大学 経済情報学部教授 田中康仁

〒675-01 兵庫県加古川市平岡町新在家2301

TEL 0794 (24) 0052 FAX 0794 (26) 2365