D-002

値の範囲を検出可能な数値情報抽出

Extraction Numerical Information with Value Range Detection

伊加田 恵志 †　　**濱口 佳孝** †
Satoshi Ikada　Yoshitaka Hamaguchi

1　はじめに

近年, 新聞記事など大量のテキストが電子化されるにともない, テキストから重要な情報だけを抽出したいという要求が高まってきている. 電子化されたテキストから, 情報抽出を行うために, 固有表現抽出といった文章中の重要な要素を抜き出す技術が研究されてきた.

本研究では, 特にテキスト中の数値に注目して情報を抽出することを試みた. 数値の記述は多くのテキストに現れ, それらは, 製品の値段やサイズなどを表しているため, 文書検索などにおいて, ユーザは, そのような数値に着目して目的の文書を探したいことがある. そのためには, テキスト中に現れる数値の記述から, その大きさや単位の種類といったものを正しく抽出する必要がある. そのような技術として, 斉藤らの研究 [3] や, 山田らの研究 [1, 2] が提案されている.

山田らの研究では, 数値部分だけの抽出だけでなく, あらかじめパターンを用意しておくことにより, 「2 万円以上 3 万円以下」のような範囲を示す表現を抽出している. しかし, 全てのパターンを事前に用意することは, 現実的ではないと考える. そこで, 本研究では, 上記のような数値範囲を示す表現に対して, 範囲に関わる語のみを用いて数値情報を抽出する手法を提案する.

2　数値情報抽出処理

本節では, テキストから抽出する数値情報と, その処理について説明する.

2.1　数値情報の定義

まず, 処理で抽出する情報について述べる. 本研究では, 数値に関わる情報 (数値情報) として以下のものを抽出することとする.

- **種別** 数値単位の種類. 長さ, 重さ, 速度など.
- **基本単位** 種別ごとの大きさの基本となる単位. 各種別ごとに任意に決定しておく.
- **下限値** 範囲の下限値. 基本単位における大きさに換算.
- **上限値** 範囲の上限値. 基本単位における大きさに換算.

なお, 範囲を持たない数値は, 下限値と上限値とを同じ値にするとする.

2.2　数値の検出

以下, 数値情報の抽出処理を順に説明する.

形態素解析の結果, 品詞を調べ数値文字列を検出する. 数値の表記には, 算用数字, 漢数字の混在, 分数や, 「1 万円の 50%」のように句や文全体で数値を表現するなど様々なものが存在する.

本研究では, 山田らの研究とほぼ同様に, 算用数字や, 漢数字の混在, 分数には対応するが, 「1 万円の 50%」という句からの計算結果に当たる数値 (この場合は, 5000) は抽出しないこととする. また, 「約 1 万円」や「1 万円程度」のように, ある数値を基点とした曖昧な範囲を表す表現は, 下限値, 上限値を適正なものを求めるのは難しいため, 範囲を持たない数値として扱うこととする.

2.3　種別の推定

前節で検出した数値がどの種別に属するか, 数値の周辺の語を元に推定する. 種別ごとに, その種別と共起するような語にあらかじめ重みを与えておく (表 1). 単位ごとに, 出現した語の重みを足しあわせて, 合計の大きい種別をその数値の種別として推定する. なお, 重みが 0 の場合は, どの種別にも属さないとして以降の処理対象としない.

種別	共起語 (重み)
長さ	メートル (20), インチ (20), マイル (20), 寸 (20), キロ (2), 距離 (5), 長さ (5),...
重さ	グラム (20), トン (20), ポンド (3), 貫 (20), キロ (2), 重み (5), 重量 (5),...
時間	秒 (10), 分 (10), 時間 (20), 日 (5), 週間 (10), ヶ月 (10), 年 (5), ...
速度	秒 (3), 分 (3), 時 (5), 毎時 (10), 毎秒 (10), 毎分 (10), 秒速 (20), 分速 (20), 時速 (20), ...

表 1　種別とその共起語と重み

2.4　基本単位への換算

次に, 数値を単位文字列, 種別をもとに, 種別ごとの基本単位に換算した数値を求める. なお, 単位換算のための係数は, 単位辞典 [6] に記載されている値を用いた. 例えば, 「3 インチ」の場合, 属性「長さ」の基本単位を「メートル」とすると, $3 \times 0.0254 = 0.0762$ と換算する.

2.5　数値範囲抽出

提案方式では, 数値の範囲を, 前節で抽出した数値の周辺の単語を含めた領域から検出する. 領域は, 隣り合う 2 つの数値の間の単語などの情報によって決定する. 例えば, 「長さが 1m になるものもあるが, 25cm 以上の長さで 30cm までのものがほとんどです.」という文の場合, 句点や, 数値と数値の間の単語数といった条件により, 図 1 のように領域を決定する.

長さが 1m になるものもあるが, 25cm 以上の長さで 30cm までのものがほとんどです .

図 1　範囲を抽出する領域

次に, 各領域内で数値に関わる表現を検出し, 下限値と上限値を決定する. 検出は, 範囲の先頭から行う. 検出する表現は数値のほか, 表 2 に示したような数値関係表現を検出

† 沖電気工業株式会社, Oki Electric Industry Co., Ltd.

数値解釈	数値関係表現
$+\infty$	以上, から, 〜, より大きい, 超える, ...
$-\infty$	以下, 未満, まで, より小さい, 下回る, ...

表 2　数値関係表現と数値解釈

する. 検出した数値及び, 数値関係表現により, 下限値と上限値を以下の規則により更新していく. なお, 数値関係表現の数値は, 表 2 で示したように解釈する. また, 下限値と上限値には, 最初は初期値が入っているものとする.

1. 下限値, 上限値が初期値の場合, 下限値, 上限値の双方を, 検出した数値で更新する.
2. 上記条件以外で数値を検出した場合,
   a. 下限値, 上限値が $+\infty$, $-\infty$ 以外の数値の場合,
      i. 検出した数値が下限値より小さい場合, 下限値をその数値で更新する.
      ii. 検出した数値が上限値より大きい場合, 上限値をその数値で更新する.
   b. 下限値が $-\infty$ であり, 検出した数値が上限値より小さい場合, 下限値をその数値で更新する.
   c. 上限値が $+\infty$ であり, 検出した数値が下限値より大きい場合, 上限値をその数値で更新する.
3. 上記条件以外で数値関係表現を検出した場合,
   a. 数値関係表現が $-\infty$ と解釈される場合,
      i. 下限値, 上限値が $+\infty$, $-\infty$ 以外の同じ数値の場合, 下限値を $-\infty$ で更新する.
      ii. それ以外の場合は何もしない.
   b. 数値関係表現が $+\infty$ と解釈される場合,
      i. 下限値, 上限値が $+\infty$, $-\infty$ 以外の同じ数値の場合, 上限値を $+\infty$ で更新する.
      ii. それ以外の場合は何もしない.

例えば, 先ほどの例の「25cm 以上の長さで 30cm までのものが」の場合, まず, 「25cm」を検出, 規則 1 より, 下限値, 上限値は 0.25 になる. 次に, 「以上」を検出, 規則 3-b-i により, 上限値が $+\infty$ となる. さらに, 「30cm」を検出, 規則 2-c より, 上限値が 0.30 となる. 最後に「まで」を検出, 規則 3-a, 3-b のどちらにも当てはまらないので, 下限値, 上限値はそのままとなる. これ以上は検出されないので, 最終的に, この領域では下限値 0.20, 上限値 0.30 が決定される.

3　精度評価実験

我々は, Web 上の日本語のテキストに対して, 数値情報の抽出精度を評価した. Web 上のテキストは, 新聞記事に比べ数値を多様な表現で記述しているので, そのような文書に対してどの程度抽出が可能かを測定した. 抽出を行った数値情報の種別は, 長さ, 重さ, 時間, 速度, 金額 (円のみ), 電圧, 電流, 周波数, バイト, ビットの 10 種類である. 評価はこれらの再現率, 適合率, および F 値を求めた. なお今回, 形態素解析には Sen[1] を用いた.

評価結果を表 3 に示す. 提案手法の全体の抽出精度としてとして, 再現率 84.3%, 適合率 90.0%, F 値 87.0 という結果が得られた.

また, 数値範囲の抽出精度についても評価した. 提案手法との比較のため, 山田らの手法を実装し, 同じテキストに対して実験を行った.

[1] https://sen.dev.java.net/

	再現率	適合率	F 値
提案手法	84.3%	90.0%	87.0
数値範囲 (提案手法)	94.0%	86.3%	90.0
数値範囲 (山田ら)	85.1%	93.4%	89.0

表 3　評価結果

結果として, 山田らの手法に比べ, 再現率で 9 ポイントの向上が見られた. 一方, 適合率については, 7 ポイント低下することとなった. これは, さまざまな範囲表現に対しても, より多く範囲として抽出が可能になった反面, 範囲表現でないものも範囲として多く抽出してしまったためである. 例えば, 「価格は 1 万円あるいはそれ以上」という範囲表現に対しても, 抽出が可能となった反面, 「価格は 1 万円である以上」という表現に対しても, 「以上」という数値関係表現が存在するために, 範囲としてしまった. このように, 語や品詞だけでは判別できない違いを認識する手法が今後の課題の 1 つとなるといえる.

4　まとめ

本論文では, テキストからパターンによらずに, 数値と数値関係表現から数値範囲の下限値と上限値を決定するような数値情報抽出の手法を提案した. Web 文書に対する抽出精度評価の結果, 適合率で 91% という精度で抽出できた. 一方, 範囲の抽出においては, 従来の手法に比べ, 再現率で向上することができた. しかし, 単語の出現だけでは判別できない範囲表現があることが確認でき, 今後, このような表現を正しく判別する手法について検討してきたい. また, 近年では, 機械学習による情報抽出の研究が盛んであり [4, 5], 範囲表現について, このような手法が適用できないか検討していきたい.

参考文献

[1] 山田, 福島: 数値情報を用いたテキスト検索方式の提案と評価, 情報処理学会研究会報告 FI-53-3, 1999

[2] 山田, 福島: インターネット多角的検索システム OTROS -数値情報の抽出と検索-, 情報処理学会第 57 回全国大会 3L-02, 1998

[3] 斉藤, 迫田, 中江, 岩井, 田村, 中川: 数値情報をキーとした新聞記事からの情報抽出, 情報処理学会研究会報告 NL-125-6, 1998

[4] 山田, 工藤, 松本: Support vector machines を用いた日本語固有表現抽出, 情報処理学会論文誌 43(1):44-53, 2002

[5] T. Hasegawa, S. Sekine, and R. Grishman: Discovering Relations among Named Entities from Large Corpora, In *proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, 2004

[6] 国際単位研究会編: SI 単位ポケットブック, 日刊工業新聞社, 1991