D-038

技術文書からの数値属性の抽出
Extraction of Numerical Attributes from Technical Document

小林 義行†
KOBAYASI, Yosiyuki †

1. はじめに

数値属性とは、「事物」の特徴を「属性」に関する「数量」で表したものである。多くの技術文書では、製品仕様や技術的な特徴を数値属性として記述している。文書中の数値表現を抽出することは情報収集における重要課題の一つである。本報告では、テキストデータから収集した共起データや単位辞書を使い、英文技術文書から数値属性を抽出する方法について述べる。

2. 数値属性抽出方法

2.1 課題

本研究では、数値属性を＜*事物, 属性, 数量*＞という 3 つ組で表現されるものとする。ここで、「数量」は、数字と単位で表現する。「属性」とは、数量が表す量に関するカテゴリである。「事物」は、物理的な物や時空間であり、その性質が「属性」と「数量」で表される。例えば、「乾電池」という事物には、「電圧」という属性があり、その数量は「1.5V」のように数値「1.5」と単位「V」で表される。3 つ組としては、＜*乾電池, 電圧, 1.5V*＞と記述する。

本報告の課題は、英語文書から数値属性を抽出することである。まずは、1 文からその文に含まれる文字列を使って表現可能な 3 つ組を抽出することを考える。例えば、“This unit controls an **upper limit voltage** of the **lithium secondary battery** to **4.1 V**.”という文から、< *lithium secondary battery, upper limit voltage, 4.1V*>を抽出する。

数値属性抽出方法に必要な機能は以下の 2 点である。

(1) 事物、属性、数量の可能性がある名詞句を同定
(2) 同定した名詞句が事物、属性、属性値であることを判定

2.2 処理の流れ

処理の流れを図 1 に示す。

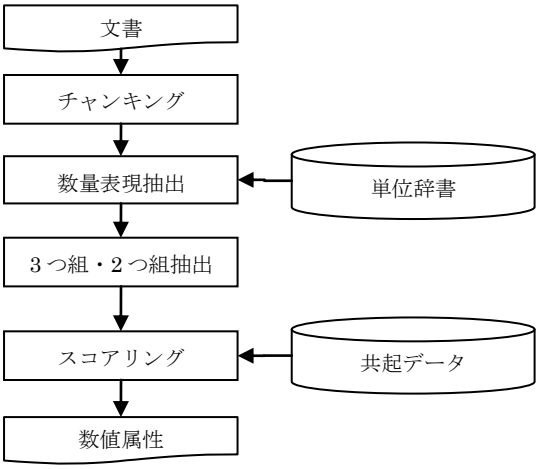

図 1. 数値属性抽出処理

各処理の内容は以下の通りである。

(1) チャンキング
本報告では OpenNLP[1]のチャンキングモジュールを利用した。以下の処理では名詞句のみ利用する。

(2) 数量表現抽出
名詞句に対して数値表現抽出を行なう。数値文字列の抽出、実数値への変換、単位の換算を単位辞書(2.3 節に詳細説明)を利用して行なう。

(3) 2 つ組・3 つ組抽出
名詞句について、以下の 3 つの制約を満たす 3 つ組と 2 つ組をすべて抽出する。
(a) 数値表現をひとつだけ含む。
(b) 文・節の境界を越えて抽出しない。文・節の境界を表す記号はピリオド、セミコロン、コロンとする。
(c) 2 つ以上の動詞句をまたぐペアを含まない。

(4) スコアリング
3 つ組と 2 つ組のスコアを計算し、あらかじめ指定した上位 n 位以上を出力する

2.3 単位辞書

物理的単位などをまとめた資料は数多くあるので、簡単に単位辞書を作成できそうである。しかし、実際の文章を調査すると専門的な文書以外では見られないような組立単位がいくつもあることが分かる。例えば、電力関係の文書には、“mΩcm2”という単位が使われている。この単位は面抵抗(単位面積あたりの導電率の逆数)を表す“Ωm2”に対して、“Ω”に“m”(ミリ)、“m2”に“c”(センチ)を表す接頭辞がついたものである。

このような特定ドメインで使われる組立単位は数多くあるので、本報告では、以下に示す手順でテキストデータから半自動的に単位を収集した。

(1)数字列の直右に出現する文字列を収集する。1 文字から 20 字までとした。
(2)文字列に対して単位構造を解析する文法にしたがって構文解析を行なう。構文解析器には再帰下降パージングを行なう Perl モジュール RecDescent を利用した。
(3)解析に成功した文字列を単位として辞書に登録する。

文法として接頭辞と基本単位(長さ単位や電圧単位など)はあらかじめ登録する必要があるが、組立て単位とは異なり、基本単位の数が限られているので困難ではない。文法を使うことで、正誤判定に加え、単位の構造から単位の換算式も自動的に取得できる。

米国特許公報を使い単位辞書を作成した。使用したのは 2007 年度の米国公開特許公報(U.S. Patent Application ACE RedBook Data XML TEXT ONLY-2007)である。国際特許分類を使い、本手法の適用先と想定するドメインに制限した。国際特許分類は第八版を使い、分類番号は、H02 とした。H02 に分類されている公開特許は 4780 件である。抽出さ

† (株)日立製作所中央研究所 Hitachi, Ltd. Central ResarchLaboratory

れた単位数は、560 個であった。約 1 割にあたる 50 個をサンプリングし、正解率を評価したところ、誤りが 7 個あった。正解率は、43/50 = 86%である。

2.4 共起データ

共起データは、構文解析結果から得られる係り受け関係と、チャンキング結果から得られる一文内の共出現を利用した。チャンキングには、OpenNLP のチャンキングモジュール、構文解析には OpenNLP の構文解析モジュールを使用した。

係り受け関係を利用することで、ある単語が、属性として使われるのか事物として使われるのかという知識を収集できる。例えば、“The minimum input voltage is 1.5V”から be 動詞“is”の主語“The minimum input voltage”と補語“1.5V”が抽出され、単位“V”の属性になりうる語“minimum input voltage”が取得できる。また、“the fixed output voltage of 3.3V”から前置詞“of”の目的語“3.3V”と前置詞句が係っている名詞句“the fixed output voltage”が抽出され、単位“V”の属性になりうる語“fixed output voltage”が取得できる。さらに、“A single li-ion of laptop battery has a voltage of 3.7V”から動詞“has”の主語“single li-ion”と目的語“a voltage”が抽出され、“single li-ion”が属性として“voltage”を持つことが取得できる。本報告では、係り受け関係として、上記の例で示した be 動詞の主語と補語、動詞“have/has”の主語と目的語、前置詞“of”の目的語と前置詞句が係っている名詞句の三種類を抽出した。

ただし、係り受け関係だけでは、データのスパースネスが大きい。チャンキング結果による共出現データは、スパースネスを軽減するために利用している。

2.5 スコアリング

3 つ組の要素のひとつは、数量あると分かっている。したがって、残る 2 つの要素について、それぞれが属性名なのか、事物なのか、それ以外なのかのスコアを計算することになる。スコアは、構文的なスコアと、語彙的なスコアの線形和として計算する。構文的なスコアは、文中における名詞句のあいだの距離の逆数によって評価する。語彙的なスコアな共起データから評価する。係り受け関係があるときのスコアを頻度とし、共出現データのスコアは条件つき確率値とした。

3. 評価

数値情報を含む米国特許公報からランダムに 6 本の明細書を抽出し、数値情報の抽出実験を行なった。

評価に用いた公開特許を以下に示す。

	公開番号
1	2007/0024126
2	2007/0064357
3	2007/0063668
4	2007/0121261
5	2007/0187954
6	2007/0228883

抽出結果を表 2 に示す。

表 2　特許公報からの数値属性抽出結果

特許データ / 評価項目	1	2	3	4	5	6
正解数	9	5	7	4	1	7
抽出した数	8	5	16	4	1	7
正しく抽出した数	7	3	5	4	1	2
再現率(%)	78	60	72	100	100	18
適合率(%)	88	60	31	100	100	18

4. 関連研究

数値情報に注目している研究動向として動向分析[2]があげられるが、これらの研究では、数値表現の数値を特徴とする事物名(本研究で事物と属性としているものに対応)を抽出することに関心が払われており、数値情報との組み合わせ方法はあまり考慮していない。

ブログからの評判情報抽出や、製品説明文からの属性情報抽出についていくつかの研究報告[3][4]がある。これらの研究では、事物は自明として、属性と属性値の 2 つ組の抽出を対象としている。

英語を対象にした属性情報抽出に、Ghan らの研究がある[5]。彼らは、属性と属性値の検出を分類問題に帰着している。再現率は 76%、適合率は 44%である。

事物、属性、属性値の 3 つ組を抽出対象としたものに、藤畑ら[6]、高橋ら[7]、五十嵐ら[8]の研究がある。藤畑らは係り受け規則を使って属性情報を抽出し、再現率は 64%、適合率は 87%である。高橋ら、テンプレートを用いて 3 つ組を抽出し、再現率は 85%、適合率は 82%である。五十嵐らは確率モデルにより評価し、再現率は 85%、適合率は 16%である。

5. おわりに

テキストデータから収集した情報を使い、英文技術文書から数値属性を抽出する方法について報告した。「事物」の抽出精度が不十分であるが、これは 1 つの文内で「事物」「属性名」「数量」がすべて記述されるとは限らないことに起因する。複数の文を対象にした抽出方法を今後検討する予定である。

参考文献

[1] OpenNLP. The OpenNLP Home Page. (オンライン) http://opennlp.sourceforge.net/.
[2] 森辰則ほか, 動向情報編纂のためのテキストからの統計量表現の自動抽出. 人工知能学会論文誌 : vol.23, No. 5 , 2008 年
[3] 飯田龍ほか. 意見抽出を目的とした機械学習による属性-評価値対同定. 情報処理学会研究報告 : NL-165, 2005 年.
[4] 西村純ほか. ネットオークションにおける属性検索のための出品情報からの属性抽出. 言語処理学会 第 14 回年次大会, 2008 年.
[5] Ghani, R., and etc.“Text mining for product attribute extraction”, ACM SIGKDD Explorations Newsletter, Vol. 8, Issue 1, 2006.
[6] 藤畑勝之ほか. 係り受けの制約と優先規則に基づく数量表現検出. 情報処理学会研究報告 : NL-145, 2001 年.
[7] 高橋哲朗ほか. テキストから属性関係を抽出する. 情報処理学会研究会 : NL-164, 2004 年 11 月
[8] 五十嵐力ほか. 枝分かれ同時確率モデルを用いた対象―属性―属性値関係の抽出. 情報処理学会研究会 : NL-189, 2009 年