機械加工用語の関連性抽出

Association Extraction of Machining Terms

増田　和浩*　　寺本　一成**　　古谷　克司*　　佐々木　裕*

Kazuhiro Masuda*　Kazunari Teramoto**　Katsushi Furutani*　Yutaka Sasaki*

*豊田工業大学　　**㈱豊田中央研究所

*Toyota Technological Institute　　**Toyota Central R&D Labs., Inc.

1. はじめに

例えばオペレーションズリサーチなど，経営分野において数学モデルを用いて意思決定を行う手法は長年研究されてきた．この研究は現在多くの企業が利用するほどの有用性へと発展を遂げ，また幅広い分野に対応できるポテンシャルを持っている．

しかし，工学分野の機械加工では，慣例的に用いられてきた入力値を使って機械加工を行っている現場がほとんどである．このような現場で入力値の変更による出力結果の改善や，体系的な経験の浅い機械加工の導入に対しては手探りになってしまうことが多い．その一方で，これらの現場には機械加工の入出力データの記録や，工学分野の知識を内包する技術文書が資源として存在する．もしこれらの資源から有益な情報を適切に抽出できれば，乱雑なデータの山と漠然としたパラメータの関係性が整理，明確化され，そこから得られる知識は業務に携わる各人の理解を深め，今後の加工条件の模索においても有用な資料となる．また人工知能や自然言語処理を機械加工に取り入れる際に，重要となる知識ベースとしての役割も持つことができる．

本研究はこのような目的のために，技術文書を解析することで各パラメータ間の関係性や，機械加工の入力値，出力値との因果関係を明らかにし，人的な意思決定を支援することを目標とする．MeCab[1]で技術文書を解析する場合を想定し，最初に機械加工分野の MeCab ユーザ辞書を作成することで自然言語処理の足場を作る．その後に専門用語間の共起性を調べ，関係性を調べている．

技術文書に対し自然言語処理を行うにあたって問題となるのは，専門用語についての情報不足である．どの専門用語同士が共に出やすいかという共起知識が足りないため，正確な構文解析を行うことは難しく，更に自然言語処理に通常用いられる，一般文章向きのコーパスには専門用語の知識が不足し，形態素解析でさえも満足に行えない．

最終的には意味解析によりパラメータ間の関係性を探ることを想定しているが，現状では形態素解析を重点的に行い専門用語の知識収集に努めている．

2. 形態素解析

本研究では形態素解析エンジンに MeCab を採用している．これは，MeCab がシステムと辞書が独立しており，辞書に後から単語を付け加えやすいという特徴に基づくものである．前述の通り，機械加工分野に自然言語処理を適応することを考えるならば，最初に専門用語に特化したユーザ辞書を作成し足場を固めなければならない．MeCab のユーザ辞書のフォーマットは，登録する単語に対し 12 の情報を付加する．その内最低限必要となるものは，コストと品詞情報になる．

コストの自動生成

MeCab は複数の解析候補に対し，コストの大小で結果の選定を行う．候補全体のコストは各形態素の単語生起コストと形態素間の連接コストによって決定される[2]．現状ではまだ十分な共起知識が得られていないため連接コストを設定するには至っていないが，単語生起コストに関しては適切な値を設定しなくては正しく分かち書きを行えず，誤判定の元になってしまう．MeCab ユーザ辞書フォーマットでのコストのクラスは，その単語生起コストを規定するものである．

単語を辞書化する際に，以下の式を用いてコストの自動生成を行う．

$$\mathrm{Cost} = -4 * L^3$$

ここで Cost は辞書化する専門用語の単語生起コスト，L は専門用語の文字数となる．例えば「被削材硬さ」が「被削材」と「硬さ」というように解析されるのを避けるため単語長の累乗を取り，長い単語であるほど低いコストとなるようにしてある．コストが負の値になっているのは，一般の単語より優先的に判定されるようにするためである．

```
切り        名詞,一般,*,*,*,*,切り,キリ,キリ
くず        名詞,接尾,一般,*,*,*,くず,クズ,クズ
の          助詞,連体化,*,*,*,*,の,ノ,ノ
排出        名詞,サ変接続,*,*,*,*,排出,ハイシュ…
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
、          記号,読点,*,*,*,*,、,、,、
すくい      動詞,自立,*,*,五段・ワ行促音便,連用…
角          名詞,接尾,一般,*,*,*,角,カク,カク
など        助詞,副助詞,*,*,*,*,など,ナド,ナド
によって    助詞,格助詞,連語,*,*,*,によって,ニヨ…
パターン    名詞,一般,*,*,*,*,パターン,パターン,パ…
が          助詞,格助詞,一般,*,*,*,が,ガ,ガ
異なる      動詞,自立,*,*,五段・ラ行,基本形,異なる
…
。          記号,句点,*,*,*,*,。,。,。
EOS
```

```
切りくずの排出      名詞,一般,*,*,*,*,*,切屑排出
は          助詞,係助詞,*,*,*,*,は,ハ,ワ
、          記号,読点,*,*,*,*,、,、,、
すくい角    名詞,一般,*,*,*,*,*,*,*
など        助詞,副助詞,*,*,*,*,など,ナド,ナド
によって    助詞,格助詞,連語,*,*,*,によって,ニヨッ…
パターン    名詞,一般,*,*,*,*,パターン,パターン,パ…
が          助詞,格助詞,一般,*,*,*,が,ガ,ガ
異なる      動詞,自立,*,*,五段・ラ行,基本形,異なる…
。          記号,句点,*,*,*,*,。,。,。
EOS
```

図 1　MeCab の形態素解析出力例

品詞情報の登録

扱う専門用語は全て名詞であるが，後々の構文解析等の段階を考えてサ変名詞であるかどうかの判定を行っている．対象の専門用語が，技術文書においてどのような文脈で出現するかを調べ，直後に続く単語が動詞の「する」の活用形であればサ変名詞と判定する．単純なマッチングでは，特定の名詞において誤判定が生じる場合がある．例えば「切削仕上げ代」の表記揺れである「切削仕上げしろ」などは，「しろ」が「する」の活用形であると間違われやすい．そのため判定には品詞情報を利用する必要があり，その際に MeCab-Python[3]による品詞 ID 取得を利用している．図 2 に示すのは，専門用語を辞書化する際の手順である．専門用語表は，機械加工分野の専門家により作成された．技術文書を MeCab-Python によって前処理し，品詞 ID を付与している．

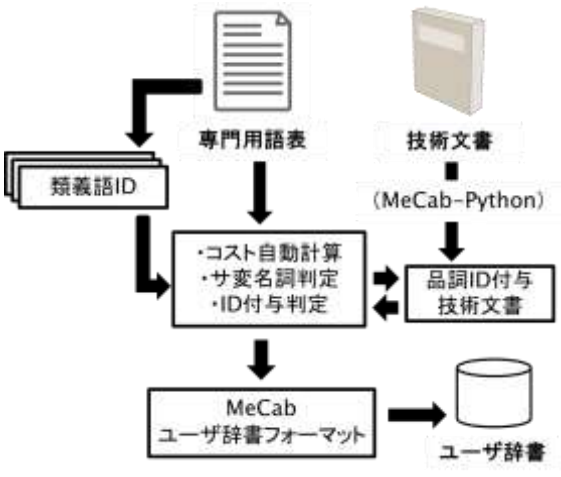

図 2　専門用語の辞書化の手順

類義語 ID の付与

MeCab のユーザ辞書のフォーマットには 12 の情報クラスが設けられているが，拡張し更なる情報を付与することも可能である．本研究では 13 番目の情報として類義語クラスを作成し，形態素解析に利用している．

類義語 ID は同義の関係にある専門用語や，言い換えパターンなどを一まとめにするために活用される．後述する関係性抽出の処理において，同じ類義語 ID を持つ単語群は同一視される．作成される類義語 ID の種類，及び類義語 ID を付与される専門用語については，前述の専門用語表に基づき，また研究途中で障害となった単語などから選出した．

用語抽出

図 1 に示すのが，「切りくずの排出は，すくい角などによってパターンが異なる．」という文章を MeCab に形態素解析させた出力結果例になる．左側の結果が標準的な一般辞書を用いたものであり，右側が作成した機械加工分野特化のユーザ辞書を使用した出力結果になる．

図 1 の「切りくず」や「すくい角」などの専門用語に着目すると，一般的な辞書では 2 単語に分割され正確に分かち書きできていないことが分かる．一方，ユーザ辞書を用いた場合では，「すくい角」と専門用語を抽出できている．更に切りくずについては，後続の排出という単語と共に「切りくずの排出」というパラメータで抽出されている．これは正確な形態素解析ではないが，最終的にはパラメータの抽出を想定するが故の処理である．また，「切りくずの排出」については最後に「切屑排出」という類義語 ID が付与されている．この ID は他に「切り屑排出」「切屑の排出」「切りくず排出」等の単語に付与され，技術文書での表記揺れに対応している．

3. 関係性抽出

前述の通り，本研究の目的は専門用語の関係性の抽出である．そのためにはまず，2 単語間の関係性を評価しなければならない．本研究ではその評価にあたって，Association Score[4]を採用している．

Association Score

２つの名詞に対する Association Score は，本研究では以下のように定義される．

$$A(c1,c2) = P(c1,c2)log\frac{P(c1,c2)}{P(c1)P(c2)}$$

ここで　A(c1,c2)は 2 単語間の Association Score，P(c1,c2)は 2 単語間の共起確率，P(c1)，P(c2)は各単語の生起確率である．

Association Score の式では第 1 項の共起確率が 2 単語の共起の一般性を示し，第 2 項は相互情報量，2 単語の共起の強さ，珍しさを示す．関係性の評価をするにあたって，各単語の生起確率を加味しなければ文中に出やすい単語，出にくい単語で差が出てしまう．Association Score の採用は生起確率を考慮し，単語の出現率の差をならすことを目的としている．

共起確率P(c1,c2)及び生起確率P(c1)，P(c2)は以下の式で計算される．

$$P(c1,c2) = \frac{f(c1,c2)}{N}$$

$$P(c1) = \frac{f(c1)}{N} \qquad P(c2) = \frac{f(c2)}{N}$$

ここで f(c1,c2)は共起頻度，f(c)は頻度，N は総単語数になる．

上記の式より，技術文書の解析にあたって Association Score を出すには，総単語数N，各単語の出現頻度 f(cn)及び任意の単語ペアの共起頻度 f(c1,c2)を数える必要がある．

処理手順

Association Score の計算，そして 2 節で作成されたユーザ辞書を用い，現在の関係性抽出のための処理手順は図 3 のようになっている．

図におけるユーザ辞書は，2 節において作成したものである．技術文書を形態素解析する際に，ユーザ辞書から専門用語の単語生起コストを活用し，分かち書きを行う．また，技術文書内に含まれる専門用語で，類義語 ID が登録されているものがあれば単語の置き換えを行い，同義語や言い換えの是正を行っている．形態素解析中に総単語数と各専門用語の頻度を計測するため，登場する単語数と専門用語ごとの出現回数をカウントしていく．

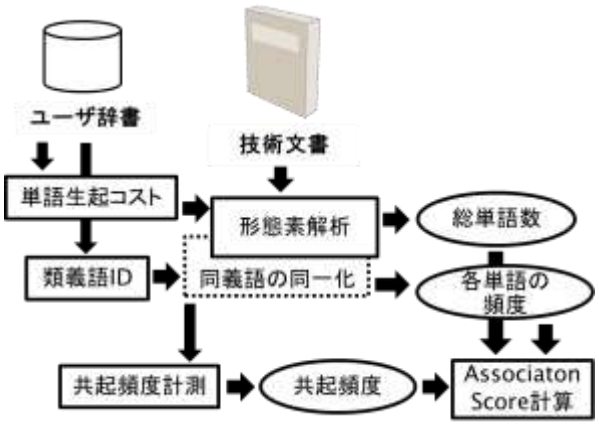

図 3　スコア計算の処理の流れ

共起頻度は，ある単語のペアに対し，特定の大きさの枠内で同時に出現した回数である．本研究では 1 文をその枠とし，同文中に出現した専門用語のペアには共起ありとして共起頻度を測定する．一般名詞と専門用語の共起関係も有用となり得るが，現在は辞書に登録されている専門用語同士の関係性抽出のみを行っている．これらの手順によって算出された総単語数，各専門用語の頻度，及び出現した専門単語ペアの共起頻度から，各専門用語ペアについて Association Score の計算を行い，スコア順に並べて出力する．最終的には，既定された値以上のスコアを持つ専門用語ペアを関連性ありと見なす．

4. 実験

2 節で作成したユーザ辞書の有用性，及び 3 節で述べた関係性抽出の評価を行うため，以降で述べる用語抽出，関係性抽出の実験を行った．

両方の実験で使用することとなるユーザ辞書及び関係性抽出の実験で解析対象となる技術文書のデータを表 1 に示す．

表 1　ユーザ辞書及び技術文書の内容

ユーザ辞書		技術文書	
収録単語数	847	ページ数	174
類義語 ID の種類数	109	行数	5287
類義語 ID 登録単語数	241	総単語数	89904
サ変名詞 ID 登録単語数	33		

用語抽出

一般辞書を使用する MeCab と，作成したユーザ辞書を使用する MeCab の 2 つを用いて，精度の比較を行った．解析対象は技術文書から無作為に選出した 50 の文とし，各文には解析の正解データを付与して双方の MeCab の解析結果と比較する．

表 2　用語抽出の実験結果

(A)50 文の解析結果		
	辞書なし	辞書あり
正答文数	20	41
正答率	40	82
(B)名詞 248 種の解析結果		
	辞書なし	辞書あり
正答単語数	189	225
正答率	76	91

表 2 は用語抽出の実験結果であり，(A)は 50 文中何文正答できたかを，(B)は 50 文中に登場した名詞 248 種に対し何種正答できたかを示す．ユーザ辞書を用いた方が用いない場合の 2 倍の文章量を正しく解析できたことが分かる．一方，正しく解析できなかった残り 9 文は，全名詞 248 種のうちユーザ辞書が未学習であった 23 種が登場した文である．これらの失敗は積極的に辞書への登録を行い，精度を高めていく．

また，例えば「切りくず形状は工具摩耗が進むとともに破砕されたりカールが少なくなる．」という文を解析したときには，本来巻きを意味する「カール」が人名のカールであると判定された．これは切りくず等のような単語が出る文では，人名より巻きの意味のカールの方が出やすい傾向にあるという共起知識の不足に起因している．次に述べる関係性抽出は，このような誤判定を無くすことにも役立つ．

関係性抽出

4 節で示した関係性抽出の結果の評価を行うため，専門家によりあらかじめ集計対象とする 58 種の専門用語の関係性についての正解データが作成された．その後，技術文書をシステムに解析させ，精度を測った．下記の 3 通りの方法で解析を行った．

① 共起頻度によって順位付けし，類義語 ID を使用しないもの
② Association Score によって順位付けし，類義語 ID を使用しないもの
③ Association Score によって順位付けし，類義語 ID を使用するもの

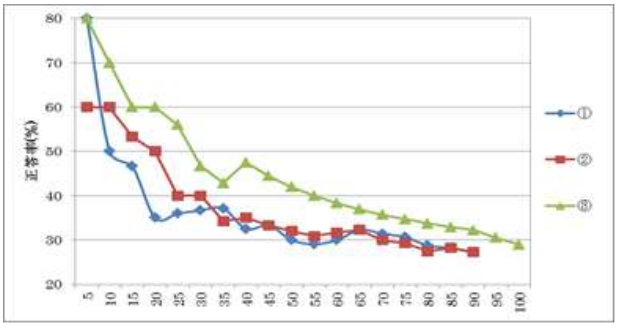

図 4　上位 n 位までの正答率の推移

図 4 が関係性抽出の実験結果になる．グラフはスコア順に並べられた出力結果の内，上位 n 位までで正答率を取った結果である．①と②は解析の内容は全く同じで，最後の関係性スコアの順位付けの仕方のみが異なっている．この２つの結果を比較すると，共起頻度の順位付けは 5 位以内の正答率は優れるものの，それ以降は急激に下降している．これは各単語の出現しやすさ，生起確率の差からくる不安定性によるもので，30 位以内であれば Association Score での順位付けの方が安定した評価を行えている．

また③においては，解析の工程に類義語 ID を活用し，同義語や表記揺れに対応した．これによって前２つの方法では抽出できなかった専門用語が解析結果に含まれ，グラフが上向きに，全体の正答率が向上する結果を得た．

5. おわりに

図 1 に示すような正答率では，未だ実用的とは言い難い．今後正答率を挙げていくには，知識ベースの拡張，整備やより複雑な解析が必要となる．前者は単純に学習単語数や同義語，言い換えの幅を広げるだけでなく，カテゴリ分けや階層構造のような単語の分類を行い，カテゴリ間の共起関係を測ることも重要である．また，解析対象として幅広い文章を読ませることも，学習や問題の探索に役立つ．後者は係り受け関係や，前述のカテゴリ分けされた単語を使うことでスコア付けの際に考慮するパラメータを増やし，より確実性のある解析を目指すことが考えられる．現状では形態素解析しか行っていないため，構文解析，意味解析[5]などより発展的な解析を目指すべきではあるが，そのためにはもっと知識ベースを拡大させる必要がある．

参考文献

[1]MeCab(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)
[2]工藤拓，山本薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析(2004)
[3]MeCab-python(http://mecab.googlecode.com/svn/trunk/mecab/doc/bindings.html)
[4]田中穂積： 自然言語処理　－基礎と応用―，社会法人電子情報通信学会(1999)
[5]長尾真： 岩波講座ソフトウェア科学 15 自然言語処理，岩波書店