SVM を用いた機械加工文書からの直接的因果関係の抽出

Direct Causal Relation Extraction from Machining Documents with SVMs

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

機械加工分野では，経験的に用いられてきたパラメータ値を使って目的としている機械加工を行っている現場がほとんどである．このような現場ではパラメータの変更による出力の改善や，体系的な経験の浅い機械加工の導入を考えた時には手探りになってしまうことが多い．また機械加工の専門知識は，マニュアルや本，データの形で現場に散在し，統一化がなされていない．得たい知識を探す場合は１人１人が本を読み，情報を探すという手間を負わなければならず，時間と労力を要する．

もしこれらのデータ群から自動で適切に情報を抽出できれば，特に入力条件と出力の因果関係を明らかにすることができれば，こういった労力を軽減できるだけでなく，知識は各人が共有できる形で整理・体系化される．そこから作られた資料によって業務に携わる各人の理解を深め，認識の共有を図ることで協議の円滑化を補助できるだけでなく，加工条件の見直しの際に人的意思決定を補助する資料としての役割も負うことができる．

本研究はこのような背景を元に，機械加工文書に自然言語処理を行い，専門用語間の関係を抽出することを目的としている．また，現在では不足している機械加工分野の言語処理基盤の構築も行う．

自然言語処理については，文章中の専門用語を識別する固有表現抽出と，識別した専門用語間の関係を調べる関係抽出を行っている．関係抽出は最初の段階として因果関係を対象とする．

本稿では機械加工文書に対する専門用語の辞書化と因果関係の関係抽出結果について報告する．

2. 関連研究

文脈における特徴を利用した関係抽出としては Kambhatla[1]や Zhou ら[2]の研究が挙げられる．これらの研究では構文木や bag-of-words などの特徴を用い関係抽出を行っている．また日本語における関係抽出のうち，因果関係を取り扱ったものについては格フレームを扱う佐藤ら[4]の研究や，接続標識に着目した乾ら[3]の研究が挙げられる．

また本研究が因果関係を持つ単語ペアの抽出を目的としているのに対し，新聞記事から因果関係を含む文の抽出を行った坂地ら[5]の研究では，素性ベクトルと機械学習を用いて f 値が 0.797 ほどの分類器を実現している．

3. 固有表現抽出

特定分野の自然言語処理において，専門用語を含んだ文は誤った解析結果を招きやすい．これは解析の際に使用する一般辞書に，専門用語の情報が不足していることに起因する．固有表現抽出では形態素解析ツール向けの機械加工用語に特化したユーザ辞書を作成することでこの問題に対応している．

辞書形式の採用は，関係の抽出対象となる専門用語をハードに設定できるという特徴に基づく．辞書化の手順は以下のようにした．

(1) 専門用語表の表記揺れ是正・リスト化
(2) コスト計算，類義語のタグ付け[6]
(3) サ変名詞の判定
(4) MeCab 規定形式で書き出し
(5) ユーザ辞書形式にコンパイル

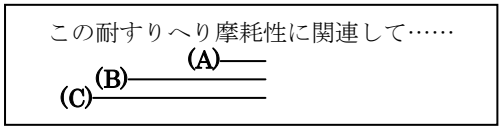

図 1 固有表現抽出の例

またアノテーションコーパスでは，専門用語が重なって存在することを許容する．例えば図 1 のような文脈の場合には，(A)摩耗，(B)すりへり摩耗，(C)耐すりへり摩耗の 3 つの専門用語を見ている．この処理には専門用語の中に含まれる部分的な専門用語を要素として汲み取る狙いがある．

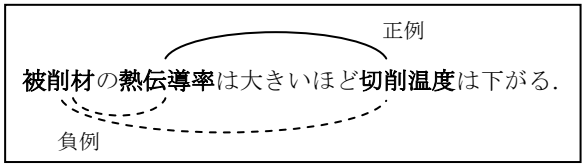

図 2 正例・負例の定義

図 2 の例文では，熱伝導率と切削温度のみが正例として扱われる．被削材と切削温度に関しては，熱伝導率という被削材の性質を挟んだ間接的な関係と捉え，今回の定義では負例として扱っている．正例は上記のような，文脈において明確に増減の影響が記述されているものを選出する．

4. 関係抽出

4.1 アノテーション

3 節で作成した辞書を用いて認識した専門用語について，１文を１枠とし，その中で作られる任意の 2 つの専門用語ペアを 1 ペアとする．

ここで，コーパス中の全ペアについて因果関係を持つか持たないかのアノテーション付けを行い，学習，精度計算に利用した．正例としたものについては，文脈的・直接的に明らかに因果関係を持つものとしている．

4.2 特徴ベクトル

表 1 に示すのは，コーパス中の全ペアに対し作られる特徴ベクトルの内容である．特徴の設定は関連研究を基に，研究途中で必要と判断した機械加工文書向けのものを加えている．

VARIABLE などは文中の変数宣言を分けるために設けている(これを加速度αとして……等)．WORDS は後述の 3 範囲で unigram から trigram までの計 9 種の N グラムを取得している．

表１　特徴ベクトルの内容

特徴名	特徴の内容	熱伝導率 ⇔切削温度	被削材 ⇔熱伝導率
DISTANCE	2 つの専門用語間の文字数	3	1
SENTENCE	2 単語が出現する文の文字数	9	9
PARSE TREE	2 単語間の構文木における係り受け数	3	1
PARSE TREE POS	2 単語間の構文木における品詞列	VP	0
MACHINING WORD	同文中の専門用語の数	3	3
VARIABLE	文中にアルファベットを含むか	0	0
DOT PHRASE	前の句が読点で終わっているか	0	0
PARENTHESES	専門用語が括弧で囲われているか	0	0
PARENTHESES VERB	上の時，括弧内に動詞を含むか	0	0
POS	専門用語の前後の品詞	助詞	助詞，文頭
CASE	直後にある助詞の種類(格)	は	の，は
WORDS	3 範囲(前，中，後)での n グラム (例は 2 単語間の unigram)	は,大きい, ほど	の

(1) 2つの専門用語より前の文脈

(2) 2つの専門用語間の文脈

(3) 2つの専門用語より後の文脈

この特徴ベクトルによって対象の2つの専門用語がどのような文脈の中に現れているのかを表現し，学習，判定に利用する．次元は16,000ほどであり，表中の値から正規化などの処理を通して特徴ベクトルへの格納が行われる．

なお，特徴ベクトルの学習と判定には，公開されているSupport Vector Machineのパッケージの1つLIBSVM[7]を使用している．

5. 評価実験

5.1 実験手順

形態素解析ツールにはMeCab[8]，構文解析ツールにはCaboCha[9]を利用した．豊田中央研究所から提供された機械加工文書に対し正例・負例のアノテーションを行い，表2のように訓練コーパス・テストコーパスに分割する．テストコーパス中のペアについて，因果関係を持つか，持たないかの分類を行い，その精度を評価した．

分類に対する評価指標にはROC(Receiver Operating Characteristic)曲線のAUC(Area Under the Curve)を用いた[10]．

表2　実験に用いたコーパス

	訓練	テスト
文数	1,196	358
専門用語数	3,852	973
用語の組数	4,833	1,366
正例数	502	146

表3　分類問題の定義

		真の結果	
		正	負
予測結果	正	TP	FP
	負	FN	TN

5.2 AUC-ROC（ROC曲線の曲線下面積）

陽性率(True Positive Rate)と偽陽性率(False Positive Rate)の相関性を示すROC曲線下の面積によって表される評価指標．ROC曲線は分離超平面からの距離に比例するスコア順に並べられた予測結果に対し，バイアスを大から小へ変化させ各点にて陽性率，偽陽性率をプロットすることで得られる．陽性率及び偽陽性率は表3より後述の式のように計算される[11]．

$$\text{陽性率} = \frac{TP}{TP + FN}$$

$$\text{偽陽性率} = \frac{FP}{FP + TN}$$

AUC-ROCは完全な分類器なら1.0，ランダムであれば0.5を取り，図2の左上へ膨らむほどいい分類器とされる．SVMの分類は最終的にバイアスによって分けられるが，最善のバイアスは文書によって異なる．AUC-ROCの採用は，正例より負例の数が多い本実験において有用である，バイアスによらず分類器を評価できるという特徴に基づくものである．

5.3 実験結果

図3に示すのがSVMによる分類結果のROC曲線である．AUC-ROCは0.76となった．結果から，当然ではあるが長い文に対する解析が弱いことが分かった．これについては構文木を見据えた処理と特徴を取り入れて対策を図りたい．また本来正例であるのにスコアが低く，負例と判定されたものの中で多かったのは，対象が括弧で囲われていた図4のようなパターンであった．太字にした温度と熱亀裂は因果関係を持っているが，スコアは低く負例と判定されており，現状では括弧を処理できる特徴が不足していることがわかった．例文では括弧は同義語を示しているが，一口に括弧と言っても様々な使われ方があり，それらを分類する必要がある．また例文では「温度」単体を専門用語として認識しているが，これを「温度変化」として取るべきなのか，固有表現抽出の範

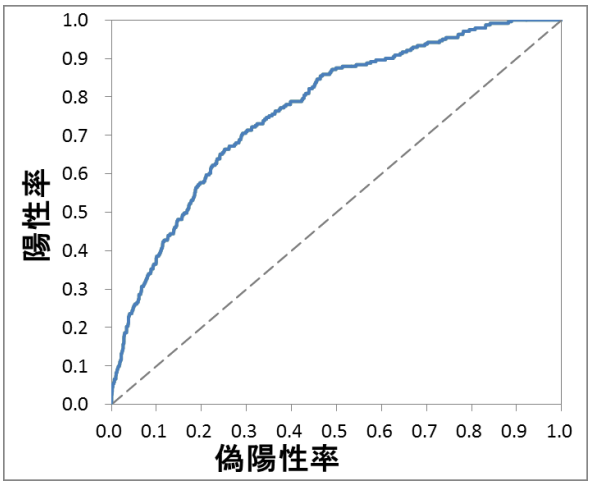

図 3 実験結果の ROC 曲線

また、断続切削に伴う温度変化によって熱膨張と収縮が起こるが、**温度**変化の上下幅が限度を超えると同様に収縮時に亀裂**(熱亀裂)**が発生し、チッピングに至る。

図 4 関係抽出の失敗例

囲の議論も一考の余地がある.

6. まとめと今後の課題

今回は機械加工文書に対する関係抽出として，形態素解析に用いる辞書を作成し専門用語を認識できるようにし，1,500 文ほどの機械加工文書に対し正例・負例のタグ付けをしたコーパスを作成した．機械加工文書向けの特徴ベクトルを定義し，SVM を用いて AUC-ROC で 0.76 程度となる分類器を作成した.

実用を考えると AUC-ROC は 0.9 以上が望ましく，現状では改善が求められる．今後精度を向上させるには，結果を精査し弱いパターンに対応・識別できる特徴の定義が必要となるため，実験，結果の評価，改善を通してより良い特徴の定義をしていく.

また前述のように，専門用語をどの程度まで認識するかも問題の１つである．温度変化など，パラメータの関係性を考える上で有用な固有表現を捉えることも将来的に役立つと予想される.

現在は因果関係のみしか扱っていないが，今後は上位語下位語，全体と部分などの関係性も扱っていき，関係情報の整理と体系化を目指していく．関係の種類ごとに捉えるべき文脈が異なることが予想できるため，種類ごとのチューニングも課題となる．どの種類の関係かという分類は多クラス分類となるので，SVM とは異なる分類手法も検討していきたい.

謝辞

本論文の執筆，及び，実験に貴重な助言を頂きました豊田工大三輪誠准教授に心より感謝します.

参考文献

[1] Kambhatla N. (2004). Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations. In Proceedings of ACL-2004. 21-26 July 2004. Barcelona, Spain.

[2] Zhou G., Su J., Zhang J., Zhang M. (2005) Exploring Various Knowledge in Relation Extraction. In ACL-05, Ann Arbor, MI, pp.427-434

[3]佐藤浩史，他(1999),テキスト上の表層的因果知識の獲得とその応用，信学技報，98(640)，pp.27-34

[4]乾孝司，他(2004)，接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得，情報処理学会論文誌，45(3)，pp.919-933

[5]坂地泰紀，増山繁(2011)，新聞記事からの因果関係を含む文の抽出手法，電子情報通信学会論文誌 D，J94-D，pp.1496-1506

[6]増田和浩，寺本一成，古谷克司，佐々木裕 (2014)，機械加工用語の関係性抽出，言語処理学会第 20 回年次大会，北海道，pp.74-77

[7]LIBSVM(http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

[8]MeCab(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)

[9]CaboCha(https://code.google.com/p/cabocha/)

[10] Charles E. Metz，訳:畑川政勝，他(1990)，ROC 解析の基礎，日放技学誌，46(6),pp.831-840

[11] 藤田広志，他(1993)，ROC 解析の基礎と最近の進歩，日放技学誌，49(9)，pp.1685-1703