Trigger Word と部分文字列を用いた機械加工用語の関係抽出

Machining Relation Extraction with Trigger Words and Partial Character Matching

増田 和浩* 寺本 一成** 古谷 克司* 三輪 誠* 佐々木 裕*
Kazuhiro Masuda* Kazunari Teramoto** Katsushi Furutani* Makoto Miwa* 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

近年，様々な社内情報がデジタルデータとして管理・保管され，利用されるようになっている．機械加工の分野でも，過去の加工情報を蓄積しておき，加工技術者が新規の加工の際に，加工条件の選定の参考にしている．現状では，加工条件の選定は加工技術者の経験によるところが大きく，もし，これらの蓄積情報から自動的に適切な情報を抽出することができれば，人手による条件設定の労力を軽減できるだけでなく，より適切な加工条件を自動的に提示することも可能となる．

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適用し，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を行うことも目的とする．

これまで，我々は，機械加工用語に対する固有表現抽出及び直接的因果関係の抽出に取り組んだ結果を報告してきた[1,2]．本稿では機械加工文書向けの素性を定義した SVM を用いて，上位下位関係を含む 4 種類の関係抽出を行った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 ら[3]や Zhou ら[4]の研究が挙げられる．これらの研究では構文木や N グラム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[5]の研究や，接続標識に着目した乾ら[6]の研究などが挙げられる．また本研究が機械加工文書からの関係抽出を目的としているのに対し，電子カルテから副作用関係という専門用語，専門的関係を対象に抽出を行った三浦らの研究[7]や三輪らの研究[8]などがある．

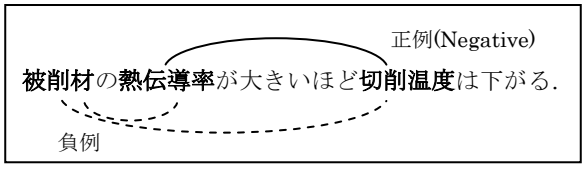

図 1 正例・負例の定義

3. 提案手法

機械加工文書及び文中に出現する関係種類について，既存の手法が有効であるか，どのような素性定義が有効であるかを考える．既存の手法としては，Support Vector Machine (SVM)[9]による教師有り学習及び表 2 に示す，n グラム，構文木や格を見る素性[3][4]を定義した素性ベクトルによる分類システムを用いた．3 節では，機械加工文書に対するアノテーション，既存の手法に新たに加えた素性について述べる．

3.1 アノテーション

専門用語及び正解の関係を機械加工文書にアノテーションした．正例として扱うものは文脈的・直接的に明らかに関係を持つとし，図 1 の例では間接的な被削材と切削温度の関係は負例として扱っている．関係の種類として定義したものは表 1 に示す 4 つとなる．

表１ ４つの関係種類とその定義

Relation		何らかの因果関係がある
	Positive	一方が上がればもう一方も上がる
	Negative	一方が上がればもう一方が下がる
Sub		上位下位関係 (is_a)

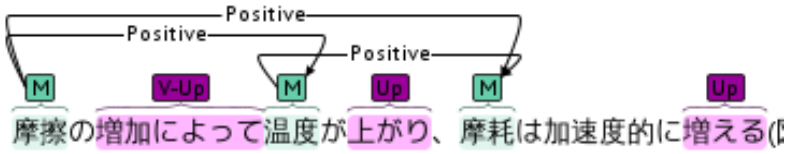

図 2 Trigger Word[10]の例

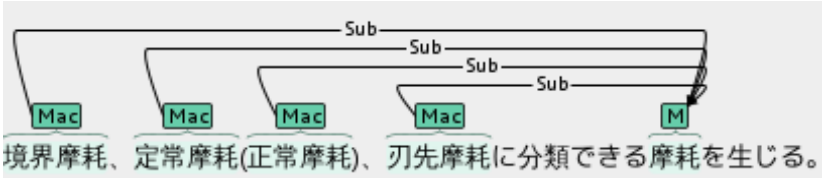

図 3 部分一致の例

Positive, Negative は入力条件と出力の因果関係，パラメータ同士の影響などの情報を記述する機械加工文書において多く出現し，Relation はこれらの上位関係となる．このような関係を抽出することで，パラメータを変更した時にどのパラメータ又は出力値に影響を及ぼすのかを明らかにする狙いがある．

表 2 ベースラインの素性定義

素性名	素性特徴の内容
DISTANCE	2 つの専門用語間の単語数 ・2 単語以上離れているか ・5 単語以上離れているか ・7 単語以上離れているか
PARSE TREE	2 単語間の構文木上の係り受け数 ・1 であるか（直接修飾） ・5 以上離れているか
PARSE LEVEL	2 単語間の構文木上の階層差 ・同じ階層か ・階層差が 2 以上か
VARIABLE	・文中に英文字を含むか
DOT PHRASE	・前の句が読点で終わっているか
PARENTHESES	・専門用語が括弧中か ・その括弧内に動詞があるか ・その括弧内に英文字があるか
FREQUENCY	専門用語の出現頻度*100
POS	専門用語の前後の品詞
CASE	直後にある助詞の種類(格)
N-Gram	2 単語の前，中間，後および文全体での 4 種類の n グラム (unigram から trigram まで)

3.2 追加素性

・Trigger Word

Positive, Negative の抽出において，特に量,性質の変動に着目するため「大きくなる」「高い」「減少する」など，物量的・性質的変動を示す語句を Trigger Word としてタグ付けする．対象語句は 45 の形容詞、動詞からなる辞書[1]によるマッチングによって検出する．タグの種類は，下記に示す 3 種類に分かれる．

・Vary-Up　：上向きの変化を表す
・Vary-Down ：下向きの変化を表す
・Vary　　：何らかの変化を表す

ここで上向きとは，「上昇/増大」を指し，その良し悪しは考慮しない．例えば「精度の向上」も「不良品率の増大」も同じ Vary-Up である．この Trigger Word が，関係抽出を行う専門用語ペアと構文木的，表層的にどのような位置関係にあるかを見る．前者は Trigger Word と専門用語が係り受けの関係にあるかどうかを判定し，後者は 2 つの専門用語の左側，中間，右側の文脈で上記三種の Trigger Word それぞれの出現を見る．

・部分一致

対象となる 2 つの専門用語が同じ部分文字列を有しているかを見る素性であり，図 3 の，摩耗の下位語にあたる専門用語が，摩耗の部分文字列を有するようなパターンを識別する．

素性では，2 つの専門用語が共通の部分文字列を持つか，最長一致文字列が片方の専門用語と等しくなるかを見る．例えば，通常摩耗と摩耗であれば最大一致文字列は摩耗であり，これは専門用語「摩耗」と等しい．

4. 評価実験

評価実験では，3.2 節で述べた素性の追加が，関係抽出精度を有意に向上させるかを評価した．実験では，コーパスを表 3 に示すサブセット A，B に分割した．最初にサブセット A を用いた 5 分割検定で精度の向上を確認し，その後サブセット A を訓練セット，サブセット B をテストセットとして行う分類で，その向上が本質的であるかを評価した．構文解析には CaboCha[11]，SVM は SVM-light [12]のパッケージを用い，不均衡データであるため-j オプション[13]を利用した．(-j 1) これは正例が負例と比べ少ない場合，学習時に正例の重みを増やす手法である．下記の 4 通りの素性定義で 4 つの関係種類を対象に抽出を行い，分類精度を比較した．

- ベースライン
  表 2 の素性を使用
- +Trigger Word
  ベースラインに素性 Trigger Word を追加
- +部分一致
  ベースラインに素性部分一致を追加
- 提案手法
  ベースラインに Trigger Word，部分一致の素性を追加

結果の F 値は PR 曲線を描いた際の最良の F 値を取り，サブセット A を用いた実験では 5 回の平均を用いている．

表 3 評価実験で用いたコーパス関係抽出

サブセット		A		B	
文数		2,414		707	
専門用語数		6,946		1,818	
専門用語ペア数		13,143		2,679	
サブセット		A		B	
関係種類		正例数	負例数	正例数	負例数
Relation		1,467	11,676	267	2,412
	Positive	182	12,961	37	2,642
	Negative	136	13,007	26	2,653
Sub		143	13,000	23	2,656

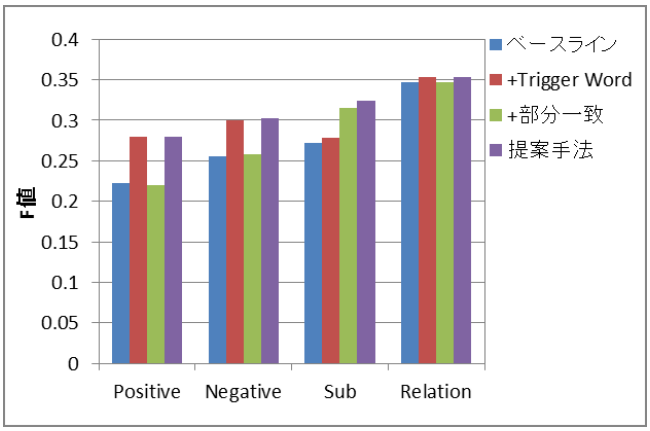

図 4 サブセット A を用いた実験結果

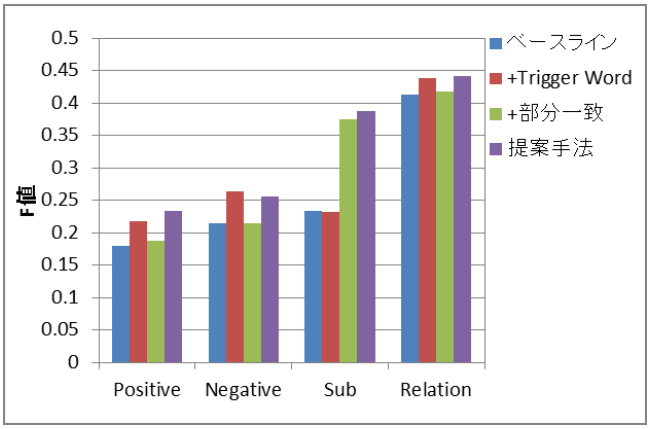

図 5 サブセット B を用いた実験結果

図 4 に示すのがサブセット A を用いた 5 分割検定の結果，図 5 に示すのがサブセット B をテストセットとした実験結果である．表 4 には実際の F 値を示す．表中の下線付きの数字は，ベースラインの精度と有意に差が見られた結果を表している．両方の実験で，素性 Trigger Word の追加で Positive, Negative の抽出精度向上が見られ，物量的・性質的な変化を示す語句を扱う素性が，Positive, Negative の抽出に有用であることが分かった．同様に，素性部分一致の追加による Sub の抽出精度向上も確認できた．これは図 3 のような，部分的に上位語を含む専門用語が多く出現したためである．このような傾向が強いと予想される，機械加工をはじめとする学術的な文章においては部分一致の素性が上位下位語の抽出に有用であると考えられる．

精度が横ばいだった Relation についてはおそらく，この関係種類がその他カテゴリであるという特徴が影響している．変化するなどの

PositiveかNegativeか判断できない関係にも，Ａを行うとＢが発生するといった発生条件関係にもRelationが付与されており，弁別は難しい．正例数が少なくなりすぎないよう注意を払いながら，更にカテゴリを細分化する等の対処が考えられる．

5. まとめと今後の課題

今回は機械加工文書中の４つの関係種類に対して，Trigger Word，部分一致の素性を提案し，SVMによる関係抽出システムを構築した．提案した手法により，F値で最大0.15の精度向上を行うことが出来た．

今後の課題としては，類義語などを利用したTrigger Word 及び専門用語の固有表現抽出が考えられる．加えて専門用語を工具，加工条件などのカテゴリに分類することができれば，関係抽出もそれを利用した精度向上が可能である．

図４　関係抽出の実験結果

サブセットA	Rel	Pos	Neg	Sub
ベースライン	0.35	0.22	0.26	0.27
+Trigger Word	0.35	**0.28**	**0.30**	0.28
+部分一致	0.35	0.22	0.26	**0.32**
提案手法	0.35	**0.28**	**0.30**	**0.32**
サブセットB	Rel	Pos	Neg	Sub
ベースライン	0.41	0.18	0.21	0.23
+Trigger Word	0.44	**0.22**	**0.26**	0.23
+部分一致	0.42	0.19	0.19	**0.38**
提案手法	0.44	**0.23**	**0.26**	**0.39**

参考文献

[1] 増田和浩，寺本一成，古谷克司，佐々木裕，機械加工用語の関係性抽出，言語処理学会第 20 回年次大会講演論文集，pp.936-939 (2014).

[2] 増田和浩，寺本一成，古谷克司，佐々木裕，SVM を用いた機械加工文書からの直接的因果関係の抽出．第21回言語処理学会年次大会，P4-12，pp.936-939 (2015).

[3]Kambhatla N., Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations, in Proc. of ACL, pp.21-26 (2004).

[4]Zhou G., Su J., Zhang J., et al., Exploring Various Knowledge in Relation Extraction, in Proc. of ACL, pp.427-434 (2005).

[5]佐藤浩史，笠原要，松澤和光，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 (1999).

[6]乾孝司，乾健太郎，松本裕治，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，(2004).

[7]三浦康秀,荒牧英治，大熊智子，外池昌嗣，杉原大悟，増市博，大江和彦，電子カルテからの副作用関係の自動抽出，言語処理学会 第 16 次年次大会発表論文集，pp.78-81 (2010).

[8]Miwa M., Sætre R., Miyao Y., et al., A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora, in Proc. of Conf. in EMNLP, pp.121-130 (2009).

[9] V. N. Vapnik, Statistical Learning Theory, Wiley (1998).

[10] Stenetorp P., Pyysalo S., Topić G., et al., BRAT: a web-based tool for NLP-assisted text annotation, in Proc. of EACL, pp.102-107 (2012).

[11]工藤拓，松本裕治，チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌,43(6),pp.1834-1842 (2002).

[12]SVMlight(http://svmlight.joachims.org/)

[13]Morik K., Brockhausen P., and Joacihms T., Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring, in Proc. of ICML, pp268-277 (1999).