入れ子構造を考慮した機械加工用語抽出

稲熊 陸[1]　小島 大[2]　東 孝幸[2]　三輪 誠[1]　古谷克司[1]　佐々木 裕[1]

[1]豊田工業大学　[2]株式会社ジェイテクト

[1]{sd19406, furutani, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp
[2]{hiroshi_kojima, takayuki_azuma}@jtekt.co.jp

1　はじめに

機械加工分野において技術者は加工条件の決定や工程の見直しの際に現場で得た知見と技術文書から得た知識を元に，加工条件・工程を策定している．この際，考慮するパラメータ条件を網羅することは難しいため，本研究では，現場で得られる加工条件と技術文書中の機械加工パラメータを対応させた機械加工に関する知識ベースを作成する．要求される知識ベースは機械加工のそれぞれの工程を根ノードとする複数の木構造で構成される．木のそれぞれのノードは各工程における加工条件や加工パラメータなどの機械加工因子で構成される．これらの機械化加工因子が，木構造の親子関係と木構造に依存しない関係によって結ばれ，階層的な知識ベースを構成する．木の階層をたどることで兄弟ノードや親ノードにあたる機械加工因子の把握が容易になるとともに，木の関係をたどることで機械加工因子の変化に付随して変化する他の機械加工因子の把握が可能となる．結果として以下のような貢献が考えられる．

- 策定に関わるパラメータの網羅
- 従来の工程の引き継ぎのスムーズ化
- 新人技術者の加工知識に関する学習の援用
- 熟練技術者の高度な気付きの補助

機械加工分野における知識ベースの作成の第一歩として，技術文書から機械加工用語とその関係を抽出する必要がある．技術文書中には加工方法やパラメータ，加工条件などの機械加工ベースを構成する因子を表す様々な機械加工用語が存在している．これらは複雑に関係で結ばれたり，用語自体が入れ子構造を形成したりすることが多く全ての機械加工因子と関係を人手でタグ付けすることは負荷がかかり難しい．

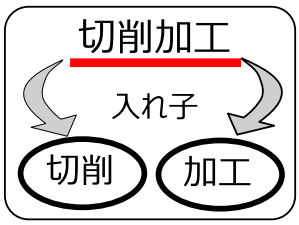

図1　入れ子構造の例

そこで本研究では，機械加工分野における情報抽出に向けて，入れ子構造を考慮した用語抽出を行う．

2　関連研究

2.1　機械加工分野における情報抽出

機械加工分野における情報抽出に向けて，増田ら[7]は形態素解析器MeCabのユーザ辞書機能を用いた用語抽出とSupport Vector Machine（SVM）を用いた関係抽出を行った．増田の用語抽出は機械加工用語をユーザ辞書にあらかじめ登録する必要があり，未知の機械加工用語の抽出精度が悪い．

2.2　入れ子構造からの用語抽出

入れ子構造は用語が用語が内包する構造のことをいう．図1に入れ子構造をとる機械加工用語の例を示す．「切削加工」という機械加工用語が「切削」と「加工」という機械加工用語が内包している．内包する用語と内包されている用語は主に上位下位の関係や属性関係，主述関係などを持っており，入れ子構造が階層的な情報を持っていると考えることができる．図1の例だと切削加工は「切削」という工程における「加工」という方法であることが分かる．

用語抽出の研究は入れ子構造を対象としない研究が盛んに行われているが，深層学習の進展に伴い，入れ子構造からの用語抽出が近年，複数提案[2, 3, 5]されている．この中で，従来のIOB2に代表されるトークン単

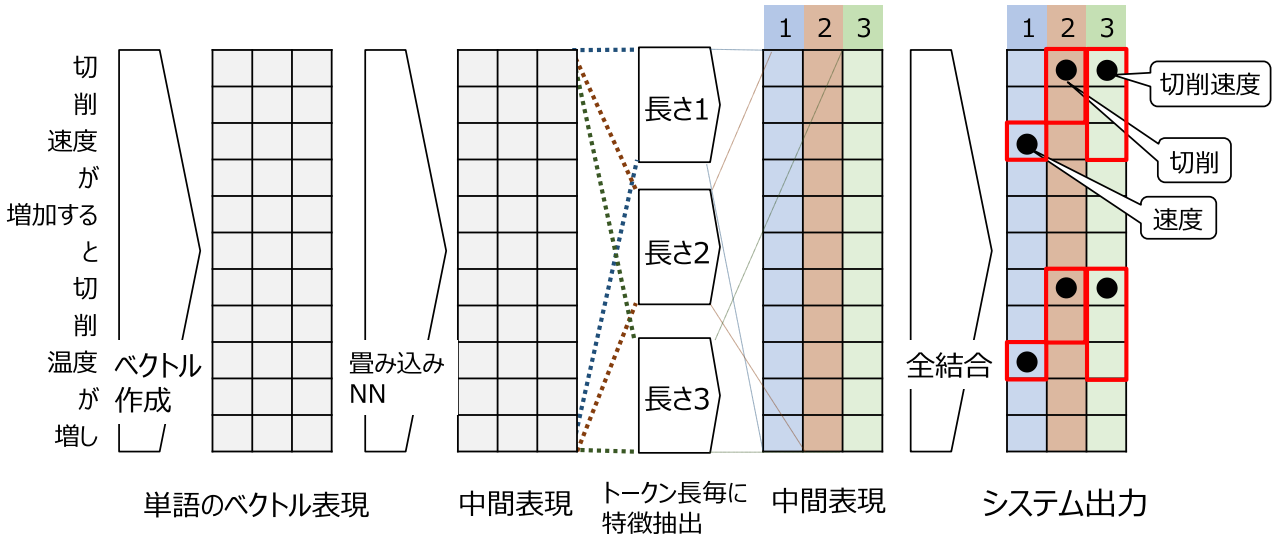

図2　提案モデルの概要図

位のタグ付けを行わず，用語の領域全体のスパンを対象として分類を行う手法が提案[3, 5]されており，用語レベルの特徴を捉えられることから，高い性能を示している．

3　提案手法

本研究で提案する入れ子構造を考慮した用語抽出モデルの概要を図2に示す．入れ子内部の用語と入れ子外部の用語の違いが用語を構成するトークンの数であることに注目し，本提案モデルはトークンの構成数毎に2値[*1]の出力をするモデルとなっている．図における格子状の四角はベクトルを表し，行方向に各トークンのベクトルの次元を，列方向にトークンの並びを表現している．五角形は演算を表現している．長さnという演算は，長さnのトークン列の表現を得る演算であり，以下に詳しく述べる．

3.1　BERT

日本語版Wikipediaで学習済みのSentence piece[6]を用いてsubword単位で分割を行って得られたトークンに対して，同じ日本語版Wikipediaで学習済みのBERT (Bidirectional Encoder Representations from Transformers)[1][6]を用いて埋め込みを行う．

3.2　畳み込みNN

前層で得られた埋め込みに対して，フィルタをかけて，要素の積和演算を行う畳み込み処理を行う．行列Xに対する畳み込み処理は以下の式で表される．

$$a_i^{(k)} = \sum_{s=0}^{n-1} \sum_{t=0}^{m-1} w_{st}^{(k)} x_{(i+s)(t)} + b^{(k)} \qquad (1)$$

ここでkはフィルタのインデックスを表す．a，xはA，X内に存在する各要素，nとmはカーネルサイズ，wはフィルタを構成する各要素の重み，bはバイアスを表す．カーネルサイズは入力データに対して畳み込む範囲を表す．

3.3　トークン数毎の用語抽出

図における長さnの演算はカーネルサイズnのフィルタによる畳み込み演算を表す．nは用語を構成するトークン数に対応している．前層で得られた中間表現に対してm枚フィルタをそれぞれ作用させ，カーネルサイズ毎のローカルな特徴を持つm個の中間表現を獲得する．次にm個の中間表現に対して1つの全結合層をそれぞれ作用させる．この全結合によってm個のローカルな特徴同士がお互いに影響し合うことを意図している．全結合の出力はm個のカーネルサイズ毎に対応したトークン数長のシーケンスになる．出力はカーネルサイズの用語を構成するトークンの開始位置に1が立つ．例として「切」「削」「速度」という3つのトークンから「切削」と「速度」と「切削速度」という機械加工用語を抽出することを考える．カーネルサイズ1に対応する出力では「速度」のトークンに対応する位置に1が出力される．これは「速度」という1トークンで用語を

[*1] 本研究で対象とするコーパスの用語タイプは1種類であるため，2値としている．

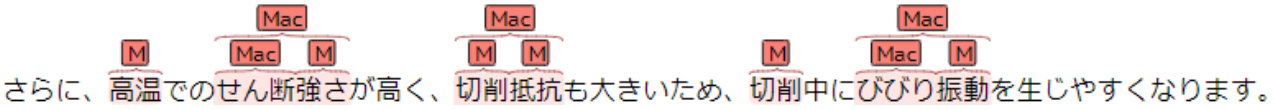

図3　システムの抽出した用語の例．表示にはbrat[4]を用いた．

構成することを表している．カーネルサイズ2に対応する出力では「切」に対応する位置に1が出力される．これは「切」と1トークン後ろの「削」の2トークンで用語を構成することを表している．カーネルサイズ3に対応する出力では「切」の位置に1が出力される．同様に考えて「切」「削」「速度」の3トークンで用語を構成することを表している．このようにトークン数という観点で出力数を増やし，開始位置を予測することで入れ子構造内部の用語も抽出が可能となる．

3.4　損失

損失は以下の式で表される．

$$L = \sum_{i=1}^{m} \sum_{j=1}^{l} (y_i^j \log(\hat{p}_i^j)) \tag{2}$$

ここでmはモデルの出力数すなわち用意するフィルタ数を表し，lは1文におけるトークン数を表す．各トークンに対する予測$\hat{p}$と正解yとの差を交差エントロピーで定義する．各フィルタに対応する出力の和をとったものを損失とする．

4　実験

提案モデルを用いて，入れ子構造を考慮した機械加工用語の用語抽出実験を行った．また入れ子内部を抽出することによる入れ子外部の抽出精度の比較実験も行った．

4.1　データセット

今回用いたデータセットは増田が用いた日本語で記述された機械加工分野の教科書である．加工技術について広く教示しており，実際に現場で働く技術者や初めて加工技術を学ぶ新人技術者も広く学習を行うことができる教科書となっている．この機械加工分野における教科書に対して専門家の指導の下でアノテーションを行い，機械加工用語には“Machining”のラベル付けを行った．対象の文は合計2,881文で学習用に2,304文，開発用に288文，評価用に289文として分割して検証を行った．

表1　用語抽出の実験結果

学習データ	F値
内部・外部をラベル付け	0.7921
外部のみラベル付け	0.7824

4.2　実験設定

最適化アルゴリズムは学習率0.001のAdamを用いた．抽出するための用語のフィルタ数は4とした．これは予備実験において，対象のデータにおける用語を構成するトークン数を計算したところ，4トークンまでで全体の用語の機械加工用語を構成するトークン数が全体の96%を越えることが分かったためである．

4.3　入れ子を考慮した用語抽出実験

図3に提案モデルが出力した機械加工用語をアノテーションツールbrat[4]で表示したものを示す．赤色のハイライトが抽出した機械加工用語を表す．せん断強さにおけるせん断と強さといった用語が入れ子内部の用語として抽出できている．稲熊のモデルでは抽出できなかった入れ子内部の用語も抽出できたことが確認できた．

4.4　入れ子構造内部の影響の評価

ここでは入れ子内部の用語の抽出による入れ子外部の用語抽出の精度の評価について述べる．具体的には，ラベルを入れ子内部と外部に付けたものと，入れ子外部のみに付けたものの2種類の教師データを作成し，入れ子内部もラベル付けした教師データによる，入れ子外部の抽出精度の変化を評価した．

コーパス中には部分的に重なった用語がないため，入れ子外部のみへのラベル付けは次のように行った．最初に1文中の全ての用語を入れ子によらず，文中において何文字目に開始されて何文字目に終了するかという情報とともに保存する．次に保存された用語全てについて他の用語の内部のものであれば，入れ子内部と判断し，除去するという処理を行う．除去対象の用語かどうかは文中における開始位置と終了位置が他の用語の開始位置と終了位置の範囲の中に入っているかど

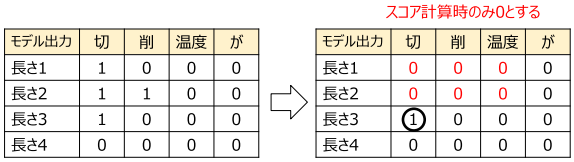

図4　評価方法

うかで判断する．

評価は，入れ子外部のみの評価は2種類の教師データで学習したモデルに対して，外部のみにラベル付けされた評価データで行った．この時，入れ子内部も予測するように学習したモデルの入れ子外部のみの評価は，1つのトークンに対して1が出力された場合，出力に対応するフィルタサイズより小さいフィルタサイズの出力を評価時のみ0とすることで行った．図4に例を示す．「切」トークンにおいて長さ1，2，3に対応する出力で開始位置を表す1が立つとする．この時長さ3に対して出力があるので「切」トークンから始まる長さ1と長さ2に対応する出力は入れ子内部の用語を示していると考える．長さ3の出力は「切」，「削」，「温度」として用語を抽出しているので，この入れ子内部となる「切削」，「削温度，「切」，「削」，「温度」に対応する出力を評価時に0とする．「温度が」に対応する出力についてであるが，「温度が」の一部である「温度」が「切削温度」に対して入れ子構造を構成することになる．今回用いた文書のアノテーションの際には，このような用語の一部のみ入れ子構造を構成するものは作成されていないため，「温度が」についての評価も0として行う．よって，図4に示す「切」「削」「温度」の長さ1と長さ2に対応する出力を評価時には全て0とした．

表1にラベル付けの違いによる用語抽出精度の比較を示す．外部のみで学習した場合がF値0.7824という結果に対して，内部も外部も学習した場合はF値0.7921を記録した．入れ子内部の予測が入れ子外部の予測に効果があることが分かった．これは学習データが増加したことによってモデルがより多くの特徴を獲得したことに起因すると考えられる．また，各トークンの出力毎の評価を表2に示す．1トークン，2トークン，4トークンで構成される用語は内部も外部も学習したことによって抽出精度が向上した．

表2　トークン数毎の用語抽出のF値

数	内部・外部をラベル付け	外部をラベル付け
1	0.7906	0.6936
2	0.8276	0.8105
3	0.8408	0.8509
4	0.7906	0.7525

5　おわりに

機械加工分野における入れ子構造を考慮した用語抽出タスクに対して，用語を構成するトークン数に注目し，複数の出力をすることで入れ子内部の用語を抽出するモデルを提案した．評価において，入れ子構造を考慮することで用語の抽出精度が向上することがわかり，入れ子構造を考慮することが有用であることがわかった．今後は機械加工ベースのための関係抽出を行うことを目標に研究を進める．

参考文献

[1] Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[2] Meizhi et al. A neural layered model for nested named entity recognition. In *Proceedings of NAACL-HLT*, pp. 1446–1459, 2018.

[3] Sohrab et al. Deep exhaustive model for nested named entity recognition. pp. 2843–2849, 2018.

[4] Stenetorp et al. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of EACL*, pp. 102–107, 2012.

[5] Zheng et al. A boundary-aware neural model for nested named entity recognition. In *Proceedings of EMNLP-IJCNLP*, pp. 357–366, 2019.

[6] Kikuta. BERT pretrained model trained on Japanese Wikipedia articles. `https://github.com/yoheikikuta/bert-japanese`, 2019.

[7] 増田ら．Trigger wordと部分文字列を用いた機械加工用語の関係抽出．言語処理学会第22回年次大会発表論文集, pp. 573–576, 2016.