2T-06

漸進的係り受け解析における未入力文節との構文的関係の同定

相津 徹也[†,a)]　大野 誠寛[†,b)]　松原 茂樹[‡]

東京電機大学未来科学部[†]　名古屋大学情報連携統括本部[‡]

1. はじめに

同時通訳や字幕生成，音声対話システムなどの音声言語システムでは，入力と同時的に処理を行うことが求められる．このようなシステムにおいて構文的情報を利用するためには，音声入力の途中で随時，構文構造を出力できる必要がある．

このような要請に答えるため，文節が入力されるごとに解析を実行し，係り先が入力されていない文節に対して，その係り先は未入力であることを明示した係り受け構造を出力するという漸進的係り受け解析手法（以下，大野らの手法）が提案されている[1]．この手法は，一定の構文情報を後段のシステムに随時提供することを実現しているが，更に豊かな構文情報を提供できる可能性が残されている．

そこで本論文では，より豊かな構文情報を後段のシステムに提供することを目的に，大野らの手法の出力構造を入力として，係り先が未入力である文節が複数あるときは，それらの係り先が同一か否か（すなわち，未入力文節との構文的関係）を同定する手法を提案する．

2. 漸進的係り受け解析の出力構造

大野らの手法は，文節が入力されるごとに解析を実行し，係り先が入力されていない文節に対して，その係り先は未入力であることを明示した係り受け構造を出力することを目的としている．図 1 は，大野らの手法が，文「さっき入って参りましたら机の上に旗が立っているのでこれは国連に来てしまったのかなという感じが致しました．」の「旗が」までが入力された段階で出力する構造を示しており，「入って参りましたら」，「上に」，「旗が」の係り先が未だ入力されていないということを示している．これにより，既入力文節内の「さっき入って参りましたら」や「机の上に」が構文的まとまりを構成することがわかる．

一方，係り先が未入力である文節が複数存在したとき，それぞれの文節が異なる未入力文節に係ることもあれば，同一の未入力文節に係ることもある．各文節の係り先が同一か否かを同定できれば，構文的なまとまりをより詳細に捉えることが可能となる．

本研究では，大野らの手法による漸進的係り受け解析の結果を入力として，係り先が未入力である文節が 2 つ以上存在したとき，それらの係り先が同一であるか否かを決定することにより，図 2 の係り受け構造の同定を試みる．図 2 の文節「上に」と「旗が」は同一の未入力文節（未入力文節 A）に，文節「入って参りましたら」は「上に」や「旗が」とは異なる文節（未入力文節 B）に係る．このような係り受け構造を同定できれば，「机の上に旗が」と未入力文節 A からなる文節列が構文的まとまりを構成し，「上に」や「旗が」はそのまとまりの中に含まれるということがわかる．

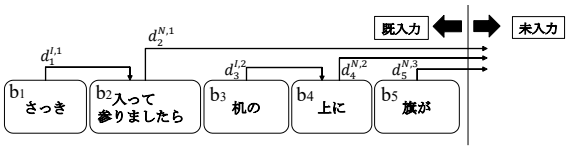

図 1　大野らの手法が出力する係り受け構造

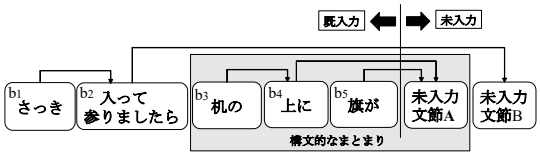

図 2　本手法により同定される係り受け構造

3. 漸進的係り受け解析における未入力文節との構文的関係の同定

本手法は，文節列$b_1 \cdots b_m$からなる文を解析する際，文節列$b_x(1 \leq x \leq m-1)$が入力されるたびに，それまでに入力された文節列$B_x = b_1 \cdots b_x$と大野らの手法の出力する係り受け構造D_xとを入力とし，係り先が未入力の係り受け関係が複数ある場合は，それらの係り先が同一か否かを出力する．

ここで，大野らの手法が出力するD_xは，文節列B_xに対する図 1 の形をした係り受け構造であり，係り先が未入力の係り受け関係$d_k^{N,\alpha}$（$1 \leq k \leq x-1$，$1 \leq \alpha \leq x$）と，係り先が既入力の係り受け関係$d_k^{I,\beta}$（$1 \leq k \leq x$，$1 \leq \beta \leq x-1$）の集合として定義されるものとする．kは係り元文節の番号を，Nは係り先が未入力であることを，Iは係り先が既入力であることを意味する．αは係り先が未入力の係り受け関係の中で，また，βは係り先が既入力の係り受け関係の中で，それぞれ係り元文節の番号で昇順に並べた際の順番を示す．例えば，図 1 の係り受け構造D_5は，$D_5 = \{d_1^{I,1}, d_2^{N,1}, d_3^{I,2}, d_4^{N,2}, d_5^{N,3}\}$と表記される．

本手法のアルゴリズムを以下に示す．

① D_xにおいて，係り先が未入力の係り受け関係$d_k^{N,\alpha}$の数$L(\alpha$の最大値$)$を集計し，$L = 1$の場合は終了する．$L \geq 2$の場合は手順②の判定を$\alpha = 1$

Identification of Syntactic Relations with Non-Inputted Words in Incremental Dependency Parsing

Tetsuya Aizu[†,a)], Tomohiro Ohno[†,b)], Shigeki Matsubara[‡]

† School of Science and Technology for Future Life, Tokyo Denki University.

‡ Information and Communications, Nagoya University.

a) 16fi002@ms.dendai.ac.jp

b) ohno@mail.dendai.ac.jp

から$\alpha = L-1$まで$L-1$回繰り返す．

② 係り先が未入力の係り受け関係$d_{k'}^{N,\alpha}$ $(1 \le k' \le x-1)$と，$d_{k''}^{N,\alpha+1}$ $(2 \le k'' \le x)$の両者の係り先が同一であるか否かを機械学習により判定する．

なお，機械学習については，SVM，ロジスティック回帰，最大エントロピー法を試すこととし，それぞれを用いて判定した．素性には，文献[2]の素性のうち，語彙情報に関するものを使用した．

図 1 を例にすると「旗が」が入力された段階での大野らの手法の出力構造は$D_5 = \{d_1^{I,1}, d_2^{N,1}, d_3^{I,2}\ d_4^{N,2}, d_5^{N,3}\}$となる．このうち，係り先が未入力である係り受け関係は$d_2^{N,1}, d_4^{N,2}, d_5^{N,3}$の 3 つである($L$=3)．まず$d_2^{N,1}$と$d_4^{N,2}$の判定，次に$d_4^{N,2}$と$d_5^{N,3}$の判定を行う．

4. 評価実験

本手法の有効性を確認するために，日本語講演データを用いて評価実験を行った．

4.1. 実験概要

実験データとして，同時通訳データベース[3]に収録されている日本語講演音声の書き起こしデータを使用した．本データは，形態素情報，文節境界情報，節境界情報，係り受け情報が人手で付与されている．なお，係り受け情報をテスト時に使う際は大野らの手法の出力結果に置き換えて使用した．実験は全 16 講演を用いた交差検定により実施した．すなわち，1 講演をテストデータとし，残りの 15 講演を学習データとする実験を 16 回繰り返した．ただし評価では，大野らの研究[1]における評価用データと同じ 14 講演(1,714 文，20,707 文節)を使用した．

評価では，係り先が未入力の文節に対する係り先の同定性能を再現率，適合率により評価した．ここで，本手法は係り先が未入力である文節について，その係り先文節を具体的に決めるわけではないため，正解と出力結果の係り先が一致するかを単純には判定できない．そのため，本手法の出力から擬似的な係り先文節(例えば，図 3 における「未入力文節 A」)を用意し，一致する係り受け関係の数が最も多くなるように正解と出力の係り先文節を動的計画法を用いて対応付け，その結果をもとに，正解と出力結果の係り先が一致するか否かを判定した．例えば，図 3 の出力結果では，「未入力文節 A」と「ありがたいと」が対応付けられ，再現率が 1/1，適合率が 1/2 となる．

比較手法として，本手法の手順②における 2 値判定を等確率でランダムに行う手法(Chance Rate)を用意した．

各機械学習のツールとして，SVM は LIBSVM[1]を，ロジスティック回帰は Scikit-learn[2]を，最大エントロピー法は NLTK[3]を，それぞれ使用した．いずれもデフォルトのパラメータとした．

4.2. 実験結果

表 1 に実験結果を示す．再現率及び F 値において，

表 1　実験結果

	再現率	適合率	F 値
Chance Rate	61.02% (31,301/51,299)	63.35% (31,301/49,406)	62.16%
SVM	62.49% (32,055/51,299)	64.88% (32,055/49,406)	63.66%
ロジスティック回帰	62.58% (32,104/51,299)	64.98% (32,104/49,406)	63.75%
最大エントロピー法	62.86% (32,249/51,299)	65.27% (32,249/49,406)	64.05%

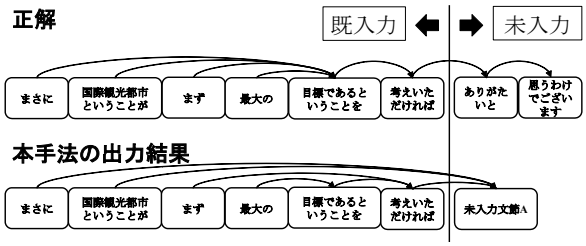

図 3　本手法の失敗例

本手法は全ての機械学習法で，Chance Rate を上回っており，本手法の実現可能性を確認した．なお，本論文で試した 3 つの機械学習法の間で比較すると最大エントロピー法が最も良い結果を示した．

図 3 に本手法の失敗例を示す．この例のように，大野らの手法の出力構造が正しくないために，本手法の同定が失敗した例が見られた．そこで，本手法の同定性能を単独で評価するため，日本語講演データに付与されている正解の係り受け構造から，大野らの手法の出力構造を抽出し，本手法への入力とした場合の実験を実施した．その結果，機械学習法を最大エントロピー法とした場合において，再現率と適合率はともに 73.36% (37,632/51,299)となった．大幅に性能が向上しており，大野らの手法の解析性能の向上が望まれる一方で，入力を正しい情報としても 70%程度であり，本手法単独での更なる性能向上も行う必要がある．

5. おわりに

本論文では，漸進的な係り受け解析における未入力文節との構文的関係を同定する手法を提案した．実験の結果，本手法の実現可能性を確認した．今後は，語彙情報以外の素性を追加するなどして更なる性能向上を図りたい．

謝辞　本研究は，一部，科学研究費補助金基盤研究（B）No. 26280082 及び（C）No.16K00300 により実施した．

参考文献

[1] 大野，松原，“文節間の依存・非依存を同定する漸進的係り受け解析，”信学論，Vol.J98-D，No.4，pp. 709-718，2015.

[2] 内元ら，“最大エントロピー法に基づくモデルを用いた日本語係り受け解析，”情処学論，Vol.40，No.9，pp.3397-3407，1999.

[3] S. Matsubara et al.,“Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research,” Proc．LREC2002，pp.154-159，2002.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm
[2] https://scikit-learn.org/stable/
[3] http://www.nltk.org/