冠詞推定のための情報構造仮説の検討

乙武 北斗 †　　永田 亮 ‡
福岡大学工学部 †　　甲南大学知能情報学部 ‡
ototake @fukuoka-u.ac.jp,　nagata-nlp @hyogo-u.ac.jp.

1 はじめに

冠詞の用法は，英語において習得が最も難しい文法項目の一つである．この難しさの解消を目指して，冠詞の自動推定に関する研究が盛んに行われている．冠詞推定の重要な応用の一つに，英語学習支援を目的とした冠詞誤り検出／訂正を挙げることができる．また，機械翻訳において，冠詞が存在しない言語から英語に翻訳する場合にも，後処理として冠詞推定は重要な役割を果たす [4].

機械学習の利用により，冠詞推定性能は大きく向上したものの，従来手法では十分に対応できていない冠詞の決定要因も依然存在する．冠詞の決定要因は，局所的文脈，文書外知識，大域的文脈の 3 種類に大別される．局所的文脈や文書外知識は従来手法で考慮されているが，大域的文脈を有効利用できている手法はほとんどなく，この部分に大きな問題が残る．大域的文脈は，冠詞決定において，定／不定の決定に大きく関わる．文献 [9] で，大域的文脈として前方照応と連想[*1]を冠詞推定に利用する手法が提案されているが，その効果は大きくない．その理由として，二度目の名詞は代名詞で置き換えられる傾向にあることが挙げられる．また，初出でも定冠詞が選択されることも多い．以上をまとめると，冠詞推定性能の更なる向上には，大域的文脈の効果的な利用が重要であると言える．

このような背景を受け，本稿では，大域的文脈を冠詞推定に効果的に利用する枠組みとして**情報構造仮説**を提案する．情報構造仮説では，人のコミュニケーションの性質を冠詞推定に利用する．コミュニケーションの究極の目的は，何らかの新しい情報を受け手に伝えることである．このことは，ある談話のまとまりは，通常，新情報を含むことを意味する．しかしながら，新情報のみの提示では，受け手の理解が困難になる．円滑なコミュニケーションのためには，新情報に加えて，旧情報も適宜受け手に提示することが必要となる．その結果として，受け手は旧情報を新情報の理解に利用することができる．このことは，円滑なコミュニケーションでは新旧情報が織り交ぜられることを意味する．本稿では，この新旧情報の織り交ぜのことを情報構造と呼ぶ．より正確には，新旧情報の連なりを情報構造と呼ぶことにする．新旧情報の連なりは複数文にまたがるため，情報構造は局所的文脈に加え大域的文脈を含むことになる．情報構造仮説では，情報構造には特定のパターンが存在すると予想する．もし，そのようなパターンが存在すれば冠詞推定に重要な手がかりを与える．通常は，旧／新情報は定／不定冠詞で表わされるからである．本稿では，コーパスを利用した実験により，情報構造仮説が実際に成り立つかどうかの検討を行う．

2 情報構造仮説

2.1 情報構造とは

本稿では，文および文章には新情報・旧情報（以後，それぞれ新・旧と省略する）が含まれると考える．これら新・旧を**情報ラベル**と定義し，この情報ラベルの連なりを**情報構造**と定義する．

情報構造の最小単位を厳密に規定することは困難であるが，本稿では名詞句を情報構造の最小単位とする．情報構造の最小単位は名詞句とは異なる単位の可能性も考えられるが，以下に示す理由により取り扱いが良いため，本稿ではそのように仮定する．

理由 1： 既存の解析器で容易に認識可能
理由 2： 名詞句の定／不定を旧／新へ容易にマッピング可能

ある名詞句に対して新／旧のどちらの情報ラベルが付与されるかは，基本的には名詞句が定であれば旧，不定であれば新と仮定する．代名詞と指示詞の場合，これらは既知の情報を指し示すため，旧と仮定する．例

[*1] 冠詞と前方照応／連想の関係については文献 [1] が詳しい．

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 1　名詞句の情報ラベルマッピング規則

	条件	ラベル
1	主名詞が固有名詞/代名詞/指示詞	旧
2	名詞句に所有限定詞を含む	旧
3	名詞句に定冠詞を含む	旧
4	1〜3 のどれも該当しない場合	新

えば以下の文 (i) では，*The book* が旧，*a man* が新となり，情報構造は “旧—新” となる．

(i) The book$_{(旧)}$ was given to a man$_{(新)}$.

以下の文 (ii) において，*I* は旧であり，情報構造は “旧—旧” となる．

(ii) Can I$_{(旧)}$ play on the swing$_{(旧)}$?

また，固有名詞は *the* が省略されていると考えることができる [1]．したがって，固有名詞も常に旧情報であると見なすことが可能である．以下の文 (iii) において，*Tokyo* は旧であり，情報構造は “旧—新” となる．

(iii) Tokyo$_{(旧)}$ is a very expensive place$_{(新)}$.

以上の点を考慮し，名詞句の情報ラベルマッピング規則をまとめると表 1 のようになる．表 1 で示す規則は，上部の規則ほど優先的に適用される．

2 つの情報ラベルの組み合わせは以下の 4 つのパターンが考えられる．

新—新　A cat likes rats.
新—旧　I bought a book. The book is interesting.
旧—新　The book was given to a man.
旧—旧　My girlfriend really likes this.

2.2　情報構造仮説の提案

円滑なコミュニケーションのためには，新情報と旧情報を適切に織り交ぜて受け手に提示する必要がある．しかしながら，単純に交互に新旧情報を繰り返すのではなく，旧—旧—旧—新のようにすることで，適切な旧情報の繰り返しが新情報の理解を助ける働きが期待でき，新情報を理解しやすいと考えられる．

母語話者エッセイコーパスの NICE-NS [8] に含まれる 210 文書を対象に情報ラベルの出現確率を調査したところ，旧の出現確率が 0.56，新が 0.44 であり，全体としては旧情報ラベルの方が現れやすい傾向となった．このことは，新情報よりも旧情報が多いほうが文書は理解しやすくなるという直感に合う．一方で，表 2 に示すように，単独の情報ラベル出現確率と比較して，より出現しやすい特徴的なパターンが存在する．表 2 において，“<d>” は文頭を表す．また，比率は該当する情報ラベルの条件付き確率と単独の出現確率の比である．表 2 から，<d>の後は旧が来やすいことがわかり，文書頭は旧情報の方が理解しやすいと解釈できる．同様に，<d>旧の後は新が来やすく，旧情報をもとに新しい情報を導入していることが伺える．

表 2　NICE-NS の情報ラベルの出現確率

	出現確率	比率
P(新)	0.44	1.00
P(旧)	0.56	1.00
P(<d>旧 | <d>)	0.62	1.11
P(<d>旧新 | <d>旧)	0.53	1.22

そこで本稿では，新旧情報の連なりである情報構造には特定のパターンが存在するという仮説を提案する．

2.3　情報構造仮説を利用した冠詞推定手法

2.2 で述べた仮説が成立する場合，冠詞推定に有益な情報となる．従来手法では各インスタンス（各名詞句）を独立として冠詞の推定を行う．しかしながら，情報構造仮説が成立すると，各インスタンスを独立に扱うのではなく，インスタンス間の関係性を積極的に冠詞推定モデルに組み込むべきである．

条件付き確率場 (CRF) を用いることで，新旧のラベルの連鎖を自然に考慮することができる．さらに，より広範囲の情報構造を考慮するために，2 パスによる推定を行う．具体的には，1 パス目で各インスタンスの新・旧の推定を行い，その推定結果を連結したものを 2 パス目の新たな素性として加える．

表 3 に本手法で用いる具体的な素性を示す．表 3 の w_i はそれぞれ，推定対象の冠詞の位置を基準に前後 i 番目の単語を表す．また，構文素性の *syntactic* はインスタンスの構文上の位置を表しており，主語の位置 (subj)，目的語の位置 (obj)，前置詞句 (pp)，その他 (other) のいずれかの値をとる．既知情報素性の $infoIsGiven$ は，インスタンスが前述した定／不定が既知である名詞句（代名詞など）であることを表すものである．定／不定が既知であるインスタンスは既知情報（$infoIsGiven$）のみを素性とし，それ以外のインスタンスは既知情報素性以外を素性とする．ただし，一部の限定詞 (another, any, some, many, much, no, all, every) はストップワードとし，単語 n-gram 素性の対象とはせず，既知情報素性にも設定しない．

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 3　冠詞推定に用いる素性

素性タイプ	素性
主名詞	$headNoun$
単語 1-gram	$w_{-3}, w_{-2}, w_{-1}, w_1, w_2, w_3$
単語 2-gram	$w_{-3}w_{-2}$, $w_{-2}w_{-1}$, $w_{-1}w_1$, w_1w_2, w_2w_3
単語 3-gram	$w_{-3}w_{-2}w_{-1}$, $w_{-2}w_{-1}w_1$, $w_{-1}w_1w_2$, $w_1w_2w_3$
構文	$syntactic \in \{\mathrm{subj}, \mathrm{obj}, \mathrm{pp}, \mathrm{other}\}$
既知情報	$infoIsGiven$

表 4　2 パス目の冠詞推定で追加する素性

素性タイプ	素性
自身	$l_0 \in \{$ 新, 旧 $\}$
周辺	$l_{-n}...l_{-2}l_{-1} - l_1l_2...l_n$ $(n \geq 1)$

表 4 に，1 パス目の推定結果をもとに追加する素性を示す．表 4 の l_i は，対象インスタンスの前後 i 番目 ($i < 0$ の場合は前，$i = 0$ は対象インスタンス自身) のインスタンスにおける 1 パス目の大規模 ME によって推定されたラベルを表す．周辺素性は対象インスタンスの前後 n 個のインスタンスを範囲とする．$n = 0$ の場合は，周辺素性は使用されない．また，自身素性 l_0 は 1 パス目の推定確率値を重みとして用いる．

3　評価実験

3.1　実験対象

本実験では母語話者による文書を対象に，データ中の名詞句の定冠詞の有無を正解として，定／不定の推定性能を評価した．評価の際，**2.3** で述べた既知情報素性を持つインスタンスは対象から外した．

評価データとして，母語話者エッセイコーパスの NICE-NS [8] を用いた．エッセイコーパスは文法誤り訂正においてよく使用される点から，本実験の対象とした．NICE-NS に含まれる 210 文書を対象に，34,970 個のインスタンスが得られた．

訓練データとして，記事コーパスの Reuters-21578 [5]，エッセイコーパスの ICNALE V2.1 [3] を用いた．記事コーパスは大規模なものが容易に入手できることから，訓練データに加えた．Reuters-21578 からは 720,992 個のインスタンスが得られた．ICNALE からは英語母語話者によるエッセイ 200 文書のみを対象に，24,236 個のインスタンスが得られた．

3.2　実験方法

本実験では **2.3** で述べた情報構造仮説を利用した手法を評価した．比較対象として，情報構造を考慮しない最大エントロピー分類器（ME）に関しても評価を行った．また，本稿で提案する 2 パス手法において，1 パス目と 2 パス目の訓練データのドメインを異なるものにした場合にドメイン適応と見なすことができるため，その比較対象として Daume III のドメイン適応手法 [2] を追加した．

ME の実装として，Classias [7] の L2 正則化ロジスティック回帰モデルを用いた．また，CRF の実装として CRFsuite [6] の L2 正則化モデルを用いた．

CRF によるモデル生成の際に用いるインスタンスの系列の範囲は，ICNALE，NICE-NS ともに文書単位とした．ME によるモデル生成の際は，**2.3** で述べた既知情報素性を持つインスタンスを訓練データから除外した．

3.3　実験結果

表 5 に各手法の評価結果を種類別に分類したものを示す．表 5 の $c2$ は L2 正則化のハイパパラメータを，n は表 4 で示した周辺素性の範囲を表す．$c2$ は $0.1 \leq c2 \leq 1.5$ の範囲を 0.1 刻みで，n は $0 \leq n \leq 10$ の範囲を 1 刻みで調整し，最も精度が高い結果を表 5 に示した．

最も精度が高い結果は Daume III のドメイン適応手法となった．また，表 5 の【通常の手法】中の 2 パス ME-ME は，周辺素性の範囲が $n = 0$ のとき，すなわち情報構造を考慮しない場合が最良の結果となった．ME-ME（$n = 0$）の場合でも性能向上は認められるが，これはドメイン適応の効果が効いていると考えられる．

情報構造を考慮した CRF による手法は，2 パスではない通常の手法と比較して性能向上が認められたが，こちらも周辺素性範囲が $n = 1$ と狭いため，ドメイン適応による性能向上の側面が大きいことが伺える．

4　考察

情報構造を利用した冠詞推定によって精度は改善しなかったが，他手法では推定できなかったものを正しく推定できた事例も存在した．例えば，以下の文

(iv) The death penalty$_{(旧)}$ is used in many countries$_{(新)}$ as a punishment$_{(新)}$ for...

において，文頭から続く “penalty”，“countries” の情報構造は，表 2 と同様の形式で表すと「<d>旧新」となる．表 2 で示したように，文頭の後に旧，および<d>旧

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 5　評価結果

手法	精度	適合率	再現率
【通常の手法】			
ME ($c2 = 0.6$) ICNALE	76.27%	77.11%	**94.02**%
ME ($c2 = 0.2$) Reuters	76.91%	79.74%	89.86%
ME ($c2 = 0.1$) Reuters+ICNALE 単純結合	78.16%	**80.52**%	90.78%
(1 パス目)ME ($c2 = 0.2$) Reuters - (2 パス目)ME ($c2 = 0.3, n = 0$) ICNALE	78.48%	80.33%	91.75%
【ドメイン適応】			
ME ($c2 = 0.3$) Reuters+ICNALE ドメイン適応	**78.58**%	79.63%	93.26%
【情報構造仮説（提案手法)】			
CRF ($c2 = 0.3$) ICNALE	76.09%	77.59%	92.62%
CRF ($c2 = 0.2$) Reuters	76.71%	80.00%	88.99%
(1 パス目)ME ($c2 = 0.2$) Reuters - (2 パス目)CRF ($c2 = 0.6, n = 1$) ICNALE	78.19%	80.20%	91.43%

の後に新が出現する確率は高くなっている．2 パスの提案手法では “penalty” と “countries” の両者とも正しく推定することに成功しており，情報構造の利用が効果的だった事例と言える．その他の手法では，どちらか一方のインスタンスしか正しく推定することはできなかった．

しかしながら，全体の性能としてはドメイン適応の方が，情報構造仮説を利用した冠詞推定手法よりも高い結果となった．その原因の一つとして，本稿で提案した情報構造仮説では情報構造の近似（名詞句単位）が粗いことが考えられる．構文構造など，名詞句よりも複雑な単位で情報構造仮説を検証する必要がある．また，大規模 ME による冠詞推定性能が既に高いことから，情報構造を利用した冠詞推定手法の性能向上の余地が少なかった可能性も考えられる．

5　おわりに

本稿では情報構造仮説の提案し，それを利用した冠詞推定手法について述べた．性能評価実験の結果，冠詞推定手法の有効性は確認できなかったが，ドメイン適応が一定の性能向上に寄与することが確認された．情報構造仮説に関しては分析の結果，特定のパターンにおいて情報ラベルの出現確率の偏りが存在することがわかり，仮説を支持する結果となった．情報構造を冠詞推定の性能向上に活かすためにはさらなる工夫が必要だと考えられ，これを今後の課題としたい．

参考文献

[1] Francis Bond. *Translating the Untranslatable*. CSLI publications, Stanford, 2005.

[2] Hal Daume III. Frustratingly easy domain adaptation. In *ACL-2007*, pp. 256–263, 2007.

[3] S. Ishikawa. Icnale: The international corpus network of asian learners of english.

[4] Kevin Knight and I. Chander. Automated postediting of documents. In *Proc. of 12th National Conference on Artificial Intelligence*, pp. 779–784, 1994.

[5] D. Lewis. Reuters-21578 text categorization test collection, 1997.

[6] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[7] Naoaki Okazaki. Classias: a collection of machine-learning algorithms for classification, 2009.

[8] 杉浦正利．言語習得研究のための学習者コーパス．藤村逸子, 滝沢直宏（編）, 言語研究の技法, pp. 123–140. ひつじ書房, 2011.

[9] 竹内裕己, 河合敦夫, 細田直見, 永田亮．前方文脈を考慮した冠詞の推定．言語処理学会第 19 回年次大会発表論文集, pp. 717–720, 2013.

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.