E_024

補強対象を選ばない英語名詞の可算性判定補強手法

A Target-Independent Method for Reinforcing Countability Prediction

永田 亮† 河合 敦夫‡ 森広 浩一郎† 井須 尚紀‡
Ryo Nagata Atsuo Kawai Koichiro Morihiro Naoki Isu

1. はじめに

英語名詞の可算性判定は，様々な自然言語処理タスクで重要な役割を果たす．特に，日本語などの冠詞がない言語から英語への機械翻訳で重要となる[1, 2]．なぜなら，可算性は，冠詞を生成するための重要な情報となるからである．例えば，日本語文：

「トイレットペーパ（toilet paper）が必要です．」

を，英文：

"We need ___ toilet paper."

に翻訳する場合，日本語文には冠詞がないため下線部に入る冠詞を何らかの方法で生成しなければならない．その際に"paper"が不可算名詞であることがわかれば，不定冠詞は非文法となり，下線部に入る冠詞は無冠詞／定冠詞に限定できる．実際，多くの従来手法（例えば，文献[6]）でも，冠詞生成のために可算性を利用している．

冠詞の生成に関連して，可算性は文法誤りの検出でも重要となる．例えば，"He has many furnitures."という英文で"furniture"が不可算名詞であることがわかれば，不可算名詞を複数形にした誤りとして下線部が検出できる．

このような応用を目的として，オントロジーを利用した手法[2]やコーパスに基づいた手法[1, 5, 7, 8]など様々な可算性判定手法が提案されているが，その性能は，まだ十分でない．可算性は，名詞の意味や文脈によって決定されるため，従来手法では完全に対応できていないのが現状である．高品質な日英機械翻訳システムや文法誤り検出システムの実現を考えた場合，従来手法の更なる性能改善が必要不可欠である．

そこで，本稿では，従来手法の判定結果を補強し，判定性能を向上させる手法を提案する．提案手法の最大の特徴は，補強対象となる従来手法の種類を選ばず，可算性判定の補強が行える点である．提案手法では，One Countability per Discourse（OCpD）という仮説を導入し，可算性判定の補強を実現する．OCpD とは，可算性判定対象の名詞（以後，ターゲット名詞と表記する）が一文書内に複数回出現した場合，全て同じ可算性を示すという仮説である．補強の基本アイデアは，従来手法で判定に失敗した可算性を，OCpD を利用して正しい可算性に書き換えるというものである．なお，大部分の名詞が可算／不可算の 2 種類の可算性を取る[5]ことから，以降ではこの 2 種類の可算性のみを判定対象とする．

2．One Countability per Discourse

One Countability per Discourse（OCpD）は，One Sense per Discourse [9]を拡張した仮説である．One Sense per Discourse とは，ある単語が一文書内に二回以上出現した場合，その単語は全て同じ語義を持つという性質である．Yarowsky [9]は，文書内で二回以上出現した単語のうち99.8%の単語についてその語義が一致したと報告している．

One Sense per Discourse に基づいて，一文書内に出現したターゲット名詞は同じ可算性を持つという仮説を導くことができる．すなわち，OCpD である．なぜなら，可算性は，文献[2]に示されるように，多くの場合，名詞の意味により決定されるからである．例えば，"paper"が「紙」という意味である文書に出現したとする．ここで，「紙」という意味の"paper"は不可算名詞である．このとき，One Sense per Discourse に従うと，同一文書内の他の"paper"も全て「紙」という意味を持つことになる．したがって，全ての"paper"が不可算名詞となる．

OCpDを検証するために，Yarowsky [9]と同様の実験[1]を我々も実施した．文献[4]で，可算／不可算の両方で使用される名詞の例として示されている 23 種類の名詞をターゲット名詞とした．British National Corpus（BNC [3]）中で，ターゲット名詞が二回以上出現する文書に対して，各文書でのMajority Countability（MC）で可算性を判定し，文書集合全体での判定精度を求めた． MCとは，文書内で頻度が高いほうの可算性のことである．仮にOCpDが常に成り立つとすれば，文書内のターゲット名詞の可算性は全てMCに一致するため，MCを用いた判定精度は 100%になる．言い換えると，実験の結果得られた判定精度はOCpDがどの程度成り立つかを表す．

表 1 に実験結果を示す． 表 1 の"MC"は，MC での判定精度を表す．また，"Baseline"とは，ターゲット名詞が可算名詞及び不可算名詞として使用された頻度を，コーパス全体に渡って数え，頻度が高かったほうの可算性を用いて，判定を行ったときの精度である．

表 1 から， MC の値を利用すると，平均判定精度 85%以上を達成でき，"Baseline"を約 10%改善することがわかる．すなわち，ターゲット名詞の可算性の大部分が MC に一致し，OCpD が良く成立するといえる．

以上の結果を踏まえ，提案手法では，MC を可算性判定の補強に利用する．しかしながら，MC を可算性判定の補強に利用するためには，次の 2 つの問題を解決しなければならない．第一に，どのように MC の値を得るかということが挙げられる．本章では，実験のため，あらかじめ MC の値を与えて可算性を判定したが，実用上は，未知文書を対象として可算性の判定を行うため，個々のターゲット名詞に対する可算性も MC の値も与えられない．第二に，仮に MC の値が得られたとしても，どのように MC を可算性判定の補強に利用するかという問題が残る．表 1 に示されるように，OCpD の傾向にはターゲット名詞間でばらつきがある．そのため， MC を効果的に利用するためには，工夫が必要となることが予想される．

†兵庫教育大学，Hyogo University of Teacher Education
‡三重大学，Mie University

[1] 使用したツールなどの実験環境は，4.の実験と同様である．

表1 Majority Countability を利用した判定精度

ターゲット名詞	MC	Baseline
advantage	0.772	0.618
aid	0.943	0.671
authority	0.864	0.771
building	0.850	0.811
cover	0.926	0.537
detail	0.829	0.763
discipline	0.877	0.652
duty	0.839	0.714
football	0.938	0.930
gold	0.929	0.929
hair	0.914	0.902
improvement	0.735	0.685
necessity	0.769	0.590
paper	0.807	0.647
reason	0.858	0.822
sausage	0.821	0.750
sleep	0.901	0.765
stomach	0.778	0.778
study	0.824	0.781
truth	0.783	0.724
use	0.877	0.871
work	0.861	0.777
worry	0.871	0.843
Average	0.851	0.754

3. 提案手法

3.1 補強の基本アイデア

第一の問題は，MC の値を推定することで解決する．未知文書では MC の真の値を知ることはできないが，少なくとも推定することは可能である．1. で述べたように，可算性を判定する手法が既に提案されているので，MC の推定に利用できる．すなわち，任意の従来手法を用いて，未知文書中のターゲット名詞の可算性を判定し，その結果の多数決をとることで，MC の値を推定できる．

第二の問題は，推定された MC の値を素性として扱うことで解決する．推定された MC の値を，従来手法で利用されている，可算性の判定に有効と思われる他の素性と共に，機械学習アルゴリズムで学習する．MC の値をどの程度重視して可算性を判定するかは，その機械学習アルゴリズムと学習データによって自動的に決定される．学習された可算性判定モデルは，従来手法の素性に加え MC を利用するため判定精度がより高くなると期待できる．

3.2 補強モデルの学習

既に述べたように，提案手法では，機械学習アルゴリズムを用いて可算性判定モデルを学習し，可算性判定補強モデルとして利用する．素性を扱える機械学習アルゴリズムであれば可算性判定補強モデルに利用可能であるが，本稿では，様々な自然言語処理タスクで有効性が確認されている Maximum Entropy Model（MEM）を用いる．

MEMを学習するためには，学習データが必要となる．言い換えると，可算／不可算のラベルが付与されたコーパスが必要となる．幸い，可算／不可算のラベルを自動的に付与する手法[8]が提案されているので，この手法を利用する．手法[8]では，限定詞や単数形／複数形などの表層情報に基づいて，可算／不可算のラベルを付与する[2]．例えば，“He read a paper”では，“paper”は不定冠詞“a”に修飾されているので，可算のラベルが付与される．

可算／不可算のラベルが付与されたコーパスから素性を抽出し学習データとする．抽出する素性として，従来手法で利用されている任意の素性が利用可能である．本稿では，次に述べるターゲット名詞周辺の単語を素性として用いる：（i）ターゲット名詞が存在する名詞句内の単語（np と省略），（ii）その名詞句から左 3 単語（-3 と省略），（iii）その名詞句から右 3 単語（+3 と省略）．ただし，冠詞は素性として用いない．また，機能語（前置詞は除く）やターゲット名詞自身など可算性の判定に重要でないと予想される単語はストップワードとして除外する．全ての単語は，小文字かつ原形に変換する．これらの素性に加えて，MC の値を素性として用いる．MC の値は，文書ごとに可算／不可算の多数決を取ることで得られる．可算と不可算が同数であった場合には，MC の値を“*unknown*”とする．

ターゲット名詞を“paper”とした，学習データの生成の例を以下に示す．いま，ある文書内で可算／不可算のラベルが次のように与えられているとする：

He read a new *paper*／可算 in the morning ... read another *paper*／可算 in the afternoon.

このとき，MC の値は可算になり，

MC=可算 -3=read np=new +3=in +3=morning L=可算
MC=可算 -3=read +3=in +3=afternoon L=可算

という学習データが得られる．ただし，“L=可算”は，可算性のラベルが可算であることを表す．

学習データ生成の際に，ターゲット名詞が 1 度しか出現しない文書では，OCpD が常に成り立つことに注意しなければならない．すなわち，ターゲット名詞が 1 度しか出現しない文書から生成された学習データでは，MC の値とターゲット名詞の可算性が常に一致することになる．このような学習データから学習された可算性判定モデルは，MC を重視しすぎることが懸念される．この傾向を緩和するために，ターゲット名詞が 1 度しか出現しない文書では，MC の値を“*unknown*”と近似する．

以上の手順で生成された学習データに，Maximum Entropy アルゴリズムを適用し MEM を学習する．学習された MEM は MC とターゲット名詞周辺の単語を考慮した可算性判定補強モデルとなる．

3.3 可算性判定の補強

可算性判定の補強の流れを図 1 に示す．いま，図 1 の（1）のような文書が与えられており，ターゲット名詞は“paper”であるとする（この例では，機械翻訳や文法誤りの検出を想定して，冠詞や単数／複数の情報を除去している）．可算性の真の値は全て可算とする．

[2] 手法[8]は，冠詞や単数形／複数形などの情報に基づいて可算性を判定するため，1. で述べた機械翻訳での冠詞生成や文法誤りの検出には利用できない．

まず，この文書に対して，任意の従来手法を用いて可算性を判定する（図1の（2））．図1の（2）では，三番目のターゲット名詞の可算性を「不可算」と誤判定している．判定に失敗した原因として，三番目の例のように，判定の手がかりとなる単語が，ターゲット名詞周辺に存在しないことが挙げられる．また，学習データの不足により，類似した例が学習データに出現しなかったことなどが想定できる．

次に，従来手法の判定結果の多数決を取り，MC の推定を行う．このとき，従来手法の判定結果に，何らかの確信度が付与されていれば，確信度を考慮して多数決をとる．すなわち，閾値よりも高い確信度が付与された判定結果だけを対象にして多数決を取り，MC の推定精度を高める．例えば，MEM や Naive Bayes 分類器では，判定結果が確率で与えられるので，確信度として利用可能である．図1の（3）では，判定結果に確信度が付与されていないので，全ての判定結果を対象にして多数決を取り，MC の値を可算と推定する．

次に，素性の抽出を行う（図1の（4）[3]）．素性の抽出は，3.2 と同様の手法で行う．ただし，3.2 とは異なり，可算性の真の値は未知であるので，可算性のラベル（3.2 の例の“L=可算”）は抽出しない．

最後に，抽出した素性に，3.3 で得られたモデルを適用して（図1の（5）），可算性判定の補強を行う（図1の（6））．ここでは，MC が素性として利用できるので，仮にターゲット名詞周辺に手がかりとなる単語が存在しなくても，正しく判定が行える可能性が高い．図1の場合でも，三番目のターゲット名詞の周辺に手がかりとなる単語は存在しないが，素性“MC=可算”によって従来手法の判定結果が正しく可算に書き換えられている．

4. 実験と考察

4.1 実験条件

本実験では，文献[8]の実験でも使用されている 23 種類の名詞をターゲット名詞として選んだ．この 23 種類の名詞は，文献[4]に可算／不可算のどちらでも使用される例として示されているものである．

学習データと評価データは，BNC[3]から作成した．BNC の書き言葉コーパス中のテキストタグに囲まれているテキストを一文書とした．全文書の約 10%にあたる 314 文書を評価データに利用した．評価データの作成は，文献[7]の手法を利用した．残りの 90%の文書を学習データとして利用した．名詞句の抽出には，OAK System[4]を用いた．長すぎるためにOAK Systemが解析できなかった文は除外した．

可算性判定手法の評価尺度には精度を用いた．精度は，正しく可算性の判定が行えた名詞数を判定した名詞数で除した値と定義した．

[3] 3.2 で述べたように，機能語は素性として抽出しない．したがって，三番目の例では，“where” と “is” は抽出されず，素性は “MC=可算” のみとなる．

[4] http://nlp.cs.nyu.edu/oak/

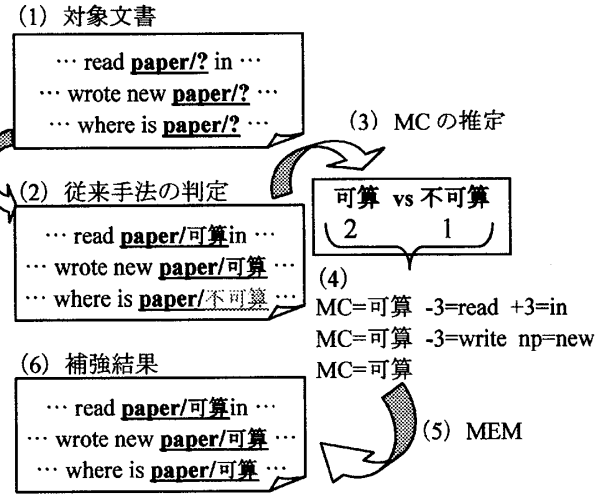

図1 可算性判定補強の流れ（提案手法）

4.2 実験手順

まず，補強モデルの学習を行った．3.2 で述べた手法を用いて学習データを生成しMEM[5]を学習した．また，学習データから，素性MCを削除し，別のMEMを学習した．このMEMは補強対象として利用した．

次に，評価データ中のターゲット名詞の可算性を判定した．はじめに，素性 MC なしの学習データから学習された MEM を用いて可算性の判定を行った．その後，この判定結果を提案手法で補強した．MC の値を推定する際の閾値は 0.95 とした．

最後に，素性 MC なしの MEM の判定精度と補強後の判定精度を求めた．また，比較のため，補強前の判定結果を，推定された MC の値で全て書き換えた場合の判定精度も求めた（以後，この手法を単純補強法と呼ぶ）．更に，ベースラインとして，学習データ中で頻度が高いほうの可算性で判定を行った場合の判定精度を求めた．

4.3 実験結果と考察

表 2 に実験結果を示す．表 2 の“補強後 MEM”は，提案手法によって，“補強前 MEM”を補強した結果を表す．

表 2 から，提案手法は，“補強前 MEM ”の判定結果を効果的に補強していることがわかる．補強による改善は，最大で 5.9%（ターゲット名詞“discipline”）となった．また，補強前と補強後では，平均精度に有意水準 1%で有意差が見られた（paired *t-test*）．一方で，推定された MC の値で，補強前の判定結果を書き換えた単純補強法では，補強前よりも判定精度が低下している．これらの結果から，OCpD（MC）は可算性判定の補強に有効であるが，効果的に補強を行うためには提案手法のようにその利用方法を工夫しなければならないといえる．

3.3 で述べたように，提案手法は MC を考慮するため，学習データの量が少なくても，その性能を発揮すると予想できる．そこで，学習データの量を 100 例と極端に減らして，補強前と提案手法による補強後の平均精度を求めた．その結果，補強による改善が大きくなり（補強前 0.759，

[5] 本実験では，MEMの学習にopennlp.maxent packag（http://maxent.sourceforge.net/）を使用した．

補強後 0.784），学習データの量が少ない時に，提案手法は特に有効であることが確認できた．言い換えると，提案手法は，比較的出現頻度が低いターゲット名詞に対して，特に有効であるといえる．

表2　実験結果

ターゲット名詞	頻度	Baseline	補強前 MEM	補強後 MEM	単純補強法
advantage	570	0.604	0.921	0.933	0.835
aid	385	0.665	0.873	0.909	0.909
authority	1162	0.760	0.851	0.857	0.849
building	1114	0.803	0.842	0.848	0.829
cover	210	0.567	0.757	0.790	0.771
detail	1157	0.760	0.904	0.906	0.831
discipline	204	0.593	0.745	0.804	0.789
duty	570	0.700	0.877	0.879	0.835
football	281	0.907	0.907	0.925	0.932
gold	140	0.929	0.929	0.929	0.921
hair	448	0.902	0.902	0.908	0.908
improvement	362	0.696	0.715	0.735	0.704
necessity	83	0.566	0.843	0.831	0.831
paper	1266	0.642	0.836	0.859	0.811
reason	1163	0.824	0.893	0.885	0.857
sausage	45	0.778	0.733	0.778	0.822
sleep	107	0.776	0.897	0.925	0.888
stomach	30	0.633	0.800	0.800	0.800
study	1162	0.779	0.819	0.832	0.791
truth	264	0.720	0.777	0.761	0.773
use	1390	0.869	0.863	0.879	0.872
work	3002	0.778	0.842	0.858	0.809
worry	119	0.798	0.840	0.874	0.832
Average	662	0.741	0.842	0.857	0.835

更に，追加実験として，MCを推定する際に使用する閾値を 0.50～0.95 まで 0.05 刻みで変化させ，平均判定精度との関係を調べた[6]．その結果を図2に示す．図2の横軸は閾値，縦軸は平均判定精度を表す．

図2から，閾値の変化に伴って，平均判定精度が 1%弱変動することがわかる．閾値が 0.65 のとき，平均判定精度は最大の 0.859 となった．これは，閾値を設定しない場合の平均判定精度 0.855 を 0.4%改善したことになる．閾値 0.80～0.90 での落ち込みは，閾値が高くなり閾値を超える判定結果の数が少なくなり MC の推定精度が低下したためと考えられる．

以上の結果を踏まえ，（a）補強対象となる従来手法が判定結果と共に確信度を出力できる，(b) 学習データが豊富にあり適切な閾値を学習データから決定できる，の2つの条件が満たされる場合は，閾値を設定するべきであると結論付ける．

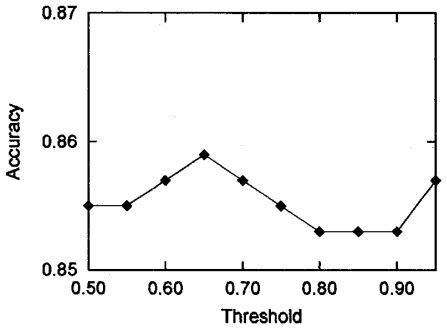

図2　閾値と判定精度の関係

5. おわりに

本稿では，One Countability per Discourse（OCpD）という英語名詞の可算性に関する仮説を導入し，可算性判定を補強する手法を提案した．実験の結果，提案手法は，可算性判定を効果的に補強できることを確認した．また，提案手法には，次の 3 点の特徴があることを確認した：（1）提案手法はどのような従来手法に対しても適用可能である，（2）閾値を設定することで，更に効果的な補強が可能である，（3）提案手法は学習データの量が少ない時に特に有効である．

参考文献

[1] T. Baldwin and F. Bond. 2003. Learning the countability of English nouns from corpus data. In *Proc. of 41st Annual Meeting of ACL*, pp.463-470.

[2] F. Bond and C. Vatikiotis-Bateson. 2002. Using an ontology to determine English countability. In *Proc.of 19th COLING*, pp.99-105.

[3] L. Burnard. 1995. *Users Reference Guide for the British National Corpus*. Oxford University Computing Services.

[4] R. Huddleston and G.K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

[5] M. Lapata and F. Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1):pp.1-31.

[6] G. Minnen, F. Bond, and A. Copestake. 2000. Memory-based Learning for Article Generation. In *Proc. of CoNLL-2000 and LLL-2000*, pp.43-48.

[7] R. Nagata, F. Masui, A. Kawai, and N. Isu.. 2004. A Method for Distinguishing Mass and Count Nouns Based on Contextual Information. In *Proc. of 4th International Symposium on Human and Artificial Intelligence Systems*, pp.516-521.

[8] R. Nagata, F. Masui, A. Kawai, and N. Isu. 2005. An unsupervised method for distinguishing mass and count noun in context. In *Proc. of 6th IWCS*, pp.213-224.

[9] D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of 33rd Annual Meeting of ACL*, pp.189-196.

[6] 可算性判定のように2クラスの分類問題では，判定結果に対応する確率は 0.5 以上になるので，0～0.45 の範囲は省略した．