英語学習者の発話における誤りの検出

齋賀豊美[†‡] 内元清貴[‡] Thepchai Supnithi[†‡※] 和泉絵美[‡#] 井佐原均[†‡#]

通信放送機構[†]
通信総合研究所自然言語グループ[‡]
Information Technology R&D Division, NECTEC, Thailand[※]
神戸大学大学院[#]

1. はじめに

我々は、日本人が最も苦手とする英語スピーキングの能力に着目し、英語の学習を支援するシステムを開発している。そのための基礎となるデータとして、これまで、日本人英語学習者の発話コーパスを作成してきた[1,2,3]。

学習支援環境では、学習者が各自の発話の誤りを克服できるように、個々の学習者に適した学習の指針が示されるようにしたい。そのためには、まず、発話の誤っている箇所を自動検出する必要があると考えている。我々がこれまでに作成してきたコーパスの一部には、現在、文法的および語彙的誤りについて 45 種類に分類した情報が人手で付与されている。今回は、その誤りの中で、近傍の文脈から判断できそうな文法・語彙の誤りを選び、検出の対象とした。

本稿では、英語学習者の発話誤りを検出する方法を提案し、コーパスを用いてどの程度誤りの検出が可能かを示す。

2. 英語学習者発話コーパスの概要

2.1 対象データ

対象は、㈱アルクが行なっているテスト SST(Standard Speaking Test)である。このテストでは、各受験者に対し 15 分程度のインタビューが行なわれる。インタビューは自己紹介に始まり、その後、受験者は、イラスト描写、ロールプレイ、ストーリーテリングの3つのタスクをこなす。このインタビューテストにより、受験者はSST独自の基準に基づいて9段階でレベル判定される。英語学習者発話コーパスは、このインタビューテストにおける試験官と受験者との会話を書き起こしたもので、総データ量は、約 1200 件(約 300 時間、100 万語以上)である。

2.2 基本情報の付与

音声から書き起こしたテキストには、繰り返しや言い直し、あいづちやフィラー、言いかけて終わっている文、いわゆる非文などの基本的な情報が付与されている。これにより、フィラーや言い直し、非文などを除いた文をコーパスから抽出することができる。下記に基本情報が付与された文の例をあげる。なお、テキストに付与する情報は、XML をベースとしたタグで表わされる。

<B><F>Well</F> <SC>when</SC> <R>why</R> why do you think you are not a good driver? <CO>You can't park or</CO>?</B>

ここで、<B></B>で囲まれた文は受験者の発話を、<F></F>はフィラー、<R></R>は繰り返し、<SC></SC>は言い直し、<CO></CO>は非文、<.></.>はポーズを表わす。

2.3 誤り情報の付与

学習者の誤りのうち、文法的、語彙的誤りについて、できる限り網羅性の高いタグセットを作成し、書き起こしテキストの一部に人手でそのタグを付与した。以降で、この誤りに関するタグをエラータグと呼ぶ。

エラータグの付与の仕方には大きく二つの方法が考えられる。誤りを含む文に対し、正しい文を考え得る限り記述するという方法と、誤りを含む文を正しい文に置き換えていく過程を説明的に記述する方法である。我々は、後者の方法を選択した。この方法には、エラータグに従って訂正した文をつなげると正しい会話が得られるという利点がある。また、文法的な誤りや語彙的な誤りを対象としたエラータグが付与しやすく、利用しやすい。同一の文に複数回表れる同種の誤りも明確に表現できるという利点もある。下記は、従属前置詞の誤りに関するエラータグ(<prp_lxc2></prp_lxc2>)を付与した例である。この例では、「to」がいらないことを示している。

<B><F>Mm</F>. <R>I</R> <SC>I <F>mm</F> <.></.> go</SC> I will <.></.> go <prp_lxc2 odr="1" crr="">to</prp_lxc2> there by train. <.></.> <CO>So</CO>. <..></..></B>

3. 誤り検出の方法

3.1 誤りのタイプ

誤りを記述方法の観点から、次の2種類に分けて考えた。ひとつは、ある単語の前に、必要な単語(列)が抜けている誤りである。これは、単語間にエラータグを挿入することによって表わす。もうひとつは、ある単語(列)が誤りであり、これを削除もしくは他の単語(列)に置換しなければならない誤りである。これは、単語(列)をエラータグで挟むことによって表わす。ここで、前者を挿入タイプの誤り、後者を置換タイプの誤りと呼ぶことにする。こ

の2タイプそれぞれについて別の検出方法を適用する。

3.2 挿入タイプの誤りの検出

挿入タイプの誤りの検出は、図1のように、デリミタを含む各単語の前に、必要な単語列が抜けていないかどうかを推定することにより行なう。同時に、誤りの種類も推定する。誤りの種類が複数ある場合は、2種類の方法で推定する。誤りの種類がNのとき、ひとつは、N種類の誤りそれぞれについて、抜けているかいないかを推定する方法である。これは、各種類の誤りに関して、各単語の前に抜けがあるかいないかの二つのラベルを付与する問題と考えることができる。もうひとつは、N種類の誤りに単語が抜けていない場合を加えたN＋1種類のどれであるかを推定する方法である。これは、各単語の前にN＋1種類のラベルのうちひとつを付与する問題と考えることができる。同じ位置に複数のエラータグが挿入されている場合は、その組み合わせを新たなエラータグとする。

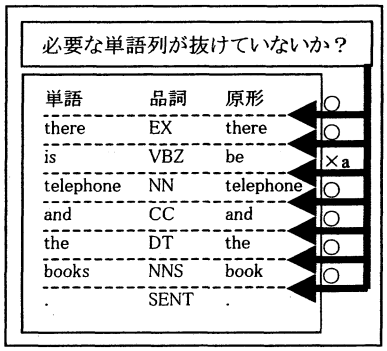

図1：挿入タイプの誤り検出方法

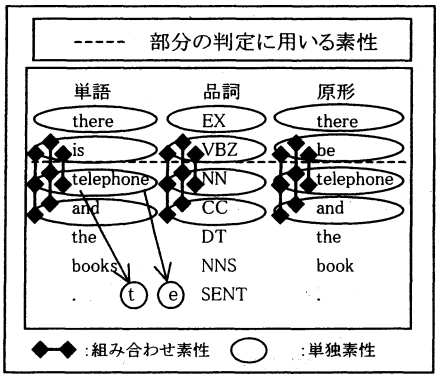

図2：挿入タイプ誤り検出に用いる素性

エラータグ推定の際に参照する情報としては、図2のように、前後2つの単語、品詞、単語の原形、そして、それらの組み合わせとして、前1つ後ろ1つ、前2つ後ろ1つ、前1つ後ろ2つの3種類、直後の単語の頭と末尾の各1文字の合計 23 種類を与えた。なお、品詞、原形はTreeTagger [4] を用いて求めた。

3.3 置換タイプの誤り検出

置換タイプの誤りの検出は、図3のように、各単語に対し、削除もしくは他の単語(列)と置換しなければいけないかどうかを推定することにより行なう。同時に、誤りの種類も推定する。誤りの種類が複数ある場合は、2種類の方法で推定する。誤りの種類がNのとき、ひとつは、N種類の誤りそれぞれについて、置換するかしないか、置換する場合は誤りの先頭の要素であるか先頭以外の要素であるかの3種類に分け、そのうちいずれであるかを推定する方法である。これは、各種類の誤りに関して、各単語に対しその単語が誤りの先頭の要素であるか先頭以外の要素であるか、もしくは誤りでないかの3種類のラベルを付与する問題と考えることができる。もうとつは、N種類の誤りを誤りの先頭の要素とそれ以外に分け、さらに削除も置換もする必要がない場合を加えた2N＋1種類を考え、対象の単語がそのうちどれであるかを推定する方法である。これは、各単語に2N＋1種類のラベルのうちひとつを付与する問題と考えることができる。これらのラベリングのスキームとしては Ramshaw らのIOBによるスキーム[7]を用いる。同じ単語に複数のエラータグが付与されている場合は最も広範囲にわたってタグが付与されているもののみを用いた。

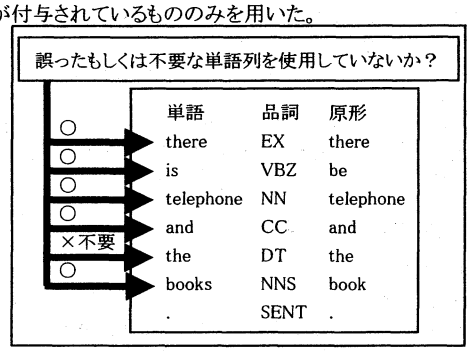

図3：置換タイプの誤り検出方法

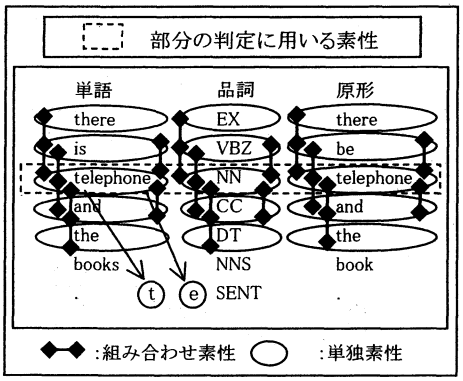

図4：挿入タイプ誤り検出に用いる素性

エラータグ推定の際に参照する情報としては、図4のように、対象の単語を含む前後2つの単語、品詞、単語の原形、そして、それらの組み合わせとして、対象の単語と前1つ後ろ1つ、対象の単語と前1つ、対象の単語と後ろ1つ、対象の単語と前2つ、対象の単語と後ろ2つの5種類、単語の頭と末尾の各1文字の合計 32 種類を与えた。

3.4 機械学習モデルの利用

エラータグの推定には、機械学習モデルのひとつである最大エントロピーモデルを用いた。このモデルでは、素性の集合を $f_j(1 \leq j \leq k)$ とするとき、式(1)を制約とし、式(2)で表される目的関数つまりエントロピーを最大にするような確率分布を求め、その確率分布にしたがって求まる各クラスの確率のうち、最も大きい確率値を持つクラスを最適なクラスとする[5,6]。

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad (1)$$
$$for\ \forall f_j\ (1 \leq j \leq k)$$
$$H(p) = -\sum_{a\in A,b\in B} p(a,b)\log(p(a,b)) \quad (2)$$

ただし、A,B はそれぞれクラスと文脈の集合を意味し、$g_j(a,b)$は文脈 b に素性 f_j があってかつクラスが a の場合 1 となりそれ以外で 0 となるとなる 2 値関数である。また $\tilde{p}(a,b)$ は、既知データでの(a,b) の出現の割合を意味する。

3．2節、3．3節に述べたラベルを上述のモデルのクラスとして推定する。ラベルの推定は文の先頭から決定的に行なう。

4. 実験

4.1. 誤りの種類

検出の対象として、表1にあげる13種類の誤りを選択した。選択の判断基準は、誤りの出現数が比較的多いこと、近傍の文脈から判断できそうであることとした。

◆名詞	名詞の単複の誤り
	名詞の語彙選択誤り
◆動詞	主語・動詞の人称・数の不一致
	時制の誤り
	動詞の補語の誤り
	動詞の語彙選択誤り
◆形容詞	形容詞の語彙選択誤り
◆副詞	副詞の語彙選択誤り
◆前置詞	従属前置詞以外の誤り
	従属前置詞の誤り
◆冠詞	冠詞の選択誤り
◆代名詞	代名詞の語彙選択誤り
◆その他	2 単語以上の語彙・表現誤り

表1：検出対象誤りの種類

4.1 タグ付きデータのみを用いた実験

現在使用可能なエラータグ付きデータは56件である。このうち、50 件(5599文)を学習に 6 件(617文)をテストに用いた。

それぞれの誤りについて3．2節、3．3節に述べた各方法で3．4節のモデルを用いて誤りを検出したところ、学習データの少ないものについては検出できなかったものが多かったが、学習データの最も多い冠詞誤りの結果は以下のようになった。

冠詞の誤り		
挿入	再現率	23/71 * 100 = 32.39(%)
タイプ	適合率	23/44 * 100 = 52.27(%)
置換	再現率	4/43 * 100 = 9.30(%)
タイプ	適合率	4/18 * 100 = 22.22(%)

また、13 種類すべてに対する検出結果は、下記の通りであった。

全種類の誤りを扱った場合		
挿入	再現率	21/ 93 * 100 = 22.58(%)
タイプ	適合率	21/ 38 * 100 = 55.26(%)
置換	再現率	5/224 * 100 = 2.23(%)
タイプ	適合率	5/ 56 * 100 = 8.93(%)

十分な精度が得られたとは言えないが、学習データの量が少なかったことが主な原因であると考えられる。そこで、正しい文を追加して精度の変化を調べた。

4.2 正しい文の追加

2．3節に述べたように、エラータグには正しい文に関する情報が盛り込まれており、その情報に基づいてタグが付与された部分を変換すると正しい文が得られる。そこで、学習データとして用いた 50 件からエラータグを元に正しい文を抽出して、誤りが全くないデータとして学習データに加えた。さらに、エラータグが付与されていないものも含めた全コーパスデータ 1202 件の内、テストに用いた 6 件を除く 1196 ファイルから試験官の発話部分を抜き出し、これも正しい文として学習データに加えた。追加した文は全部で 104925 文であった。これらの正しい文を追加して学習し、テストを行なった結果を下表にあげる。

冠詞の誤り		
挿入	再現率	8/71 * 100 = 11.27(%)
タイプ	適合率	8/11 * 100 = 72.73(%)
置換	再現率	0/43 * 100 = 0.00(%)
タイプ	適合率	0/ 1 * 100 = 0.00(%)

この表から正しい文を学習データに追加することにより再現率は低下したが適合率が高くなることが分かる。

次に、単純な規則により人為的に誤りを生成し、誤りを

含むデータとして学習データに追加することによる効果を調べた。

4.3 人為的に生成した誤りを含む文の追加

冠詞の誤りに対象を絞り、学習データを参照しながら誤りの傾向を調べた。a/an/the/冠詞なしが混同して用いられることが多いため、正しく使用されている冠詞を他の3種類に置き換えるという単純な規則で誤りを生成させた。4．2節に述べた正しい文と人為的に生成した誤りを含む文データ7578文を追加した結果は以下の通りであった。

冠詞の誤り		
挿入	再現率	24/71 * 100 = 33.80(%)
タイプ	適合率	24/30 * 100 = 80.00(%)
置換	再現率	2/43 * 100 = 4.65(%)
タイプ	適合率	2/ 9 * 100 = 22.22(%)

挿入タイプについてはタグ付きデータのみを用いた場合に比べて再現率、適合率ともに良くなっている。今回は冠詞のみを対象としたが、他の種類の誤りに対しても、人為的に誤りを生成した文を追加することにより、検出精度が向上すると期待できる。

置換タイプについては、あまり変わらない結果となった。a と the のいずれを用いるべきかについては前の文脈の情報など、より広範囲の情報を考慮する必要があると考えられるため、今回用いた素性のセットでは検出が難しかったと考えられる。今後、同じ単語が前の文脈に現れているかどうかなどの情報を考慮することにより、置換タイプの誤りに対してもどの程度検出精度が向上するかを調べたい。

5. まとめと今後の課題

本稿では、英語学習者の発話誤りを検出する方法を提案した。英語学習者発話コーパスを用いた実験では挿入漏れのタイプの誤りに対し、再現率約30%、適合率約50%であった。さらに、正しい文や人為的に誤りを生成した文を追加することにより、再現率は保ったままで、適合率が80%まで向上することが分かった。

誤り情報の付与が終了しているデータはコーパス全体の0．5%と少ないため、今後、残りのデータに関しても誤り情報を付与していきたい。その際には、本稿で提案した方法がタグ付けの支援に使えると考えている。

また、本実験を行なうにあたって、コーパスのタグの誤りも発見された。本稿で提案した方法は、このようなコーパスの誤りの検出に生かせると考えている。

Minnenらは名詞句に冠詞を付与するべきかどうか、付与する場合にはどの冠詞を付与するべきかを学習モデルを用いて推定する方法を提案した[8]。彼らが対象としたのは誤りがほとんどない新聞記事であった。また、推定の対象を名詞句に限定している。一方、我々が対象としたデータは冠詞以外の様々な誤りを多く含み、名詞句以外にも誤りが生じ得る。そのため、我々の方法では、対象を名詞句に限定せず、すべての単語に関して誤りと関連するかどうかを推定する。Minnen らの方法はより豊富な情報を利用しており、今後、彼らの用いた情報も我々のモデルに採り入れ、冠詞以外の検出にも有効であるかを検証したい。

謝辞

本研究において、データの収集と利用に際して、平野琢也(㈱アルク)、金子恵美子(㈱アルク)、金子朝子(昭和女子大学)、投野由紀夫(明海大学)、成田真澄(㈱リコー)の各氏の協力が不可欠でした。ここに感謝いたします。

参考文献

[1] 井佐原均、投野由紀夫、平野琢也：日本人学習者のレベル別英語発話コーパスの作成、言語処理学会第6回年次大会発表論文集、pp.32-34、2000

[2] 齋賀豊美、井佐原均：日本人学習者の英語発話コーパスの作成―概要と開発環境―、言語処理学会第7回年次大会発表論文集、pp.541-544、2001

[3] 和泉絵美、井佐原均：英語学習者発話コーパスにおける誤り分析―エラータグとその応用―、言語処理学会第7回年次大会発表論文集、pp.545-548、2001

[4] Helmut Schmid Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing. pp.44-49, 1994

[5] Jaynes, E. T. "Information Theory and Statistical Mechanics" Physical Review, 106, 620-630, 1957

[6] Jaynes E. T. "Where do we Stand on Maximum Entropy?." In Levine, R.D and Tribus, M.(Eds.), The Maximum Entropy Formalism, p15. M.I.T Press, 1979

[7] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In Proceedings of the Third ACL Workshop on Very Large Corpora, pp 82-94, 1995

[8] Guido Minnen, Francis Bond, Ann Copestake, "Memory-based Learning for Article Generation" In Proceedings of CoNLL-2000 and LLL-2000, pages 43-48, Lisbon, Portugal, 2000