# チャンツに基づいた英語教材生成のためのストレス位置自動判定

水本　智也† 永田　亮†† 船越孝太郎†††

† 甲南大学理工学部 †† 甲南大学知能情報学部 ††† ホンダ・リサーチ・インスティチュート・ジャパン

E-mail: †ss664084@ center.konan-u.ac.jp., ††rnagata@ konan-u.ac.jp., †††funakoshi@ jp.honda-ri.com.

## 1. はじめに

ジャズ・チャンツ（以下，単にチャンツと表記）は英語の話し言葉のリズムと伝統的なアメリカンジャズのリズムを結びつけるリズム表現法である [4]．チャンツの主な利用方法の一つに，チャンツに基づいた英語学習法がある．この学習法では，英語のストレスに合わせて英語を音読することで学習を行う．その結果，英語の自然なリズムとイントネーションを身につけることができる．また，音読とともに絵や動作（手拍子，踊りなど）を用いることで，楽しみながら語彙と基本的な表現も覚えることができる．このような特徴のため，チャンツに基づいた学習法は特に小学校における英語学習に適している（小学校英語活動におけるチャンツの利用については文献 [2], [3] が詳しい）．なぜならば，小学校では英語の音声と基本的な表現に慣れ親しむことを目標の一つとしており [5]，中学，高校で使用される文字中心の教材が適さない小学生でも，音声，絵，動作で楽しみながら学習できるためである．

小学生を対象としたチャンツに基づいた学習法では，教師は英文にストレス位置を示した補助教材（以下，チャンツ教材と表記）を用いることが多い．例として，チャンツ教材 “Frank, Hank” を以下に示す．

```
  *     *     *          *
Frank, Hank, walk to the bank.

 *    *   *          *
Jill, Phil, run up the hill.

 *     *     *          *
Mike, Spike, ride your bike.

*      *     *          *
Andy, Sandy, eat your candy.
```

アスタリスク（以下，ストレス記号と表記）がストレスのある単語と音節を示している．チャンツに基づいた学習法はストレス記号に合わせて英語を音読することで学習を行う（多くの場合，ストレス記号に合わせて手拍子などの動作をつける）．英語母語話者でない日本人にとって，ストレス記号は非常に重要な役割を果たす．なぜならば，ストレス記号がない場合，ストレスの位置を間違ってチャンツを利用してしまうことが多いためである．典型例として

```
  *     *     *   *   *   *
Frank, Hank, walk to the bank.
```

のように全ての単語にストレスをつけてしまう間違いが挙げられる．実際，有馬 [2] は，小学校で英語を担当する教師でも同様の間違いをすると報告している．以上の背景を考慮すると，日本でチャンツに基づいた学習法を利用するためには，ストレス位置が明示されているチャンツ教材が必要となる．

しかしながら，現状では，ストレス位置が明示されている教材は非常に少ない．そもそも，英語母語話者はストレス位置がわかるため，（英語母語話者にとっては）ストレス記号は必要がないことが大きな理由である．したがって，非母語話者がチャンツに基づいた学習法を用いるためには，なんらかの方法でストレス位置を判定しなければならない．

そこで，本研究ではチャンツ教材自動生成を目指して，ストレス位置を自動で判定する手法を提案する．ストレス位置が自動判定できれば，教師は任意の英文を用いてチャンツに基づいた学習法を実践できる．更に，ストレス位置に合わせて手拍子，動作，発声をし，英語を教授するチャンツロボットへの応用などにも繋がる．このような応用を念頭に置き，本研究では 3 種類の基礎的なストレス位置自動判定法を検討する．第一に，ベースラインとして内容語にストレスを付与するというルールベース手法を検討する，加えて Hidden Markov Model (HMM) [1] を用いた 2 種類の手法の検討を行う．

以下，2. で，提案手法について説明を行う．3. で，実験とその結果について述べる．4. で，実験結果を考察する．

## 2. 提案手法

### 2.1 対象とするストレス位置

チャンツ教材のストレス位置はどの単語のどの音節に付くかまで明示される．例えば，

```
    *       *          *          *
He likes apples. She likes bananas.
```

のように，He，apples，She は 1 音節目にストレスが付き，bananas は 2 音節目に付くことが明示されている．

しかしながら，提案手法のストレス位置判定では，単語の何音節目に付くかという判定は行わず，英文中の各単語のストレスの有無の判定のみを行う．これは，何音節目にストレスが付くかということは単純に辞書引きで解決できるからである(注 1)．また，なかには 1 単語に複数のストレスがつく

(注 1)：例えば，音声合成ソフトウエア festival の辞書などが利用可能である．

という可能性もある．しかしながら，そのようなケースは稀であるため，本研究では対象としない．実際，3. の実験に用いたチャンツ教材の中には，そのようなケースは存在しなかった．

### 2.2 ルールベース手法

ストレス位置判定を行うには，どのような単語にストレスが付きやすいのかを考える必要がある．そこで上述のチャンツ教材の例を見てみると，

```
  *     *     *          *
Frank, Hank, walk to the bank.
```

のように名詞や動詞といった内容語に付いていることがわかる．このことからストレスの付きやすい単語は名詞，動詞，形容詞など内容語であると推測できる．

このことを考慮して，全ての内容語をストレス位置と判定するルールベース手法を提案する．ルールベース手法では，まず，入力英文の各単語に対して既存の品詞タガーで品詞を付与する．その後，付与された品詞から内容語を特定しストレス位置とする．ただし，内容語は，名詞（普通名詞，固有名詞），動詞（動名詞，現在分詞，過去分詞を含む），形容詞，副詞，数詞とする．

### 2.3 HMM を用いた手法

内容語にストレスがつきやすいという傾向はあるものの，ルールベース手法では上手くいかない場合がある．言い換えると，内容語でもストレスの付かない場合や，機能語でもストレスが付く場合があるということである．例えば，

```
  *          *     *              *
Cats love the sun. Flowers love the rain.
```

では，動詞 love にはストレスはついていない．一方，

```
   *         *      *       *
Where's my hat? It's on the door.
```

では，前置詞 on にストレスがついている．これらの例からわかるようにストレス位置は前後の単語とストレス，関係する音節数など複数の要因が関係して決まる．例えば，上述最初の例の二つ目の love については直前の二つの単語にストレスが付いているためだと考えられる．

そこで，この問題を解決する手段として，ストレス位置判定問題を品詞タグ付け問題として解くことを考える．品詞タグ付けでは当該の単語とその前後の情報を使って品詞を決定することが多い．そのため，品詞タグ付け問題とストレス位置判定問題は類似した問題として捉えることができる．これを説明するために例文

```
Frank, Hank, walk to the bank.
```

を考える．この英文にストレスをつけると，

```
  *     *     *          *
Frank, Hank, walk to the bank.
```

のようになる．これは表現を変えると，

```
Frank/S ,/N Hank/S ,/N walk/S to/N the/N bank/S
./N
```

のように表すこともできる(注 2)．ただし，S がストレス位置であることを表し，N が非ストレス位置であることを表す（以下，S がストレス位置を表し，N が非ストレス位置であることを表す）．一方で，同じ文に対して品詞を付与すると

```
Frank/NN ,/, Hank/NN ,/, walk/VB to/PP the/DT
bank/NN ./.
```

のように表すことができる．ただし，NN が名詞，VB が動詞，PP が前置詞，DT が冠詞を表す．これらの例より，ストレス位置判定も品詞タグ付けも，単語のラベル推定問題として捉えることができることがわかる．よって，ストレス位置判定問題は品詞タグ付け問題として解くことができる．

そこで本研究では品詞タグ付けによく用いられている HMM を利用したストレス位置判定手法を提案する．HMM を用いた手法として，単語を入力とする HMM（以下，HMM-Word と表記）と品詞を入力とする HMM（以下，HMM-POS と表記）の 2 種類の手法を提案する．

HMM-Word は入力を英文とし，出力を S または N とする．まず，入力文をトークンに分割する．例えば，入力文を

```
Frank, Hank, walk to the bank.
```

とすると，

```
Frank , Hank , walk to the bank .
```

というトークンに分割される．これを学習済みの HMM に入力して，S，N のタグを付与する．例えば，上述のトークンを HMM に入力すると

```
Frank/S ,/N Hank/S ,/N walk/S to/N
the/N bank/S ./N
```

という出力が得られる．HMM の学習は人手によりストレスが付与されたチャンツ教材を用いて行う．

HMM-POS は五つのステップから成る．入出力は HMM-Word と同じで，入力を英文，出力を S または N とする．第 1 ステップでは HMM-Word のときと同様に入力文をトークンに分割する．第 2 ステップではトークンに分割された英文に既存の品詞タガーを用いて品詞タグを付ける．例えば，HMM-Word のときと同じ例文では，

```
Frank/NN ,/, Hank/NN ,/, walk/VB
to/PP the/DT bank/NN ./.
```

---

（注 2）：通常は ',' や '.' はストレス位置とは関係ないが，品詞タグ付け問題との比較のため，このように表記する．

のようになる．第 3 ステップとして，英文から単語を取り除き，品詞列にする．その結果，

NN , NN , VB PP DT NN .

のような品詞列が得られる．第 4 ステップでは，品詞列に対して，学習済みの HMM を用いて S，N のタグを付与する．上の例文では，

NN/S ,/N NN/S ,/N VB/S PP/N DT/N
NN/S ./N

となる．HMM の学習は，人手によりストレスが付与されたチャンツ教材を用いて行う．ただし，単語列を既存の品詞タガーで品詞列に変換してから学習を行う．最後に，第 5 ステップとして，得られたストレス位置の情報と最初の英文をマージすることで，最終的な出力を得る．上の例からは，

Frank/S ,/N Hank/S ,/N walk/S to/N
the/N bank/S ./N

を得る．

## 3. 実　　験

### 3.1 実験条件

各手法の性能を評価するため実験を行った．実験にはテキスト [4] より集めた 17 のチャンツ教材（153 文，627 単語）を用いた．比較のため，全ての単語にストレスを付与するベースラインを用意した．そのベースラインを含め，実験は HMM-Word，HMM-POS，ルールベースの 4 種類の手法の性能を比較した．

実験には，trigram ベースの HMM を用いた．ただし，bigram，unigram で線形補完を行った．未知語については，語彙生成確率として S と N の生起確率を割り当てた．S と N の生起確率は学習データから推定した．

品詞タガーは独自に開発したものを使用した(注 3)．品詞タグセットは Penn TreeBank を基にして 44 個の品詞タグを用いた．

評価尺度として，recall，precision，$F$-measure，accuracy の 4 種類を用いた．それぞれ

$$R = \frac{\text{正しく } S \text{ と判定した数}}{S \text{ の数}}, \quad (1)$$

$$P = \frac{\text{正しく } S \text{ と判定した数}}{S \text{ と判定した数}}, \quad (2)$$

$$F = \frac{2PR}{P+R}, \quad (3)$$

$$A = \frac{\text{正しく判定した数}}{\text{全単語数}} \quad (4)$$

で定義した．それぞれの評価尺度の値は，leave-one-out を用いて求めた．取り出す単位は 1 チャンツ教材とした．

(注 3)：品詞解析および句解析ツールを http://nlp.ii.konan-u.ac.jp/tools.html にて公開している．

### 3.2 実験結果

実験結果を表 1 に示す．表 1 から HMM を用いると，precision を下げることなしに，recall を改善できることがわかる．特に，HMM-POS では，recall も precision も改善できていることがわかる．

表 1：　実験結果

| Method | $R$ | $P$ | $F$ | $A$ |
|---|---|---|---|---|
| ベースライン | **1.00** | 0.344 | 0.507 | 0.344 |
| ルールベース | 0.685 | 0.750 | 0.716 | 0.659 |
| HMM-Word | 0.739 | 0.746 | 0.742 | 0.678 |
| HMM-POS | 0.832 | **0.777** | **0.804** | **0.745** |

## 4. 考　　察

3. の実験で，単語を入力とする HMM-Word より品詞を入力とする HMM-POS のほうが性能がよいことが明らかになった．この一つの理由として，HMM-Word では，単語そのものを素性にしているため未知語となるケースが多かったことが挙げられる．特に，固有名詞は問題となる．HMM-Word は未知語の影響を大きく受けたため，HMM-POS より性能が低くなったと分析できる．未知語に起因する失敗の例として，

*　　*　　*　　*
Gail likes sail. But Lee loves to ski.

があった（これは HMM-Word の実験結果を表す）．この例では，固有名詞 Lee にストレスが付かず，loves にストレスが付いてしまっている．これは Lee が未知語である一方で，loves が既知語かつ学習データでストレスが付いていたことが原因である．この問題は，学習データを増やすことである程度改善されると考えられる．なお，正解は

*　　*　　*　　*
Gail likes sail. But Lee loves to ski.

である．

一方で，HMM-Word のほうが優れている点もあることが明らかになった．HMM-POS では，品詞の分類が大きすぎて失敗することがあった．例えば，HMM-POS では，you と it は代名詞として同一視されるが，you のほうがストレスが付きやすい傾向にある．例えば，

*　　*　　*　　*
Who told you? You told Tim.

*　　*　　*　　*
Where's my sweater? It's in the drawer.

のように，you にはストレスが付いているが，it にはストレスが付かないことがわかる，この問題を解決する方法として，

ある特定の単語（例えば，機能語）に関しては，1 単語 1 品詞を割り当てることが考えられる．これにより上述の例では you と it は区別されることになり，性能改善が期待できる．

逆に，品詞の分類が細かいので，まとめてしまったほうがよいものもある．例えば，今回実験で使用した品詞タガーでは，名詞は 4 種類（NN:単数名詞; NNS:複数名詞; NNP:単数固有名詞; NNPS:複数固有名詞）に分類される．しかしながら，ストレス位置の決定に，単数／複数，普通／固有の情報は寄与しないと予想されるので，単に名詞として扱うべきである．同様に，動詞の現在／過去なども一つの品詞にまとめるべきである．以上をまとめると，HMM-POS は，ストレス位置判定に有効であるが，品詞の分類を適切な粒度に設定する必要があるといえる．

また，分析の結果より，遠隔依存関係を考慮するとよいことも示唆された．例えば，上述した，

*　　*　　*　　*
Where's my hat? It's on the door.

の on にストレスが付くのは Where という疑問詞と呼応関係にあるからだと考えられる．この問題は，Conditinal Random Fields (CRF) を用いて遠隔依存関係を素性として取り入れることで解決できる可能性がある．

更に，実験結果の分析により，チャンツにはストレスを決定するうえで重要な制約があることが明らかになった．その制約とは一つのチャンツの中では，ストレス数が 4 の倍数になるということである．1. で例として挙げた “Frank, Hank” を見るとストレスの数は 16 となり 4 の倍数になっていることがわかる．提案手法では，この制約を明らかに考慮していない．その結果，HMM-Word，HMM-POS ともにストレス数が 4 の倍数になっていない判定結果が多くなり，失敗の原因となった．更なる改善のためには，この制約を考慮することが必須である．

## 5. おわりに

チャンツ教材自動生成のため，ストレス位置自動判定手法の提案を行った．ストレス位置判定問題を品詞タグ付け問題として解くことを考え，HMM を用いた 2 種類の手法を提案した．実験の結果，提案手法の一つである HMM-POS は $F = 0.804$，$A = 0.745$ を達成した．実験結果の分析により，改善のためのアイデアとして，(1) 特定の品詞（例えば，名詞の単数複数）を一つの品詞にまとめること，(2) 逆に特定の単語（機能語など）には 1 単語 1 品詞を割り当てること，(3) 遠隔依存関係を考慮すること，(4) ストレス数に関する制約を考慮することを得た．今後は性能向上のため，これらの改善のためのアイデアを検討していく予定である．また，今回は HMM を用いた手法を提案したが，CRF を用いることも考えている．

## 謝　辞

チャンツに関する有益な助言をいただいた有馬千香子先生に感謝します．

## 参考文献

[1] J. Allen, Natural Language Understanding (2nd Edition), The Benjamin/Cummings Publishing Company, 1995.
[2] 有馬千香子, “小学校英語活動における教師のチャンツ使用に関する研究,” 兵庫教育大学修士論文, 2008.
[3] 有馬千香子，佐藤真, “小学校英語活動におけるチャンツに関する一考察 — 我が国の英語教育におけるチャンツの変遷から —,” 日本基礎教育学会紀要, vol.13, pp.7–14, 2008.
[4] C. Graham, Creating Chants and Songs, OXFORD, 2006.
[5] 文部科学省, 小学校学習指導要領, 2008.