音素から形態素へ
—Harrisの仮説の英語、中国語コーパスを用いた検証—

田中久美子　　靳志輝
東京大学情報理工学系研究科
{kumiko, zhihui}@i.u-tokyo.ac.jp

概 要

Zellig S. Harris は 1955 年の「From phoneme to Morpheme」の論文の中で、音素列における「後続音素の種類数」の変化の極大点が形態素や単語の境界と一致するとの仮説を示している。この仮説は、言語学上は二重分節における音素列と形態素列の関係を示す興味深いものであると共に、言語工学上も unsupervised segmentation を示唆する有用なものである。 本稿では、この仮説を英語と中国語のコーパスを用いて検証する。

1　はじめに

Harris は 1955 年の論文「From Phoneme to Morpheme」の中で、「後続音素の種類数」を観察することにより、形態素や単語の境界点がわかると述べている [4]。たとえば、/hiyzklevər/ (“He is clever”) の分節点を調べるには、/h/から始まる発話 (たとえば “hot coffee” や “How are you” など)、つぎに/hi/から始まる発話 (“hit it”, “he is good”)、続いて/hiy/, /hiyz/, /hiyzk/などと接頭部分の長さを徐々に伸ばし、直後に来る音素の種類を調べる。すると、その種類数は接頭部分の長さに応じて上下を繰り返し、極大点が単語あるいは形態素境界となるというのである。

1955 年当時はコーパスはまだ整備されていなかったので、Harris は各音列から始まる文を実際に人に答えてもらい、音列の直後の音素を人手で解析することで検証した。そして、単語境界は “accords very well” 形態素境界は “accords quite well” と結論付けている。この実験法は人間の認知上の分節境界を調べる上で有効な方法の一つではあるが、実験規模が限定されてしまうという問題がある。昨今はコーパスが十分に整備されているので、コーパスを用いてこの仮説を検証することができる。

Harris の仮説の意義の一つは Martinet の二重分節 [7]—言語は音素列と形態素列に二重に分節されるとの説—における二層の関係を示し、有意味な単位がどのように発生しうるのかの一つのメカニズムを示唆する点にある。同時に、Harris の仮説が大規模データ上で成り立つならば、意味単位への分節をうながす構造が言語データに内在することも示唆する。さらに、意味単位は形態素のみならず、単語、句と、小さな単位からより大きなものがあるが、Harris の仮説がそれらの分節過程一般のメカニズムをある程度説明するものである可能性もある。実際、Harris の仮説を漢字列に適用することで、漢字列から単語境界を同定することはほぼ９割近くの精度でできることが示されているし [9][6]、単語列から同様の方法で句境界を得ることもできることも [3] や [10] において示されている。

本稿では、音素から形態素あるいは単語へ、Harris の仮説をコーパス上で検証する。本稿の詳細は [11] に記述されている。

2　Harrisの仮説の若干の改定

実際に検証に入る前に、Harris の仮説を現代風にアレンジする。Harris は「後続音素の種類数」を用いたが、別の指標の方がよいかもしれないと述べている。Harris の論文からは「複雑さ」の指標として「要素種類数」を用いていることを読み取ることができるため、別の指標の候補としてエントロピーを考えることができる。実際、1955 年当時に複雑度の指標としてエントロピーが用いられなかったことは無理もないが、現代ではそれはもちろん可能である。また、他の類似する境界判定においては要素種類数とエントロピーで、

判定される境界がかなり共通することも示されている[10]。そこで、本稿では、指標として「後続音素のエントロピー」を用いることにする。

具体的には、ある音素集合 χ に対して、その要素 $x \in \chi$ の生起確率が $p(x)$ として与えられるとき、ある具体的な長さ n の χ の要素の列 x_n がすでに起こったという条件下で、後続する χ の要素 x の生起確率を条件付き確率 $p(x|x_n)$ として記述すると、x_n に後続する要素のエントロピーは、

$$H(X|x_n) = -\sum_{x \in \chi} p(X = x|x_n) \log p(X = x|x_n), \quad (1)$$

で与えられる。この $H(X|x_n)$ が本稿で計測する値である。そこでより簡単に $h(x_n)$ と記述し、本稿の残りの部分で用いる。

Harris の論文では仮説は端的に与えられているに過ぎない。しかし、$h(x_n)$ を用いて、言語の大域的な特性に関連させて彼の仮説を根拠付けることができる。特定の x_n ではなく、長さ n の任意の文字列が与えられた場合のエントロピー $H(X|X_n)$ を考える。

$$H(X|X_n) = -\sum_{x_n \in \chi_n} p(X_n = x_n) h(X_n = x_n) \quad (2)$$

$H(X|X_n)$ は n が長くなれば、小さくなり、予測が簡単になるという性質がある [1]。つまり、どのような長さ $(n+1)$ の要素列 y_{n+1} をとってきても、平均的には $h(y_{n+1}) < h(x_n)$ が成り立つ。すると、y_{n+1} が、x_n を prefix にもつ x_{n+1} である場合は、同じ文脈をより多く与えられているのだから、ますます $h(x_{n+1}) < h(x_n)$ が成り立ちそうである。この対偶をとると、$h(x_{n+1}) \geq h(x_n)$ となるところは、文脈が切れ、分節境界となると考えることができるのである。

また、Harris は種類数の極大点を分節境界としているが、以上の論は、$h(x_n)$ の極大点ではなく、むしろ増大点の方が境界として妥当であることも示唆する。実際、増大点を境界とみなすことにより、極大点をとる場合に見落としてしまう境界も捉えることができるため、工学的応用上もより適切である。

まとめると、本稿で検証する仮説は

> x_n が x_{n+1} の prefix を成す (つまり、x_{n+1} は x_n の後に何か別の要素が後続した部分列である) とき、$h(x_{n+1}) > h(x_n)$ が成り立つならば、n は境界である。

となる。

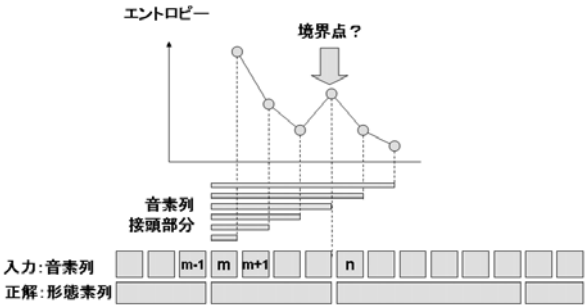

図 1: 検証方法

3 検証実験方法

検証に際しては、形態素境界、単語境界が知られているテストデータ、また、エントロピーを計算するための大規模な学習データが必要である。さらに、これらのデータは共に音素列のデータである必要がある。英語、中国語のデータの具体的な内容については、§4.1 および §4.2 にて述べる。

テストデータは、句読点単位で事前に文の断片に切っておき、断片ごとに以下の単純な処理を行って境界点を得た上で、精度を計測する。$x_{m,n}$ を、断片 x の m から n 番目の音素の直前まで示すものとし、以下の処理を各断片について行う（図 1 参照)。

1. $m := 0$, $n := m + 1$.
2. $h(x_{m,n})$ を学習データで計測する。
3. その値を $h(x_{m,n-1})$ と比較。もし $h(x_{m,n}) - h(x_{m,n-1})$ が決められた閾値より大きい場合には n を境界点として出力。
4. $n > m+maxlen$ ならば、$m := m+1$, $n := m+1$。それ以外は $n :=n + 1$。2 へ戻る。

閾値は、適宜変更して精度を計測する。また、断片の全部分列について後続する音素のエントロピーを計測することを避けるために $maxlen$ を設けてあり、言語別にエントロピーが 0 となってしまう長さとした (後述)。

この処理は、前から後方へと走らせる後ろ向き処理が通常処理であるが、後方から前へと走らせて、ある音素列の直前の音素の複雑さを計測にすることにより、境界点を推測する前向き処理を行うこともできる。本稿では、前向き・後ろ向きの両方について、各 m について得られる推定境界の集合和をとり、正解と比較する。

検証方法は、N_{test} を検出された境界数、N_{true} をテストデータに事前に付与されている正解境界数、

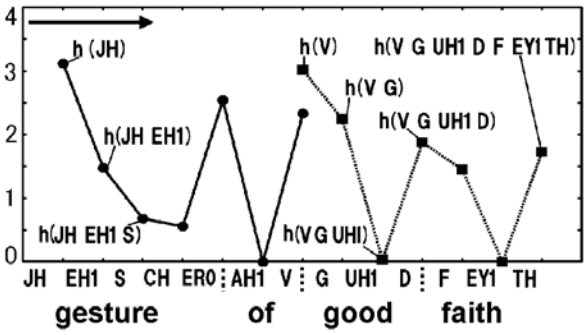

図 2: 後続文字列のエントロピーの変化の例

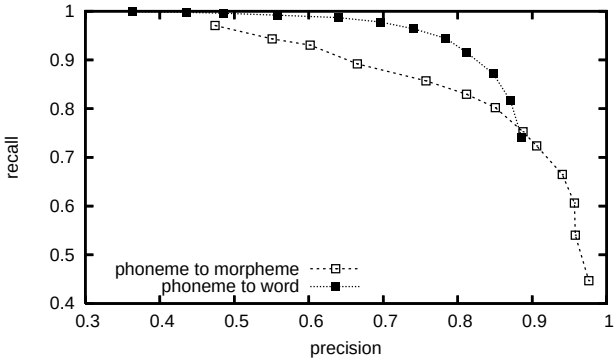

図 3: 英語における検証

$N_{correct}$ を検出された境界のうちの正解数として、Precision=$\frac{N_{correct}}{N_{test}}$, Recall=$\frac{N_{correct}}{N_{true}}$ として 算出する。

4 検証結果

4.1 英語

データは WSJ 100MB を利用し、うち 1MB をテストデータとした。データはすべて CMU Pronouncing Dictionary [2] を用いて、音素列に変換した。たとえば、

he is clever
→ HH IY1 IH1 Z K L EH1 V ER0

などと変換される。音素列への変換後は当然のことながら単語間のスペースは存在しない。単語が CMU 辞書に含まれない場合には、その断片全体を削除したため、用意したデータはテスト・学習共に実際は 23.4%分削減されている。処理上のパラメータである *maxlen* は、10 とした。

正解は、単語境界については英語の元テキスト中にある単語境界をそのまま利用した。形態素境界については、PC-KIMMO[8] を用いて形態素に分解し、さらに人手で正誤を確認して修正をして、正解とした。この作業は手作業であるため、形態素に関する正解データは 50 断片程度の非常に小さなものにとどまった。

例文 gesture of good faith という英語の境界を実際に推定したものを図 2 に示す。この入力の音素列は JH EH1 S CH ER0 AH1 V G UH1 D F EY1 TH であり、この図は、後ろ向き処理、すなわち後続音素のエントロピー変化を示している。二つの線があり、そのうち左方のものは、音素 JH から徐々に接頭部分列を後方に伸ばしていった場合の変化を表す。すなわち、h(JH), h(JH EH1) ... h(JH EH1 S CH ER0 AH1 V) を表す。ER0 と AH1 の間と、V の後に増大点が伺え、正しく単語境界を示している。右方の線は同様に h(V) ... h(V G UH1 D F EY1 TH) をプロットしたものであり、D と F の間と、TH の後に増大点があり、こちらも正しい結果である。実際、この例では、前向き・後ろ向き処理の両方において各音素から始めてすべての増大点を得て集合和をとると、全単語境界が正しく得られる。

さらに大規模な結果として、1MB の全データに関し §3 の閾値を 0.0 から 2.4 まで 0.2 刻みで変化させてそのときどきの Precision/Recall を算出した結果を図 3 に示す。単語境界の精度と形態素境界の精度を現す線がそれぞれ示されている。

Harris が述べたように、単語境界の方が高い精度を示しており、特に閾値が 1.6 のときに F-score は最大値 86.1% (precision=81.2%, recall=91.5%) となった。precision の値から、18.8%は単語境界ではない場所が境界として検出されていることになるが、これには形態素境界を含む。形態素境界の F-score は同じ閾値において 80.4% (precision=90.5%, recall=72.3%) であったから、この仮説を用いると、1 割程度は誤った境界が混入することになる。形態素境界は単語境界を含むことから、前者の precision は後者の precision よりも必ず高い値になるにもかかわらず形態素境界の F-score が低いということは、形態素境界に見逃しがかなりあることを示唆する。総合するに、英語は音素から直接単語列が分節される構造を持っていることも暗示している。

4.2 中国語

データは、北京大学コーパス 200 MB を用いた [5]。このうち、7.8 Mbyte は、人手で単語境界が付されてい

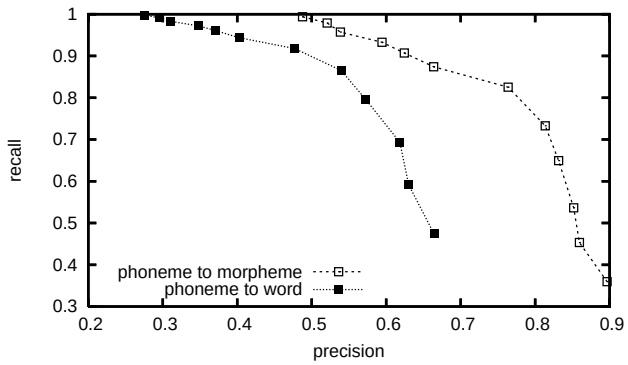

図 4: 中国語における検証

るので、これをテストデータとして用いた。一方、形態素境界は各漢字が一形態素を表すものとして近似した。中国語のテキストは pinyin にすべて変換し、pinyin 表記を音の観点から整理することにより、一音素一記号となるように音素列へと変換した。*maxlen* は 12 とした。

図 4 に、precision/recall の図を示した。英語と同様、閾値を 0.0 から 2.4 まで 0.2 きざみで変化させて、プロットしたものである。今回は、形態素境界の方が F-score は高く 79.4% (precision=76.7%, recall=82.4%, 閾値 1.2) であった。これは、英語の 80.4%と拮抗する。とはいえ、precision の値から、各漢字ごとに形態素をなすとみなす手法があまりよい近似でない可能性もある。

単語境界について, F-score が 66.9%(precision=54.7%, recall=86.1%, 閾値 1.6) と大幅に精度が下がり、この仮説を用いて音素から単語境界が定まるとは結論付けにくい結果となった。

英語との差を考察するに、中国語においては、単語と漢字のレベルで、分節階層が一段違うことを示唆する。これをさらに根拠付けるものとして、入力データを漢字列とし、単語境界を §3 に示した方法で得てみると、F-score は 83%と高い [6]。すなわち、中国語における単語は、まず音素列から漢字列へ分節し、その後漢字列から単語列へと、二つの分節過程を経て生成されている可能性がある。

5 結論

Harris の仮説を現代風にアレンジした上で、それがコーパス上では 8 割程度の F-score で成り立つことを示した。本検証からは、英語と中国語で意味単位の分節のされ方が異なることが観察された。すなわち、英語は音素列から直接単語が分節されるのに対し、中国語では音素列からまず形態素が分節され、形態素列から単語が分節されている可能性があることが示された。

今後は本検証で得られた言語差の知見の信憑性を探り、表記の差の影響などと絡めてその差の原因を探ってみたい。

参考文献

[1] T.C. Bell, J.G. Cleary, and I.H. Witten. *Text Compression*. Prentice Hall, 1990.

[2] Carnegie Mellon University. CMU pronouncing dictionary version 0.6, 2006. visited 2006, `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[3] T.K. Frantzi and S. Ananiadou. Extracting nested collocations. pages 41–46, 1996.

[4] Z.S. Harris. From phoneme to morpheme. *Language*, pages 190–222, 1955.

[5] ICL. People's daily corpus,Beijing university, 1999. `http://www.icl.pku.edu.cn/icl_res/`.

[6] Z. Jin and K. Tanaka-Ishii. Unsupervised segmentation of Chinese text by use of braching entropy. In *COLLING/ACL*, 2006.

[7] André Martinet. *Éléments de linguistique générale*. Colin, 1960.

[8] SIL. PC-KIMMO version 2, a morphologial parser, 1995. `http://www.sil.org/pckimmo/`.

[9] K. Tanaka-Ishii. Entropy as an indicator of context boundaries —an experiment using a web search engine —. In *IJCNLP*, pages 93–105, 2005.

[10] K. Tanaka-Ishii and Y. Ishii. Multilingual phrase-based concordance generation in real-time. *Information Retrieval*, 2007. Springer. Accepted, in press, to appear in 2007.

[11] K. Tanaka-Ishii and Z. Jin. From phoneme to morpheme: Another verification using a corpus. In *International Conference on the Computer Processing of Oriental Languages*, pages 234–244, 2006.