

確率的言語モデルに基づく効率的読書のモデル化

二階堂 学¹ 藤井 敦¹

概要：人間の読書に関する既存のモデルは、重要でない単語や語頭から語末が予測できる単語を読み飛ばす仕組みの解明に有効である。本研究の目的は、確率的言語モデルを応用して、与えられた文脈における単語の予測精度を向上させ、結果として読書を効率化させる点にある。

キーワード：確率的言語モデル，効率的読書

Modeling Efficient Reading based on Probabilistic Language Models

MANABU NIKAIKIDOU¹ ATSUSHII FUJII¹

1. はじめに

読書とは最も基本的な情報取得の手段である。読書には趣味で読む書籍、報道記事、業務に関連する書類などがある。さらに近年は専門性や職種を問わずに不特定多数のユーザーが発信する有用な解説記事などもある。そこで、読書に要する時間を短縮することで人生の時間をより有意義に過ごすことができる。

読書をモデル化したものとしては E-Z Reader[3] や Mr. Chips[1], [2] などがある。E-Z Reader は人間の読書を正確にモデル化することを目的としており、Mr. Chips は効率的なモデル化することを目的としている。読書に要する時間の短縮という観点においては効率的な読書のモデル化は、どうやれば理解度を落とさずに読めるか、どういった文章なら理解度を落とさずに早く読めるかを知る重要な一歩となるため重要だ。

しかし Mr. Chips では、単語認識において誤認を全く認めていないことや、文脈情報を使っていないことなど、読み方や利用する情報を変えることで読書をさらに効率化する余地が残されている。本研究では Mr. Chips これらの点に対して変更を加えることで、より効率的な読書をモデル化することを目的とする。

2. 関連研究

2.1 RSVP

効率的な読書の支援手法としては、文章を表示するインターフェースを変えるという手法もある。インターフェースを変える既存の手法としては、文章中の単語を1つずつ逐次的に画面に表示していく RSVP(rapid serial visual presentation) という手法を用いたもの [4] がある。

この手法は文章を最後まで読む速度は向上しているが、表示単語が指導的に遷移してしまうため、読み返しが難しい、長い文章は集中力が続かない、理解度が落ちるなどの問題がある。

2.2 Mr. Chips

Mr. Chips は、人間が利用可能な情報や実行可能な動作という制限の中でできるだけ効率的に読書を完了するということで、人間の読書を評価するための基準を提供することや、人間の行動の原因を解明する手助けとなることを目的として作られた。Mr. Chips では文章中の単語を最初から順に認識していく。その過程において効率的なサッカードを行うために Mr. Chips は視覚と語彙の情報を利用して、単語を認識するのに最適なサッカードを計画している。

この計画には、視覚により得られる情報、語彙、サッカード計画時の計算アルゴリズムを定義する必要がある。

¹ 東京工業大学
Tokyo Institute of Technology

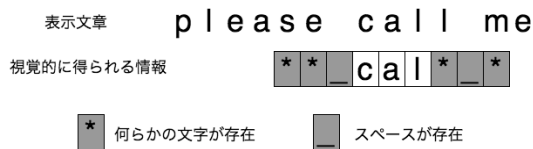


図 1 視野の二つの領域と得られる情報

2.2.1 視覚

Mr. Chips では視野を二つの領域に分類し、それぞれに対して得られる情報を定義している。一つ目は視野の中心にある文字を識別することが可能な高解像度領域で、二つ目が視野の周辺にある文字とスペースを区別可能な低解像度領域だ。

例えば図 1 のように「please call me」という文があり call の a の位置を注視しているとす。またこの図では中心視野(白背景部分)幅と周辺視野(灰背景部分)幅がどちらも 3 文字幅としているが、これは任意に変更可能である。この場合、左側の周辺視野領域に入っている文字は「se」の 3 文字であり、周辺視野では文字とスペースの区別のみが可能なため、その位置に何らかの文字が 2 つとスペースが 1 つあるという情報が得られる。中心視野領域に入っている文字は「cal」の 3 文字であり、中心視野では文字の識別が可能なため、その位置に「cal」という文字があるという情報がえられる。右側の周辺視野領域は左側と同様に、何らかの文字・スペース・何らかの文字という情報がえられる。

2.2.2 語彙

Mr. Chips の語彙知識は単語とそれらの頻度のリストで構成されていて、任意の辞書が使用可能となっている。

2.2.3 サッカー計画時の計算アルゴリズム

サッカー計画は、文章を閲読中の n 回目までの注視で得られた情報により認識の終了した最後の単語の次の単語(最初の未認識の単語)のエントロピーが最小となるような位置を計算し、その位置へサッカーを行い、 $n+1$ 回目の注視を行う。

計画手順としては以下の通りで、これを繰り返すことにより閲読が進行していく。

- (1) 文章中の文字を注視する
- (2) 周辺の文字情報を視覚の設定 2.2.1 節に応じて得る
- (3) 得られた情報と語彙知識から一意に特定できた単語を認識終了としていき、認識の終了しなかった最初の単語を「次に認識する単語」とする
- (4) 「次に認識する単語」に関して n 回目の注視までに得られた情報と語彙知識から単語の候補一覧を抽出する
- (5) 単語の候補一覧とその頻度から、次の注視位置を計算する
- (6) 求めた注視位置へサッカーを行う

(5) の計算では、次のサッカー長候補全てに対して、そのサッカーを行なった場合の予想エントロピーを求め、

エントロピーが最小となるサッカー長を次のサッカー長として選択している。

このアルゴリズムを変更することで、より少ない注視回数で文章の閲読を完了させることを目指す。

3. 提案手法

3.1 概要

本研究で提案する変更点は以下の 4 つである。

- (1) bigram 利用
- (2) 逆行なし
- (3) 単語認識終了条件緩和
- (4) サッカー長選択基準変更

(1)bigram 利用とは、利用する語彙知識を unigram から bigram に変更することで、単語の予測精度の向上を目的としている。(2) 逆行なしとは、逆行が必要となった場合には、認識中の単語を「単語候補中で最も頻度の高いもの」として次の単語認識に移ることで、効果の薄い注視を減らすことを目的としている。(3) 単語認識終了条件緩和とは、単語認識の終了条件を「単語を一意に特定する」から「単語のエントロピーを一定値未満にする」に変更することだ。この値は任意に変更できる。これは、Mr. Chips で発生していた「単語候補が 2 つしかなく頻度の差が大きい時のように、単語がほぼ確定しているにもかかわらず、その単語を次に認識する単語としてサッカーを計画する」という行動をなくすことを目的としている。(4) サッカー長選択基準変更とは、2.2.3 節で示したサッカー計画の最後に行うサッカー長の選択時において、「エントロピーが最小のもの」ではなく「エントロピーが言って入り未満で最も長いもの」を選択するように変更することだ。この値は任意に変更できる。

4. 評価実験

4.1 個別適用

4.1.1 論点

3.1 節で上げた変更点の有用性を示すために、まずは個別に適用したものでそれぞれ評価を行なった。

4.1.2 実験方法

それぞれの手法と何も変更していない手法のそれぞれに対して同じ文章を閲読させその効率性を比較した。ただし、サッカー長選択基準変更に関しては個別適用だと逆行が多発するため、逆行なしと併用して実験を行なった。全ての設定に共通して、中心視野幅を 9 文字幅、周辺視野幅を 4 文字幅とした。利用する語彙知識は、COCA コーパスの n -gram データ [5] を利用した。閲読させる文章は、COCA コーパス内で多く利用されているため NewYorkTimes の記事を利用した。期間は COCA コーパス内に入っていないものから選択し、2015 年 11 月 5 日以降のものから古い順に 200 件を選び、さらにその中から 1000 単語以上の記

	平均スコア
変更なし	1.655194749
逆行なし	1.665448441
bigram	1.741997534
bigram+逆行なし	1.757436679
unigram+認識終了条件緩和(0.4)	1.686312789
unigram+認識終了条件緩和(0.8)	1.70048948
unigram+認識終了条件緩和(1.2)	1.704394179
unigram+認識終了条件緩和(0.16)	1.706307758
unigram+認識終了条件緩和(0.20)	1.703588042
unigram+サッカーカード長選択基準変更(0.2)	1.710939795
unigram+サッカーカード長選択基準変更(0.4)	1.73454496
unigram+サッカーカード長選択基準変更(0.6)	1.738057034
unigram+サッカーカード長選択基準変更(0.8)	1.742253174
unigram+サッカーカード長選択基準変更(1.0)	1.750348816
unigram+サッカーカード長選択基準変更(1.2)	1.733638508
unigram+サッカーカード長選択基準変更(1.4)	1.737497159
unigram+サッカーカード長選択基準変更(1.6)	1.737381683

図 2 実験結果

事である 86 件を利用した。

評価基準は「1 注視あたりに正しく認識できた単語数」とした。

4.1.3 結果

結果は図 2 のようになり、それぞれの手法に対して評価値の向上が見られた。() 内は任意に設定可能な閾値を表している。

参考文献

- [1] Legge, Gordon E., Timothy S. Klitz, and Bosco S. Tjan. "Mr. Chips: an ideal-observer model of reading." *Psychological review* 104.3 (1997): 524.
- [2] Legge, Gordon E., et al. "Mr. Chips 2002: New insights from an ideal-observer model of reading." *Vision research* 42.18 (2002): 2219-2234.
- [3] Reichle, Erik D., Tessa Warren, and Kerry McConnell. "Using EZ Reader to model the effects of higher level language processing on eye movements during reading." *Psychonomic bulletin & review* 16.1 (2009): 1-21.
- [4] Beccue, Barbara, and Joaquin Vila. "Assessing the impact of rapid serial visual presentation (RSVP): A reading technique." *International Symposium and School on Advancex Distributed Systems*. Springer Berlin Heidelberg, 2004.
- [5] N-grams: based on 520 million word COCA corpus 入手先 (http://www.ngrams.info/samples_coca1.asp) (参照 2016-9-30).