機械学習による自動辞書引きを利用した英文の読解支援システム

江原 遥　　二宮 崇　　中川 裕志
東京大学情報理工学系研究科
{ehara,ninomi,nakagawa}@r.dl.itc.u-tokyo.ac.jp

1　はじめに

本稿では，英文読解を支援するために，自動的に辞書を引いて語義のアノテーションを行うシステムを提案する．事前に与えられた利用者の語彙力に関する情報を訓練データとして機械学習に与えることにより，利用者にとって未知の語彙を特定し，利用者の英語習得レベルに合わせたアノテーションを行う．

英語学習者 (English as Second Language Learner, ESL 学習者) の英文読解を支援する方法として，語義のアノテーションを行うことが有効であることは，主にコンピュータ支援語学学習 (Computer Assisted Language Learning, CALL) の分野における研究によって示唆されてきた．しかし，CALL 分野における語義のアノテーションに関する研究は，アノテーションの教育効果を確かめることに焦点があてられているため，ESL 学習者のレベルに応じて自動的に語義のアノテーションを行うシステムの研究は少ない．自動的に語義のアノテーションを与えるシステムとしては，マウスカーソルでテキスト中の単語を選択することにより辞書を自動的に引く手法が存在する．この手法は，計算機上でテキストを扱っている間は有効であるが，利用者は常にコンピュータと向き合っている必要があり，ポータブルな媒体としての紙に印刷する時には使用できない．

本研究は利用者の英語レベルに合わせて語義のアノテーションを与えるべき箇所を自動的に判別する手法を提案する．アノテーションを行う必要があると判別された箇所にのみアノテーションを与えるため，余白等に語義のアノテーションが付与された印刷イメージを作成することが可能である．しかしながら，この手法は語義の既知/未知を自動的に判別するために，誤判別する可能性がある．本研究では，(Method 1) 単語頻度および SVL12000 に登録された単語難易度表のみを用いた判別手法，(Method 2) 利用者 ID を用いて利用者に適応した判別手法，(Method 3) 数百人規模の利用者に対する語彙情報が得られた場合の判別手法の比較実験を行った．利用者に適応するために利用者情報が必要となるが，本研究では利用者の 5 語から 600 語の語彙に対する既知/未知の情報が得られることを想定した実験を行った[1]．また，語義の曖昧性解消は本研究では扱わないこととする．

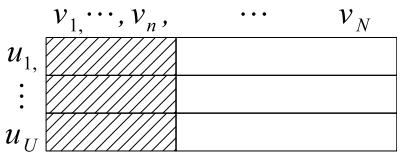

図 1: 利用者-単語行列

2　問題設定

単語は全部で N 個あるとし，利用者には事前に $n(\leq N)$ 個の単語集合 $\{v_1, \ldots, v_n\}$ を提示し，単語の既知/未知の情報を入力してもらうものとする．この n 個の利用者の単語に関する知識の情報から，利用者が残りの $N-n$ 個の単語について既知か未知かを判別することを目的とする．

以上の問題設定を，図 1 に示すように，利用者 u が単語 v を知っている時 $(u,v)=1$，知らない時 $(u,v)=-1$，情報がないときに $(u,v)=0$ であるような，“利用者-単語行列” で表す．$n(\leq N)$ 列からなる図 1 の斜線部 $V_0 = \{(u_i, v_j) | i \in \{1, \ldots, U\}, j \in \{1, \ldots, n\}\}$ が，利用者に事前に既知か否かを入力してもらう行列の要素である．つまり，本稿の問題は，図 1 の利用者-単語行列の斜線部 V_0 を訓練データとして機械学習で学習し，残りの白抜きの部分の要素値を推定する問題である．

本研究では，iKnow[2] というシステムからの情報，SVL

[1] 外部試験 (TOEIC, TOEFL 等) の結果を利用する手法も考えられうるが，外部試験を受けたことが無い人はシステムを利用出来なくなるため，本研究では外部試験の結果を想定しないこととした．
[2] http://www.iknow.co.jp/

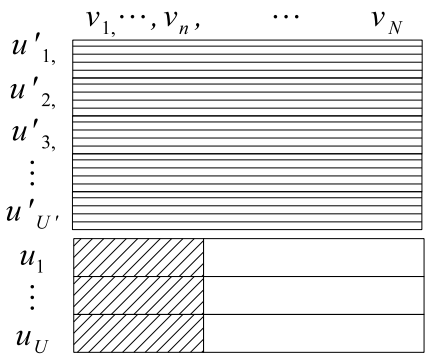

図 2: iKnow のデータを含む利用者-単語行列

単語難易度および Google コーパス[3]からの 1-gram 確率の三つを外部リソースとして用いた．iKnow は，CALL の分野において CAVOCA(Computer Assisted VOCabulary Acquisition, コンピュータ支援語彙習得) システムと呼ばれるシステムの一例であり，語学学習者が一定期間にできるだけ多くの語彙を学習することを目的として，語学学習者の忘却を考慮しながら単語を繰り返し提示し，語彙の習得を促すシステムである．CAVOCA システムは，学習者に対して語彙を習得させることが目的であり，本稿で扱う読解支援とは目的が異なる．iKnow は，CAVOCA システムとしては，著者らが知る限り利用者数が最大のシステムであるため[4]，このデータを外部リソースとして利用した．

iKnow を外部リソースとした情報を用いた場合の問題設定を，図 2 に示す．図中の斜線部と iKnow からのデータである横線部を，機械学習を用いて学習し，白抜き部分を推測する問題となる．ここで，斜線部の n 単語は全て既知/未知が付いており密であるのに対し，iKnow からのデータでは必ずしも全 N 単語について既知/未知をつけているとは限らないため，横線部は疎である．iKnow からは 10,526 人分のデータを取得したが，iKnow では単語の既知/未知を必ずしもつける必要がなく，既知と申告しなければ未知となるため，データに偏りが大きい．そこで，学習した単語の 15%以上を既知としており，かつ，100 単語以上の単語を学習している利用者 675 人を絞込み，外部リソースとして利用した．

本研究では，単語に関する素性として，SVL 単語難易度と Google コーパスからの 1-gram 確率の二つを用いた．SVL[5]単語難易度は，単語頻度やネイティブスピーカーによる感覚的判断を参考に，12,000 語の単語に対して 12 段階の難易度を割り振ったものである．また，Google コーパスは，1 兆語分の Web ページにおける単語の出現頻度を 5-gram まで記録したデータであり，この 1-gram 確率の対数を素性に用いた．

3 手法

本稿では，以下の手法の比較実験を行った．

Method 1 単語頻度および SVL12000 に登録された単語難易度表のみを用いた判別手法

Method 2 利用者 ID を用いて利用者に適応した判別手法

Method 3 数百人規模の利用者に対する語彙情報が得られた場合の判別手法

利用者 u の語彙 v に対する既知/未知の判別には，$(u, v) = 1$ であるか $(u, v) = -1$ であるかを判別する二値分類を行う識別器を用いる．判別に用いた素性は，“Method 1”, “Method 2”, “Method 3” によって，それぞれ異なるので次に説明する．

Method 1 単語頻度および SVL 単語難易度のみを用いた判別手法であり，最も素朴な方法である．この手法では，図 1 中の斜線部で示される訓練データ（V_0 に対応する利用者-単語行列の要素）に対して，単語頻度および SVL 単語難易度のみを素性に入れ，利用者 ID を素性に入れないため，利用者の区別は行わない．どの利用者に対しても，同じ単語 v の既知/未知の判断には，同じ閾値が用いられる．

Method 2 “Method 1” の単語頻度および SVL 単語難易度の素性に加えて，利用者 ID を素性に入れることにより，利用者に適応した判別手法である．“Method 1” と同様に，図 1 中の斜線部で示される V_0 の訓練データに対して学習を行うが，利用者 ID も素性に入れているため，同じ単語 v の既知/未知の判断に使われる閾値が利用者 u によって異なる．

Method 3 “Method 2” の素性に加えて，数百人規模の利用者に対する語彙情報が得られた場合の判別手法である．図 2 に示すように，V_0 の訓練データに加え，外部リソースとして，iKnow から取得した 675 人の利用者に対する語彙情報も訓練に利用する．

[3]Web 1T 5-gram Version 1, LDC Catalog No.: LDC2006T13, ISBN: 1-58563-397-6

[4]2008 年 10 月時点で 27 万人
http://www.iknow.co.jp/blog/ja/2008/10/1/88178-yu-za-ga-1-nen-de-2-oku-30-man-goi-wo-gakushuu

[5]http://www.alc.co.jp/goi/PW_top_all.htm

以上の “Method 1” から “Method 3” は，全て，単語頻度および SVL 単語難易度を素性として用いている．これらの単語素性を除き，利用者-単語行列のみを素性として用いた場合の性能を評価するため，次の “Method A” についても比較実験を行った．

Method A “Method A” は，“Method 3” と同様，多くの利用者に対する語彙情報が得られた場合の判別手法であるが，“Method 3” で使用していた単語頻度および SVL 単語難易度の素性を除き，利用者 ID と単語 ID のみを素性に用いている．

4 評価実験

§3 に挙げた判別手法の精度を測定するために，学生 10 人に対し，SVL 中の 12,000 語について，単語を知っている度合いを次の 5 段階から選択してもらうことにより評価用データを作成した．

1. 見たこともない
2. 見たことがある気がする
3. 確実に見たことはあるが意味は知らない/覚えたことがあるが意味を忘れている
4. 意味を知っている気がする/意味が推測できる
5. 意味を確実に知っている

このうち，5. のみを「利用者 u が単語 v について既知である」場合，すなわち $(u, v) = 1$ とみなし，1. から 4. は「利用者 u が単語 v について未知である」場合，すなわち $(u, v) = -1$ であるとした．

この評価用データから，10,000 語のテストデータ，1,400 語の Development Set を固定し，残りを最大 $n = 600$ 語の訓練データとして分割した．訓練データは図 1 や図 2 の斜線部に相当し，Development Set とテストデータはこれらの図中の白抜きの部分に相当する．

訓練データ数 n が小さいほど，利用者の初期の労力が減るため利用者にとっては望ましい．なぜなら，利用者はシステムを最初に利用する際に，n 個の単語について語彙の既知/未知をシステムに教えなければならないからである．そこで，n を $5, 30, 100, 300, 600$ と増加させた時に，“Method 1”, “Method 2”, “Method 3” について比較実験を行った．n の最大値は 600 とした．これは，利用者がこの初期作業にかけられる時間を最大 20 分程度と想定したためである．1 単語の既知/未知を入力するのに 2 秒程度かかるので，20 分では 600 語程度を入力できると考えられる．

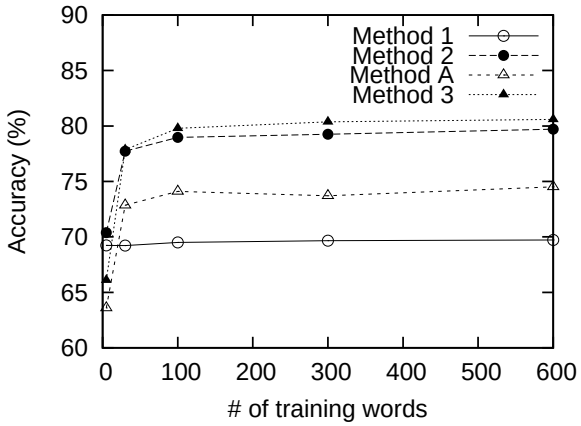

図 3: 評価実験結果グラフ

二値分類を行う識別器として，RBF カーネルを使った SVM を用いて行った．SVM の実装としては，libsvm[1] を用いた．評価実験の結果を図 3 と表 1 に示す．

	$n = 5$	30	100	300	600
Method 1	69.23	69.22	69.50	69.65	69.72
Method 2	**70.38**	77.73	78.98	79.25	79.72
Method 3	66.10	**77.87**	**79.79**	**80.37**	**80.59**
Method A	63.60	72.86	74.10	73.70	74.51

表 1: 評価実験結果表

図 3, 表 1 より，訓練データ数 n を 100 語より多くしても，精度は余り向上しないことがわかる．n は利用者が事前にシステムに与えなければならない既知/未知を付与した語彙数であるので，利用者にとっては少ない方がよく，この結果は 100 語程度入力すればよいことを示している．利用者が 100 語程度の語彙数の既知/未知を入力するために要する時間は長くても 5 分程度なので，5 分程度の準備で本システムを使用可能であることを示唆している．

“Method 1” と “Method 2” を比較すると，“Method 2” の方が精度が高い．これは，“Method 2” には利用者 ID の素性が含まれているため，“Method 1” では一律に決められていた単語に関する素性の閾値が，利用者ごとに決定されるようになったためであると考えられる．

“Method 2” と “Method 3” を比較すると，わずかではあるが，“Method 3” の方が精度が高い．“Method 3”

で追加した数百人規模の CAVOCA システム利用者に対する語彙情報には，欠損値が多く含まれているが，このような情報を追加しても精度が向上することが示された．

“Method A” は 74%程度の精度を得ている．通常，利用者の語彙の既知/未知を判別するためには，単語難易度や単語頻度の素性が重要な役割を果たすと考えられる．しかし，この結果は，単語素性を用いず利用者 ID と単語 ID だけを素性に用いても，多くの CAVOCA システム利用者に対する語彙情報が得ることができれば，利用者の語彙の既知/未知を判別することが可能であることを示唆している．

5 関連研究

英文の読解を支援する目的における語義のアノテーションの有効性は，CALL 分野における研究により示唆されている．[3] は，英語力と辞書の使用の有無と文章の理解度の関連について調査した．その結果，英語力の低いグループでは，辞書を使用した学生の方が，辞書を使用しなかった学生と比較して有意に高い文章の理解度を示した一方，英語力の高いグループでは，有意な差は見られなかったとしている．アノテーションの種類として語義を用いることの有効性は，[2] に詳しい．この文献では，ESL 学習者は英語の習熟度にかかわらず，アノテーションとして語義を発音や画像より好むことを実験により示している．

また，利用者にとって未知の箇所を判別するための指標として単語難易度が考えられるが，[4] は，ESL 学習者にとっての単語難易度の指標としてコーパス中の単語頻度を用いることが妥当であることを報告している．

6 結論と課題

本稿では，英文の読解を支援するために，自動的に辞書を引いて語義のアノテーションを行うシステムを提案した．学生 10 人に，実際に 12,000 語について単語を知っている度合いを入力してもらい，このデータを用いて利用者の語彙の既知/未知を判別する精度を評価した．評価実験の結果，利用者が最初に 5 分程度をかけて 100 語について単語を知っている度合いを入力すれば，80%程度の精度で，その利用者の語彙の既知/未知を判別することが可能であることが分かった．また，欠損値が多く含まれていても，数百人規模の CAVOCA システム利用者に対する語彙の情報を用いると判別精度がわずかながら向上することが分かった．

今後の課題としては，単語に関する素性を工夫し精度を向上することや，Web 上のテキストや PDF に対して実用に耐えるシステムを作成することが挙げられる．

参考文献

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] G. Ercetin. Exploring esl learners' use of hypermedia reading glosses. *CALICO Journal*, Vol. 20, No. 2, pp. 261–283, 2003.

[3] S. Knight. Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal*, Vol. 78, No. 3, pp. 285–299, 1994.

[4] J. M. Tamayo. Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. *Educational and Psychological Measurement*, Vol. 47, No. 4, pp. 893–902, 1987.