機械学習による語学用空所補充問題の自動生成

星野 綾子　　中川 裕志
東京大学 大学院学際情報学府　　東京大学 情報基盤センター
hoshino@dl.itc.u-tokyo.ac.jp　　nakagawa@r.dl.itc.u-tokyo.ac.jp

1　はじめに

この研究は、語学テストのための４択空所補充問題を自動的に生成するシステムの構築を目指す。このシステムにより、オンデマンドで語学問題を提供することができるようになる外、現存の語学テストより詳細な受験者の言語運用能力の評価を目指す。

自然言語処理技術の応用として、多肢選択問題の生成は近年まで皆無に等しかった [7]。しかしながら、ごく最近になり報告が見られるようになった。

Mitkov らは計算言語学の用語集を入力とし、用語の知識を問う問題を半自動的に作成した。Sumita らのシステムは、任意のテキストから空所補充型の動詞の語彙選択問題を全自動で生成する [9]。Liu らは、曖昧性解除を用いて大量コーパスの中から指定された語が指定された意味で用いられている例文を抽出し、選択式空所補充問題を得るシステムを作成した [5]。

既存の研究に共通点には以下のような特徴がある。1) 全自動のものに関しては、空所位置をルールで決定する。典型的には、動詞・群動詞を空所とする。2) 誤り選択肢はシソーラスの類義語を利用する。これによって生成される問題は、動詞の語彙問題というタイプのものに限定されることになる。

本研究の方法の特色は、人手で作成された問題を集めて分析し、機械学習の手法を用いて問題作成をする点である。これにより、限られたタイプの問題しか作成できないという先行研究の問題点の克服を目指す。

2　自動的問題作成システム

提案システムの構成は図 1 のようになっている [1]。

2.1　空所補充・多肢選択問題

本研究の対象とする問題形式は空所補充・多肢選択問題と呼ばれるものである [8]。

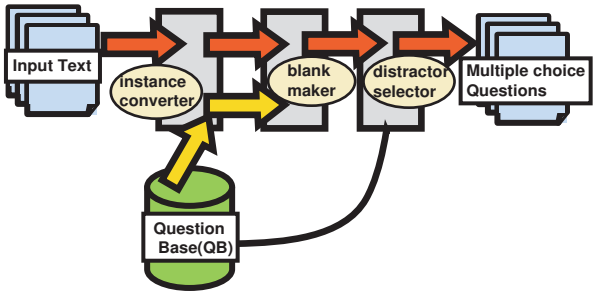

図 1: システム概要図

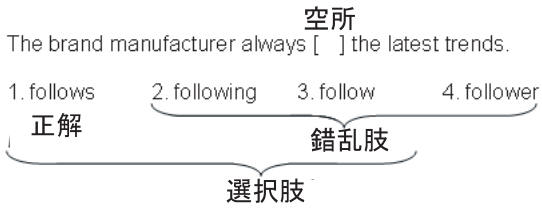

図 2: 空所補充・多肢選択問題

問題ベース (Question Base) には同様の形式の、人手で作られた問題 [1]が集められている。入力テキストは文に区切られ、さらに品詞などの素性抽出が行われ、機械学習の対象となるインスタンスに変換される。この後、問題生成器（空所位置の決定、誤り選択肢の決定）を経て空所補充型選択問題が自動生成される。

2.2　トレーニングインスタンスの生成

問題ベースの中の問題はトレーニングデータとして用いられる。正解語で空所補充された問題文は以下のように空所位置を移動させることで機械学習の対象となるインスタンス群（図 1）に変換される。

このようにして得たインスタンスに素性（空所位置の語の品詞、前後の語の品詞、文中での位置、WordNet

[1]TOEIC 対策問題集 [6] 150 問を利用した。

負例	[ ] sentence is converted into multiple instances.
負例	A [ ] is converted into multiple instances.
負例	A sentence [ ] converted into multiple instances.
正例	A sentence is [ ] into multiple instances.
負例	A sentence is converted [ ] multiple instances.
負例	A sentence is converted into [ ] instances.
負例	A sentence is converted into multiple [ ] .
負例	A sentence is converted into multiple instances [ ]

表 1: 一文から生成されたインスタンス群

における、BNC コーパスにおける頻度 [4]、文の長さ等）を付与する。元の空所位置を正例、移動した位置を負例としたが、移動した位置にも問題作成に適している空所位置があるかもしれないので、半教師あり学習を行い、この負例への偏りを是正した。

半教師あり学習の手順は以下の通りである。まず、仮分類されたトレーニングデータに基づいて学習した分類器を用いてトレーニングデータを再分類する。この結果、仮分類が負例であったものの中から再分類において正例とされたもののラベルを正例へと変更する[2]。この過程により、人手を介することなく、負例への偏りを是正したトレーニングデータを得ることができる [2]。

2.3 空所位置の決定

このようにして得たトレーニングデータを用いて、ナイーブベイズ、kNN の 2 種類の分類器を学習させた [3]。これらの分類器によって入力テキストから得たインスタンスを分類させることにより、空所位置を決定する。

2.4 錯乱肢の決定

錯乱肢の決定は、問題ベース (QB) 中の錯乱肢群をラベルとし、生成された空所位置インスタンスに与えることで行われる。ナイーブベイズ、kNN の分類器を用いることにより、QB 中の錯乱肢群の中で最も適当な錯乱肢群を与える。素性は、空所位置決定タスクの際に付加された素性を利用する。全く同じ素性列が入力テキストから得られた場合、このシステムでは以下のように類題を生成することが保障されている。

[2]学習・ラベル変更の過程を繰り返す。実験データにおいて正例の増加は収束することが確認された。

QB 中の問題：

The University will [ ] the student with the scholarship.
1. provide　2. offer　3. give　4. present

テキストから作成された問題：

The government will [ ] the couple with one-time payment.
1. supply　2. offer　3. give　4. present

1 が正解で、2 から 4 がラベルとして与えられた錯乱肢群である。以上のような方法で作成された問題を次の章で評価する。

3 評価

長さ 219 語の新聞記事から 15 人の英語教師あるいは英語教師経験者に同形式の問題を作成してもらい、問題作成システムの評価をした。

3.1 人手作成の問題の収集

問題の収集は web アプリケーションを通じて行われた。問題作成器と同様の方法で問題作成をしてもらうため、1 番目の画面で空所位置の選択、2 番目の画面で錯乱肢の入力をしてもらった。1 人 5 問以上作成するよう指示を与えた。

3.2 空所位置について

空所位置の選択 131、全 58 題が集められた。異なり空所位置の数は 64 箇所であった。参加者間の一致は充分とはいえなかったが、これは記事の長さに対し各参加者の選択された空所位置の個数が充分ではなかったことによる。

1 人でも選択した空所位置を妥当な空所位置とし、以下のように精度 (P)、再現率 (R)、F 値 (F) を計算した。

$$\text{精度} = \frac{\text{システム、参加者が共に選んだ空所位置}}{\text{システムが選んだ全空所位置}}$$

$$\text{再現率} = \frac{\text{システム、参加者が共に選んだ空所位置}}{\text{参加者が選んだ空所位置}}$$

$$F\ \text{値} = \frac{2 * \text{精度} * \text{再現率}}{\text{精度} + \text{再現率}}$$

N	0	1	2	3	4	5	6
kNN P	.382	.435	.426	.414	.415	.413	.405
kNN R	.306	.435	.541	.565	.576	.588	.600
kNN F	.340	.435	.477	.478	.483	.485	.483
NB P	.400	.518	.418	.426	.418	.425	.430
NB R	.071	.341	.447	.541	.541	.565	.576
NB F	.120	.411	.432	.477	.472	.485	.492

表 2: 半教師あり学習の繰り返し回数 N による各アルゴリズムのスコアの変化

半教師あり学習の効果

本研究では、「文章中の全ての動詞を空所とする」というルールと比較して、英語教師の選択に近い空所箇所選択を行うことを目指す。ルールによって空所箇所を決めた場合の精度は 0.67、再現率は 0.28、F 値は 0.39 であった。表 2 に半教師あり学習の繰り返し回数 N による各アルゴリズムの精度、再現率、F 値の変化を示す。半教師あり学習を行わなかった場合、F 値においてルールによる方法を下回るが、1 回以上行った場合、再現率のみではなく精度も向上し、ルールによる方法を上回った。繰り返し回数を増やした結果、精度を大きく損なうことなく再現率が上がることが確認された。

QB 中の問題数を増加させた場合の変化

次に、QB 中の問題数を増加させた場合の各アルゴリズムの精度、再現率、F 値の変化を調べた。QB 中の問題数を全 150 問のうちからランダムに選んだ 10 問、30 問、50 問、100 問、全 150 問と変化させた。図 3 に QB 中の問題数の変化と各アルゴリズムの精度、再現率、F 値の変化を示す。

ナイーブベイズの F 値は 100 問までは単調に改善しているが、150 問においては悪化している。kNN 法の精度、再現率、F 値は問題数の変化に対し安定しなかった。

3.3 錯乱肢について

4 名の英語教師の 5 段階評価により、提案手法の錯乱肢選択を評価した。

先の実験で英語教師が 2 名以上選択した空所位置 9 箇所において、英語教師が作成した錯乱肢群 2 組 (Human)、150 問の問題の錯乱肢からランダムに 1 語ずつ錯乱肢を選ぶ方法 (Random)、150 組の錯乱肢群からランダムに 1 組の錯乱肢群を選ぶ方法 (Random altset)、ナイーブベイズによって選ぶ方法、kNN によって選ぶ方法を比較した。図 4 に合計点数（Human の場合は 2 組の平均点）の比較を示す。

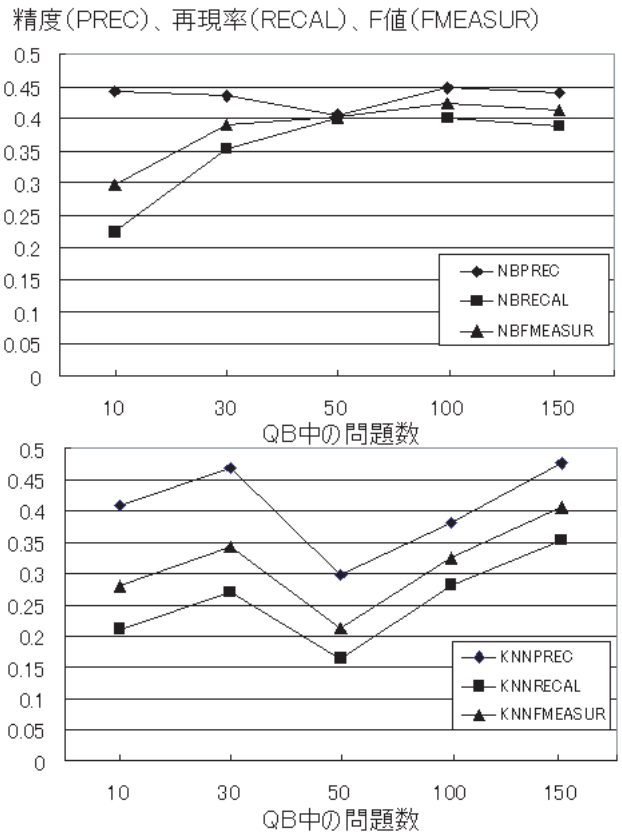

図 3: QB 中の問題数の変化と各アルゴリズムの精度、再現率、F 値の変化

この評価の結果、人手作成の問題もあまり高い点は得られなかった。これは、人手で作成した問題にも複数正解があったことなどによる。Random と Random altset を比較すると、Random altset の方が 11 点低い。これにより錯乱肢群を選択する提案手法の設定は 150 という限られた数の錯乱肢群においては、より困難な設定であることが分かる。しかしながら、kNN の結果は Random altset より 14 点ほど高く、Random より高い点を得た。ナイーブベイズについては、Random を 3 点上回るのみであった。

以上の評価の結果、空所位置決定のタスク・錯乱肢選択のタスクの両方において充分な性能は得られなかった。これは、特に空所位置決定のタスクにおいては、QB の問題数の不足によって解決されるものではないことがスコア向上の伸び止まりから分かる。また、錯乱肢決定の方法も工夫が必要である。

興味深かった点は、半教師あり学習によって明らかな性能の向上が見られた点である。移動してできた位

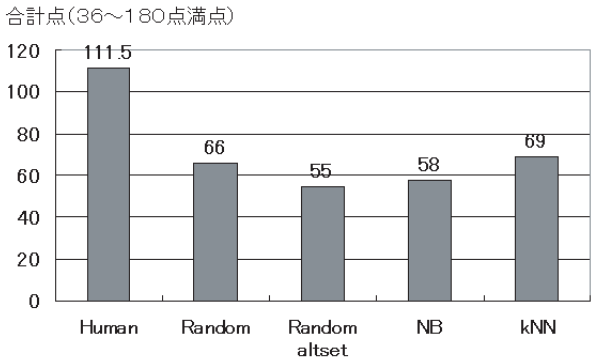

図 4: 錯乱肢についての評価

置を負例とした仮分類には大きな誤差が存在したこと、このような誤差を是正するために半教師あり学習が有効であることが分かった。

4 おわりに

以上、機械学習の手法を用いた自動問題作成システムを提示した。現状では提案手法よって妥当な問題が生成できないが、トレーニングデータの質と量によっては。今回は類似研究とは異なり、問題集の問題を用いた。問題集の問題に付加された情報を用いることで、より質の高い問題を生成するシステムの実現が可能かもしれない。対象とする文法規則など、問うべき項目を限定して生成を行うことなどが考えられる。

謝辞

東京工業大学の仁科先生には、ご協力と貴重なディスカッションに感謝いたします。東海大学の松本先生をはじめとして、実験に参加してくださった 15 名の先生方に厚く感謝いたします。

参考文献

[1] Ayako Hoshino and Hiroshi Nakagawa. A real-time multiple-choice question generation for language testing: A preliminary study. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pp. 17–20, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[2] Ayako Hoshino and Hiroshi Nakagawa. Webexperimenter for multiple-choice question generation. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pp. 18–19, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[3] Ian H.Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, October 1999.

[4] Adam Kilgarriff. BNC database and word frequency lists. University of Brighton, 1995. http://www.itri.brighton.ac.uk/ Adam.Kilgarriff/bnc-readme.html.

[5] Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pp. 1–8, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[6] Shuhou Matsuno, Tomoko Miyahara, and Yoshi Aoki. *STEP-UP Bunpo mondai TOEIC TEST.* Kirihara Publisher, 2000.

[7] Ruslan Mitkov and Le An Ha. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, pp. 17 – 22, Edmonton, Canada, May 2003.

[8] T Shizuka, K Yoshizawa, and O Takeuchi. 外国語教育リサーチとテスティングの基礎概念 - Basic Concepts in Foreign Language Education Research and Testing. Kansai Daigaku Shuppanbu, 2002.

[9] Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pp. 61–68, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.