2H2-1

「ロボットは東大に入れるか」プロジェクトにおける英語科目の到達点と今後の課題

Current status and future challenges for the English subject in the "Can a Robot Get into the University of Tokyo?" project

東中 竜一郎 *1 Ryuichiro Higashinaka
杉山 弘晃 *1 Hiroaki Sugiyama
成松 宏美 *1 Hiromi Narimatsu
磯崎 秀樹 *2 Hideki Isozaki
菊井 玄一郎 *2 Genichiro Kikui
堂坂 浩二 *3 Kohji Dohsaka
平 博順 *4 Hirotoshi Taira
南 泰浩 *5 Yasuhiro Minami
大和 淳司 *6 Junji Yamato

*1NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories
*2岡山県立大学 Okayama Prefectural University
*3秋田県立大学 Akita Prefectural University
*4大阪工業大学 Osaka Institute of Technology
*5電気通信大学 The University of Electro-Communications
*6工学院大学 Kogakuin University

We have been working on the English problems in the "Can a Robot Get into the University of Tokyo?" project. This paper summarizes what we have achieved during the past few years and mentions future challenges we identified. We describe our approaches to individual problems in the English exam, some of which were successful while others performed rather poorly because of the lack of the understanding of "meanings". We point out that, although deep learning can be a key to better results, currently, to solve difficult problems such as entrance exams, we require more data as well as more technical advances.

1. はじめに

2016 年 11 月に行われた東ロボ成果報告会において,「ロボットは東大に入れるか」プロジェクト（東ロボ）では，これまでのように毎年模試を受けるという営みは当面行わない旨が発表された．これは，主に，読解問題の難しさが依然大きな課題であり，短いスパンでは解決できそうにないと関連研究者が判断したためである．この判断には，英語科目における読解問題へのチャレンジの結果が大きく関わっている．今年の東ロボ君の英語の模試（ベネッセ模試）の成績は 95 点（受験者平均 92.9 点）と伸び悩んだ．我々は，従来の統計的手法だけでなく，我々は近年さまざまな分野でよい性能を発揮している深層学習に基づく手法を利用し，得点の改善を図ってきた．しかし，類似度ベースの手法にいずれも及ばなかった．以下は，英語科目の結果から得られた知見である．

- 会話文完成や不要文除去は「文の並びの自然さ」を測る問題である．一文問題（問題文が一文程度からなる問題）の結果（2. 節を参照）によると，自然な「単語」の並びを十分理解するには 500 億単語のデータ（1.9G 文）が必要であることが分かった．ここから，自然な「文」の並びを十分理解するには，少なくとも 500 億文規模の複数文データ（たとえば対話データ）が必要と考えられる．このようなデータは世の中に存在せず，十分な統計情報が現状得られない．
- 意見要旨把握のような問題は「文」と「文章」が特定の関係性（この場合は要約）に当たるかどうかを判断する問題である．先行研究 [Hermann 15] によると，要約に含むべき単語を一つ選択する問題でも 100 万程度の高品質な事例を利用している．よって，単語を複数選択する問題では，100 万以上の高品質な事例が必要と考えられるが，そのようなデータは現状存在せず，構築するにもコストがかかりすぎる．また，機械的に作成した事例では質が不十分である（詳しくは 3.2 節を参照）.

本稿は，我々の東ロボにおける英語科目へのアプローチの詳細をまとめたものである．英語問題はおよそ一文問題，複数文問題（問題や選択肢が複数文からなるもの），長文問題からなり，それぞれ難しさが異なるため，この単位で章立てを分け，問題と解法を述べている．なお，紙面の都合ですべての問題種別の解法は記載していない．一文問題はおおむね解けることが分かったが，複数文問題以降は難しい．英語チームとして，今後も複数文問題の解決に取り組んでいく予定である．

連絡先: higashinaka.ryuichiro@lab.ntt.co.jp

2. 一文問題

2.1 文法・語法・語彙

文法・語法・語彙問題とは，図 1 のように，文中に開いている空欄に最もふさわしいものを、4 つの候補の中から選ぶ穴埋め問題である．

Most of the students voted [12] Tom's proposal, and it will be put into practice soon.

① at ② for ③ into ④ to

図 1: 文法・語法・語彙問題の例

我々はこの問題に回答するため言語モデルを適用した．空欄に選択肢のいずれかを埋め込み，言語モデルで対数尤度を計算することで，単語の並びとして尤もらしい選択肢を選ぶ．2015 年度までは，SRILM のデフォルト設定（頻度カットあり）を用いて，One billion word corpus から 5gram で構築した言語モデルを利用していたが，この方法で訓練した言語モデルでは，特定の事物を表す名詞などの出現数の小さい表現で頻繁に Ngram が途切れてしまい，文の流暢さを正しく評価できない問題があった．そのため，2016 年度では，Ngram の頻度による足切りの影響と，表 1 に示す，コーパスサイズの影響について検証を行った．また，出現単語の頻度などによる，対数尤度の正規化についても検証を行った．なお，本検証のモデル作成には KenLM を用いた．モデルの検証には，大学入試センター試験の本試験及び追試験の過去問, 代ゼミセンター模試, ベネッセ模試, 独自に収集したその他の問題を合わせた, 合計 552 問をベンチマークデータとして用いた．

まず，Ngram 頻度による足切りの有無について比較する．5gram の言語モデルを One billion word corpus で訓練すると，SRILM デフォルト設定の頻度足切りあり（$p > 1e-8$）の場合，ベンチマークの開発データ上で 0.650, 頻度足切りなしでは 0.779 であった．

コーパス名	説明	文数	サイズ
1billion	One billion word corpus(ニュース)	32,541,199	4.0GB
Gutenberg	Gutenberg corpus（フリー小説）	105,724,676	12GB
UMBC	UMBC text corpus（ニュース系）	134,000,311	18GB
Enwiki	English wikipedia corpus	146,768,004	15GB
LDC2011	Gigaword corpus	164,676,799	21GB
Common Crawl(1G)	Common Crawl から抽出	1,009,716,842	108GB
0.9G sent	Common Crawl(1G) を除いて合算	917,741,427	70GB
1.9G sent	全て合算	1,927,458,269	178GB
6G sent	Common Crawl を増やしたもの	6,001,232,913	637GB

表 1: コーパスとサイズ．単語頻度 10 以下は UNK に置換．

正規化方法	精度
正規化なし	0.822
文長正規化	0.833
空欄頻度正規化	0.833
文長正規化+空欄頻度正規化	0.853
文長正規化+空欄頻度正規化+数詞変換	0.857

表 2: 正規化方法による正答率の変化

次に，対数尤度計算時の正規化の方法について，7gram, 1.9G sent で Ngram 頻度足切りなしで訓練したモデルについて，表 2 の項目の検証を行った．空欄頻度正規化は，空欄に入る選択肢中の単語の 1gram 出現頻度の平均を，全体の対数尤度から引いたものである．数詞変換は，数字を桁数のみがわかるように変換する．文長正規化は，文全体の単語数で，得られた対数尤度を割るものである．表 2 より，各正規化を全て行った場合が最もよいことがわかる．

図 2 に，頻度足切りを行わずに 7gram 言語モデルを訓練した場合の性能について，訓練に用いたコーパス中の文数との関係を示す．おおむねコーパスの文数の対数に比例して，正答率が上昇していることがわかる．ただし，1.9G sent から 6G sent へ増やしても増分は見られなかったため，おおよそ 2G sent 付近が本アプローチの上限となっている可能性もある．

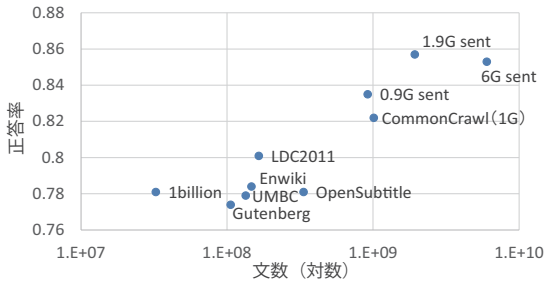

図 2: コーパスサイズと正答率の関係

2.2 語句整序完成

語句整序完成問題は，空所を含む文に対して，与えられた数個の単語列を適切に並べ替えて，文法的・意味的に正しい文を完成させる問題である．

我々は，語句の並べ方を全通り作成して元の文に当てはめ，言語モデルによって文の流暢さを評価するアプローチを採用した．言語モデルは文法・語法・語彙問題と同一の，1.9G sent で訓練した 7gram の言語モデルを利用した．なお，本問題では，使用される単語が全て同一であるため，正規化はしない．本手法のベンチマークでの正答率は 0.962（152/158）であった．2015 年度に用いていた，UMBC コーパスで構築した Ngram 言語モデルと，opensubtitles で学習した LSTM とを組み合わせたアプローチでは 0.91（143/158）であった．また，本言語モデルと 2015 年度の LSTM を組み合わせた場合でも，正答率はほぼ同一であった．このことから，語句整序問題においても，コーパスのサイズが非常に重要であることが分かる．

2.3 発話文生成

発話文生成問題は，空所を含む文に対して，与えられた数個の単語列の組を適切に並べ替えて，直前の発話に対する応答として，文法的・意味的に正しい文を完成させる問題である．

本問題においても，語句の並べ方を全通り作成して元の文に当てはめ，言語モデルによって文の流暢さを評価するアプローチを採用した．言語モデルは文法・語法・語彙問題と同一の，1.9G sent で訓練した 7gram の言語モデルである．なお，選択肢によって単語数が異なるため，単語長対数尤度を割ることで正規化を行った．ベンチマークにおける正答率は 0.875（105/120）であった．

3. 複数文問題

3.1 会話文完成

会話文完成問題は，二人の話者の会話の空所に相応しい文を 4 つの選択肢から選び，会話文を完成させる問題である．会話文完成問題を解くため，4 つの選択肢の各場合について会話の流れの自然さを推定し，最も自然な流れとなる選択肢を選ぶという解法を開発した．会話の流れの自然さのスコアは，隣接発話らしさのスコア，感情極性の流れの自然のスコアの重み付き和として算出する [堂坂 16]．

隣接発話らしさとは，2 つの発言が会話の中で隣り合って現れる確からしさを表す．隣接発話らしさを認識するために，NTT シチュエーション対話コーパスと Movie-DiC コーパス [Banchs 12] から SVM 認識器を学習した．隣接発話らしさを特徴づける素性として，2 つの発言それぞれに含まれる単語 n-gram のペア (フレーズペア) を用いた．NTT シチュエーション対話コーパスは，会話の場面と話題を指定した上での会話を収集したものであり，68,020 発言から成る．Movie-DiC コーパスは映画のスクリプトからなり，本研究ではその一部のデータ (277,184 発言) を使った．

感情極性の流れの自然さとは，連続する発言の間では感情極性の確率分布の変化が小さいほど，自然な会話の流れであるという考え方に基づくスコアであり，感情極性コーパス [Pang 04] から学習した SVM 認識器を使った．

評価のため，大学入試センター試験の本試験及び追試験の過去問，代ゼミセンター模試，ベネッセ模試，独自に収集したその他の問題を合わせて，合計 241 問のうち，163 問を開発データセット (dev set)，78 問をテストデータセット (test set) とした．開発データセットでパラメタを調整し，テストデータセットでの正解率を評価した．

学習データが隣接発話らしさの認識に与える影響を評価するため，NTT シチュエーション対話コーパスのみから認識器を学習した場合 (situ), Movie-DiC コーパスのみから認識器を学習した場合 (movie), 2 つのコーパスを混合したデータから認識器を学習した場合 (situ-movie-mixed)，各コーパスから学習した 2 つの認識器によりスコアを計算し，その重み付き和を総合スコアとする場合 (situ-movie-coord) を比較した．評価結果を表 3 に示す．2 つのコーパスから学習した認識器を組み合わせた場合 (situ-movie-coord) が最も高い正解率 0.45 であった．今後，解法の性能の向上を図るためには，隣接しない発言の間の首尾一貫性を捉えることが必要となると考える．

解法	dev set	test set
situ	60/163(0.37)	31/78(0.40)
movie	58/163(0.36)	31/78(0.40)
situ-movie-mixed	59/163(0.36)	28/78(0.36)
situ-movie-coord	64/163(0.39)	35/78(0.45)

表 3: 会話文完成問題解法の正解率

3.2 意見要旨把握

意見要旨把握は，対話中の一人の話者の発話に対して，相手の意見内容に対して，もう一人の話者が提示する四つの要旨候補の中から最も相手の意見に近いものを選択するという問題で

手法	Word2vec	CNN	attentive reader	WMD	End-To-End MN
正解率	48/120 (0.4)	45/120 (0.375)	45/120 (0.375)	47/120 (0.391)	38/120 (0.317)

表 4: 各手法のベンチマークでの評価（正解率）.

ある．この問題に対して，DNN を用いる手法の有効性を調べた．具体的には，以下の 5 種類のモデルを比較した.

Word2vec 提示意見と要旨候補の Word2vec [Mikolov 13] に基づくコサイン類似度を用い，最も類似した要旨を正解とするもの.

CNN 文献 [Wang 16] に示す CNN を利用したもの.

Attentive reader 文献 [Hermann 15] に示す attentive reader のモデルを実装したもの.

WMD WMD(Word Mover's Distance) [Kusner 15] を利用したもの．なお，fasttext [Joulin 16] を用いた極性判定の情報も付加した.

End-To-END MN end-to-end memory networks [Sukhbaatar 15] を利用したもの．ただし，最終的な出力に識別器の NN を付加することで，選択肢の中から正解を選べるようにした.

学習データ：各モデルで若干ことなるが，(2),(3) に関しては NTT 言い替えコーパス（88,000 言い替え）を用いた．また，評価データ（ベンチマーク）にはセンター模試（過去問題）を用いた．表 4 にベンチマークの評価結果を示す.

この結果から，今年度の模試では，最も性能のよい Word2vec による手法を採用した．結果は，3 問中 1 問正解の 33%の正解率であった．以上のように，各種の複雑な DNN による手法を用いても，高い成果率を示すことはなかった.

4. 長文問題

4.1 読解（情報処理）

センター試験の 4A, 4B では，図表を読みとるツールが必要である．ここでは，表やグラフを読むツールの作成について述べる．オープンソースの画像処理ソフト OpenCV (`http://www.opencv.org/`) と，文字認識ソフト Tesseract OCR (`https://github.com/tesseract-ocr`) を利用して，図表の読み取りツールを作成した.

表の読み取り [磯崎 15] では，すべてのセルが罫線で囲まれている表は，輪郭線検出により，簡単に読めることが判明した．しかし，複数行にまたがるセルや複数列にまたがるセルがあると，その例外処理が複雑になる.

グラフの読み取り [磯崎 17] では，グラフの種類ごとに読み取るためのソフトを開発した．図 3(a) のような「マークのある折れ線グラフ」では，凡例領域からマークを切り出してテンプレートとしたテンプレートマッチングにより，それぞれのマークの位置を検出することでグラフを数値化できる.

しかし，図 3(b) のようなマークのない折れ線グラフはマークに頼った読み取りができない．そこで，黒いピクセルの塊を求めて，それぞれの塊を分ける．実線は一つの塊なのですぐに読めるが，破線・点線・一点鎖線等は多数の塊でできているので，グルーピングしてから読み取る.

図 3(c) のような円グラフは，Hough 変換で円の中心座標と半径を推定する．次に円の中心から伸びる扇型の境界線を検出．最後に凡例の塗りつぶしパターンをテンプレートとしたテンプレートマッチングを行い，数値化している.

棒グラフは矩形なので簡単そうだが，図 3(d) のような「積み上げ棒グラフ」はかなり面倒である [中野 17].

これらのツールにより，表やグラフを読み取ることはできるようになってきたが，グラフにはこれ以外の種類もある．また，得られた情報を文章の情報と付き合わせて答を導くところも今後の課題である [磯崎 16].

4.2 内容一致問題

内容説明問題（もしくは，読解（論説文））は，与えられた長文本文全体や特定のパラグラフの内容に関して問う各小問に対して，最も合致する選択肢を解答する問題である．センター試験では主に問 6A で出題され，5〜8 パラグラフ，30〜50 文程度の本文に対し，4 択問題が 5 問程度出題される.

内容説明問題については，指定パラグラフに含まれる文と問題選択肢の文のペアで最も意味的類似度が高い選択肢を解答として選ぶという方針で自動解答手法を開発した．最終的にとった方法では，文同士の意味的類似度は各文を構成する単語同士の意味的類似度の合計として計算した．ただし、類似度を計算する単語のペアは総当りにはせず，述語項構造を考慮し，述語動詞，3 つの項タイプ（Arg1〜Arg3）それぞれでの単語のペアについての意味的類似度を計算した（詳しくは [尭天 16] を参照のこと）．この手法について事前に評価実験を行ったところ，ベネッセ模試の過去問 3 回分の計 15 問については 9 問正解し 60%と比較的高い正答率が得られた.

そのため，この手法をフォーマルランでも採用したが，結果としてはフォーマルラン小問 5 問でいずれも不正解となった．不正解の原因としては，過去問では，文の言い換えレベルで選択肢と同様の意味を持つ本文の文を探せば正答出来る問題が多く出題されていたのに対し，フォーマルランでは 5 問すべてがそのような方法で解けない問題になっていたことが挙げられる．具体的には，大幅に構文が異なっていたり，熟語が用いられていたり，本文にない内容を問われていたり，人間の価値判断を問う問題などであった．今後このような問題に対応するためには，述語項構造が異なってくるような言い換えの考慮，複数文の要約と選択肢との類似度の評価,「本文の内容と合致しないもの」を問う設問への対応などが必要であると考えられる.

4.3 段落タイトル付与

「段落タイトル付与問題」は長文の指定された 4-5 個の段落に対してタイトルとして適切な言語表現を選択肢から選ぶ問題である．選択肢の数はタイトルを付与すべき段落と同数であり，各段落に対して重複なく選ばなければならない．本問は 4-5 個の段落の全てについて完答した場合にのみ得点となることから，ランダムに選択した場合に得点できる確率 (完答率) は 1/4!=0.04 ないし 1/5!=0.008 と非常に低い.

我々は「適切なタイトルは当該段落と意味的な類似性が高い」と考え，可能な解答候補（段落に対して重複なく選択肢を割り当てたもの）のうち，段落とこれに割り当てられた選択肢との間の「意味的類似度」の合計が最大となるようなものを選ぶこととした（詳しくは [井内 17] を参照のこと）．ここで，段落と選択肢の類似度はそれぞれをベクトル化したものの間のコサイン値である.

段落に対するベクトルは段落全体を bag-of-words と考えて，含まれる全ての単語を word2vec [Mikolov 13] でベクトル化し IDF を重みとして足し合わせたもの ([Pershina 15] を参考にした) をベースとする．なお，この時に段落全体の単語を使うのではなく，その一部を使うことも試した．具体的には，i) 段落の最初の 1 文，ii) 段落の最後の 1 文，iii) 段落の最初の 1 文と最後の 1 文，iv) 段落中で最も選択肢との類似度が高い文，である．なお，iv については類似度を計算する選択肢ごとに異なる文が選ばれる可能性がある．選択肢に対するベクトルも段落と同様に含まれる単語を word2vec と IDF でベクトル化した．なお，選択肢間で共通する単語を削除する方法も試した.

実験の結果を表 5 に示す．表の段落正解率とは段落単位で求めた正解率であり，正しい選択肢と対応づけられた段落数を解答すべき段落の総数で割ったものである．表から，段落のベクトル化に際しては選択肢との類似度が一番高い文 (iv) を使い，選択肢のベクトル化については選択肢間の共通単語を削除

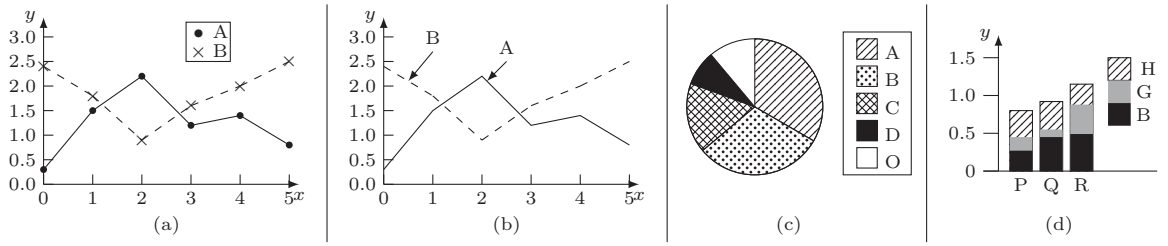

図 3: 色々なグラフ

段落の文選択	選択肢の共通単語	完答数 (42 問中)	段落正解率 (%)
iv	削除	19	67
全体	削除	17	64
全体	残す	17	62
iv	残す	16	68
iii	削除	16	61

表 5: 段落タイトル付与問題の実験結果

模試	2014 年 6 月	2014 年 9 月	2014 年 11 月
Kaldi (2015)	15.6%	13.0%	15.1%
Kaldi (2016)	7.7%	9.9%	8.7%
Google	5.7%	4.7%	6.1%

表 6: 各模試での音声認識率の比較 (単語誤り率).

したものが一番高い完答数となった．完答率は 45%とランダムな解答より有意に高いものの十分とは言えない．

誤り理由としては段落の表現と大きく異なる正解選択肢の類似度がさほど高くならないこと，正解でない選択肢に当該段落に含まれる単語（に類似した単語）が出現するとそちらの類似度が高くなることなどがあげられる．前者に対して特に段落のタイトルが属性名を含む時 (例：○○の歴史)，段落が当該属性に関するものかどうかを文書分類で推定する方法が考えられる [井内 17].

5. その他の問題

5.1 リスニング

リスニングに対して，本年度は，音声認識精度を向上させることを重視した．昨年の音声認識では，Kaldi に標準で添付されている Librispeech タスクの音声認識を利用した [Panayotov 15] が OCR によるエラーや対話データが少ないと言う問題があった．本年度は，これらの問題を改善するため，OCR エラーと思われる箇所に対する前処理を行い，さらに，対話データとして 24 億単語からなる opensubtitle データと 130 万単語からなる moviedic データ [Banchs 12] を学習データに加えた．この音声認識システムの評価のため，過去のリスニング問題に対する単語誤り率を比べた．また参考のため Google の音声認識結果も評価に加えた．表 6 に認識結果を示す．この結果，昨年の音声認識システムより，改善が見られた．

問題解法では，リスニング問題のうち，意見要旨把握や対話文完成問題などに変換できる問題だけを選び，他の章で説明している解法を用いて解答した．この枠組みでセンター模試を行った結果，14 点 (50 点満点) であった．受験者の全国平均は 26.3 点となっており，かなり低い正解率となった．リスニングでは，文が簡単であるにも関わらず,「植物が一週間で 15 ｃｍ伸びる」＝「成長が早い」という一般常識を問う問題が多いため，低い正解精度となっていると考えられる．このような問題は，データを増加させたからといって，必ず解ける問題とは限らない．

5.2 イラスト理解

リスニング問題の中で，例年 2 問程度，選択肢がイラストで示された出題がある．英語班では 2015 年度までは取り組めていなかったが，2016 年度に取り組みを開始した．紙面の関係で詳細は省くが，イラスト選択肢を英文テキストに変換することで，他のリスニング問題と同様に解くアプローチを検討している．本アプローチは，画像キャプショニングの手法と選択肢の差異に着目した説明文生成からなる．イラスト理解の学習データには，Manga109 データベース [Fujimoto 16] を利用している．現状の認識精度は 40%程度であり，今後さらなるイラストデータの追加と，問題文語彙とキャプション生成のより密な結合の両面から，イラスト問題への取り組みを行っていく．

謝辞

本研究を推進するにあたって，大学入試センター試験問題のデータをご提供下さった独立行政法人大学入試センターおよび株式会社ジェイシー教育研究所に感謝いたします．実験データをご提供くださいました学校法人高宮学園，株式会社ベネッセコーポレーションに感謝いたします．CommonCrawl データをご提供いただいた NTT コミュニケーション科学基礎研究所の鈴木潤氏に感謝します．

参考文献

[Banchs 12] Banchs, R. E.: Movie-DiC: A Movie Dialogue Corpus for Research and Development, in *Proc. ACL*, pp. 203–207 (2012)

[Fujimoto 16] Fujimoto, A., Ogawa, T., Yamamoto, K., Matsui, Y., Yamasaki, T., and Aizawa, K.: Manga109 Dataset and Creation of Metadata, in *Proc. Internaional workshop on coMics ANalysis, Processing and Understanding* (2016)

[Hermann 15] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P.: Teaching machines to read and comprehend, in *Proc. NIPS*, pp. 1693–1701 (2015)

[Joulin 16] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T.: Bag of tricks for efficient text classification, *arXiv preprint arXiv:1607.01759* (2016)

[Kusner 15] Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q.: From Word Embeddings To Document Distances, in *Proc. ICML*, Vol. 15, pp. 957–966 (2015)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013)

[Panayotov 15] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books, in *Proc. ICASSP*, pp. 5206–5210, IEEE (2015)

[Pang 04] Pang, B. and Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, in *Proc. ACL* (2004)

[Pershina 15] Pershina, M., He, Y., and Grishman, R.: Idiom Paraphrases: Seventh Heaven vs Cloud Nine, in *Proc. EMNLP Workshop on Linking Models of Lexical,Sentential and Discourse-level Semantics*, pp. 76–82 (2015)

[Sukhbaatar 15] Sukhbaatar, S., Weston, J., and Fergus, R.: End-to-end memory networks, in *Proc. NIPS*, pp. 2440–2448 (2015)

[Wang 16] Wang, Z., Mi, H., and Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition, *arXiv preprint arXiv:1602.07019* (2016)

[井内 17] 井内 健人, 菊井 玄一郎, 杉山 弘晃, 但馬 康宏：表現の類似性と文書分類を併用したセンター試験英語段落タイトル付与問題の解答手法, 第 23 回言語処理学会年次大会, p. to appear (2017)

[磯崎 15] 磯崎 秀樹, 伊藤 圭汰, 荒木 良元：論文 QA のための画像処理～表を読む～, 第 21 回言語処理学会年次大会 (2015)

[磯崎 16] 磯崎 秀樹, 小村 祐介：センター試験英語の計算文章題の自動解答器構築に向けて, 2016 年度人工知能学会全国大会 (2016)

[磯崎 17] 磯崎 秀樹, 中野 仁登, 浅川 護, 荒木 良元：論文 QA のための画像処理 ～ グラフを読む, 情報処理学会第 79 回全国大会 (2017)

[尭天 16] 尭天 貴之, 植田 佳文, 東中 竜一郎, 杉山 弘晃, 平 博順：述語項構造解析を用いた英語長文読解問題の自動解法, 第 22 回言語処理学会年次大会, C3-3 (2016)

[中野 17] 中野 仁登, 磯崎 秀樹：英語センター試験を自動で解くための棒グラフの自動読み取り, 2017 年度人工知能学会全国大会 (2017)

[堂坂 16] 堂坂 浩二, 坂本 祐磨, 高瀬 惇：隣接発話らしさを利用した英語会話文完成問題の回答手法, 2017 年度人工知能学会全国大会, 1K3-4 (2016)