CE-012

講演音声認識の言語モデル適応のための Doc2vec によるフィルタリングを活用した自動コーパス構築

Automatic Corpus Construction with Doc2vec Filtering for Language Model Adaptation Towards Lecture Speech Recognition

和田 蒼汰[1)]　早川 大智[1)]　岩田 憲治[1)]
Sota Wada　Daichi Hayakawa　Kenji Iwata

1　はじめに

近年，ディープラーニング技術の導入により，音声認識の性能は飛躍的に向上し，会議やオンライン授業の字幕システムとして利用されるなど，一般的な技術として普及しつつある．音声認識システムでは，通常，大容量のテキストコーパスから作成された汎用言語モデルが用いられる．汎用言語モデルは，一般的に用いられる単語に対しては，適切に単語同士のつながりの傾向を表現できる．しかし，会議やオンライン授業など音声認識システムを活用する現場では，特定の分野や業界でしか使用されない言い回しや専門用語が頻出する．そのような言い回しや専門用語は，大容量コーパスに含まれていない場合があるため，汎用言語モデルではカバーできず，認識性能の劣化を引き起こすことがある．

特定のドメインに対する認識性能を向上させる方法の一つとして，そのドメイン特有の言い回しや専門用語を含むコーパス（以下，ドメインコーパス）で言語モデルを適応する方法が考えられる [1, 2]．例えば，ドメインを大学の数学の講義とする音声認識を考えた場合，講義の音声を書き起こしたテキストから言語モデルを学習することで，数学の証明などのドメイン特有の言い回しや，数学の用語などの専門用語を含む音声に対して，高い認識性能が期待できる．この方法を実現するためには，十分な量のドメインコーパスを用意しなければならない．しかし，講義の音声を書き起こす作業など，手作業によるドメインコーパスの収集は時間コストが大きい．従って，ドメインコーパスを自動収集する枠組みの構築が必要となる．

ドメインコーパスを自動収集する方法の一つとして，講演のスライド資料などのドメインに関する資料から抽出した特徴語に基づき，Web テキストを収集する技術がある [3]．文献 [3] では，Web から収集したテキストに対して汎用言語モデルによりパープレキシティを測定し，この値が小さい文を選択することにより，話し言葉や文章のスタイルになっていないもの（単語の羅列など）をフィルタリングする．しかし，この手法では特定のドメインと関係のある文章かどうかを考慮したフィルタリングはできず，ドメインコーパスの構築を効率的に行えない可能性がある．

本稿では，収集した Web テキストから，ドメインと関係のないテキストをフィルタリングすることにより，ドメインコーパスを効率的に構築する手法を提案する．具体的には，ドメインに関する資料と構築したコーパス内のテキストを Doc2vec[4] を用いて固定長のベクトルに変換し，cos 類似度を測定することで対象のドメインとテキストとの類似度を求め，類似度が低いテキストをフィルタリングする．12 種の講演音声で性能を評価した結果，パープレキシティに基づくフィルタリングを用いる場合と比較して，音声認識精度と専門用語の認識精度が向上した．

1）　東芝 研究開発センター

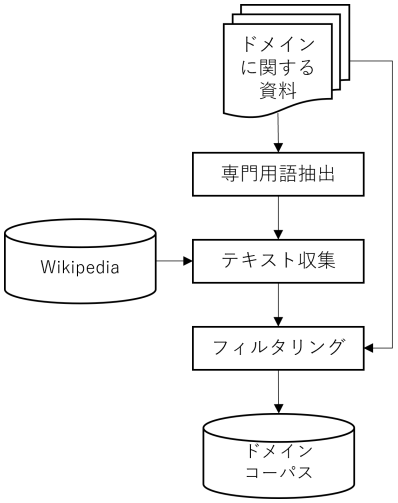

図 1　ドメインに関する資料からドメインコーパスを構築する手順

2　ドメインに関する資料からドメインコーパスを構築する手順

本稿では，ドメインに関する資料から専門用語を抽出し，専門用語を含むテキストを Web から十分な数だけ収集することで，ドメインコーパスを構築する．

2.1　全体像

ドメインに関する資料からドメインコーパスを構築する手順の全体像を図 1 に示す．手順は以下の 3 ステップに分けられる．

1. ドメインに関する資料から専門用語を抽出
2. 専門用語を含む Wikipedia 記事を収集
3. 専門用語とドメインに関する資料に基づき，収集された Wikipedia 記事からドメインと関係のない内容の記事をフィルタリングし，ドメインコーパスを構築

次節以降にて，各ステップの詳細を説明する．

2.2　専門用語抽出

ドメインに関する資料から専門用語を抽出する．本稿では，以下に述べる一定の基準を満たす単語列を専門用語とする．一定の基準とは，C-value[5]，汎用言語モデルを使用したパープレキシティ，誤認識のしやすさ度の 3 つの指標から構成される．以下，それぞれの指標について詳細を説明する．

2.2.1 C-value

C-value は，ドメインに関する資料に含まれるコロケーション（連続した単語列）のうち，どのコロケーションが高い重要度を持つかを判定する指標の一つであり，あるコロケーション a の C-value は次式で定義される．

$$\text{C-value} = \begin{cases} 0, & \text{if } n(a) = n(a') \\ (|a|-1) * n(a), & \text{if } c(a) = 0 \\ (|a|-1) * \left(n(a) - \frac{t(a)}{c(a)}\right), & \text{otherwise} \end{cases} \quad (1)$$

ここで，$|a|$ は a の構成要素単語数，$n(a)$ は a の出現頻度，$t(a)$ は a を内包するコロケーションの出現頻度の合計，$c(a)$ は a を内包するコロケーションの種類数である．これらは，以下のような指標で単語列 a の専門用語性を判定しているということになる．

- a の構成要素単語数が多いと専門用語性が高い
- a の出現頻度が高いと専門用語性が高い
- a を内包する単語列の出現頻度が高く，それらの単語列の種類数が少ないと専門用語性が低い

2.2.2 汎用言語モデルを使用したパープレキシティ

パープレキシティは，汎用コーパスから作成された汎用言語モデルを用いて，次式で求めることができる．

$$log_2 PP = -\frac{1}{N} log_2 P(w_1, w_2, \cdots, w_N) \quad (2)$$

ここで，PP はパープレキシティ，N は専門用語の構成形態素数，$P(w_1, w_2, \cdots, w_N)$ は形態素列 $w_1, w_2, \cdots, w_N$ のモデル中での出現確率である．一般に汎用言語モデルのコーパスに頻出する表現であれば，パープレキシティは小さくなり，モデルに出現することが少ない表現はパープレキシティが大きくなる．つまり，パープレキシティが大きい用語（形態素列）は，一般的な文書で使われることが少なく，専門用語性が高いということになる．

2.2.3 誤認識のしやすさ度

誤認識のしやすさ度とは，音声認識システムが誤認識する可能性が高い単語列を抽出する指標である．誤認識のしやすさ度の測定方法を以下に示す．

1. 漢字仮名交じりのドメインに関する資料を読み仮名列に変換
2. 読み仮名列を入力し，音声認識エンジンがどのような単語列を出力するかを推定．推定には，文献 [6] の手法を用いた．
3. 音声認識エンジンが出力した単語列とドメインに関する元文書の形態素列を比較し，差分を検出することで，誤認識しやすい形態素列（差分）を元文書から抽出
4. 差分が検出された元文書の形態素列，文書中でその形態素列が誤認識された回数，閾値以上の C-value が算出された単語列に基づき，次式で誤認識のしやすさのスコア AEscore を算出

$$\text{AEscore}(w) = \sum_x \text{score}(w, x) \quad (3)$$

$$\text{score}(w, x) = \begin{cases} \text{counts}(x), & \text{if } w = x \\ \text{counts}(x) * \frac{len(w)}{len(x)}, & \text{if } w \subset x \\ \text{counts}(x) * \frac{len(sub(w))}{len(w)} * \frac{len(sub(w))}{len(x)}, & \text{if } sub(w) \subset x \end{cases} \quad (4)$$

ここで，w は閾値以上の C-value が算出された単語列，x は差分が検出された元文書の形態素列，counts(x) は形態素列 x が文書中で誤認識された回数，sub(x) は形態素列 x の部分形態素列，len(x) は形態素列 x の文字列長である．つまり，閾値以上の C-value が算出された単語列のうち，誤認識しやすい形態素列と一致する部分が多い単語ほど，その単語の誤認識のしやすさ度のスコアは大きくなる．

2.2.4 3 つの指標を用いた専門用語の抽出方法

以下，3 つの指標を用いてどのように専門用語を抽出するかについて説明する．まず，ドメインに関する資料を形態素単位へ分割し，単語列のみ抽出する．次に，単語ごとに C-value を測定し，閾値以上の C-value を持つ単語列（以下，専門用語候補）を抽出し，専門用語候補のパープレキシティと誤認識のしやすさ度を算出する．最後に，専門用語候補を C-value，パープレキシティ，誤認識のしやすさ度の 3 つスコアを約 5.0:3.5:1.4 の比率で重みづけ和したスコアを用いてソートし，上位の M（M は 1 以上の整数）単語を専門用語として出力する．

2.3 専門用語を含む Wikipedia 記事の収集

2.2 節に示した手法で専門用語を抽出した後，専門用語を含むテキストを Web から収集する．本稿では，MediaWiki[7] を用いて，Wikipedia から専門用語に関する記事を検索することで，専門用語および専門用語の構成語の一部を含むテキストを収集した．

収集された Web テキストには，タグや記号が含まれていたり，平仮名，片仮名，もしくは漢字のいずれかの割合が低いものが含まれることがある．これらは，言語モデルの学習に不適切なため，あらかじめ排除した．

2.4 Doc2vec によるフィルタリング

2.3 節の手法により収集されたテキストには，ドメインと関係のない内容のものが含まれる場合がある．例えば，情報処理学会の講演資料から抽出された専門用語「プログラミング教育指導助手養成」を含む記事を収集した時，スーパー戦隊シリーズの「獣拳戦隊ゲキレンジャー」の記事が収集された．この記事は明らかにドメインと関係のない内容のものであり，ドメインコーパスには含まれるべきではない．そこで，専門用語とドメインに関する資料に基づき，収集された Wikipedia 記事からドメインと関係のない内容の記事をフィルタリングする．本稿では，フィルタリング手法として Doc2Vec[4] によるフィルタリングを提案する．以下にその詳細を示す．

Doc2vec とは，任意の長さの文章を固定長のベクトルに変換する技術である．Doc2vec によるフィルタリングでは，まず 2.3 節に示した手法で収集したテキストとドメインに関する資料を学習データとして Doc2vec を学習する．次に，学習したモデルを用いて記事とドメインに

関する資料を固定長のベクトルに変換して，cos 類似度を測定し，類似度の高い上位 N(N は 1 以上の整数) 記事をドメインコーパスとして採用する．

3 評価実験

3.1 実験データと条件

12 種の情報処理学会の講演を対象に評価実験を行った．各講演の時間は約 20 分から 90 分であり，合計 7 時間である．講演資料それぞれに含まれる単語数は 0.2K から 3.7K である．

使用した音響モデルは，CSJ[8] のフルデータを学習データとして用い，音節を基本単位として Connectionist Temporal Classification(CTC)[9] により学習されたものを用いた．詳細については文献 [10] を参照されたい．汎用言語モデルは，CSJ のフルデータ（7.8M 単語）から学習した 4-gram モデルを用いた．

ドメインコーパスは，講演ごとに専門用語を抽出し，それらが含まれる記事を 100 記事ずつ収集した上で，Doc2vec によるフィルタリングを適用することで構築した．Doc2vec によるフィルタリングでは，講演資料と類似度の高い上位 10 記事をドメインコーパスとして採用し，Python ライブラリの gensim[11] を用いてモデルを学習した．この時，特徴ベクトルの次元数は 400 とし，学習アルゴリズムには PV-DM を指定した上で，100 エポック学習した．その他のパラメータは gensim のデフォルト値を用いた．

比較手法としては，先行研究より，パープレキシティに基づくフィルタリング [3] を用いた．パープレキシティに基づくフィルタリングとは，汎用言語モデルにより，収集したテキストのパープレキシティを計算し，この値が小さい文を選択することで，話し言葉や文章のスタイルになっていないもの（単語の羅列など）をフィルタリングする手法である．本実験では，2.3 節に示した手法で収集したテキストを基に，パープレキシティに基づくフィルタリングによって構築されるコーパスと Doc2vec によるフィルタリングによって構築されたコーパスの単語数が同等程度になるように，パープレキシティが低いテキストから順にドメインコーパスとして採用した．

ドメイン言語モデルは，講演ごとのドメインコーパスを用いて 4-gram を計 12 個作成した．それらを汎用言語モデルと重み付け和し，WFST 変換した．重み付け和の重みは，汎用言語モデル：ドメイン言語モデル = 7 : 3 の割合になるように指定した．

音声認識の評価指標には，文字誤り率（Character Error Rate, CER）を用いた．文字誤り率は，以下のように定義される．

$$\mathrm{CER}(\%) = \frac{D+S+I}{N} * 100(\%) \tag{5}$$

ここで，N は正解文に含まれる文字数，D, S, I はそれぞれ，脱落誤り（Deletion error），置換誤り（Substitution error），挿入誤り（Insertion error）の文字数である．本稿では，できる限り多くの表記ゆれをカバーするように正解データを作成し，表記ゆれによる CER の劣化が起きないようにした．また，専門用語の認識性能は以下の手順で評価した．まず，資料から全単語を抽出し，クラウドソーシングにより専門用語であると判定された単語をオラクル専門用語として抽出した．次に，認識結果と正解文を比較し，オラクル専門用語が正しく出力されているかを F 値で評価した．

3.2 評価結果

音声認識性能と専門用語認識性能の 12 講演平均の結果、各講演での結果をそれぞれ表 1、表 2 に示す．各手法は以下の通りである．

- Baseline：CSJ で作成した汎用言語モデル
- PPL：パープレキシティに基づくフィルタリングを用いて作成した適応言語モデル
- D2V：Doc2vec によるフィルタリングを用いて作成した適応言語モデル
- D2V+PPL：Doc2vec によるフィルタリングを用いてフィルタリングした記事を基に，更にパープレキシティに基づくフィルタリングを用いて，パープレキシティが低い上位 90%のテキストを抽出することで構築したコーパスから作成した適応言語モデル

まず，表 1 より，Baseline と PPL, D2V を比較すると，CER, F 値共に改善していることから，ドメインコーパスで言語モデルを適応することで，そのドメインに対する認識性能が向上したことがわかる．また，PPL と D2V を比較すると，12 講演の平均の CER, F 値はどちらも D2V の方が良い結果となったことがわかる．

続いて，表 2 より，講演ごとの評価結果は，多くの講演において D2V の方が CER, F 値共に良かったことがわかる．ここで，それぞれの手法の講演 3 に対する音声認識例を表 3 に示す．講演 3 は，アジャイルソフトウエア開発に関する講演であり，講演資料から専門用語として「アジャイル向け」や「コスト管理」が抽出された．また，講演資料には，「顧客との契約交渉において重視すべき項目」が含まれていた．表 3 の上段は，専門用語「アジャイル向け」と「コスト管理」を含む音声に対する音声認識例であり，表 3 の下段は，「顧客満足」というドメインに関する単語 (以下，ドメイン単語) を含む音声に対する音声認識例である．結果，PPL はどちらも正解表記を出力できなかったが，D2V は正しく出力できたことがわかる．

一方，表 2 より，一部の講演は PPL の方が CER, F 値が良かったことがわかる．認識誤りの種類を分析したところ，例えば講演 2 には、表 4 のような置換誤りが多かった．また，講演 2 は，博士課程や研究に関するドメインの講演であることから，Doc2vec によるフィルタリングで構築されたコーパスには学者や研究物の記事が多く含まれていた．これら研究関連のドメイン単語は，「ボルクム島灯台海岸局」のように固有名詞と紐づき複合名詞として使われることが多く，それを含む文はパープレキシティが大きくなりやすいため，パープレキシ

表 1 認識性能評価における 12 講演の平均結果

	CER(%)	F 値 (%)
Baseline	16.44	66.2
PPL	15.20	74.5
D2V	**14.69**	**77.7**
D2V+PPL	14.73	77.6

表 2　認識性能評価における講演それぞれの結果

	PPL		D2V		D2V+PPL	
	CER (%)	F 値 (%)	CER (%)	F 値 (%)	CER (%)	F 値 (%)
講演 1	**9.29**	**87.1**	9.79	84.9	9.82	84.9
講演 2	**10.04**	87.0	10.31	**87.3**	10.31	**87.3**
講演 3	17.30	64.3	**15.34**	**75.9**	15.73	73.3
講演 4	16.14	50.0	14.93	**66.3**	**14.90**	**66.3**
講演 5	12.86	70.4	12.68	**76.5**	**12.67**	**76.5**
講演 6	20.48	75.6	**20.46**	**77.2**	20.48	**77.2**
講演 7	17.51	81.1	**17.13**	**83.8**	17.16	**83.8**
講演 8	10.50	87.9	**9.11**	**92.9**	9.13	**92.9**
講演 9	15.34	61.7	14.90	**70.8**	**14.84**	**70.8**
講演 10	14.42	**81.0**	**14.23**	80.2	14.32	80.2
講演 11	22.71	61.7	**21.62**	61.0	21.65	**62.1**
講演 12	**15.77**	**86.7**	15.81	75.9	15.79	75.9

表 3　PPL と D2V の音声認識例

(正解表記)	アジャイル向けのスケジュールコスト管理
(PPL)	はじゃえ MK のスケジュールコスト感る
(D2V)	アジャイル向けのスケジュールコスト管理
(正解表記)	における顧客満足の実現方法を
(PPL)	における各蛮族の実現方法を
(D2V)	における顧客満足の実現方法を

ティに基づくフィルタリングで構築したコーパスには出現する回数が少なく，PPL の認識結果には出力されなかったと考えられる．

従って，Dov2vec によるフィルタリングは，パープレキシティに基づくフィルタリングよりも，専門用語やドメイン単語を含むテキストをドメインコーパスに多く含めることができるが，認識時におけるそれらの出現確率は別途調節を行う必要があるなど，改善の余地があることがわかった．この原因としては，Doc2vec によるフィルタリングは，Wikipedia の記事単位でフィルタリングしていることが挙げられるため，文単位でフィルタリングすることで改善される可能性がある．

最後に，D2V と D2V+PPL を比較すると，CER, F 値共に平均値はほぼ同じだったが，講演 3 のみ D2V+PPL の方が大幅に性能が低いことがわかる．認識結果を分析した結果、D2V+PPL は「ウォーターフォール」と「フィボナッチ」という専門用語の認識精度が低かった．ここで，パープレキシティに基づくフィルタリングによって構築されたコーパスを分析したところ，これらの専門用語を含む文は，パープレキシティが高かったことでフィルタリングされ，ドメインコーパスから除外されていた．特に，「ウォーターフォール」は 1 単語が 1 文となっているものが多く，1 単語のみの文はパープレキシティが大きくなりやすいことから，多くのテキストがフィルタリングされたことで，適応言語モデルの出現確率が下がり，認識結果が悪化したと考えられる．つまり，パープレキシティに基づくフィルタリングでは，ドメインとは関係のある単語だが，文ではなく単語列で出現したものをドメインコーパスに含むことができない場合があることがわかった．

表 4　D2V の認識誤り例

正解表記	D2V の誤認識例
東大	灯台
文科省	も歌唱法
インターンシップ	IN 端子

4　終わりに

本稿では，収集した Web テキストから，ドメインと関係のないテキストをフィルタリングする手法を提案した．具体的には，ドメインに関する資料と収集した Web テキストを Doc2vec を用いて固定長のベクトルに変換し、cos 類似度を測定することで対象のドメインとテキストとの類似度を求め、類似度が低いテキストをフィルタリングする手法を提案した．12 種の情報処理学会の講演を対象に評価実験を行った結果，パープレキシティに基づくフィルタリングを用いる場合と比較して，音声認識精度と専門用語の認識精度が向上した．今後は，より大規模のテキストコーパスを用いた言語モデル適応や Doc2vec によるフィルタリングを文単位で行うといった対策を検討する．

参考文献

[1] A. Park, T. J. Hazen, J. R. Glass, “Automatic processing of audio lectures for information retrieval: vocabulary selection and language modeling,” in 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2005, pp. 497-500

[2] I. Trancoso, R. Nunes, L. Neves, C. Vianan, H. Moniz, D. Caseiro, A. Isabel, Mata, “Recognition of Classroom Lectures in European Portuguese,” In Proc. Interspeech, 2006, pp. 281–284.

[3] 河原 達也, 根本 雄介, 勝丸 徳浩, 秋田 裕哉, “スライド情報を用いた言語モデル適応による講義音声認識,” 情報処理学会論文誌, Vol.50, No.2 (2009).

[4] Quoc V. Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents,” arxiv:1405.4053, 2014.

[5] K. T. Franzi, S. Ananiadou, “Extracting nested collocations,” in Proc. COLING, 1996, pp. 41-46.

[6] 藤村 浩司, “音声認識結果出力装置、音声認識結果出力方法及び音声認識結果出力プログラム,” 特許第 6580882 号, 2017.

[7] https://www.mediawiki.org/wiki/MediaWiki/ja (2022 年 6 月 23 日閲覧)

[8] K. Maekawa, H. Koiso, S. Furui, et al., “Spontaneous speech corpus of Japanese,” in Proc. LREC, 2000, pp. 947–952.

[9] Alex Graves, Santiago Fernandez, Faustino Gomez and Jur-gen Schmidhuber, ”Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proc. ICML, pp. 369-376, 2006.

[10] H. Fujimura, M. Nagao and T. Masuko, ”Simultaneous Speech Recognition and Acoustic Event Detection Using an LSTM-CTC Acoustic Model and a WFST Decoder,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5834-5838

[11] https://radimrehurek.com/gensim/ (2022 年 6 月 23 日閲覧)