能動学習と局所対話への特化を考慮した対話文における言い淀み検出

中島 寛人[1]　嶋田 和孝[2]

概要：話し言葉に含まれる言い淀み表現は，可読性を下げるのみならず，機械学習モデルで話し言葉を使用する際のノイズにもなり得る．したがって，事前に自動的な整形を行うことが望ましい．既存の研究では，深層学習モデルをファインチューニングによってタスクに適応させるものが多い．通常，ファインチューニングでは，タスクへの汎用的な学習のみを行う．したがって，方言話者を含む対話のように，特定の 1 対話単位に存在する局所的傾向にも頑健なモデルを作成することは難しい．そこで，本論文では能動学習の考え方に着目する．対象の 1 対話内から追加の学習データを獲得することで，局所的傾向に特化した検出を試みる．

キーワード：冗長表現, 言い淀み, 能動学習

Dialogue Disfluency Detection with Active Learning based on Local Tendency

Abstract: Disfluency in dialogue not only decreases the readability of the transcript but also introduces noise in Machine Learning data. Recent works in the disfluency detection task use deep-neural models fine-tuned by general knowledge about the task. However, dialogue sometimes includes local tendencies like dialects or specialized terminology. Such tendencies may negatively affect the detection by models with general knowledge. In this paper, we propose disfluency detection with Active Learning (AL) to specialize in local tendencies. We also discuss the effectiveness of different sampling methods in this task.

Keywords: Redundant expressions, Disfluency, Active Learning

1. はじめに

論文のような書き言葉と異なり，実際の人の発話によって構成される話し言葉には，特有の冗長な表現が多く含まれている．冗長表現の中でも代表的なものが言い淀み表現である．「やす，安くて」のような，「安くて」を途中で一度中断して繰り返す言い直し，もしくは「こと、今年度」のように「今年 (こと<u>し</u>)」を言いかけて「今年度 (こ<u>んね</u>んど)」と別の言葉に変える言い換えなどが言い淀みに内包されるものである．これらは基本的に文意に関係しないため，単に読みづらいだけでなく機械学習に使用される場合にもノイズとして影響を与えることがある．したがって，言い淀み表現の自動的な検出や整形は可読性・応用性の両観点から必要不可欠である．

言い淀みの検出は系列ラベリングとして解かれることが多く，固有表現認識 (NER: Named Entity Recognition) タスクに類似した枠組みが用いられている．近年では，深層学習モデルを使用した研究が主流となっており，ファインチューニングによって従来より高い精度で検出を行うことが可能になっている．しかし，専門用語や方言といった頻出しない単語を含む対話や，珍しいテーマで行われている対話にも頑健なモデルを作成することは難しい．例えば以下のような事例を考える．

- …これはきゅう、九で、九州の九ですね。でその次が

[1] 九州工業大学 大学院情報工学府
Department of Creative Informatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[2] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN

こう、こうは工事の工で…

この事例では，一見中断や繰り返しが多数発生しており，言い直しが何度も生じているように考えられる．しかし，実際にはこの対話では漢字の書き取りが行われており，この発話内に言い淀み表現は存在しない．このような，レアケースへの過度な適合はモデルの汎化性能を下げ，結果的に精度が低下することもある．しかし，レアケースをターゲットとする場合に 1 からモデルを訓練・ファインチューニングすることは，学習そのものやデータセット作成などにかかるコストの面から現実的ではない．

本研究では，このような問題を解決するために能動学習 (Active Learning) の考え方に着目する．能動学習では，大規模な元データのすべてに対して正解ラベルのアノテーションを付与する代わりに，モデルに予測させる中で人間に一部データのアノテーションを要求し，再度モデルを学習する．これにより，大規模データの中から効率的に学習可能なデータのみを選出することで，アノテーションにかかるコストを削減できる．つまり，言い淀みを検出したい対話から小規模な事例を学習データとして選出し，モデルを追加学習させることで，検出対象である対話における検出精度を向上できる可能性がある．

本論文では，タスクにファインチューニングされたモデルを更に能動学習によって追加学習することで，検出対象のデータに特化させる手法を提案する．実験では，提案手法の有無による検出精度の変化や，能動学習における訓練データ選出手法による精度の差について論じる．

2. 関連研究

本節では，本論文のタスクである言い淀み検出と能動学習について関連する研究を中心に述べる．

2.1 言い淀み検出

言い淀み表現の機械的な検出においては様々な研究がなされてきた．言い淀みの出現に周囲の文脈が関連していることは，その中でも多くの研究で論じられている．この周囲文脈に関する研究の一例として，RIM(Repair-Interval-Model) [1] が存在する．言い淀みの一種である自己修復 (所謂言い直しに近い表現) が出現する際に特定の文法的構造を伴うとされるもので，同じく冗長表現のひとつである，「えー」「あー」などで表されるフィラーとの関わりも示唆されている．文脈情報を用いた言い淀み検出についても尾嶋ら [2] や Tanaka ら [3] など多くの研究が存在する．文脈情報を考慮できる深層学習モデルの登場後はこれを用いた検出が盛んになっており [4]，生成モデルによる仮文脈の有効性 [5] なども示されている．自身の先行研究 [6] においては，前後文脈の補完によって文長の短い対話文における精度向上を示し，GPT2 [7] による補完についても検証した．本論文では，文脈情報を直接的に補完する手法は用いられていないが，能動学習時の訓練データは特化する特定の 1 対話内にある文脈そのものであるため，その対話内における言い淀みの出現傾向のような事実上の文脈情報を学習することを期待している．

2.2 能動学習

能動学習は NER タスクでも有効性が示されてきた [8] 手法である．能動学習において基本的な論点となるのは，「アノテータに対してどの事例を提示するか」である．Query By Commitee (QBC) [9] は Seung らによって提案された，複数のモデルによる合議が最も難しかった事例を選択する手法である．このような高難度事例を発見する手法は難易度の高い事例を発見できる一方で，実際にはアノテータに負担を預けている可能性が存在する．Gabriel ら [10] は簡単な一部の事例について提示の際は二値分類タスクにすることで，アノテータの負担軽減を行っている．BERT をはじめとして RoBERTa [11]，DistillBERT [12] など複数の深層学習モデルからなる Committee によって精度を改善しつつ 20 %弱の労力削減を達成している．また，Kile ら [13] はエンティティに対するラベル付けを行う NER タスクで，自動推測機能を組み込んだ能動学習によってアノテーションに掛かる時間的コストを削減している．このように，能動学習においてはアノテータの負担軽減も論点となることがある．系列ラベリングはタスクとしては比較的複雑なため，アノテータの負担は重要な要素であると言える．本論文では，選出してアノテータに提示する能動学習用訓練データをごく小さな規模に留めることで，アノテータの量的な負担を軽減するアプローチを取った．

一方で，QBC は前提として複数のモデルを要求するため，先述のような大きなモデルを複数同時に動かす必要があり動作コストの面では望ましくない場面も存在する．最小確信度スコアは最も基本的な不確実性サンプリング手法の 1 つで，系列ラベリングにおいても研究が存在する [14]．本論文では，NER タスクに類似したタスクである言い淀み検出タスクにおいて，最小確信度スコアに基づいた能動学習について検証する．

3. データセット

本節では，本論文で使用するデータセットについて説明する．本論文では，対話文のデータセットとして話し言葉コーパスの書き起こし文を使用する．『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese，以降 CSJ) [15] は，日本語で収録された講演や発表などの独話データ，インタビューや自由対話などの対話データを含む話し言葉コーパスである．各データには，対応する音声データの他に，音声を基にして人手で書き起こしを行った転記テキストがテキストデータとして収録されている．本論文では，対話文にあたる全 58 対話に対応する転記テキストをデー

表 1　CSJ に付与されているタグの一例．本論文で使用する言い淀み (D タグ) は言い淀んでいる部分のみを範囲としている．

タグ	例文
言い淀み	…原型と変形の **(D しょうちゅし)** 聴取時間の差…

表 2　前処理後の CSJ 対話データの一例 (ただし，話者交替を明示するために話者 ID を，結合位置を明示するために IPU 単位での区切りだった位置に/(スラッシュ) を付与している)．

ID	発話
A	気質なのか
B	何だろう
A	結構 (D ん)(D き) 多分アメリカ人気質に似てるような
B	(F あ) / (D なん) そうかもしれないですね
A	(F うーん) 気がする (D う)
B	(F うん) / (F うん) 凄い主張するし
A	主張が激しい

表 3　前処理後の CSJ 対話データ内の発話内訳．発話内で複数回言い淀みが発生している場合があるため，言い淀みを含む発話の数と言い淀みの総数は異なることに注意．

全発話	19551
言い淀みを含む発話	2147
言い淀みの総数	2629

タセットとして使用する．転記テキスト中では，フィラーや言い淀みなど，話し言葉で発生する様々な言語現象に対してラベリングが行われている．本論文では，表 1 に示すような，言い淀みに対応する D ラベルの情報を基にして正解ラベルを作成する．

CSJ では発話毎の区切りに「200 ミリ秒以上のポーズ」を基準とした転記基本単位 (IPU：Inter-Pausal Unit) を使用している．しかし，この区切り単位での発話長は短くなっており，深層学習モデルにおいて極端に短い入力では文脈情報を活かしにくい．したがって今回は転記基本単位を使用せず，「話者が交代するか，終止形文末が出現するまで」を基準として同一話者による連続した発話を結合する前処理を行っている．前処理後の対話データの例を表 2 に，前処理後の対話データ全体における発話数と言い淀みを含む発話数を表 3 に示す．

4. 提案手法

本節では，まず本論文で提案する手法の概要を説明し，続く各節で手法内で使用される事前学習モデルや学習法について説明する．

提案手法の概要を図 1 に示す．提案手法では 2 段階に分けてモデルの学習を行うことで，特定の対話へ特化した言い淀み表現の検出を試みる．1 つ目は事前学習済みモデルを言い淀みタスクに適応させるファインチューニング段階で，言い淀みを含む事例群の学習によって言い淀み検出モデルを作成する．2 つ目は言い淀み検出モデルを更に検出対象に特化させる能動学習段階で，検出するデータから事

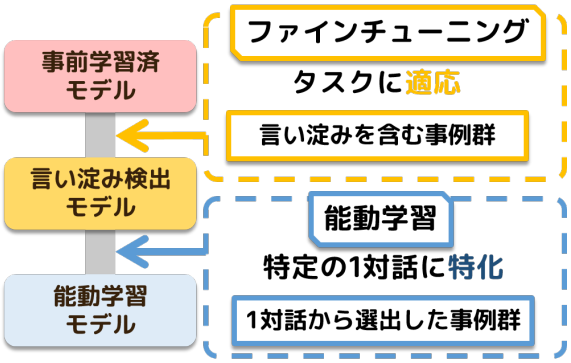

図 1　提案手法の概要．ファインチューニングによってタスクに適応したモデルを，能動学習によって更に検出対象となるデータに特化させる．

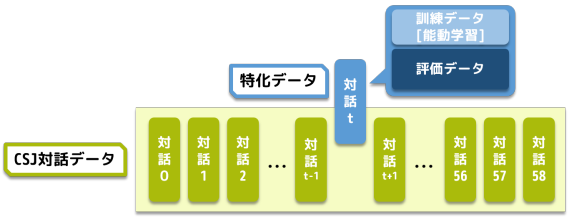

図 2　提案手法における特化対象．データセットに含まれる対話の中から任意の 1 対話を選出し，特化データとする．

例を選出して追加で学習を行い，能動学習モデルを作成する．なお，本論文における検出対象とは，図 2 に示すように，3 節で示したデータセット全 58 対話の中から 1 対話選出されたものである．したがって，能動学習によって学習を期待するのは，対話の話題や参加話者の言い淀み方といった，対話内に存在する大きな傾向である．

4.1 言い淀み検出モデル：BERT

本論文では, BERT をベースとして言い淀み検出モデルを作成する．BERT(Bidirectional Encoder Representations from Transformers) [16] はファインチューニングを行うことで様々なタスクに適応可能な汎用言語モデルであり，自然言語処理に関する多くのタスクで活用されている．本論文では，東北大学によって公開されている BERT モデル[*1]を事前学習済みモデルとして使用する．日本語 Wikipedia から抽出された約 2.6GB 相当のテキストデータによって学習が行われているモデルであり，日本語の汎用的な知識を事前学習している．

ファインチューニングでは，事前学習済み BERT を系列ラベリングによる言い淀み検出タスクに適応させ，言い淀み検出モデルとする．言い淀み検出モデルの概要図を図 3 に示す．ファインチューニング後のモデルでは，トークンと呼ばれる単語のような小単位で区切られた文章を入力として，各トークンにおける BERT の出力ベクトルに softmax, argmax の各関数を適用したものを推測ラベル列

[*1] https://huggingface.co/cl-tohoku/bert-base-japanese

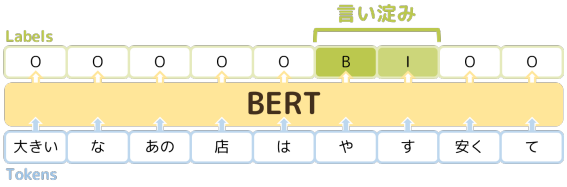

図 3 ファインチューニング後の言い淀み検出 BERT モデル．各トークンに対して IOB2 形式のラベルを付与し，言い淀みのチャンクを推測する．

とする．

正解ラベルには，系列ラベリングによる言い淀み検出で多く用いられているチャンクタグによるラベルを採用した．チャンクタグでは複数のトークンにまたがる事例を考慮してラベルを付与するため，単純に各トークンそれぞれについて言い淀み表現であるかないかを判定する場合より系列情報を考慮しやすい．本論文では，チャンクの先頭に B タグ，先頭以外のチャンク範囲に I タグ，チャンクに属さない部分に O タグを付与する，IOB2 (Inside-Outside-Beginning) 方式によってラベリングを行う．図 3 に示す例では，B タグの付与された「や」と I タグの付与された「す」の部分がまとめて一つの言い淀みチャンク「やす」として検出されている．

4.2 能動学習モデル

能動学習では，4.1 節で作成したモデルに対して，検出対象となる特化データ内の発話の中から小規模な事例を選出して能動学習用データとし，特定の 1 対話に特化する学習を行う．この際，通常テストデータの正解ラベルを知ることはできないため，人間のアノテータに選出したデータを提示し，付与されたアノテーションを正解ラベルとして使用する．

能動学習ではアノテータを経由してラベリングを行うため，アノテータの作業コストの観点から提示する事例群は可能な限り小規模となることが望ましい．また，小規模なデータでモデルの学習を行う観点では，各事例が「検出対象に特化する学習」のためにより効率的なものであることが望ましい．このような条件を満たす，「良質な」能動学習用データを選出する戦略は多岐にわたる．ランダムサンプリング，不確実性サンプリングなどが基本的な戦略群である．

不確実性サンプリングは，機械学習モデルが推測に悩んだ際の予測確率の混乱に着目しており，「モデルがデータ内で最も推測を困難としている事例が最も効率的な訓練データである」という考え方による戦略である．本論文で用いるのは不確実性サンプリングの中でも一般的な，最小確信度 (Least Confidence) に基づく選出法である．予測毎の最小確信度スコアは，モデルの出力ベクトルを p として式 1 で表され，100 %の確信度 (1.00) とモデルが推測したラベルの予測確率との差分である．

$$LCScore = 1.00 - max(softmax(p)) \quad (1)$$

最小確信度スコアはモデルが推測したラベルの予測確率が他ラベルより飛び抜けているほど小さく，また各ラベルの予測確率が拮抗しているほど大きくなる．したがって，最小確信度スコアの大きい順にデータを選ぶことで，モデルの学習にとって効率的なデータを抽出できる．

文章分類のように，ひとつの事例に対してひとつのラベルがモデルに推測されるタスクでは，事例全体の最小確信度スコアはモデルの推測ラベルを用いた最小確信度スコアと同等であると見なせる．しかし，本論文のタスクである系列ラベリングにおいては，モデルは各トークンに対してそれぞれラベルを推測する．つまり，ひとつの事例に対して入力されたトークン列の長さと同数のラベルが推測される．したがって，各トークンが最小確信度スコアを持っており，事例全体の最小確信度スコアを決定するための手法が必要である．以下では，本研究で使用する各スコア決定手法について取り上げる．各手法が事例全体の最小確信度スコアを決定する過程を図 4 に示す．

事例内平均値による最小確信度スコア
この手法では，トークン毎の最小確信度スコアを平均したものが事例全体の最小確信度スコアとなる．平均を取ることによって事例全体の混乱度合いを見ることができ，外れ値的な 1 トークンに事例全体のスコア決定を左右されない．したがって，事例全体を満遍なく考慮するという側面では，対話中にある大きな傾向を学習するという本論文での目的に近い選出を期待される点が，平均値手法の利点である．

事例内最小値による最小確信度スコア
この手法では，各トークンの中で最も小さい最小確信度スコアが事例全体の最小確信度スコアとなる．この手法による最小確信度スコアを基に選ばれる事例は，「事例全体で最も確信度の高いトークン」の確信度が最も小さい事例である．したがって，事例全体で最も混乱している事例を選ぶことができる．平均値を取った場合，トークン長の短い文ほど各トークンが最小確信度スコアに与える影響は大きくなる．トークン長によって最小確信度スコアの決定にギャップが存在することは好ましくないが，本論文で扱う対話文では，短い発話も多く存在している．トークン長による影響なく事例全体での混乱度を測ることが可能な点が，最小値手法の利点である．

事例内最大値による最小確信度スコア
この手法では，各トークンの中で最も大きい最小確信度スコアが事例全体の最小確信度スコアとなる．この手法により選ばれる事例は，「最もモデルが混乱しているトークンの確信度を事例全体の最小確信度スコアに

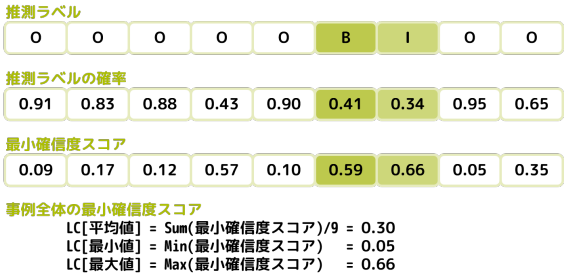

図 4 各手法で事例全体の最小確信度スコアを決定する過程の例．同じ事例に対しても，どの値を採用するかによって事例全体の最小確信度スコアは大きく変動する．

利用する．本論文のタスクである言い淀み検出では，言い淀み部分以外はチャンク外となるため，事例中の大部分のトークンが O タグと判定される．このようなチャンク外のラベルは基本的にモデルの確信度も大きくなるため，O タグを含むほとんどの事例では平均値・最小値ともに O タグの確信度に大きく左右されてしまう．これにより，トークン長の長い文に存在する B タグや I タグなど，タスクに関わるラベルの誤りのように重要な事例が見逃されてしまう可能性がある．確実にモデルが最も予測困難なトークンを含む事例を取り出せる点が，最大値手法の利点である．

5. 実験

本節では，提案手法の有効性を確認するために行った実験に関する設定や結果・考察について述べる．実験では，能動学習を行ったモデルに関する比較検証を行う．また，4.2 節で論じた能動学習における各種選出手法の有効性についても調査する．

5.1 データ分割について

本論文におけるモデルの性能評価にあたっては，「検出の対象とするデータに対してどれだけ特化できたか」を特化性能として評価するが，加えて「未知のデータに対する汎化性能をどこまで維持できたか」を汎化性能として評価する．これは，検出対象に特化させる過程で過学習に至り，モデルの言い淀み検出モデルとしての性能が悪化していないかを調べるためである．

本論文におけるデータ分割の概要を図 5 に示す．対話データ全 58 対話をファインチューニングに用いる訓練データ 39 対話，提案手法による特化の対象となる特化データ 1 対話，どちらにも含まれない未知データ 18 対話に分割する．

能動学習を行う各モデルでは，学習にあたって特化データから 5 事例を選出して追加の訓練データとして学習する．したがって，提案手法モデル群が選出するこれらの事例が特化性能評価の際に含まれることは，各提案手法モデルが自身で選出した 5 事例の系列についてすべて正解できることに等しい．このような状態では，ベースモデルと提案手法モデルの比較は対等な条件とは言いがたい．したがって，特化性能の評価に使用する評価データは特化データ内の全事例中でどの提案手法モデルからも能動学習用訓練データとして選出されなかった事例のみで構成される．

図 5 データ分割の概要．特化データは未知データを除いたデータセット内のある 1 対話で，残りをファインチューニング用訓練データとしている．能動学習用訓練データが選出されるほど評価データの総量は少なくなることに注意．

表 4 本実験で使用するモデル一覧．

手法	モデル名	能動学習用データの選出
ベースライン	*Base*	なし (ファインチューニングのみ)
ランダム	*Random*	ランダム選出
不確実性	LC_{Ave}	最小確信度スコア [事例内平均値]
	LC_{Min}	最小確信度スコア [事例内最小値]
	LC_{Max}	最小確信度スコア [事例内最大値]

5.2 実験設定

ベースラインには，ファインチューニングのみを行った言い淀み検出モデルにあたるモデル *Base* を使用する．提案手法モデルとしては能動学習モデルの中から，選出法によってランダムサンプリングから *Random* を，不確実性サンプリングから LC_{Ave}，LC_{Min}，LC_{Max} の 3 モデルを使用する．各提案手法モデルでは，能動学習時に特化データの中から対応する選出手法によって選出された 5 事例を追加の訓練データとして学習を行う．本来能動学習ではアノテータを通じて正解ラベルを得るが，本論文では擬似的なアノテーションラベルデータとして，選出されたデータに対応する正解ラベルデータを与えている．使用するモデルは表 4 の通りである．なお，訓練時のエポック数は，ファインチューニング時は 15 エポック，能動学習時は 3 エポックとした．評価指標には Precision，Recall，F 値を使用する．また，結果では特化データによる結果のブレを均す目的で特化データ 1 対話の選び方によって 8 回実験を行い，各指標の 8 回分の平均によって評価する．

5.3 実験結果

実験結果を表 5 に示す．能動学習を使用する各モデルは，LC_{Max} を除いてベースモデルである *Base* を上回る F 値を評価・未知データの双方において示している．これにより提案手法である能動学習による特化の有効性が確認できた．今回の実験結果では汎化性能を評価する未知データ

表 5 各モデルによる言い淀み検出実験結果．指標毎の最高値を太字で，2 番目に高い値を下線付きで示している．

モデル	特化性能			汎化性能		
	Pre.	Rec.	F1	Pre.	Rec.	F1
Base	0.250	**0.677**	0.351	0.235	**0.722**	0.351
Random	**0.593**	0.513	**0.509**	**0.559**	0.544	**0.514**
LC_{Ave}	0.567	0.464	0.491	0.459	0.482	0.440
LC_{Min}	0.259	0.653	0.359	0.241	0.703	0.355
LC_{Max}	0.316	0.290	0.230	0.385	0.317	0.218

においても精度の向上が見込めていることから，特化を行いながらも過適合は抑えられていると考えられる．

一方で，サンプリング手法間での比較を行うと，ランダムサンプリングのモデルが最も高い精度を示しており，次点に平均値による最小確信度を用いた不確実性サンプリング (以降，平均値手法) のモデルが続いている．結果からは不確実性サンプリングはランダムサンプリングより良い発話を選出できていなかったと考えられる．

実際に選出された発話を確認すると，表 6 に示すように不確実性サンプリングにより選ばれる発話は，最小値手法を除き発話長が極端に短い傾向があった．発話長の短い発話を選出しやすい平均値手法のみならず，トークン長に依らない最大値手法においても発話長の短い発話を選んでいた．これは，最小確信度スコアが上位 5 位であるトークンは短い発話中の事例であったことを意味している．つまり，周囲の文脈情報を活用することができる BERT モデルにおいては，文脈情報の乏しい短い発話こそが最もモデルにとって推測困難な事例であると考察できる．不確実性サンプリング手法はモデルが混乱する事例を選出するが，発話長の短い発話はそもそも文脈情報が乏しいことに原因がある．したがって，このような事例を選出しても言い淀み検出や対話内の局所的傾向に関する有効な情報を得ることは難しかったと考えられる．

6. 追加実験

本節では，5 節の実験とその結果を受けて，提案手法に関する考察を深めるために行った追加実験について述べる．

6.1 閾値付きサンプリング手法による能動学習

5.3 節で述べたとおり，多くの不確実性サンプリング手法は発話長の短い発話を選出しており，またその原因は BERT が極端に短い入力を不得手とするためであると考えられる．そこで，本実験では各選出手法にトークン長の閾値を加えた手法を作成し，モデルにとって得意な事例であるという前提を置いた能動学習での効果を検証する．閾値を越えるトークン長の事例内から，ランダムおよび不確実性の各手法に基づく選出を行ったモデル $Random_{Long}$，$LC_{AveLong}$，$LC_{MinLong}$，$LC_{MaxLong}$ を作成する．本実験では閾値を 5 トークンに設定し，これらの閾値付きモデル

表 6 各モデルにおける能動学習用訓練データの実例．

モデル	平均トークン長
Random	7.925
発話例	
それを聞かれるとやだなと思ってたの実は	
自分で思う訳ですよ	

モデル	平均トークン長
LC_{Ave}	5.85
発話例	
(D は) はい	
あはい	

モデル	平均トークン長
LC_{Min}	20.575
発話例	
本当に (D にん) うーんあんまり好きじゃないんですね	
どうして嫌いなの甘いもの何が嫌なの	

モデル	平均トークン長
LC_{Max}	2.25
発話例	
はい	
うんん	

を表 4 のモデル群に加えて，同一の実験設定で再度実験を行った．

実験結果を表 7 に示す．5.3 節と同じく，最高精度は閾値無しランダムサンプリング手法の *Random* となった．閾値付きのモデルは最小値・最大値手法においては閾値無しの同手法モデルを上回ったが，ランダム・平均値手法では逆に精度が悪化した．

全体的な傾向として，閾値付きモデルは Recall が高くなる傾向にあった．これは，ある程度長い発話の中から難しい事例を選出した結果，「ある程度長い発話には言い淀みが存在する」という仮定を持って言い淀み判定が過剰になった可能性が考えられる．5.3 節の実験において選出発話の平均文長が最も長かった LC_{Min} が Recall で高い値を示していたことも，この考察を支持している．文脈情報に富んだ，モデルにとって得意な事例であるだけでは，特定対話の局所的傾向を学習するには不十分だったと考えられる．

6.2 他コーパスの対話への特化

ここまでの実験では，使用データは特化データから未知データまですべて CSJ 内の対話データを使用していた．しかし，システムを実際に運用する際はファインチューニングに使用したデータセットとは属性の異なる対話が特化データとして入力されることもあり得る．そこで，本実験では他の対話コーパス内の対話に対する特化を行い，能動学習の特化性能がコーパス間の大きなギャップを埋められ

表 7　閾値付きモデルを加えた言い淀み検出実験結果．指標毎の最高値を太字で示している．

モデル	特化性能			汎化性能		
	Pre.	Rec.	F1	Pre.	Rec.	F1
Base	0.232	0.663	0.329	0.235	0.722	0.351
Random	**0.532**	0.494	**0.464**	0.547	0.560	**0.517**
$Random_{Long}$	0.258	0.621	0.350	0.256	0.702	0.371
LC_{Ave}	0.529	0.447	0.453	**0.588**	0.499	0.451
$LC_{AveLong}$	0.225	**0.664**	0.318	0.228	**0.728**	0.343
LC_{Min}	0.222	0.630	0.321	0.233	0.706	0.348
$LC_{MinLong}$	0.250	0.646	0.350	0.248	0.703	0.364
LC_{Max}	0.307	0.298	0.250	0.350	0.334	0.237
$LC_{MaxLong}$	0.236	0.655	0.333	0.240	0.715	0.356

表 8　Kyutech コーパス対話データの一例 (ただし，結合位置を明示するために IPU 単位での区切りだった位置に/(スラッシュ)を付与している)．実在市外や店舗の話題では英字による置換がなされている．

発話
(F まあ) うどんもありとは思う
俺の中じゃバイキングはちょっと / ちょっとないかないやでもーどうでしょう
(D 寿司) 寿司って他にないんですかね
もう、なんか結構ありそうな気がするんですよね

るか確認する．

Kyutech コーパス [17] は，複数人の話者によって行われる議論を収録したコーパスで，飲食店の誘致に関する議論など全 9 対話が映像，音声，書き起こしデータでそれぞれ収録されている．実際の書き起こしテキストデータの例を表 8 に示す．例にも見られるように，書き起こしデータ内では言い淀みのラベルも CSJ と同様の方式で付与されているため，同様の手順で入力データとして使用することができる．

実験では，ファインチューニング用の訓練データとして CSJ から対話文 39 対話，特化データに Kyutech コーパスから 1 対話，未知データに Kyutech コーパスから 1 対話を使用する．使用モデルは 6.1 節と同様閾値付きモデルを含めた 9 モデル，評価は特化データの選び方によって 8 回実験を行い，8 回分の平均によって評価する．

実験結果を表 9 に示す．*Random* が最高精度となったが，全体的に改善幅は小さく，また数値的にも悪い結果となった．*Base* の精度からも分かるとおり，CSJ と Kyutech コーパスの間には対話のスタイルや参加者の人数，話し方など様々な差があるため，一方で訓練されたモデルではもう一方の問題を解くことはできない．本実験では能動学習によって対話内の傾向を学習することでこのギャップが埋まることを期待していたが，多くのモデルでは学習に失敗しており，効果は不十分なものだったと言える．CSJ 対話文同士ではこのようなギャップが比較的小さかったため，5 件という超小規模なデータでも特化データに特化できた．しかし，大きく性質の違うコーパスを扱う場合は，追加のファインチューニングや更なる能動学習用データを要する可能性がある．

7. おわりに

本論文では，話題や話し手の特性などの要因によって局所的に変動する対話文の言い淀み検出を行った．能動学習の仕組みを利用した多段階の学習によって，対話単位に特化した言い淀み検出手法を提案し，能動学習を行わないモデルとの比較検証によって提案手法の有効性を確認した．また，能動学習においてアノテータに提示するデータの選出手法についてもいくつかの手法を比較して論じた．

実験では，能動学習モデルに使用する事例の選出手法として，ランダムサンプリングと最小確信度に基づく不確実性サンプリングのみを使用した．しかし，能動学習における選出手法は多岐に渡り，異なる手法についても検証の余地があると考える．本論文の結果では，特定の恣意性がなくランダムに選出するランダムサンプリングが最高精度を示していた．またモデルにとって解くのが難しい事例や，モデルが有効活用できる文脈情報を多く含んだ事例であるだけでは，必ずしも対話単位の傾向にリーチできていないことも示されている．したがって，ランダムサンプリングを発展させた，データ内の多様性を意識した手法こそが有効である可能性がある．

一方で，能動学習用訓練データとして選出する事例数については課題が存在する．異なる選出手法を同時に比較する場合，本論文の実験で行ったような，各能動学習用訓練データとの重複を避けたテストデータの作成が必要となる．しかし，特化データから複数の手法が訓練データを選出するほど，テストデータのサイズは小さくなってしまう．本論文では，能動学習を用いる手法を最大 8 手法同時に実験した．実際に選出された事例は必ずしも言い淀みを含んでいなかったが，もし選出される事例がすべて言い淀みを含む事例であったならば，各手法が重複なく 5 事例ずつ選出した場合 40 事例の有効なテストデータが消滅してしまう．これは今回使用したデータセットにおける 1 対話あたりの平均言い淀み事例数である 37 事例を上回るものであ

表 9 Kyutech コーパスでの言い淀み検出実験結果．指標毎の最高値を太字で示している．

モデル	特化性能			汎化性能		
	Pre.	Rec.	F1	Pre.	Rec.	F1
Base	0.053	**0.340**	0.087	0.085	**0.398**	0.140
Random	**0.311**	0.098	**0.115**	0.426	0.198	**0.219**
$Random_{Long}$	0.067	0.270	0.099	0.109	0.337	0.163
LC_{Ave}	0.125	0.004	0.007	**0.625**	0.017	0.032
$LC_{AveLong}$	0.057	0.270	0.087	0.114	0.336	0.163
LC_{Min}	0.054	0.278	0.089	0.099	0.341	0.152
$LC_{MinLong}$	0.054	0.286	0.089	0.101	0.351	0.155
LC_{Max}	0.000	0.000	0.000	0.250	0.002	0.005
$LC_{MaxLong}$	0.059	0.260	0.090	0.116	0.312	0.162

る．したがって，多くの選出法を並列で実験する場合や，各手法で選出する事例数を増やす場合は十分に言い淀み事例を含んだ対話データを特化データとすることが重要であると考える．

謝辞

本研究は科研費 23K11368 の一部です．

参考文献

[1] Christine Nakatani and Julia Hirschberg. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 46–53, 1993.

[2] 尾嶋憲治, 河原達也, 秋田祐哉, and 内元清貴. 話し言葉の整形作業における削除箇所の自動同定. **情報処理学会研究報告自然言語処理** *(NL)*, 185:85–91, 2008.

[3] Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, Takanobu Oba, and Yushi Aono. Disfluency detection based on speech-aware token-by-token sequence labeling with blstm-crfs and attention mechanisms. In *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1009–1013, 2019.

[4] Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. Disfluency Detection with Unlabeled Data and Small BERT Models. In *Proceedings of Interspeech 2021*, pages 766–770, 2021.

[5] Morteza Rohanian and Julian Hough. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3693–3703, 2021.

[6] Hiroto Nakashima and Kazutaka Shimada. Disfluency detection with context information from real utterances and generative utterances. In *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 462–467. IEEE, 2023.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[8] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition, 2017.

[9] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT92. ACM, July 1992.

[10] Gabriel Corvino, Vitor Vasconcelos Oliveira, Angelo C. Mendes da Silva, and Ricardo Marcondes Marcacini. On the use of query by committee for human-in-the-loop named entity recognition. In *Anais do X Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2022)*, KDMiLe 2022. Sociedade Brasileira de Computação - SBC, November 2022.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[13] Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. From zero to hero: Human-in-the-loop entity linking in low resource domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[14] Ankit Agrawal, Sarsij Tripathi, and Manu Vardhan. Active learning approach using a modified least confidence sampling strategy for named entity recognition. *Progress in Artificial Intelligence*, 10(2):113–128, January 2021.

[15] 前川喜久雄. 『日本語話し言葉コーパス』 の概要. **日本語科学**, 15:111–133, 2004.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL, Volume 1*, pages 4171–4186, 2019.

[17] Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. The Kyutech corpus and topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 95–104, 2016.