記事トピックのクラスタを用いた多言語ニュース推薦手法の提案

AL18036　片岡凪
指導教員　木村昌臣

データ工学研究室

目次

1. 背景

ニュース読者が**出来事の一部**しか把握できない

- 記者が出来事を解釈し、主張したい要素を切り取るため
  - 地域の文化ごとに解釈は異なる
  - 地域の政治ごとに切り取り方は異なる
- **時間的コスト**により、全ての主張を把握できないため
  - Web上の記事を全て読むのは不可能
  - 海外の記事には読解・翻訳のコスト

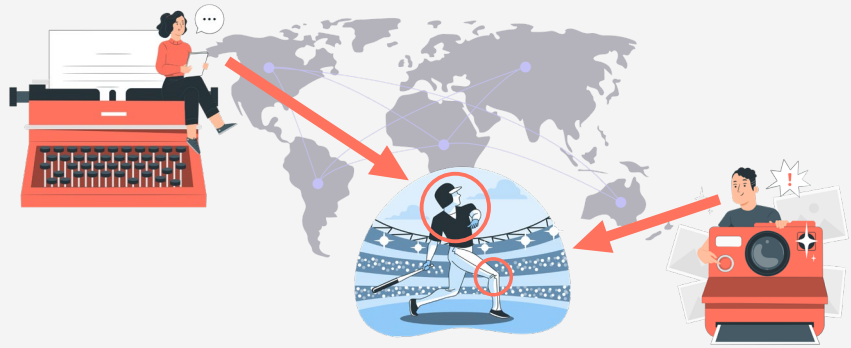


図1　地域ごとに異なる出来事の解釈と切り取り方

図2　時間がかかる出来事・主張の把握

2. 目的

ニュース読者が**出来事**と**それに対する主張**を把握できる推薦手法の提案

- 出来事 (青) := 記者の解釈に依存しない事象
- 主張 (赤) := 記者が伝えるべきだと判断した出来事の解釈

1. 閲覧記事と同じ出来事を書く記事をWebから抽出
2. 抽出した記事の主張の文章を抽出
3. 手軽に把握できる主張の文のクラスタを推薦

➢ 文章間の出来事の類似度・主張の類似度の算出が必要

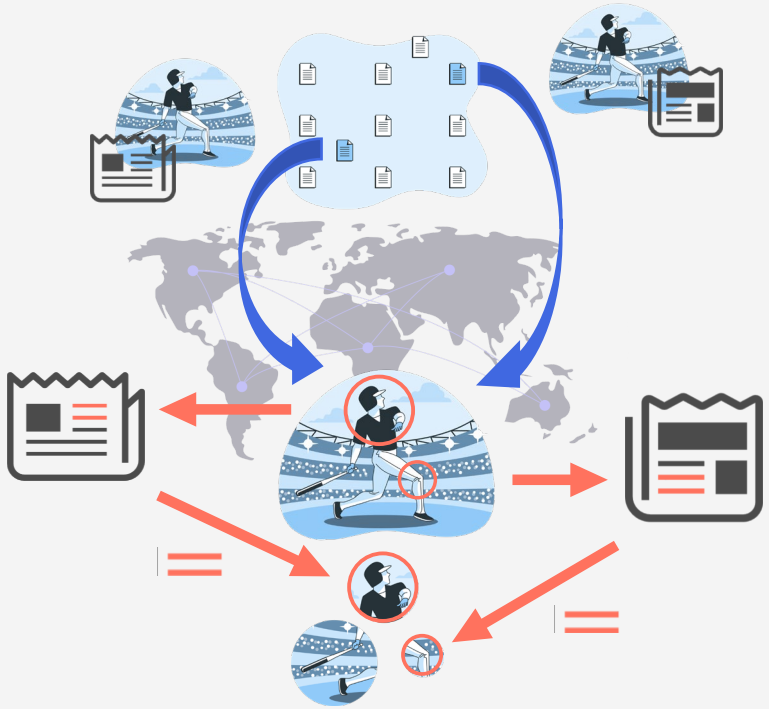


図3 出来事とその主張がより把握できる推薦手法

3. 先行研究

従来の**文章の話題[1]の定量化手法**は
出来事の類似度・主張の類似度の算出に不向き

1. 話題 := 出来事の要約

BleiらのLDA (Latent Dirichlet Allocation) [1]

- **話題を表す単語同士の関係性が得られない**

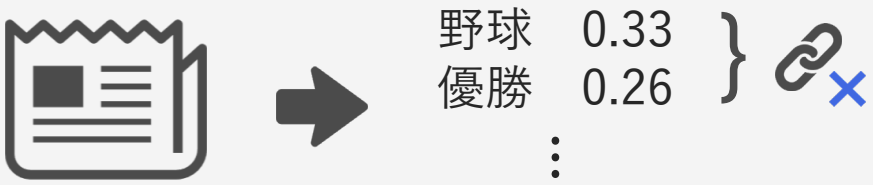


図4 記事に関連の深い単語と関連度の出力

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993--1022, January 2003.

ReimersらのSentence-BERT [2]

- **全文から出来事と主張の区別ができない**

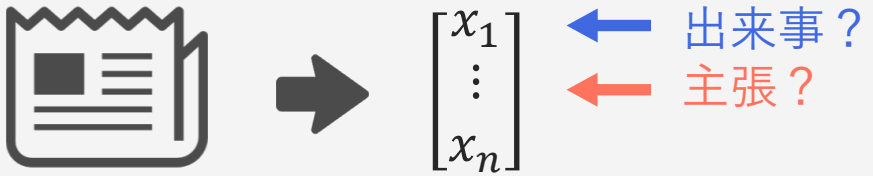


図5 文全体を考慮した文意のベクトルの出力

[2] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sen-tence Embeddings using Siamese BERT-Networks. https://arxiv.org/abs/1908.10084v1(2021年7月19日参照).

4. 提案手法 – 概要

記事の**出来事**と**主張**のクラスタを用いた
多言語ニュースの**主張の文**の推薦手法の提案

- **仮定** – 記事の文章が出来事を述べる文か主張を述べる文に分類できる

1. DeepL APIを用いて世界の記事を英訳
2. 記事の文章を出来事の文と主張の文に分類
3. 出来事の文章の類似度を基に記事をクラスタリング
4. 主張の文 を類似度を基にクラスタリング

➢ 閲覧記事とより類似した出来事の主張の文のクラスタを推薦

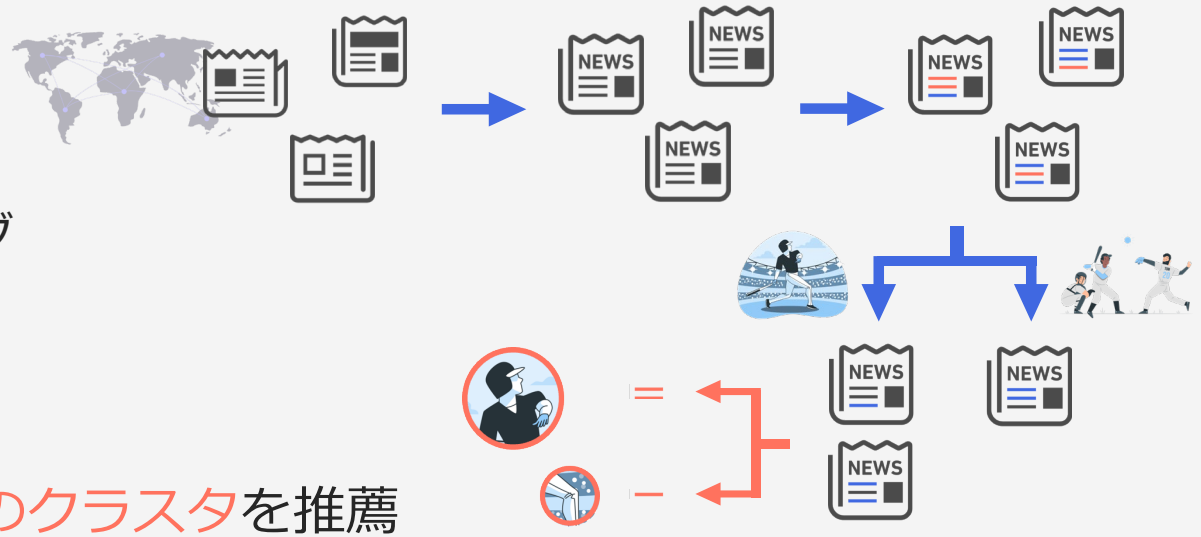


図6 提案手法の概要

4. 提案手法 – 分類の詳細

出来事の文と主張の文の分類器に**Transformer**を応用

- **RoBERTa** (Robustly optimized BERT approach) で単語埋め込み
  - 6300万件の記事などを事前学習した**文のベクトル化**
- **Transformerのエンコーダ部分**を応用して分類
  - Googleが文章表現を事前学習した汎用モデルを拡張
  - **テキスト分類**に応用するため、ラベル付きデータセットで追加学習
- IBMの**Debater Dataset**で教師あり学習
  - 英文をEvidence か Claim かでラベル付けしたもの
  - この学習で出来事と主張に分類できることを期待

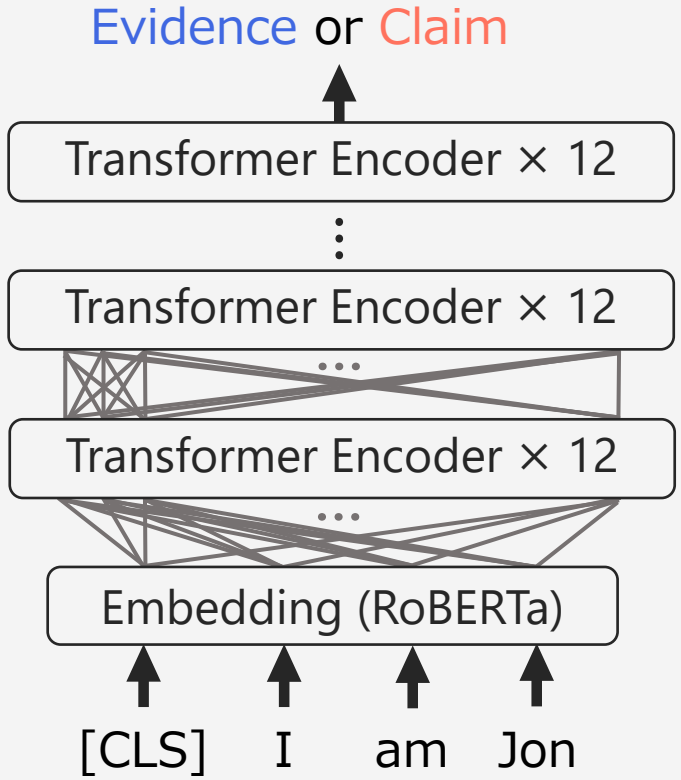


図7　出来事の文と主張の文のTransformer分類器

4. 提案手法 – クラスタリングの詳細

記事の出来事と主張の文のクラスタリングにSentence-BERTを利用

- **Sentence-BERT**で単語埋め込み
  - 高速なクラスタリングのために事前学習した文章のベクトル化
- **コサイン類似度**でベクトルを比較
  - 文章のベクトルの類似度算出に適した手法
- **Ward法**で出来事と主張の2回に分けて階層的クラスタリング
  - 出来事の意味が階層的であると考えられるため
  - 主張の文のクラスタを読者が望む粒度で推薦

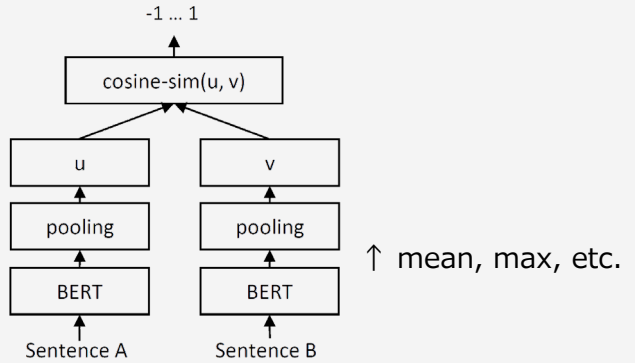


図9 Sentence-BERTとコサイン類似度 [2]

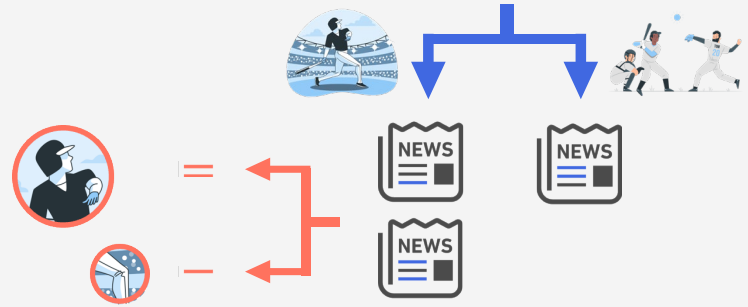


図10 出来事と主張の階層的クラスタリング

[2] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks https://arxiv.org/abs/1908.10084v1 (2021年7月19日参照).

5. 研究状況

EvidenceとClaimのデータセットで
出来事と主張の文の分類に部分的に成功した

- Transformer分類器で日本の記事の分類実験
  - 追加学習により適合率 0.994
  - Evidence に分類された文
    - 記者の解釈に依存しない出来事を示していた
  - Claim に分類された文
    - 記者が切り取った組織や人の主張を示していた
  - **誤分類**された文
    - 出来事と主張の**混合**，**出来事らしい主張の言葉**など
    - EvidenceとClaimの**連続値**での出力・クラスタリングを検討

表1　Evidenceの文とClaimの文の分類結果

分類	翻訳前の入力文
Evidence	決勝のヒットを打った23日の試合も1球だけで終わった
Claim	日本シリーズ進出を決めてうれしい
Evidence（誤）	一方、敗れた中日・落合博満監督は「今年1年は思いがけない風が吹きっぱなしだった

6. まとめ

記事の**出来事**と**主張**のクラスタを用いた
多言語ニュースの**主張の文**の推薦手法の提案

- 手軽に出来事の全容を把握できる主張の文のクラスタの推薦を目指す
  - 記事の文を出来事の文と主張の文に分類
  - 出来事の文で記事をクラスタリングし、主張の文をクラスタリング
- EvidenceとClaimを学習し、出来事と主張の文の分類実験を実施
  - 誤分類を基に分類器の精度向上を検討中

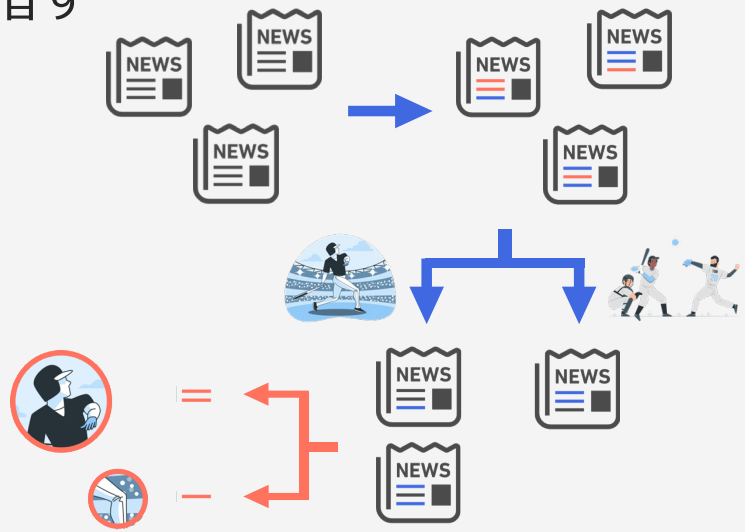


図11　提案手法の概要の一部

7. 今後の予定

実施内容	8月	9月	10月	11月	12月	1月
分類器の精度向上						
クラスタリングの実装						
評価方法の検討・実装						
一部の手法を変えて比較						
本実験						
スライド作成・発表練習						
論文執筆						

参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993--1022, January 2003.

[2] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://arxiv.org/abs/1908.10084v1 (2021年7月19日参照).

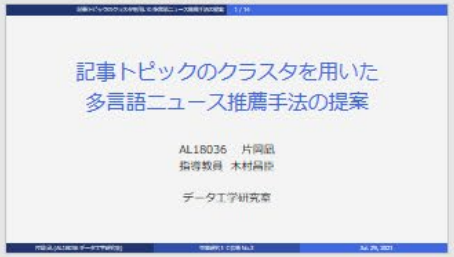
[3] RSF. 2021 World Press Freedom Index: Journalism, the vaccine against disinformation, blocked in morethan 130 countries. https://rsf.org/en/2021-world-press-freedom-index-journalism-vaccine-against-disinformation-blocked-more-130-countries (2021年7月19日参照).

[4] IBM Corporation. Project debater datasets. https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml (2021年7月19日参照).

[5] 坂本俊之. Japanese FakeNews Dataset. https://www.kaggle.com/tanreinama/japanese-fakenews-dataset (2021年7月19日参照).

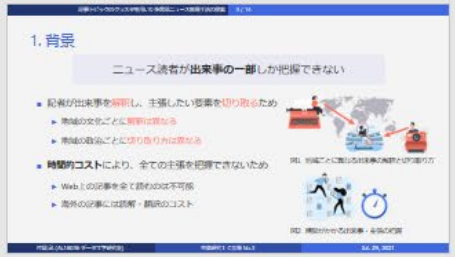
ご清聴ありがとうございました

質疑応答



1

2



3

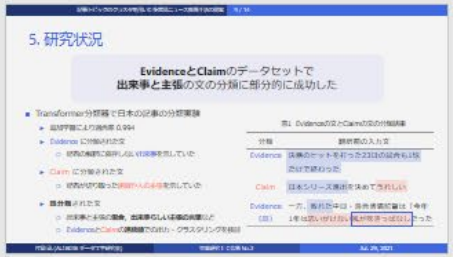
4

5

6

7

8



9

10

11

12

図解 - 1

図解 - 2

図解 - 3

以降、予備スライド

p. 3 - 出来事と主張に着目したきっかけ

- ひとつの記事に対して多様なコメントがなされてる
  - 多様さが読者の解釈がきている
    - 記者の解釈も存在するのでは
  - 誤った主張やそれを過信する読者が少なくない
  - 中には記事より有用な主張を補足するものも
    - 多角的な視点でニュースから知識を得てほしい
- デモ活動の聴取をした際に
  活動者が相手側の主張を理解できていなかった経験

図12 量子コンピュータが実用段階であるかで分かれる主張 [6]

[6] 共同通信．量子コンピューター初設置　東大とIBM、汎用型．
https://news.yahoo.co.jp/articles/74d5d745cc7c942b4a972ecac3979d701ce2855b/comments (2021年7月28日参照)．

p. 3 – 出来事の切り取り方の具体例

- 「野球の大会Aで優勝した選手B」という出来事の記事
  - 「選手の大会での活躍」を切り取る記者
  - 「選手の1年前の膝の故障」を切り取る記者

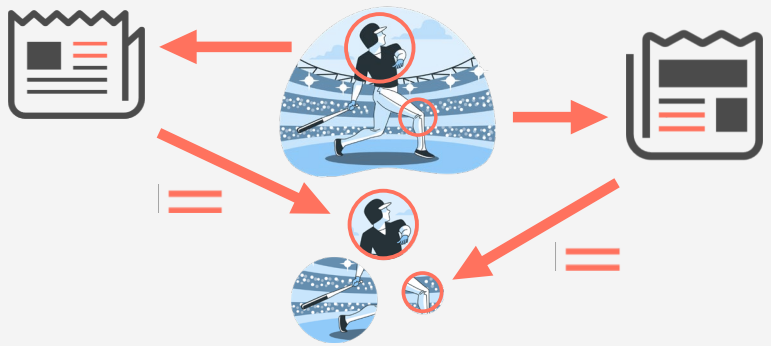


図13　出来事とその主張がより把握できる推薦手法

- 領土問題の記事
  - 所有権を主張する地域Aで生まれ育った記者の切り取り方
  - 所有権を主張する地域Bで生まれ育った記者の切り取り方
  - 中立な立場にある地域Cで生まれ育った記者の切り取り方

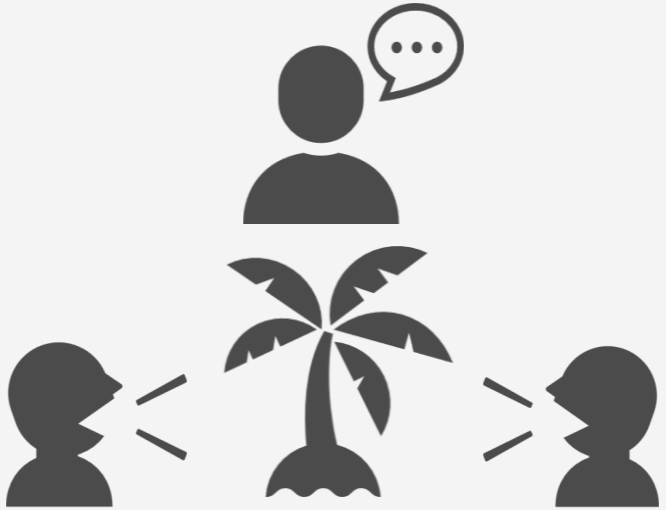


図13　領土問題で切り取り方が異なる記者

p. 3 - 文化と政治で主張が異なる具体例

- NSF (Reporters Sans Fronti`eres ; 国境なき記者団)
  - 「政治や文化などの要因によって記者の主張が制限されている」[7]
  - 文化で主張が異なる例
    - 宗教文化の違いから食文化の主張が異なる
    - 先進国と発展途上国の飢餓問題の当事者意識、支援できる可能性を踏まえて主張が異なる
  - 政治で主張が異なる例
    - 「Covid-19が存在しない」と代表が述べる地域とその他の地域との主張の違い
    - 「記者クラブという組織が記事内容に介入する日本」[8] と「海外」との主張の違い

[7] RSF. 2021 World Press Freedom Index: Journalism, the vaccine against disinformation, blocked in more than 130 countries. https://rsf.org/en/2021-world-press-freedom-index-journalism-vaccine-against-disinformation-blocked-more-130-countries (2021 年7 月19 日参照).
[8] RSF. Tradition and business interests. https://rsf.org/en/japan (2021 年7 月19 日参照).

p.3 – 出来事も記者によって解釈が異なるのでは

- 出来事の解釈が異なる可能性はある
  - 例
    - 同じ金属バットに対し「アルミ製のバット」「アルミ、銅、マグネシウムの合金バット」とする記事
    - 量子コンピュータに詳しくない記者「（精度に触れずに）実用化された」
    - 他の地域の報道を確認できない記者「最近流行している感染症はただの風邪」
  - 同じ事象は捉えている
    - Sentence-BERTによる文章表現で捉えられる可能性がある
    - 同一日時である可能性が大きく、モデルに組み込む価値がある
  - 出来事のクラスタの階層レベルを調節することで出来事の解釈の小さな差異は解消可能

p.3 – 他の地域で書かれない出来事は推薦されないのでは

- 本研究のモデルでは最適な対応とはいえない
- 本研究のモデルで対応するならば
  - 出来事の階層レベルを大きく設定
    - 少し出来事が類似していなくても読者の参考になる記事を推薦

p.5 – 話題とは

- 本研究
  - 出来事の要約
- 辞書（Weblio）
  - 話題、主題、題目
  - 陳述される中心的対象
  - 話の抽象度を最もあげた時の概念的なもの
- トピックモデルの分野
  - トピックの判断材料
  - 文章、画像、音楽などに利用

p.5 – LDAの詳細

- トピック毎の単語の分布、文章毎のトピックの分布は、ディレクリ分布に従うと仮定
  - $\phi \sim p(\phi|\beta)$
- 上記は、各トピック毎に単語分布を生成
  - $\theta \sim p(\theta|\alpha)$
- 上記は、各文章毎にトピックの分布を生成
  - $z \sim p(z|\theta)$
- 以上より、単語のトピックに該当する単語分布を選び、単語を生成
  - $w \sim p(w|\phi_z)$

p.5, 8 – Sentence-BERTの詳細

- 埋め込み表現を事前学習したBERTを用いて2文の類似度を出力
- 全記事を同時学習するBERTのクラスタリングより遥かに軽量
  - 10000文のクラスタリングに65時間 → 5秒
  - 精度は維持
- Pooling
  - ベクトルの平均、最大値、CLSに対応する要素のいずれかを設定
  - この順で精度が高いと確認されている
- タスクによって精度が異なるSentence-RoBERTaも存在

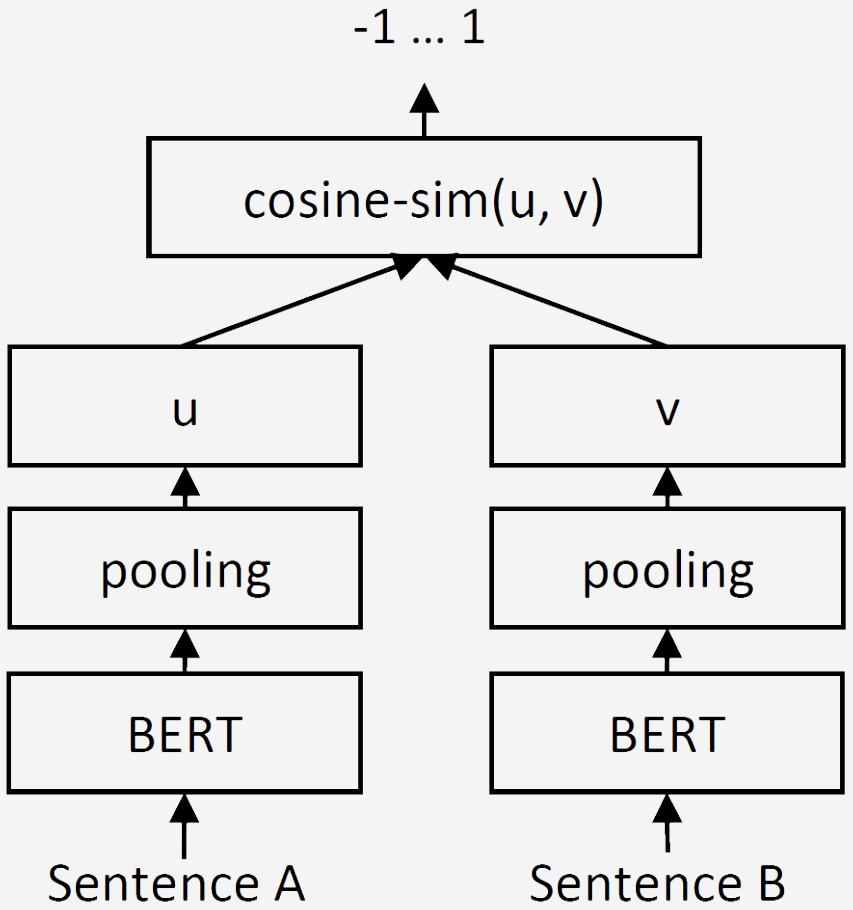


図14 Sentence-BERTとコサイン類似度 [2]

[2] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sen-tence Embeddings using Siamese BERT-Networks. https://arxiv.org/abs/1908.10084v1(2021年7月19日参照).

p.6 – 仮定の是非

- **仮定** – 記事の文章が出来事を述べる文か主張を述べる文に分類できる
  - 主張の言語表現があれば主張、なければ出来事
    - どちらでもない、は存在しない
    - 言語表現の頻度や記事中の重要度による比重は存在
      - どちらの比重も同程度の場合、クラスタリングで考慮しないことで精度が向上する可能性あり

表1　Evidenceの文とClaimの文の分類結果

分類	翻訳前の入力文
Evidence	決勝のヒットを打った23日の試合も1球だけで終わった
Claim	日本シリーズ進出を決めてうれしい
Evidence（誤）	一方、敗れた中日・落合博満監督は「今年1年は思いがけない風が吹きっぱなしだった

p.6 – 出来事のクラスタリングの例

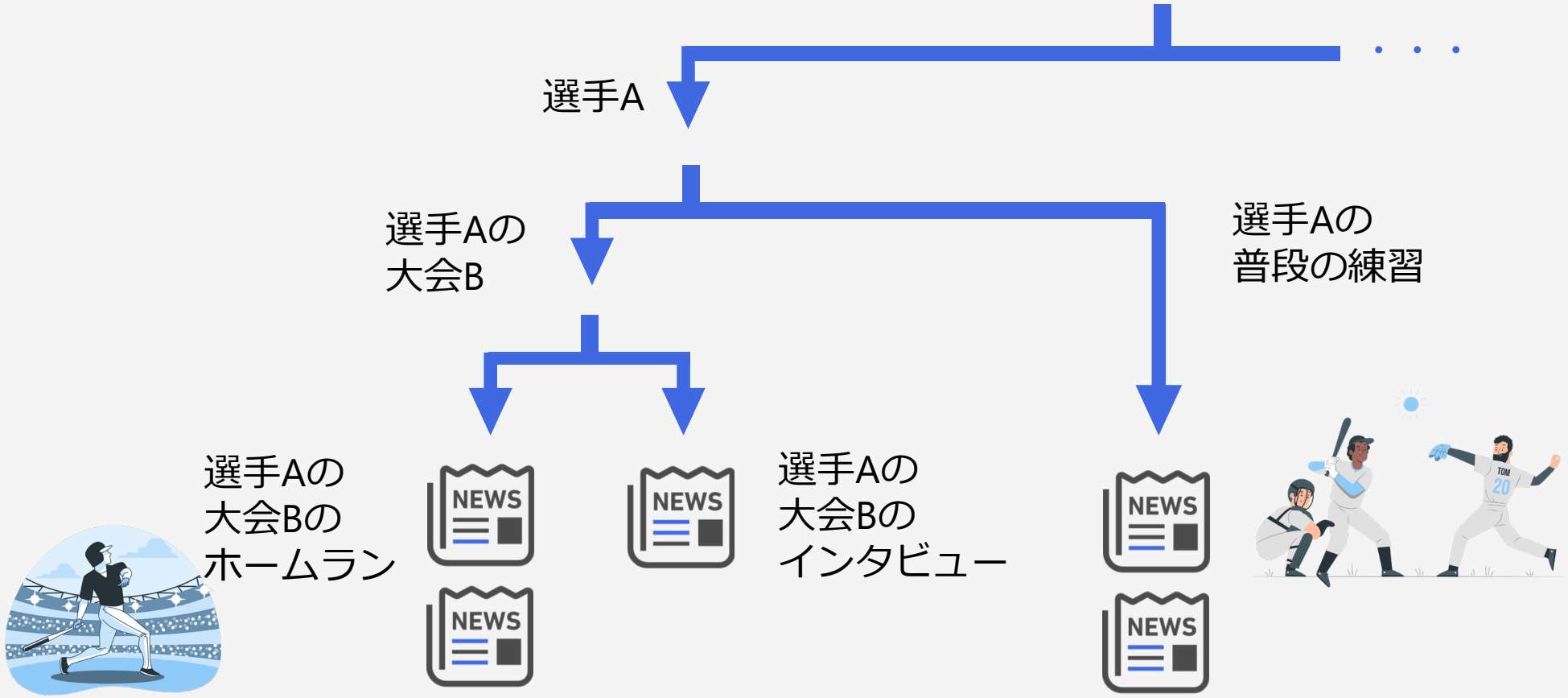


図15　出来事の類似度による記事のクラスタリングの例

p.6 – 主張のクラスタリングの例

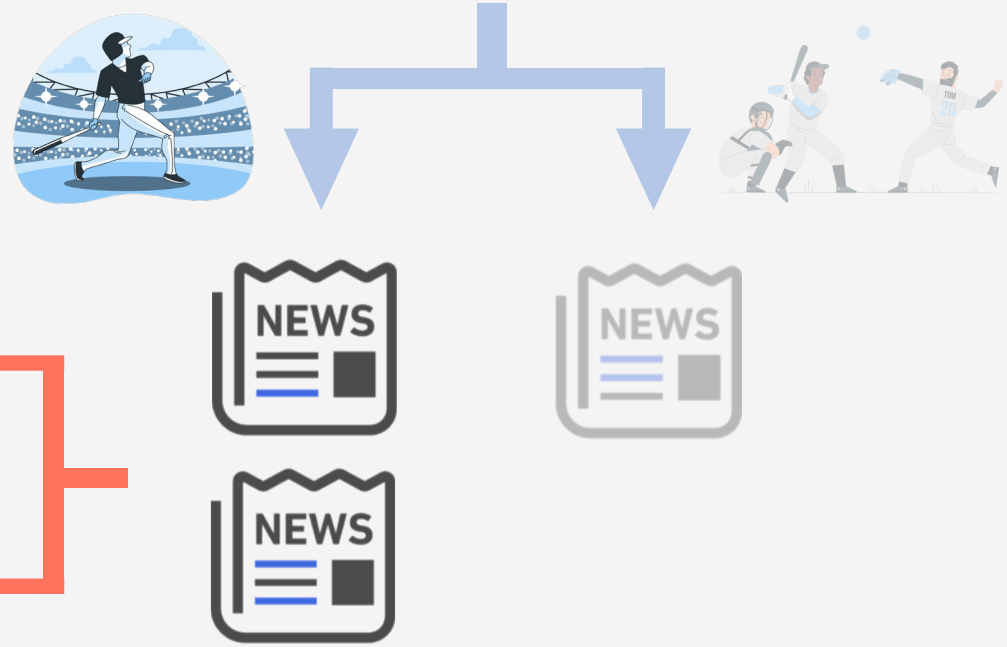


図11　主張の文の類似度を基にしたクラスタリングの例

p.7 – RoBERTa (Robustly optimized BERT approach) の詳細

- BERT（発表で紹介したGoogleのモデル）を改良した汎用モデル
- 単語埋め込み（訓練のしやすい文章表現のベクトル）に応用
- 学習量が多い
  - CC-News(76GB)、OpenWebText(38GB)、Stories(31GB)
- 学習自体の改良
  - バッチサイズの拡大
  - Next Sentence Prediction(NSP)の不使用
  - より長い文章を入力
  - 同じマスクを何度も使用せず、ランダムに指定

[10] Y. Liuほか, 「RoBERTa: A Robustly Optimized BERT Pretraining Approach」, *arXiv:1907.11692 [cs]*, 7月 2019, 参照: 7月 26, 2021. [Online]. Available at: http://arxiv.org/abs/1907.11692

p.7 – 単語埋め込みの具体例

- 'requisitions’
  - ['re', '##qui', '##sit', '##ions’]
    - 意味のまとまりで区分
    - 意味内容によって記号を付与

[9] Vicek (MIcrosoft). Deep Learning with BERT on Azure ML for Text Classification. https://techcommunity.microsoft.com/t5/ai-customer-engineering-team/deep-learning-with-bert-on-azure-ml-for-text-classification/ba-p/1149262 (2021年7月28日参照).

p.7 – Transformerの詳細

- 複数のAttentionを組み込んだ機械翻訳などに利用されるモデル
  - Attention
    - 特定の単語に注目して学習するモデル
    - 文字列の学習の忘却が少ない、30語以上の文章に対応可能
- 同じ文章を3つの見方で学習
  - 文章の処理する部分、注目の仕方、基底を変えて比較
- 並列可能な行列演算を主に利用するため高速
- 左のエンコーダ部分で文章表現を512次元のベクトルに変換

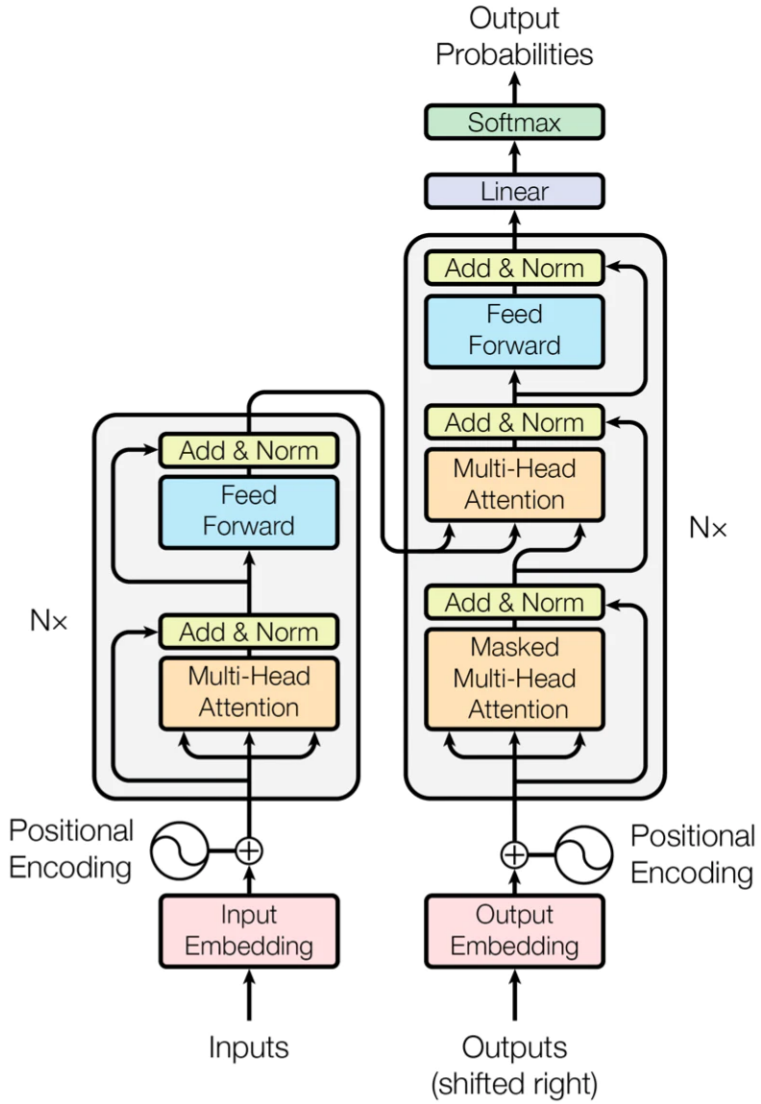


Figure 1: The Transformer - model architecture.

[11]A. Vaswani et al. Attention is All you Need.
http://papers.nips.cc/paper/7181-attention-is-all-you-%0Aneed.pdf
(2021年7月28日参照).

p.7 – BERTの詳細

- 文章表現を学習する双方向の汎用モデル
- 入力文の次の単語を予測
- 文章のマスク部分を予測
- [CLS]トークン
  - Classifyに利用する記号
  - 入力文の先頭に配置
  - 文全体の言語表現に相当

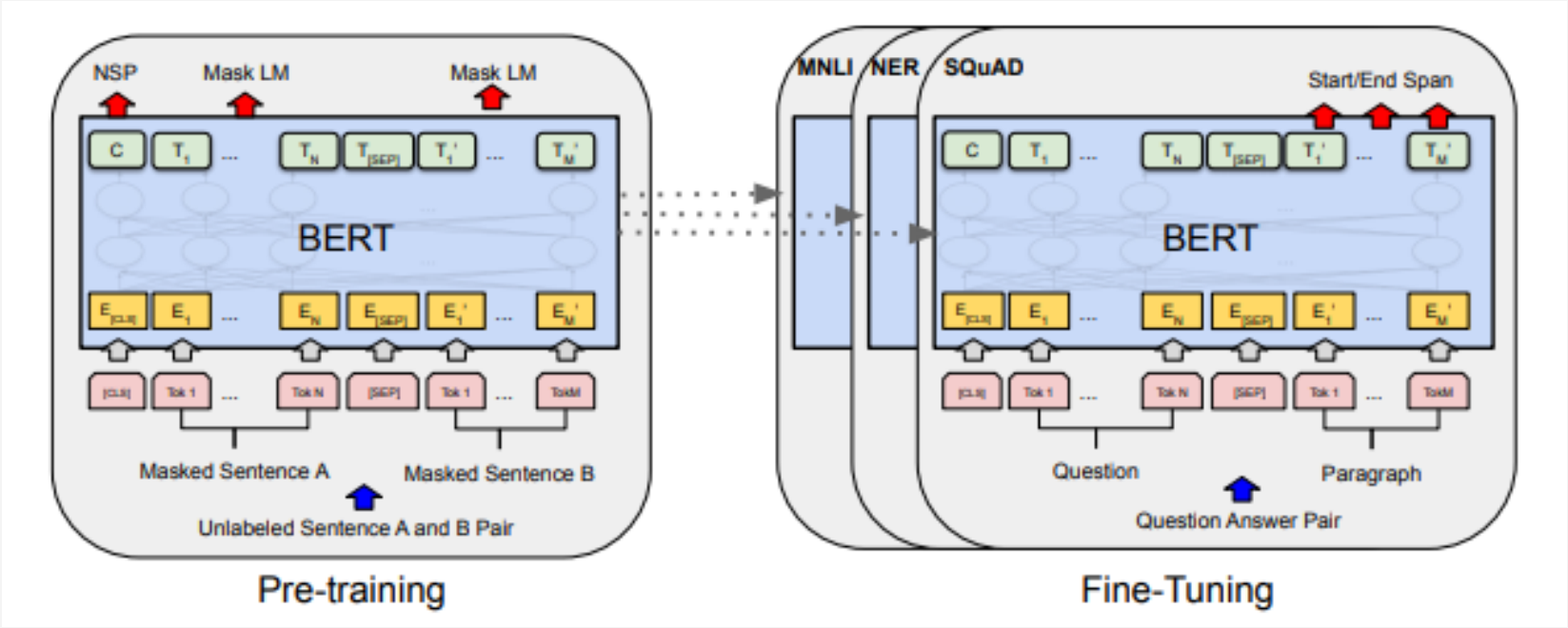


図11　BERTモデルとその転移学習・ファインチューニング

[12] J. Devlin, M.-W. Chang, K. LeeとK. Toutanova, 「BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding」, *arXiv:1810.04805 [cs]*, 5月 2019, 参照: 7月 24, 2021. [Online]. Available at: http://arxiv.org/abs/1810.04805

p.7 – IBM の Debater Dataset の詳細

- 議論のEvidenceとClaimを検出するための種々のデータセット
- 実験ではIBM Debater® - Claims and Evidenceを使用
  - ラベル付けしたWikipediaの記事58件
  - Claim：2294文、Evidence：4690文

[4] IBM Corporation. Project debater datasets. https://www.research.ibm.com/haifa/dept/vst/debating_data.shtml (2021年7月19日参照).

表2 IBM Debater® - Claims and Evidenceの例（日本訳）

Claimの文	暴力的なビデオゲームへの曝露は、少なくとも一時的な攻撃性の増大を引き起こし、この曝露は現実世界における攻撃性と相関している。
Evidenceの文	2001年の研究では、暴力的なビデオゲームへの曝露は、少なくとも一時的な攻撃性の増大を引き起こし、この曝露は現実世界における攻撃性と相関することが明らかになっている。

p.8 - 階層的クラスタリングのその他の手法

- 凝集型 (agglomerative)
  - 類似度の高いものからまとめる手法
    - 単リンク法 (single linkage method) 別名：最短距離法
    - 完全リンク法 (complete linkage method) 別名：最長距離法
    - 群平均法 (group average method)
    - セントロイド法 (centroid method) 別名：重心法
    - 重み付き平均法 (weighted average method)
    - メジアン法 (median method)
- 分割型 (divisible)
  - データ集合全体が一つのクラスタの状態から，順次クラスタを分割して，クラスタの階層を生成する.

p.8 – 「主張の文のクラスタを読者が望む粒度で推薦」とは

- 提案手法では、階層的なクラスタを提供
  - 利用するニュースサイトの目的に沿った応用
    = そのニュースサイトを好む読者の目的に沿った応用
    - 多忙な読者が多いので３つだけ推薦
    - 読者に技術者が多いので１０個推薦
    - 自ら調節したい読者が多いので、
      読者が個数を調整できるようなシステムで推薦

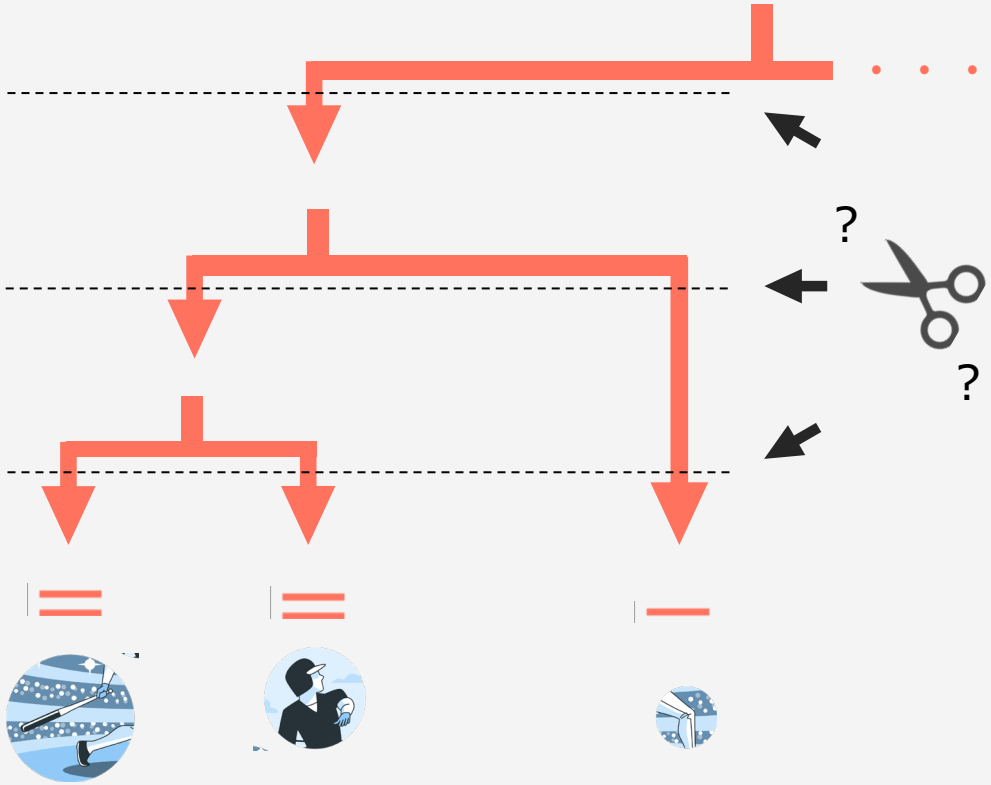


図12　読者が望む粒度のクラスタ数の選択

p.9 – 適合率の選択理由

- 適合率 $= \frac{\text{正しく Evidence と分類した数}}{\text{Evidenceと分類した数}} = \frac{TP}{TP+FP} = \frac{\text{どれだけ正解が含まれるか}}{\text{検索の中に}}$
- 再現率 $= \frac{\text{正しく Evidence と分類した数}}{\text{実際にEvidenceである数}} = \frac{TP}{TP+FP}$

表2　混同行列の一覧表

	陽性 （予測は正）	陰性 （予測は誤）
陽性 （実際は正）	TP 正しいEvidence分類	FN 誤ってClaim分類
陰性 （実際は誤）	FP 誤ったEvidence分類	TN 正しいClaim分類

表1　Evidenceの文とClaimの文の分類結果

分類	翻訳前の入力文
Evidence	決勝のヒットを打った23日の試合も1球だけで終わった
Claim	日本シリーズ進出を決めてうれしい
Evidence（誤）	一方、敗れた中日・落合博満監督は「今年1年は思いがけない風が吹きっぱなしだった

p.9 – 日本の記事に何を用いたか

- Japanese FakeNews Dataset
  - オープンデータコモンズパブリックドメイン専用およびライセンス（PDDL）
    - 商業的に利用したり、技術的な保護手段を用いたり、本データやデータベースを他のデータベースやデータと組み合わせたり、変更や追加を共有したり、秘密にしたりすることができます
  - FakeでないCC-BYのウィキニュースを使用
    - 元の作品・データの出典を明記すればどのように公開してもよい
    - 10記事の文章において、主張と出来事を正しく分類できたかを確認

[5] 坂本俊之. Japanese FakeNews Dataset. https://www.kaggle.com/tanreinama/japanese-fakenews-dataset (2021年7月19日参照).

p. 9 – その他の分類結果

- すぐに正解が判断できないが逆に言えば不正解とも判断できないレベル
- その規則として
  - 感情の単語
  - 曖昧な表現
    - 推定
    - 観測
    - みられている

表3　Evidenceの文とClaimの文の分類結果 2

分類	翻訳前の入力文
Evidence	12月12日16時19分頃、岩手県沖の深さ48kmを震源とするマグニチュード5.6（暫定値）の地震が発生し、青森県階上町で最大震度5弱を観測した
Claim	津波の心配はない
Claim	メカニズムは、東西に圧力軸をもつ逆断層型と推定されている
Evidence	気象庁は、今後約1週間は震度5弱程度の余震に注意するよう、呼びかけている
Evidence（誤？）	また、今後約2、3日は同程度の地震に注意すべきだという
Evidence	青森県で震度5弱を観測したのは2019年12月以来であり、この時も階上町で震度5弱を観測した
Claim	なおこの地震は、2011年3月の東北地方太平洋沖地震（東日本大震災）の余震だとみられている
Claim（誤？）	震度3以上を観測した地域は以下の通り

p.9 – 分類結果の成功した点、失敗した点

- 成功した点
  - 多くの分類は、人間が即断できないレベルで正解 or 不正解している
    - ラベルを連続値にすることでこの細かい違いを捉えることができる可能性あり
    - 逆に即断できないような文はクラスタリングの前に除外することでより良い推薦ができる可能性あり
- 失敗した点
  - 1文中の出来事と主張の混合
    - 出来事の部分が主張の類似度算出に悪影響となる可能性
  - 出来事らしい主張の表現（ex. 風が吹く）
    - 1文中に出来事の表現が多いため、出来事のクラスタリングには影響は少ないと考えられる
    - 本来主張として推薦したかった文が推薦されない可能性がある
    - 主張らしい出来事の表現も存在する可能性がある

p.9 - 分類器のその他の改善案

- 出来事と主張の比重が同程度の文はクラスタリングに考慮しない
- 翻訳前の教師ありデータ（主張 or 出来事）の作成
  - 世界の全ての言語への対応にコストがかかる
- 記事特有の書き方の考慮
  - 言語学的なアプローチ

新規性

- 主張と出来事で分割してクラスタリングを行う点
- 目的の達成のために複数の既存手法を組み合わせている点
- 目的がより良く達成されるように分析・工夫を行う点

ニュース読者が**出来事**と**それに対する主張**を把握できる推薦手法の提案

記事の**出来事**と**主張**のクラスタを用いた
多言語ニュースの**主張の文**の推薦手法の提案

多言語である必要性

- 今の手法は単一言語でも適用できる可能性が高い
- 容易な機械翻訳を使い、翻訳後の文章を使うだけで多言語に対応できる手法
- 社会への貢献度がより大きい
- 翻訳機のその時代の性能が異なるため、再現性がないことに注意
  - R言語で再現性を確保する研究が存在

評価方法

- 出来事の記事が類似しているかを確認
  - LibRecライブラリ、IDOMAAR、STREAMINGRECなど
- 出来事とその主張が多角的に見れているかを分析
  - Desarkarら（2014）のニュースの多様性の評価手法
    - ニュースオブジェクト間の関連性と非類似性の両方を高くすべきであるという二基準の最適化問題
  - 自身で理由をつけて説明
  - アンケートと統計
- 推薦にかかる時間の測定
  - 一部手法を変更して比較

[13] M. Karimi, D. JannachとM. Jugovac, 「News recommender systems – Survey and roads ahead」, Information Processing & Management, vol. 54, no. 6, pp. 1203–1227, 11月 2018, doi: 10.1016/j.ipm.2018.04.008.

ニュース専用でないDebater Datasetを使うのはなぜか

- ニュースを出来事と主張で分類するデータが存在しない
  - アンケートで収集するのは簡単でない
    - 人によって観点が異なる
  - 優れた既存のデータセットをどれだけ応用できるか検証した方が有意義
- ニュース専用の教師ありデータセットを作ったときに上手くいくことが示唆される

記事特有の書き方の考慮

- ニュースは5W1Hで書かれる
  - 重要度の低いものから消える

[14] I. Fang, Writing Style Differences in Newspaper, Radio, and Television News. Monograph Series No. 1. Center for Interdisciplinary Studies in Writing, University of Minnesota, 227 Lind Hall, 207 Church St, 1991. 参照: 7月 13, 2021. [Online]. Available at: https://eric.ed.gov/?id=ED377481

目的はエコーチェンバー現象とフィルターバブルから派生

- エコーチェンバー現象
  - 価値観の似た者同士で交流し、共感し合うことにより発生
  - 特定の意見や思想が増幅されて影響力をもつ現象
  - 攻撃的な意見や誤情報などが広まる一因ともみられる
  - 特定地域の記事ばかりを読む状況では、世界規模でこの現象が発生しているといえる
- フィルターバブル
  - 読者に最適化されたコンテンツばかりがサジェストされる推薦システムの罠
  - 情報の泡に包まれてその他の情報が見えにくくなってしまっている状態

画像の提供元

- Online illustrations by Storyset
  - https://storyset.com/online
- ICOON MONO
  - https://icooon-mono.com/
- unDraw
  - https://undraw.co/