QAにおける評価用データセットの役割と日本語QAデータセットの必要性についての考察

田保健士郎、小林景（慶應大）

自己紹介

- 田保　健士郎（タボ　ケンシロウ）
- 慶應義塾大学　理工学研究科　小林景研究室所属　修士2年
- **自然言語処理特化の研究室所属ではないので、業界の不文律等があれば教えてください**
- 次年度からは金融工学系の研究業務に従事（自然言語処理もあるかも）

本発表の概要

・本発表ではQAデータセット全体を対象とした、やや俯瞰的な話をします

〇前半　QAにおける評価用データセットの役割について

・評価用データセットの役割の確認と用語の整理

・QAにおける2つの評価パラダイムの紹介

〇後半　日本語QAデータセットの必要性について

・日本語QAデータセットの整理

・翻訳QAデータセットの問題の可視化

・日本語QAデータセット(JAQKET)単語分布の特徴析出

データセットについて

自然言語処理における類似概念の整理(※明確な定義はなく、意見が分かれる可能性あり)

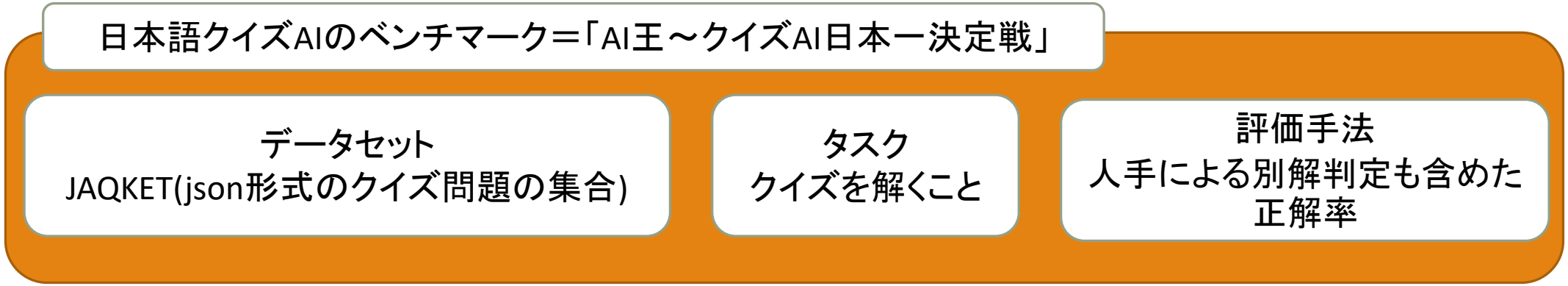
・ベンチマーク・・・データセット、タスク、評価手法のセット

・(評価用)データセット・・・特定のタスク、評価手法に対応する形で整理されたデータの集合

・コーパス・・・ある領域において集積され、構造化されたテキストのデータベース

・言語資源・・・上記を含め、研究に利用できる言語データの総体

例:「AI王～クイズAI日本一決定戦」[10]　https://sites.google.com/view/project-aio/home

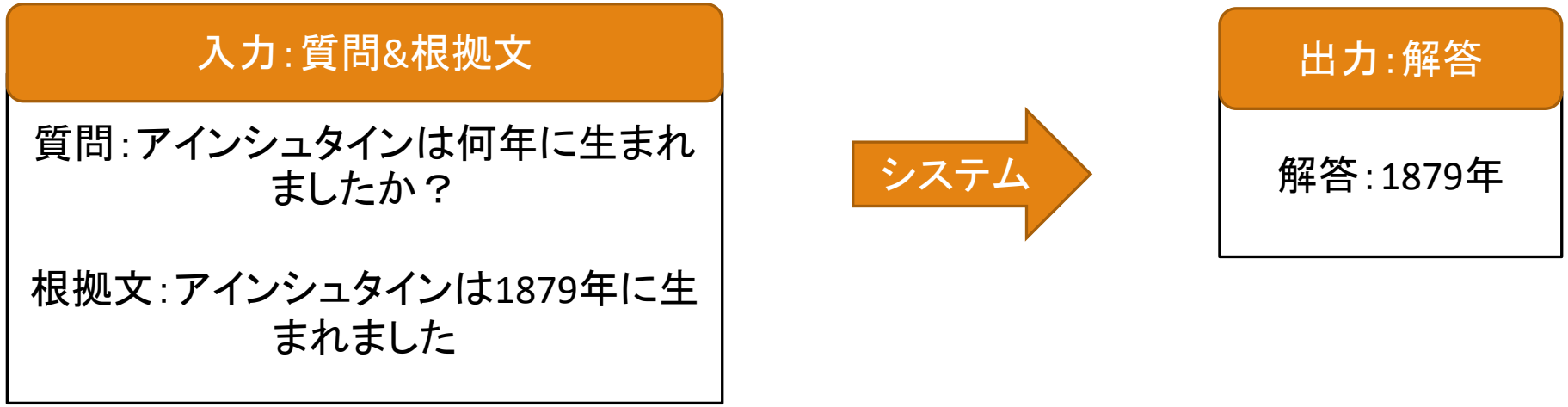


QAとは

○質問と解答があり、質問を基に解答を求めるタスク

・根拠文が与えられている場合はそれを使う

・根拠文が与えられていない場合は知識ベースやWebなどの知識源を基に解答する



※正確には、これはRetriever-Readerモデル寄りの説明で、解答を生成するモデルも存在する

QAにおける2つのパラダイム

違いはQAの目的設定

〇The Cranfield Paradigm(クランフィールドパラダイム)

・情報検索における伝統的なパラダイム

・図書館の司書のようにユーザーの役に立つことに主眼がおかれている

⇨ユーザーが教えられる側

・ネット上の質問を集積して作られたデータセットはこちらに属する

例)Natural Questions (Kwiatkowski et al., 2019)[12] 、MS MARCO (Nguyen et al., 2016)[13]

・産業界で意識されやすい

QAにおける2つのパラダイム

〇The Manchester Paradigm(マンチェスターパラダイム)

・(Rodriguez et al. 2021)[11]によって提示されたパラダイム

・情報検索以外の目的に主眼をおいている

・AIがどれだけできるのかを試す

⇨ユーザーが教える側

・コンピュータを理解するため、AIと人間の能力を比較するため、AIが人間に匹敵することを証明するため　に質問応答をする

・学術的関心に基づく場合が多い

・人々をワクワクさせるという特別な役割を担う(Jeoperdy!でクイズ王を負かしたIBM Watsonなど)

Rodriguez et al. 2021における主張

・クランフィールド=役に立つことに主眼、マンチェスター=AIの能力を試すことに主眼

・2つのパラダイムには一方が進歩すればもう一方も進歩するという相互作用が存在するが…

⇨であるがゆえに、認識されにくい問題が存在する

・一方のパラダイムにしか寄与しにくい研究が存在する
- 敵対的な例に対するロバスト性を保持する＝人工的な例になりがち⇨ユーザーに寄与するのか？

・マンチェスターパラダイムで成果を残しても、産業的に成功するまでにはギャップがある
- IBM Watsonのスピンオフの一つ「Watson Health」の売却(2022/01/21)
- SQuADは答えを知っている人間によって作られ、自然な質問とは異なる特徴を有することになり、結果的に「システムにとっての抜け道」が生まれてしまった

・ユーザーの役に立つことを意識した研究なのか、AGI(汎用人工知能)を目指した研究なのかは意識して取り組むべき

前半のまとめ

・QA研究には2つのパラダイムが存在

The Cranfield Paradigm・・・ユーザーの役に立つことが主眼

The Stanford Paradigm・・・AIの発展に寄与することが主眼

・各データセット、ないしはベンチマークにおいて、それを達成したら何ができたといえるのかを意識することが重要(QAに限らず言えることかもしれない)

・各データセットを利用する際には、どのパラダイムに属したデータセットなのか意識するとよい

Dataset	Paradigm	Domain	Author	Citation
Deep Read	Manchester	Stories	🧑	Hirschman et al. (1999)
TREC-8 QA	Cranfield	News	🧑	Voorhees (2000b)
TREC-9 QA	Cranfield	Search	🧑	Voorhees (2000a)
TREC QA 2001	Cranfield	Search	🧑	Voorhees (2001)
TREC QA 2002	Cranfield	Search	🧑	Voorhees (2002a)
TREC QA 2003	Cranfield	Search	🧑	Voorhees (2003)
TREC QA 2004	Cranfield	Search	🧑	Voorhees (2004)
TREC QA 2005	Cranfield	Search	🧑	Voorhees and Dang (2005)
TREC QA 2006	Cranfield	Search	🧑	Dang et al. (2006)
TREC QA 2007	Cranfield	Search	🧑	Dang et al. (2007)
QA4MRE 2011-2013	Manchester	Multiple	🧑	Peñas et al. (2013)
MCTest	Manchester	Stories	🧑	Richardson et al. (2013)
WebQuestions	Cranfield	Search	🧑+👤	Berant et al. (2013)
CNN/Daily mail	Manchester	News	🧑	Hermann et al. (2015)
Simple Questions	Manchester	Freebase	🤖→👤	Bordes et al. (2015)
Children's Book Test	Manchester	Stories	🤖	Hill et al. (2016)
bAbI	Manchester	Stories	🤖	Weston et al. (2016)
SQuAD 1.0	Manchester	Wikipedia	👤	Rajpurkar et al. (2016)

Rodriguez et al. 2021では、論文の最後に代表的なQAデータセットがどのパラダイムに属するかまとめられている

後半のまとめ(目次)

1. 日本語QAデータセットの整理

   ・・・QAデータセットに対する先行研究を基に分類・整理した

2. 翻訳QAデータセットの問題の可視化

   ・・・翻訳でデータセットを作成すると生じる問題を実際に確認した

3. 日本語QAデータセット単語分布の特徴析出

   ・・・JAQKETと英語QAデータセットを比較し、JAQKETの単語分布の特徴を析出した

分類の観点

先行研究(Rogers et al. 2021)[1]で紹介されている分類の観点

・IR(Information Retrieval)的かRC(Reading Comprehension)的か
　QAは検索、読解の2段階で解かれる

・形式
　質問の形式、解答の形式、エビデンス(根拠文)の形式の違いがある
⇨詳細は補足資料参照

・ドメイン
　どこから作られたか、どこを対象にしているか
　　(例:百科事典、フィクションなど)

・推論スキル(今回の分類では用いず)
　各データセットがどういった推論スキルで解けるか

・多言語か単言語か

データセット名	JAQKET(第 2 回)
year	2021
多言語か単言語か	単言語
RC 的か IR 的か	IR 的 QA
作成方法	新規作成
質問の形式	自然な質問
解答の形式	自由
エビデンスの形式	非構造化テキスト
エビデンスの量	部分的ソース
ドメイン	クイズ
元データ	“abc/EQIDEN” and クイズ作家
問題数	22,335+ 開発用 2,000

日本語QAデータセット分類結果1

データセット名	QAC-1	Japanese Slot Filling Quizzes	JQAC	JAQKET(第 1 回)	JAQKET(第 2 回)	運転 QA ドメインデータセット
year	2003	2020	2018	2020	2021	2019
多言語か単言語か	多言語内	単言語	単言語	単言語	単言語	単言語
RC 的か IR 的か	IR 的 QA	RC 的 QA	(QA 作成が目的)	IR 的 QA	IR 的 QA	RC 的 QA
作成方法	新規作成	新規作成	新規作成	新規作成	新規作成	新規作成
質問の形式	自然な質問	穴埋め, ストーリー補完	自然な質問	自然な質問	自然な質問	自然な質問
解答の形式	自由	カテゴリカル	抽出	多肢選択 (20 個)	自由	抽出 (一部解答無し)
エビデンスのモダリティ	非構造化テキスト	非構造化テキスト	非構造化テキスト	非構造化テキスト想定	非構造化テキスト	非構造化テキスト
エビデンスの量	部分的ソース	単一ソース	単一ソース	単一ソース	部分的ソース	単一ソース
ドメイン	ニュース	フィクション	百科事典	クイズ	クイズ	運転
元データ	毎日新聞記事 (1998,1999)	青空文庫	Wikipedia	“abc/EQIDEN” and クイズ作家	“abc/EQIDEN” and クイズ作家	運転
問題数	1,402		問題 1,018 解答 1,101	訓練用 13,061 開発用 995+997	22,335+ 開発用 2,000	20,007

スプレッドシート形式のものはこちらから(各データセットの提供元リンクも記載)

https://docs.google.com/spreadsheets/d/1uwuqM4i4-4UO-RxSeN1twwTaDwq8KaMB/edit?usp=sharing&ouid=113089316920471115347&rtpof=true&sd=true

日本語QAデータセット分類結果2

データセット名	常識に基づく推論のためのデータセット	SQuAD 翻訳データセット ([4] による提供)	JaSQuAD (AI Shift による提供)	SQuAD 翻訳データセット (@kumakura013 による提供)	NIILC Question Answering Dataset	Yahoo!知恵袋データ (第 3 版) (2021 年度版)
year	2020	2018	2020	2020	2003	2021
多言語か単言語か	単言語	多言語内	単言語	単言語	単言語	単言語
RC 的か IR 的か	RC 的	RC 的	RC 的	RC 的	IR 的	IR 的
作成方法	新規作成	人による翻訳	機械翻訳 (google 翻訳)	機械翻訳 (google 翻訳)	新規作成	新規作成
質問の形式	ストーリー補完	自然な質問	自然な質問	自然な質問	自然な質問 (ただし、タグが含まれる), クエリ (自然な質問をクエリ化)	自然な質問
解答の形式	多肢選択	抽出	抽出	抽出	自由 (ただし百科事典で解ける想定)	自由
エビデンスのモダリティ	非構造化テキスト	非構造化テキスト	非構造化テキスト	非構造化テキスト	構造化テキスト	非構造化テキスト
エビデンスの量	単一ソース	単一ソース	単一ソース	単一ソース	ソースなし	ソースなし
ドメイン	コーパス	百科事典	百科事典	百科事典	不明	該当なし
元データ		Wikipedia	Wikipedia	Wikipedia	作業者による	Yahoo!知恵袋
問題数	104,000	327	103.62MB (問題数未確認)	20,000	開発用 800 テスト用 200	質問 263 万 回答 670 万

スプレッドシート形式のものはこちらから(各データセットの提供元リンクも記載)
https://docs.google.com/spreadsheets/d/1uwuqM4i4-4UO-RxSeN1twwTaDwq8KaMB/edit?usp=sharing&ouid=113089316920471115347&rtpof=true&sd=true

整理により分かること

〇日本語QAデータセットとしては存在しない領域

・音声や動画、それらのハイブリッドをエビデンスとしたQAデータセット

(英語では、音声：DAQA[14]、Spoken-SQuAD[15]、Spoken-CoQA[16]、動画:MovieQA[17]、TVQA[18]、MarioQA[19]、PororoQA[20]など、ハイブリッド:HybridQA[21]、MultiModalQA[22]などが存在)

・オープンに共有された特定ドメインのQAデータセットが少ない(金融や医療のような)

(英語では、BioASQ[24]やHead-QA[25]などが存在)

※実際に日本語版が必要かどうかは作成コストと合わせて考える必要がある、翻訳によって補うかといった検討も必要

QAデータセットと言語の問題

言語資源が英語や一部の主要な言語に偏っていることが指摘されている

ACL2022のテーマ「言語の多様性：低資源言語から危機に瀕した言語まで」

会議名	英語の比率 (%)	次に多い言語	次に多い言語の比率 (%)	出典
ACL2004	87	中国語	9	Mielke 2016[158]
ACL2008	63	ドイツ語、中国語	4	Bender 2009[159]
ACL2008	87	中国語	16	Mielke 2016[158]
EACL2009	55	ドイツ語	7	Bender 2011[161]
ACL2012	86	中国語	23	Mielke 2016[158]
ACL2015	75	中国語	5	Munro 2015[162]
ACL2016	90	中国語	13	Mielke 2016[158]

[2]を参考に作成

翻訳に頼らざるを得ない現状

翻訳がなぜダメなのか(実証手順)

○実際に検証する

1. 英語版SQuAD[3]と日本語翻訳版(人手)SQuAD[4](同一問題・327問になるようにしている)を用意
   (SQuAD=Stanford Question Answering Dataset)

2. 英語版はDistilBERT、日本語版はBERTで解く(共にBERTベースのモデル、性能に大きな差はない)
   (DistilBERTはGoogle Colabratory上で学習、BERTは学習済みのものを使用)

3. その言語変化による性能の差異をみる

翻訳がなぜダメなのか(実証結果)

SQuADと日本語翻訳版SQuAD解析結果

	SQuAD(英語)	SQuAD(日本語)
EM	0.792	0.566
F1	0.871	0.566

EM=Exact Match(完全一致)
疑似=目視による修正

	SQuAD(英語)	SQuAD(日本語)
疑似 EM	0.894	0.739
疑似 F1	0.874	0.709
(疑似 EM)-(EM)	+0.0822	+0.143
(疑似 F1)-(F1)	+0.0240	+0.173

結果

・日本語においては各指標がそのままでは機能しないことを確認
・目視による修正を施しても、英語の方が高性能

目視による採点：https://docs.google.com/spreadsheets/d/1pyS7EQG4D7VGqKTgMB-vI1YvmC00TFZ8/edit?usp=sharing&ouid=113089316920471115347&rtpof=true&sd=true

翻訳がなぜダメなのか(考察)

SQuAD とSQuAD 日本語翻訳版解析結果(不具合の例)

	問題	模範解答 (日本語)	予測結果 (日本語)	模範解答 (英語)	予測結果 (英語)
1	スーパーボウル 50 は 50 周年記念を強調するために使用されたのはどの色か?	黄金, 黄金, 黄金	金色	gold,gold,gold	gold
2	ワルシャワで生まれた人のうちもっとも有名な一人は誰か?	Marie Curie(× 3 同じなので省略)	MarieCurie	Maria Sklodowska-Curie(× 3)	Maria Sklodowska-Curie
3	教育の科学を表す別の名前は何か?	教育, 教育学, 教育学	教育学、教育の科学の研究が含まれている。教師は、他の専門家のように、資格を得た後に継続して教育を受ける必要がある可能性があり、これは継続的な専門的発達	pedagogy,pedagogy,formal education	pedagogy
4	ビクトリア州の総生産国総生産はオーストラリアで何位にランクされるか?	2 位 (× 3)	2	second(× 3)	second
5	Jacksonville の市政府が統合したのは何年か?	1869 年,1869 年,1869	1869	1968(× 3)	1968

バラバラに翻訳することで生じる問題

注:時間の都合上全ては説明しません

JAQKETと各データセットの比較

○JAQKET[5]とは?

・AI王～クイズAI日本一決定戦～というコンペティションで使用されるデータセット

・学習用データ22,335問、開発用データ1,000問、評価用データ1,000問(未公開)


```
{
   "qid": "QA20QBIK-0002",
   "competition": "第1回AI王",
   "timestamp": "2020/01/27",
   "section": "開発データ問題 (dev1)",
   "number": "2",
   "original_question": "童謡『たなばたさま』の歌詞で、「さらさら」と歌われる植物は何の葉？",
   "original_answer": "ササ",
   "original_additional_info": "",
   "question": "童謡『たなばたさま』の歌詞で、「さらさら」と歌われる植物は何の葉?",
   "answers" :["ササ"]
}
```


[10]より引用

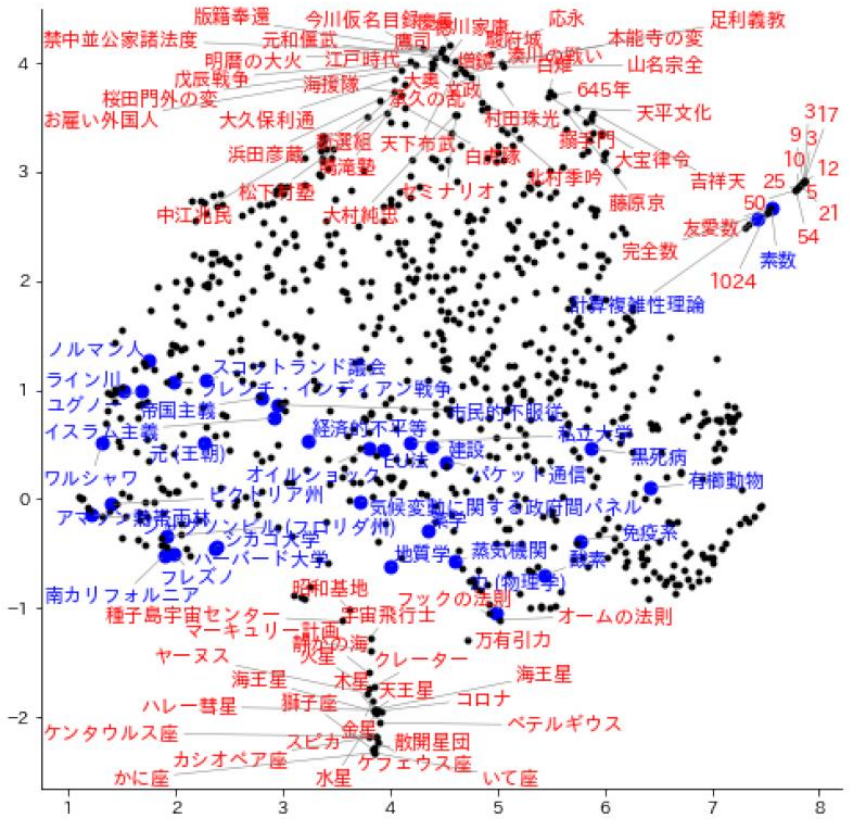
JAQKETとSQuAD2.0の単語分布

用いるもの : JAQKET 開発用データ1,000問の解答、SQuAD2.0[6] バリデーションデータのタイトル35個

手順

1. SQuAD2.0バリデーションデータのタイトル35個を翻訳(人手)
2. 学習済み日本語版Wikipedia2Vec(100次元)[8]でベクトルを取得(JAQKETは827個、SQuAD2.0は34個取得できた)
3. PCA、t-SNE、UMAP、各手法で次元を落として可視化(ここではUMAPによる可視化の結果を紹介)

JAQKETとSQuAD2.0の単語分布



JAQKET and SQuAD2.0(UMAP)

・JAQKETが日本史的単語を特徴的に有することがわかる

・天文学的単語も特徴的に析出されている

黒点：JAQKET、青点：SQuAD2.0

特徴的な単語をアノテート
赤字：JAQKET 、青字：SQuAD2.0

JAQKETとTriviaQAの単語分布

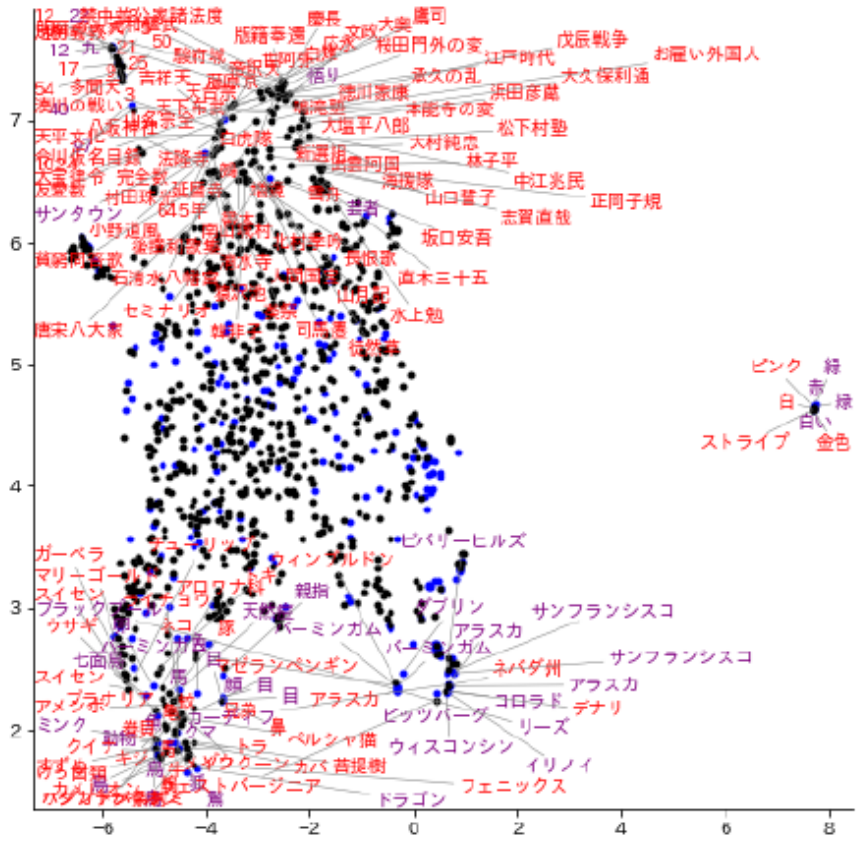
TriviaQA[7]はトリビアクイズ愛好家によって作られた95,000 の問題と解答のペアからなる英語のQA データセット

属性がJAQKETと近いので、先の違いがトリビアクイズによるものなのか言語によるものなのかわかる

手順

1. TriviaQAのバリデーションデータの解答17,944 個の中からランダムに500 個取り出す(JAQKETは先と同じ)
2. Googletrans API(Google翻訳のAPI)を利用して日本語に翻訳
3. 学習済み日本語版Wikipedia2Vec(100次元)[8]でベクトルを取得(JAQKETは827個、TriviaQAは235個取得できた)
4. PCA、t-SNE、UMAP、各手法で次元を減らして可視化(ここではUMAPによる可視化の結果を紹介)

JAQKETとTriviaQAの単語分布



JAQKET and TriviaQA(UMAP)

・先と同じくJAQKETが日本史的単語を特徴的に有する

・天文学的単語は析出されず(トリビアクイズの特徴か)

黒点：JAQKET、青点：TriviaQA

特徴的な単語をアノテート
赤字：JAQKET 、紫字：TriviaQA

全体のまとめ

〇前半

・データセットという言葉の定義を確認するとともに、QAにおいてデータセット作成の際に意識するべき2つのパラダイムを紹介した

〇後半

・日本語QAデータセットの整理と、そこから見えてくる英語QAデータセットとの差を紹介した

・日本語QAデータセットを、特に工夫せずに英語から翻訳することで作成した場合に生じる問題を確認した

・代表的なQAデータセットの単語分布を比較することにより、日本語でオリジナルに作成したQAデータセットが日本の文化に関する単語を特徴的に有することが確認された(英語もしかり)

=根拠文が付随しない場合、翻訳先の言語で、解くのに必要な情報を集めにくい可能性

参考文献

[1] Anna Rogers et al., QA Dataset Explosion:A Taxonomy of NLP Resources for Question Answering and Reading Comprehension",*CoRR*, Vol.abs/2107.12708, 2021.

[2] Emily M. Bender. The #BenderRule: On Naming the Languages We Study and Why It Matters. September 2019.

[3] Pranav Rajpurkar et al. SQuAD: 100, 000+ Questions for Machine Comprehension of Text, *CoRR*,Vol.abs/1606.05250, 2016.

[4] Akari Asai et al.,Multilingual Extractive Reading Comprehension by Runtime Machine Translation, *CoRR*, Vol.abs/1809.03275, 2018.

[5] Masatoshi Suzuki et al., JAQKET: クイズを題材にした日本語QA データセットの構築, 言語処理学会第26 回年次大会発表論文集, 2020.

[6] Pranav Rajpurkar Robin Jia, and Percy Liang, Know What You Don't Know: Unanswerable Questions for SQuAD, *CoRR*, Vol.abs/1806.03822, 2018.

[7] Mandar Joshi et al.,TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, In Proceeding of the 55th Annual Meeting of the Association for Comuputational Linguistics(Vol 1:Long Papers), pp.1601-1611, 2017.

[8] Ikuya Yamada et al., Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia, In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 23-30., 2020.

[9] Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996{5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] AI 王～クイズAI 日本一決定戦～. https://sites.google.com/view/project-aio/home. (Accessed on 01/16/2022).

参考文献

[11] Pedro Rodriguez, Jordan Boyd-Graber, Evaluation Paradigms in Question Answering, In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.

[12] Tom Kwiatkowski et al., Natural Ques tions: a Benchmark for Question Answering Research., *Transactions of the Association of Computational Linguistics*, 2019.

[13] Tri Nguyen et al., MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR*, Vol. abs/1611.09268, 2016.

[14] Haytham M. Fayek and Justin Johnson. Temporal Reasoning via Audio Question Answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2283-2294, 2020.

[15] Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *CoRR*, Vol. abs/1804.00320, 2018.

[16] Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. Towards Data Distillation for End-to-end Spoken Conversational Question Answering. *CoRR*, Vol. abs/2010.08923, 2020.

[17] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. , 2016.

[18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369{1379, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[19] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. MarioQA: Answering Questions by Watching Gameplay Videos. *CoRR*, Vol. abs/1612.01669, 2016.

[20] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. DeepStory: Video Story QA by Deep Embedded Memory Networks. In *IJCAI*, pp. 2016-2022, 2017.

参考文献

[22] Wenhu Chen et al., HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026{1036, Online, November 2020. Association for Computational Linguistics.

[23] Alon Talmor et al., MultiModal*f*qa*g*: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021.

[24] George AU Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, et al. SemEval2016 Task 3: Community Question Answering. In *An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition*. Association for Computational Linguistics, 2015.

[25] David Vilares and Carlos Gomez-Rodrguez. HEAD-QA: A Healthcare Dataset for Complex Reasoning., *CoRR*, Vol. abs/1906.04701, 2019.

分類の観点(形式について)

〇質問の形式
- ・自然な質問・・・人間が行う質問の形式に近いもの
- ・クエリとしての質問・・・検索エンジンでなされる形式
- ・穴埋め形式の質問・・・一部がマスキングされることで作成された質問の形式
- ・ストーリー補完・・・後に続く文章を予測するなどの形式で、穴埋め形式に近い

〇解答の形式
- ・抽出形式・・・証拠文(エビデンス)の中から解答スパンを抽出して答える形式
- ・多肢選択形式・・・選択肢の中から解答を選択して答える形式
- ・カテゴリカル形式・・・選択肢形式ではないが、日付や時間のように答えのカテゴリーが制限されている形式
- ・自由形式・・・特に制限を設けない形式で、抽出or生成によって解答されることが多い

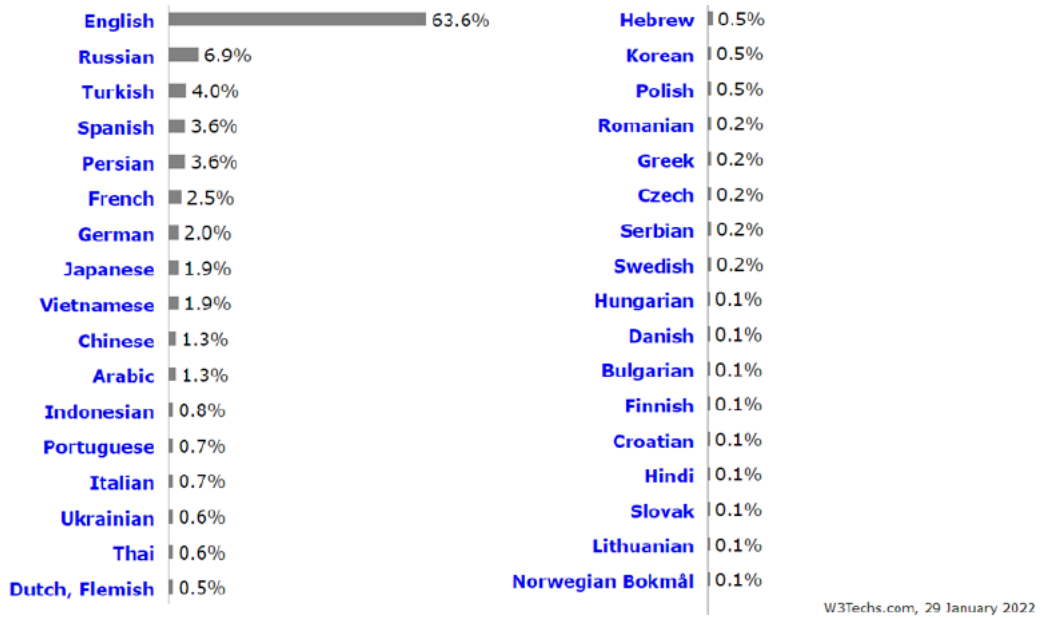
〇エビデンスの形式　主に3つの視点がある

エビデンスのモダリティ・・・テキストが構造的か、画像なのか音声なのか動画なのかなど

エビデンスの量・・・単一ソース、複数ソース、部分的ソース、ソース無し

動的か静的か・・・時間的発展がある(動的)かない(静的)か

言語資源の偏り



世界のウェブサイトにおける使用言語の割合
(W3Techs.comより;2022年1月29日時点)

総記事数順位	総記事数	言語	点数 B
8	1,746,979	スペイン語	99.94
15	1,250,918	中国語	99.80
7	1,787,968	ロシア語	99.50
1	6,441,836	英語	97.32
17	1,133,531	ウクライナ語	89.98
5	2,391,211	フランス語	84.32
12	1,310,085	日本語	83.41
4	2,655,899	ドイツ語	81.42
13	1,270,088	ベトナム語	74.70
9	1,737,066	イタリア語	74.69
18	1,081,652	ポルトガル語	72.41
16	1,154,823	アラビア語	55.37
11	1,506,021	ポーランド語	53.54
6	2,078,602	オランダ語	53.44
3	2,717,221	スウェーデン語	51.33
14	1,265,641	ワライ語	18.81
10	1,539,192	エジプト・アラビア語	9.97
2	6,110,413	セブアノ語	5.98

各言語版Wikipedia 総記事数と点数B(点数B による降順;2022 年1 月29 日時点)

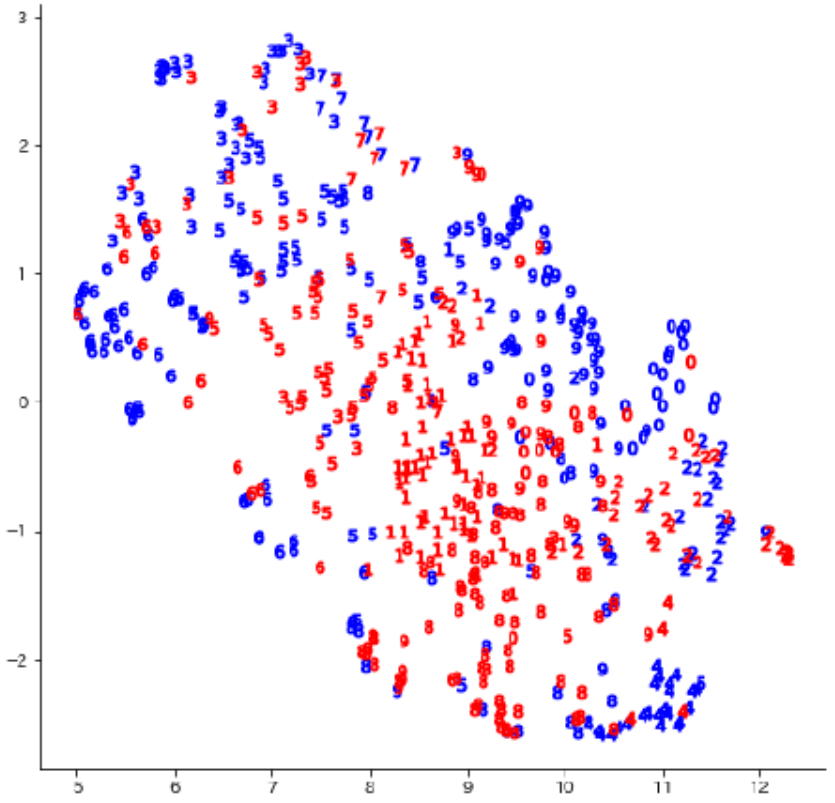
データセット間差異の定量的評価

〇kmeans法による比較

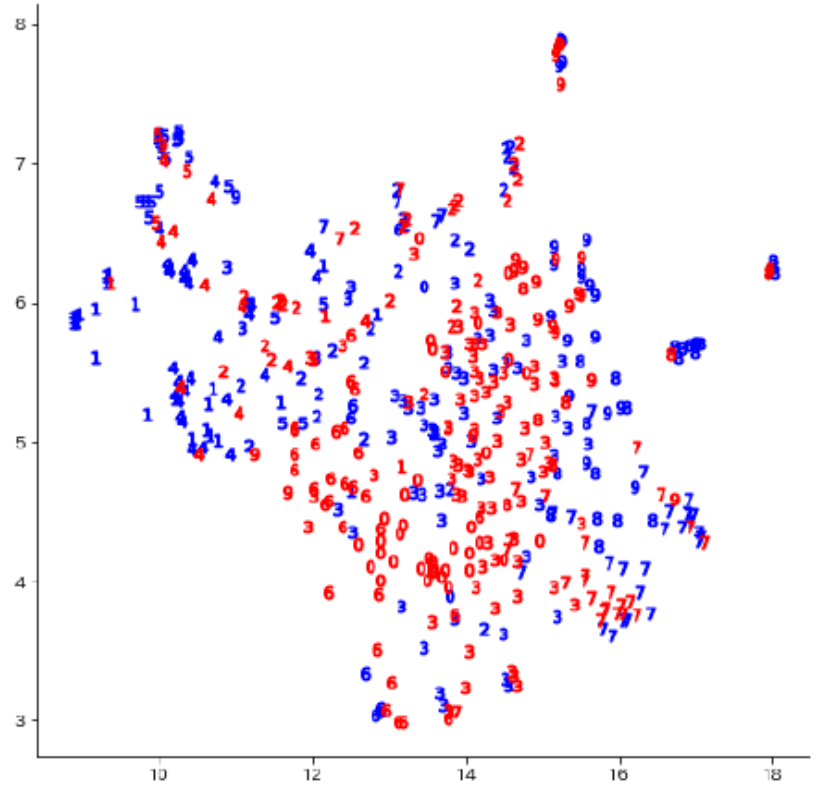
手順

1. 各データセットに対して学習済み日本語版Wikipedia2Vecを用いてベクトルを取得
   翻訳が必要な場合はgoogletrans APIで日本語に翻訳している
2. 比較対象のデータセットのサイズをサンプリングにより等しくする
   例えば、JAQKET 827個とSQuAD2.0(train+validation) 291個だったらJAQKET 827個から291個をサンプリング
3. kmeans法を用いて10個の群に分類
4. 可視化(UMAP)して確認&全体での割合と各群での割合の差をみる

データセット間差異の定量的評価



JAQKET(赤)とSQuAD2.0(青)(UMAP)



JAQKET(赤)とTriviaQA(青)(UMAP)

データセット間差異の定量的評価

JAQKETとSQuAD2.0の各群割合

群	割合	デフォルト値からの差
0	0.314	-0.186
1	0.971	+0.471
2	0.567	+0.0667
3	0.346	-0.154
4	0.138	-0.362
5	0.438	-0.0625
6	0.246	-0.254
7	0.467	-0.0333
8	0.753	+0.253
9	0.325	-0.175

JAQKETとTriviaQAの各群割合

群	割合	デフォルト値からの差
0	0.949	+0.449
1	0.111	-0.389
2	0.511	+0.0111
3	0.548	+0.0476
4	0.268	-0.232
5	0.167	-0.333
6	0.816	+0.316
7	0.473	-0.0273
8	0.258	-0.242
9	0.523	+0.0227

TriviaQAとSQuAD2.0の各群割合

群	割合	デフォルト値からの差
0	0.477	-0.0231
1	0.722	+0.222
2	0.430	-0.0698
3	0.536	+0.0362
4	0.371	-0.129
5	0.438	-0.0625
6	0.484	-0.0161
7	0.509	+0.00909
8	0.667	+0.167
9	0.459	-0.0405

※　割合は左2つがJAQKETの占める割合、右はTriviaQAの占める割合
※　デフォルト値からの差は(群での割合)-(全体の割合;ここでは0.5)

データセット間差異の定量的評価

デフォルト値からの差まとめ

データセットの組	絶対値の平均	絶対値の最大値
JAQKET と SQuAD2.0	0.202	0.471
JAQKET と TriviaQA	0.207	0.449
TriviaQA と SQuAD2.0	0.0775	0.222

・定量的にもJAQKETが特異な分布を有していることが確認された

・また、googletrans APIをdeepl APIに変更(翻訳精度を向上)して行った実験により、
TriviaQAやSQuAD2.0がアメリカ文化を象徴する単語を多く有していることも確認された。
⇨間接的に、翻訳&各言語で学習されたコーパスにより各言語圏の文化的単語が失われやすいことを示した