Vol.2021-MPS-133 No.12
Vol.2021-BIO-66 No.12
2021/6/29

クラウドソーシングを用いた字幕作成の性能評価

目良侃太郎[1]　杉村太一[1]　小板隆浩[1]

概要：本研究では，言語の異なる視聴者が動画を楽しむために必要不可欠な字幕の作成方法として，クラウドソーシングを用いた字幕作成手法を提案した．提案手法と YouTube の自動字幕作成システムによる字幕と Amazon Transcribe による字幕と比較した結果，提案手法は最も高い精度を示し，字幕作成において有用な手法となり得る可能性を示した．

キーワード：クラウドソーシング，字幕作成，多数決手法

Performance Evaluation of Making Captions by Crowdsourcing

KANTARO MERA[1]　TAICHI SUGIMURA[1]
TAKAHIRO KOITA[1]

1. はじめに

言語の異なる視聴者が動画を楽しむためには字幕が必要である．既存の自動字幕作成システム（既存システム）は，動画がアップロードされている YouTube などのプラットフォームに実装されており，動画の音声データから字幕を作成する．既存システムが扱う音声データは，複数の音の合成音である．合成音の字幕作成をするためには，合成音から個別の音を聞き分ける必要があり，既存システムでは正しい字幕が作成されない場合も多い．そこで，本研究では，個別の音を聞き分けることに適した人を，効率的に使うクラウドソーシングを用いる手法を提案する．クラウドソーシングとは，インターネットを通じて不特定多数のワーカーに仕事を依頼する仕組みのことであり，機械学習の教師データ収集などにおいて広く利用されている．先行研究では，クラウドソーシングを用いた字幕作成手法では既存システムよりも高精度の字幕を作成できることが明らかにした[1]．先行研究では，作成した複数の字幕それぞれの精度と既存システムで作成した字幕の精度を比較しており，提案手法では複数の字幕から 1 つの最良の字幕を作成していない．そこで，本研究では作成した複数の字幕から一意の字幕を作成するためにクラウドソーシングを用いた多数決手法を用いる．本研究の目的は，先行研究の手法にクラウドソーシングを用いた多数決手法を用いることにより一意の字幕を作成する手法を提案し，提案手法の有用性を示すことである．

2. 多数決手法

本研究で採用する多数決手法について述べる．クラウドソーシングはインターネットを通じて不特定多数のワーカーに仕事を依頼できるため高い利便性がある一方，悪意のあるワーカーによる回答などに起因するデータ品質の劣化が一般的な問題としてある．多数決手法とはクラウドソーシングにおける手軽なデータ品質手法の一つである．1 データあたり複数の回答を募り多数決を行う方法が挙げられる[2,3,4]．事前に十分な数の回答を集めることで回答品質は大幅に改善される[5]．本研究では，回答者をある一定の基準などでフィルタリングするなどしない単純多数決手法を用いる．

3. 実験

3.1 実験内容

本実験では 2 つのシステムと 1 つの手法を用いて字幕を作成し，それぞれの精度を比較する．用いる手法は下記の 3 つである．

- YouTube の字幕
  YouTube に実装されている字幕作成システムで字幕を作成する．
- Amazon Transcribe による字幕
  Amazon Transcribe とは Amazon Web Service が提供する Automatic speech recognition と呼ばれる深層学習プロセスを使って迅速かつ高精度に音声をテキスト変換するサービスのことである[6].
- クラウドソーシングによる字幕
  本研究の提案手法である．動画の字幕作成をクラウドソーシングで依頼し，得られた結果に対して，再度クラウドソーシングで多数決をとり最良の字幕を決定する．

3.2 提案手法を用いた字幕作成

本実験では，1 分の英語の動画に対して 5 名のワーカーに

1 同志社大学大学院理工学研究科
Graduate School of Engineering, Doshisha University

Vol.2021-MPS-133 No.12
Vol.2021-BIO-66 No.12
2021/6/29

字幕作成を依頼する．次に，得られた 5 つの字幕データの中で最良の字幕を 10 名の多数決で決定する．本実験では，字幕作成も字幕作成もすべて 1 人あたり$0.05 で作業を依頼する．

3.3 字幕の精度

本実験では，精度を動画の台本と作成された字幕を行ごとに比較し，同じ内容の行数を字幕の得点 s とし，s を台本の行数 n で割ったものに 100 をかけたものと定義する．式(1)に精度を求める計算式を示す．

$$P(\textit{精度}) = s(\textit{字幕の得点}) \div n(\textit{台本の行数}) \times 100 \qquad (1)$$

4. 実験結果

4.1 各手法で作成した字幕の精度比較

YouTube の字幕と Amazon Transcribe による字幕，提案手法による字幕の精度を図 1 に示す．精度は小数点第 2 位以下を切り捨てとする．YouTube による字幕の精度は 64.7%で，Amazon Transcribe による字幕の精度は 76.5%であった．提案手法による字幕の精度は 88.2%となり，3 つの手法の中で最も高い精度となった．

5. 考察

本実験結果を精度と作成時間，コストの 3 つの観点で比較したものを表 1 に示す．YouTube では動画をアップロードするだけで自動的に字幕が作成されるため字幕を作成するためにコストは発生しない．また，YouTube で字幕が作成されるまでに必要な時間は数分である．Amazon Transcribe では 40 秒で 1 分の動画の字幕を作成することができる．字幕作成にかかるコストは，アジアパシフィック（東京）リージョンでは$0.016 である．クラウドソーシングを用いた提案手法では字幕を作成するために 1 日必要である．提案手法の作成時間に関しては，本研究においては字幕の作成と多数決による最良な字幕の決定のために 2 度タスクを作成し依頼する手順を人手で行った。この手順をプラットフォーム上で自動化することで，人手を用いることで発生している時間を削減し，全体の作成時間を短縮することが可能である. 作成にかかるコストとしては$0.75 が必要である．精度のみで各手法による字幕を比較した際には提案手法は，他の手法と比較して高い精度であるが，精度と作成時間，コストの 3 つの観点で比較した際には Amazon Transcribe が相対的に有用と考えられる．

6. まとめ

本研究ではクラウドソーシングを用いた字幕作成手法に多数決手法を追加した新たな手法を提案し，YouTube の自動字幕作成システムによる字幕と Amazon Transcribe による字幕と比較した．精度と作成時間，コストの 3 つの観点から比較した際には Amazon Transcribe が最も有用な手法であったが，提案手法は字幕作成において最も高い精度となった．したがって，提案手法は字幕作成において有用となる可能性を示した．また，本研究では 1 人の話者の 1 分の動画に対してのみ実験を行ったため，コンピュータにとって困難な音の聞き分けを必要とする複数人の話者の動画に対しても実験を行う必要がある．また，動画に対して 1 時間あたり単語数などの観点から動画による精度の違い等についても調査することで各使用用途において，YouTube の自動字幕作成システムと Amazon Transcribe，提案手法のどの手法が適切な手法であるか，さらなる調査が必要である．本研究では，提案手法は動画の字幕作成において有用な手法となり得る可能性を示した．

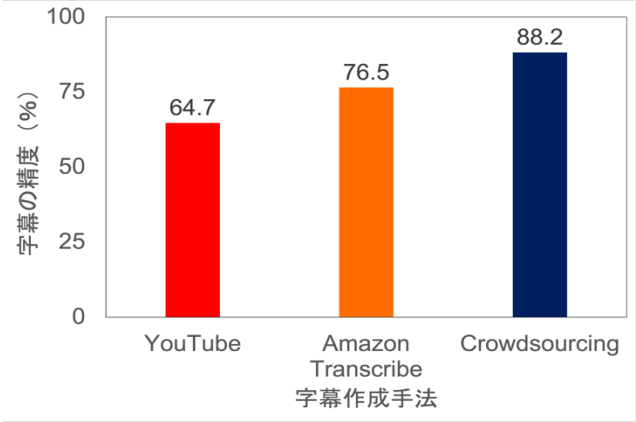

図 1　各手法の字幕精度

表 1　各手法の多角分析

	精度（%）	作成時間	コスト（$）
YouTube	64.7	数分	0
Amazon Transcribe	76.5	40 秒	0.016
提案手法	88.2	1 日	0.75

参考文献

[1] Kantaro Mera et al.: A proposed system for machine translation by crowdsourcing, Bulletin of Networking, Computing, Systems, and Software, vol 10, no1, pp.25-26, 2021.
[2] Snow et al.: Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks, Proc. of the EMNLP'08, Association for Computational Linguistics, pp.254-263, 2008.
[3] Sorokin et al.: Utility data annotation with Amazon Mechanical Turk, Proc. if the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp.1-8, 2008.
[4] Callison-Burch et al.: Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk, Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.286-295, 2009.
[5] Raykar, V. C. et al.: Ranking annotators for crowdsourced labeling tasks, Advances in Neural Information Processing Systems 24, Curran Associates, Inc., pp.1809-1817, 2011.
[6]“Amazon Transcribe 音声をテキストに自動的に変換する”. https://aws.amazon.com/jp/transcribe/, (参照 2021-05-24).