

AI 技術を使った誤テロップ自動検出に関する技術検証

The Poc on automatic detection of erroneous telop using AI technology

中野 信^{*1}
Makoto Nakano

^{*1} 日本テレビ放送網株式会社
Nippon Television Network Corporation

Keyword. Television, Telop, Broadcast, TV program

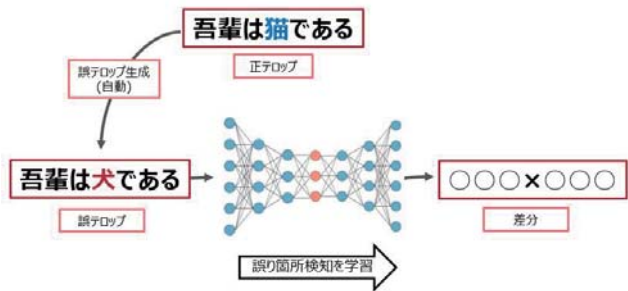
1. 背景

放送局で制作・放送しているテレビ番組では、画面上にいくつもの「文字テロップ」が、次々に切り替わりながら表示されている。これは、放送している内容を視聴者の方にスムーズに分かっていただくためであったり、演出上の工夫であったりする。しかし、時に、このテロップの文字に誤りがあり、不体裁となる場合が少なからず発生している。放送している画面にテロップが表示されるまでには、人間による校閲を経ているのではあるが、どうしても見逃しが出てしまっている。これは、熟練した校閲者であっても人手による確認では限界があること、また、テロップは毎日大量に作られており、チェック作業量は膨大である、などもあり、誤テロップを完全になくすことは困難である。そこで、このチェック作業を補助することを目的とした自動チェックの仕組みが、AI 関連技術を使ってできないか、技術検証を行った。

なお、検証に当たっては、富士通株式会社様に技術協力を頂いた。

2. 概要

基本的なアプローチは、まず、正しいテロップ<正テロップ>から、過去の校閲事例を参考にして誤りを埋め込み生成したく誤テロップ>を生成、この誤テロップを RNN (Recurrent Neural Network)へ入力し、テロップの「誤り箇所検出の学習」をさせる、という手法とした。



誤テロップ生成の方式は、実際の校閲事例を参考に、以下8つの誤りパターンに分類した。

- ・同じ読みの漢字で置換 (漢字 → 幹事)
- ・似た平仮名/片仮名を置換 (づくし → ずくし)
- ・文字削減 (Congratulations → Congratulation)

- ・文字追加 (ドラゴンズ → ドランゴンズ)
- ・文字入れ替え (ノビレチン → ノビチレン)
- ・ら抜き言葉 (見られる → 見れる)
- ・近い意味の言葉で置換 (猫 → 犬)
- ・数字を置換 (1998 年 → 1996 年)

また、この誤りパターンごとに Neural Network を分けて学習させるという「モデル分割」という手法も取り入れ、分割有り・無しでの比較も行った。

なお、RNN へ入力する学習データであるが、実際に放送で利用されたテロップのテキストデータの十分な量の入手が困難であるという事情のため、Wikipedia から抽出した文章もあわせて使い、以下のようなデータセットで検証を行った。

	学習データ	モデル分割	学習データ数	検証データ	検証データ数
A	Wikipedia	有り	640万件 (×5モデル)	Wikipedia	3.2万件 (×5モデル)
B		無し	3200万件		16万件
C	校閲事例集	有り	2500件	校閲事例集	2668件
D		無し	2500件		168件

また評価指標は以下とした

入力	吾輩は猫である	苗字はまだない	夏日漱石
	TN	TP	FP
	正常検知成功	異常検知成功	誤検知
			FN
			異常検知漏れ
出力	○○○○○○○	××○○×××	○○○○

評価値	説明	計算例	評価値
TP	異常検知 (P) に成功 (T) した件数	—	2
TN	正常検知 (N) に成功 (T) した件数	—	13
FP	正常検知 (N) に失敗 (F) した件数	—	3
FN	異常検知 (P) に失敗 (F) した件数	—	1
Precision	異常判定した中で実際に異常であった割合	$TP / (TP + FP) = 2 / (2 + 3)$	40%
Recall	異常データの中で実際に異常判定した割合	$TP / (TP + FN) = 2 / (2 + 1)$	67%
Accuracy	全体データの中で判定が正しかった割合	$(TP + TN) / (TP + TN + FP + FN)$	79%
F-measure	RecallとPrecisionの調和平均	$2 * Recall * Precision / (Recall + Precision)$	0.73

3. 検証

技術検証を進めて行った結果、以下のような課題や新たなアイデアが出てきた。

- 誤テロップ補正の学習が収束しない
- 推定原因
- ・ 学習データ生成時の自動誤テロップ生成に問題がある。
- 改善対策案

連絡先: 日本テレビ技術統括局技術開発部 中野 信
m-nakano@ntv.co.jp

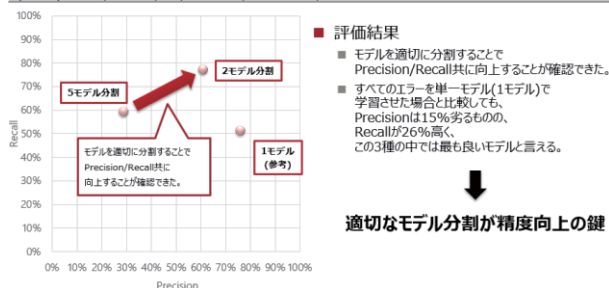
＜誤テロップ生成条件を制約する＞

- ・数字誤りは同一桁数の置換に限定
 - ・前後の文字置換は同一文字種別の交換に限定
 - ・過剰文字は同一文字種別に挟まれた箇所への挿入に限定
- ＜誤変換生成辞書の拡充＞
- ・ipadic 辞書(36 万件)⇒ipadic 辞書+SKK 辞書(52 万件)
- ＜誤りパターンの整理＞
- ・前述 8 パターンを、5 パターンへ集約

●学習済みモデルの分割について

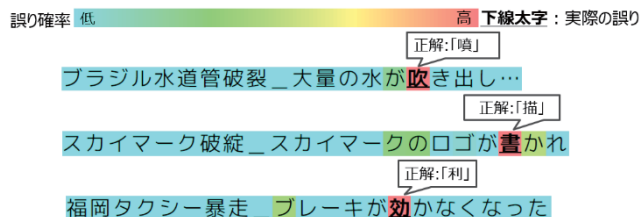
前述した「モデル分割」の手法であるが、誤りパターンごとにモデル分割した際に、個別のモデルでは高い精度が出るものの、単一文章に対してモデルの重ね掛けを行うと精度が非常に低い、ということがわかった。そこで、さらにモデルの分割数を減らす(前述の 5 パターンを、言語的誤り検出、意味的誤り検出の 2 モデルへ集約)ということを試みたのでその結果を以下に示す。

モデル数	TP	TN	FP	FN	Precision	Recall	Accuracy	F-measure
2モデル	637,852	23,508,900	416,401	192,442	61%	77%	98%	0.68
5モデル	154,875	7,800,621	370,237	106,841	29%	59%	94%	0.39
1モデル	132,246	8,130,080	40,653	129,470	76%	51%	98%	0.61



●誤り検出結果の可視化

今回の取り組みでは、実際にテロップ制作や、校正・校閲作業を担っている現場担当者に対するヒアリングも行った。その際に「検出した誤りを、校閲者に分かりやすく提示できないか」という意見が出て、それを実現するために「どれくらいの確率で誤りである、と判断したか含め分かりやすく可視化する」ということに取り組み、以下のように、誤り推定確率を色の違いで表現する、という手法を考案した。



4. まとめ

今回、DeepLearning を誤り検出に応用した事例がほとんど無い中で技術検証を行い、条件に制約はあるものの、技術的に可能であることを確認できた。

また、大量の正しいテキストに誤りを埋め込んで教師データ化し、学習を行う手法の有効性を実証した。これは、テロップの実データに限りがある中、今後の取り組みに大いに役に立つものと思われる。一方、実データを使つての学習においては、校閲事例の蓄積フォーマットが、学習用データとするには整って

おらず、クレンジングの手間が多く掛かる結果となった。そのため、今後の取り組みに備えるためにも、蓄積フォーマットを整備する、という現場運用の課題も見出せた。

また、校閲事例から作った学習用データの数を増やせば増やすほど、自動誤テロップ検出精度が向上することも確認でき、前述のフォーマットの整備とあわせて、データ整備の必要性を認識できた。

最後に、現場で使う場合を想定した手法、Neural Network による誤り推定確率を色で表現し、校閲者に分かりやすい形で出力する、ということもできることが分かり、作業量削減のための現場での適用という目的に沿う技術であるという目処も得られた。

5. 今後に向けて

① 実データ(テロップデータ及び校閲事例集)の蓄積・管理
DeepLearning を使った誤テロップ検出の精度改善には実データの教師データ化が非常に重要であることが実証できたのでデータを効率よく蓄積、管理する仕組みを整備する。

② 誤テロップ検出モデルの改善

実際に放送に使用したテロップに対しての誤り箇所検出精度が Precision 45%、Recall 34%程度であり、誤テロップ検出モデルのさらなる改善の余地がある。

③ 固有表現のモデル分離

固有名詞や数詞の誤りは、校閲事例を振り返っても約 5 割を占めていてぜひ自動チェックしたいところであるが、パターンで判断できるものではなく、現状のモデルには不適。ここを実現できるモデルをぜひ見つけたい。

③ 事実を参照できる仕組み

上記③にも書いたが、一般的な Neural Network はパターンを判断するのみで、情報を記憶しない。しかし、時々刻々変化する人の年齢や、スポーツにおける成績(勝率など)は、人間でもチェックが困難で時間が掛かる。これが自動化できるとかなりの校閲作業省力化が見込まれるため、ぜひ実現したい。

最後ではあるが、実験検証実施、本論文作成に当たり多大なご協力を頂いた富士通株式会社様へ深謝いたします。