The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Twitter URL Paraphrase Corpus に基づく要約データセットの構築

Construction of dataset for summarization based on the Twitter URL Paraphrase Corpus

永塚 光一[*1]　　渥美 雅保[*1]
Koichi Nagatsuka　　Masayasu Atsumi

[*1]創価大学理工学部情報システム工学科
Information Systems Science, Faculty of Science and Engineering, Soka University

The purpose of text summarization is to produce a condensed version of an input text which has a core meaning of the original. Most of the summarization systems are built on dataset using news articles. While Social Networking Service(SNS) such as Twitter is increasingly becoming an important information resource, the lack of dataset for SNS is one of the challenges in extending the range of summarization systems. In this paper, we address the problem by transforming the Twitter URL Paraphrase Corpus into summarization dataset. In order to extract important paraphrases for summary and increase the number of high quality paraphrases, we created the paraphrase classifier and paraphrase generator using supervised learning based on the corpus. In experiments, we evaluate paraphrase classifier by quantitative evaluation and paraphrase generator by qualitative evaluation respectively.

1. はじめに

近年, アテンションを取り入れたエンコーダデコーダモデル[Bahdanau 14]による機械翻訳における性能向上を受けて, ニューラルネットワークを用いた自動要約研究が盛んに行なわれている.自動要約研究に用いられるデータセットとして, DUC[Over 07]や, Gigaword[Graff 03]などがある. しかし, こうした要約データセットの多くが, 要約対象のデータとしてニュース記事を採用している一方で, 現在ニュース記事と並んで Twitter に代表される SNS が大きなテキスト情報源となりつつある.

本論では, URL を共有する tweet に基づいて自動収集されたパラフレーズのデータセットである Twitter URL Paraphrase Corpus[Lan 2017]を要約データセットの構築へ応用することを提案する. Twitter URL Paraphrase Corpus の一部には, 人間の評価者により各派生 tweet に対する元 tweet との意味的類似度ラベルが付与されており, これを教師データとして 2 つの tweet が paraphrase とみなせるか否かを判定する paraphrase 分類器を作成し, 良質な paraphrase ペアを収集している. しかしながら, twitter 要約データセットには, 入力としてより多くの paraphrase ペアを収集することが望ましい. そこで, 本手法では, paraphrase 分類器に加え, 元 tweet から派生 tweet を生成する paraphrase 生成器を学習により作成する. 要約データセットの構築は, 分類器と生成器により, 元 tweet に対して派生 tweet を選別, もしくは生成することによりなされる. 本論では, 分類器と生成器の性能を実験により評価することを通じて, これらを用いて, 要約データセットを構築することの実現性を示す.

2. データセット

2.1 Twitter URL Paraphrase Corpus

Twitter URL Paraphrase Corpus は,教師ツイート 1 文とそれに対する複数の派生ツイートを 1 ペアとして,51524 ペアから構成される.教師ツイートは,CNN や New York Times, BBC などのニュース機関の公式アカウントから投稿されたもので, 派生ツイートには,教師ツイートとのパラフレーズとしての意味的類似度がラベル付けされている.この意味的類似度の決定においては, 6 人の評価者により,各派生ツイートが教師ツイートに対するパラフレーズと言えるかどうかの投票が行われ,最小で 0,最大で 6 の 7 段階でラベルが付与される. これに加え,ラベル付けがされていないサブデータセットが 114,025 ペア公開されている. このサブデータセットは,人手によりラベル付けされたデータセットから作成された paraphrase 分類器により収集されている[Lan 17]. 本手法では,こうした paraphrase 分類器の作成に加え, ラベル付けのないデータセットから paraphrase 生成器も作成し.要約データセットの構築に適用させることを提案する. 図 1 にラベル付けされたデータセットにおける派生ツイートのラベルの内訳を示す.

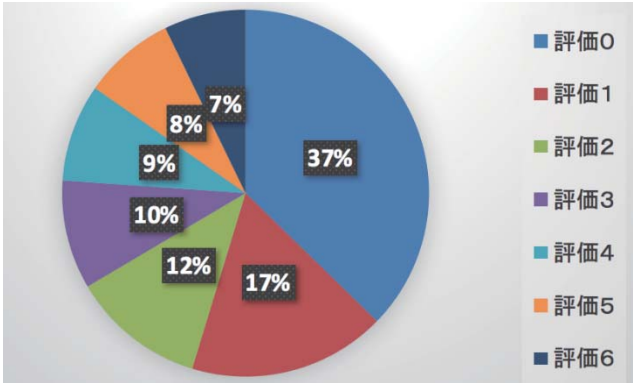

図 1 Twitter URL Paraphrase Corpus の
意味的類似度ラベルの内訳

2.2 Twitter 要約データセット

あるトピックに対する要約タスクを考える時, ニュース記事と比べて Twitter が異なる点として, 書き手が複数であり, 表現がまばらであるということが挙げられる. 特に, Twitter などにおいては,表現形式として短文が主体であり, 同じ内容の言い換え表現であることが多い. これらのことから, 本手法では, Twitter の要約の一形態を, トピックを共有する paraphrase 群からの要約文生成タス

連絡先: 永塚 光一, 創価大学理工学部情報システム工学科
〒108-0074 東京都港区高輪 1-1-11-206 080-6576-6124
e1452206@soka-u.jp

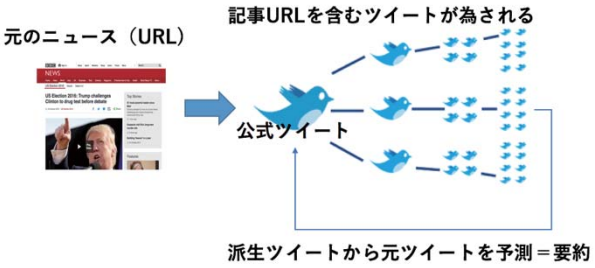

図 2 tweet 要約の流れ

クとして定義する(図 2). Twitter URL Paraphrase Corpus において, tweet に同じ URL を共有することと, 同じトピックを共有することは, 同義となっており, この前提に従うことで, 多くの paraphrase tweet を, URL のみに基づいて, paraphrase 分類器により機械的かつ持続的に収集することが可能となる[Lan 17]. 実際に, Twitter URL Paraphrase Corpus は, 分類器を用いて, 月平均で, 約 30000 の paraphrase を収集することに成功している. 提案する要約データセット収集システムもこの手法に則り, 独自に paraphrase 分類器を作成した. 加えて, 本手法では,paraphrase 生成器により, データセットを増加させることを提案する. これは, 要約モデルの学習が, 入力データとして複数文の paraphrase を必要とするため, 元 tweet に対して,通常よりも多くの paraphrase を含むことが望ましいからである.

3. 提案手法

paraphrase 分類器と paraphrase 生成器を組み合わせた要約データセット構築の流れを図 3 に示す.

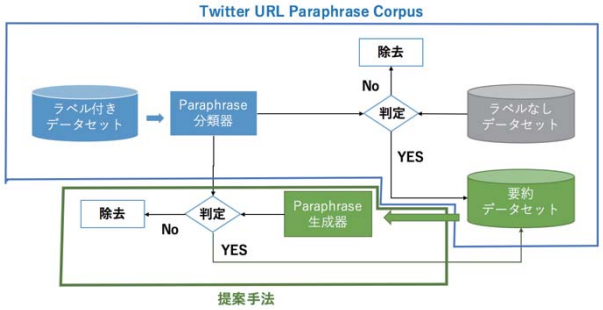

図 3. 要約データセットの構築手法

まず, ラベル付きデータセットを訓練データとして, paraphrase 分類器を作成する. その後, ラベルなしデータセットに対し, この分類器を適応させることで, paraphrase である可能性の高い tweet を自動的に選別する. 加えて, 構築したデータセットより, paraphrase 生成器を学習する. 最終的に生成された paraphrase をもう一度, paraphrase 分類器に判別させることで, 要約データセットの総量を増加させることが出来る.

3.1 Paraphrase 分類器

paraphrase 分類器学習においては, データセット構成によるバイアスを防ぐため, 全体における割合が高い 0~3 のラベルを 0.5 の確率で除去した.結果的に, 36532 ペアのトレーニングデータ, 1000 ペアのテストデータを獲得した. 図 4 に分類器ネットワークの簡略図を示す. 各 tweet を, 重みを共有した LSTM に入力したのち, 最終の隠れ層を結合し, フィードフォワードニューラルネットにかけて, sigmoid 関数により, paraphrase である確率を生成

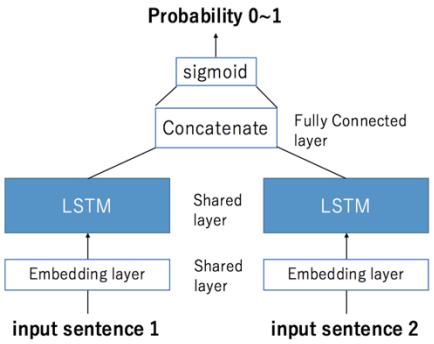

図 4. Paraphrase 分類器

する. この時, 教師データとして, 0~6 のラベル値を 6 で割り, 教師信号となる確率として正規化して与えている.

3.2 Paraphrase 生成器

生成器の学習には, attention 付きエンコーダデコーダモデル [Bahadanu 14]を使用している. ここでは, 学習速度を速めるため, LSTM よりパラメータ数の少ない GRU を用いている. また, 学習では, [Lan 17]による paraphrase 分類器によって, 事前に処理されたデータセットを用いて, トレーニングを試みている. 文の最大長は 25 であり, 最適化手法を Adam として, 1,000,000 イテレーションの学習を行った.

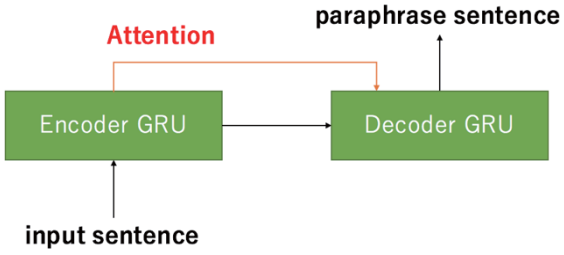

図 5 paraphrase 生成器

4. 実験

4.1 実験概要

実験では, paraphrase 分類器に対して, recall と accuracy による分類精度評価を, paraphrase 生成器に対して, 質的評価をそれぞれ行う.

4.2 分類器の実験結果

図 6, 7 にトレーニングデータとテストデータにおける分類性能を示す. 分類器のトレーニングでは, 学習高速化及び汎化性能向上のために, ミニバッチを導入し, その値を 100 としている. 文の最大長を 25 で統一し, ドロップアウト率を 0.5 とした上で, 3epoch 学習させた. 実験では, paraphrase と判断する出力確率の閾値を 0.1~0.9 まで 0.1 毎に変えて, トレーニング時とテスト時における正例と負例の recall, 及び accuracy の性能を評価する.

図 6, 7 から分かるように, トレーニングとテスト時において, 閾値を上げる程, 正例の recall は下がり, 負例の recall は増加している. また, 図 8 が示すように, トレーニング時とテスト時のどちらも, accuracy は中央値である閾値 0.5 付近において, 最も高い値であった. (トレーニング時で 85.8%, テスト時で, 71.6%). 閾値の設定は, これらの結果を踏まえた上で, 収集したい paraphrase の質に応じて, 変える必要がある.

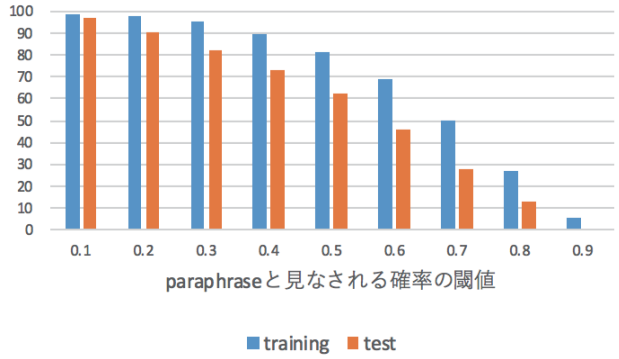

図 6 閾値の変化による正例の recall(%)

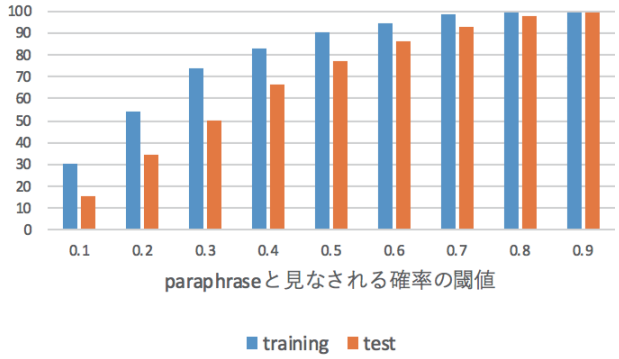

図 7 閾値の変化による負例の recall(%)

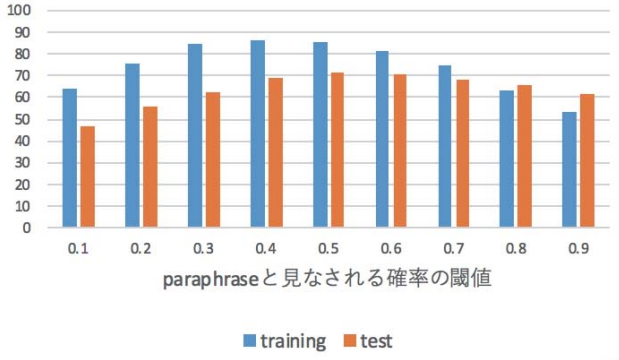

図 8 閾値の変化による accuracy(%)

4.3 生成器の実験結果

表 1 は, 生成した paraphrase の一部を示したものである. 生成 paraphrase を見ると, 中心的な意味を保ち, 語彙や付属語を変化させて文法的に正しい paraphrase を生成出来ていることがわかる. しかしながら, 中には, 単語の繰り返し表現や, 文法的かつ意味的に成り立たない paraphrase も多く含まれていた. 生成した paraphrase の末尾には, 作成した分類器によって出力した元 tweet に対する paraphrase である確率を示している. 表 1 に示したサンプルに関しては, どの文も 7 割を上回っている. これは,前述した閾値の設定において, 最大 0.7 を選択した場合においても, paraphrase としてデータセットに採用されることを意味する.

表 1 生成した paraphrase 例

example1
input
Donald Trump Agrees to Pay $25 Million in Trump University Settlement
reference
donald trump agreed to pay million to settle fraud lawsuits . about students will share in the settlement .
genetated paraphrase examples
donald trump agreed to pay million to settle fraud lawsuits . <EOS>
example2
input
Be careful what you say of others . Donald Trump Agrees to Pay $25 Million in Trump University Settlement
reference
donald trump agreed to pay million to settle fraud lawsuits . about students will share in the settlement .
genetated paraphrase examples
donald trump has agreed to pay million to settle three lawsuits against trump university . <EOS>
example3
input
Trump in March I dont settle cases . I win in court . Agrees to pay $25M to settle Trump U case
reference
donald trump agreed to pay million to settle fraud lawsuits . about students will share in the settlement .
genetated paraphrase examples
donald trump agreed to pay million to settle three lawsuits against trump university . <EOS>

5. むすび

本論文では, Twitter URL Paraphrase Corpus を活用した SNS 要約データセットの構築システムを提案した.その上で, データセット構築に用いる paraphrase 分類器と paraphrase 生成器の性能評価を行った. 分類器と生成器は共に改善の余地はあるものの, システムの実現可能性を示すことが出来た. 今後の課題として, 本手法により構築したデータセットと, 処理を加えないデータセットに対して, 要約学習を行い, 結果を比較する必要がある.

参考文献

[Over 07]Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. Information Processing & Management, 43(6):1506–1520.

[Graff 03]David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. Linguistic Data Consortium, Philadelphia.

[Bahdanau 14]Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio: Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473*, 2014.

[Lan 2017]Wuwei Lan, Siyu Qiu, Hua He, Wei Xu: A Continuously Growing Dataset of Sentential Paraphrases, *arXiv preprint arXiv:1708.00391*, 2017.