6C-07

動画コメントデータ及びブログ記事における駄洒落の抽出

谷津元樹† 荒木健治‡

†青山学院大学理工学部 ‡北海道大学大学院情報科学研究科

1 はじめに

ユーモアは生活におけるストレスを緩和する材料となり，また結果的に健康増進に寄与することが報告されている．音声対話システム等の人工知能の機能にユーモアを取り入れることにより，その恩恵をより得やすくなると考えられる．そこで著者らは，最も身近な言語的ユーモアの一つである駄洒落に関して，その生成・抽出及びコーパス化に関する研究を行っている．

これまでに，谷津ら [1] は子音の音韻類似度を考慮する駄洒落の検出手法を提案し，また，荒木ら [2] はWebから収集を行い駄洒落コーパスを作成した．しかし，前者の研究ではアノテーションに費やすことのできる労力の制約より開発データの分量が限られ．さらなる検出性能向上のためには大規模なデータからの音韻類似度行列の算出が必要となっていた．

本稿では，[1] の手法を拡張して子音・母音の双方の音素について大規模コーパス [2] より算出した音韻類似度行列を素性として用いた駄洒落の検出システムを提案し，ブログ記事及び動画コメントにおける駄洒落の検出を試みる実験の方法について説明する．

2 駄洒落コーパス

駄洒落コーパス [2] は，収集された 51,000 文の駄洒落から構成される．形式は，予め形態素解析が行われ分かち書きされた駄洒落原文に対し，その配列として種表現及び変形表現が括弧表記により指定されたものである．もし種表現・変形表現が形態素境界をまたいでいる場合，それらを全て含んでいる形態素列を指示する．また，併置型 (Perfect)・併置型 (Imperfect)・重畳型・不明の 4 類型の分類タグが付与されている．

本研究では，**5.1** 節で述べる検出器の使用する素性のうち，(1)bag-of-words 素性群，(2) 子音・母音類似度行列の構築に用いる開発データとして，及び教師あり学習の正例学習データとして利用する．

Extraction of puns in video comment data and blog posts

Motoki YATSU†and Kenji ARAKI‡

†College of Science and Engineering, Aoyama Gakuin University, yatsu@it.aoyama.ac.jp

‡Graduate School of Information Science and Technology, Hokkaido University, araki@ist.hokudai.ac.jp

3 口語文コーパス

駄洒落の検出手法を構築するにあたり，他の言語現象と同様に，できるだけ現実世界における生起頻度や生起のための条件 (駄洒落を含む会話が発生する可能性の高いトピックなど) を反映した分析を行う必要がある．本研究では，異なる媒体から構築された，以下の 2 種のコーパスを用いる．

3.1 ブログ記事コーパス

Ptaszynski ら [3] の構築した約 3.5 億文の YACIS ブログ記事コーパスを利用する．同コーパスより，人手により負例を抽出したものを検出器の負例学習データとして用いる．また，無作為抽出した文をテストデータとして用いる．ブログ記事テキスト中には駄洒落が含まれているため，人手による正例及び負例への分類を事前に行う．

3.2 動画コメントデータ

国立情報学研究所の公開するニコニコデータセット[1]における動画コメント等データより，「w」等の記号や顔文字のみのコメントを除去した文をテストデータとして用いる．同様にコメントデータ中には駄洒落が含まれているため，人手による正例及び負例への分類を事前に行う．

4 子音・母音音韻類似度行列

種・変形表現をモーラ (1 つの子音と 1 つの母音の組み合わせを単位とする情報) 列に変換し，子音・母音のそれぞれにおいて類似した音素同士の類似度を算出する．

種・変形表現は，モーラの欠落・余剰を含んでいるため，類似音素をもつモーラ同士の対応付け (アライメント) を行う必要がある．最初に子音の類似度を得るため，母音の一致するモーラを対応付けることとす

[1] https://www.nii.ac.jp/dsc/idr/nico/nico.html

[単位:文]	駄洒落	ブログ	動画コメント
開発データ	30,000	15,000	15,000
学習データ	2,000	1,000	1,000
テストデータ	200	100	100

表 1: 検出実験に用いた各種データの件数.

る．子音の類似度行列が得られれば，子音類似度を利用して子音側を対応付けるアライメントを行い，母音類似度行列を算出することが可能となる．

音韻類似度行列構築のためのアライメント手法として，Needleman-Wunsch 法 [4] を用いる．同手法はゲノム解析においても用いられる，2 記号列に対するグローバルアライメントの普遍的手法である．同法を用いた場合，種・変形表現を包含する形態素列の先頭から末尾までが対象に含まれるため，モーラの欠落や不連続的な配置に対し柔軟性が高くなる．

ここでは，子音の場合の類似度の算出方法を述べる (母音も同様の手法となる)．アライメントを行った 2 モーラ列において対応づけられた子音のペア c_1,c_2 について, 開発データ内の出現頻度 $freq(c_1, c_2)$ より, 式 (1) を用いて算出した値を子音ペア間の類似度とする．なお式 (1) は，自己相互情報量 (PMI) を元にして開発された Strength of Association 法 [5] の式と同値である．

$$\mathrm{SoA}(c_1, c_2) = \log \frac{freq(c_1, c_2) freq(\overline{c_2})}{freq(c_2) freq(c_1, \overline{c_2})} \qquad (1)$$

5 駄洒落検出器

5.1 学習に用いる素性

(1) Bag-of-words 素性群

正例及び負例の学習データより，MeCab[2]による形態素解析を行い得られた各形態素を Bag-of-words ベクトルに変換する．形態素辞書は [2] と同じく IPADic を用いた．

(2) アライメント子音・母音類似度の平均

生テキストから種・変形表現候補を検出し，同時に **4** 節にて述べた子音・母音音韻類似度行列による音韻類似度の平均値を求める．この平均値は，検出された種・変形表現候補の駄洒落らしさを表現していると考えられるため，教師あり学習による検出の素性として用いた．

具体的には, 入力された一文を形態素解析し, 各形態素を種表現候補とする．各種表現候補と文全体をモーラ列に変換し, 子音・母音のそれぞれについて Needleman-Wunsch 法によるアライメントを行う．この対応付けの際に算出されるスコアは，形態素内のモーラに関して音韻類似度行列から得られる値の総和である．各々の種・変形表現候補のもつアライメント時のスコアにおいて値域 $(0, 1]$ への正規化を行った後の最大の値を，入力文に対する本素性の素性値とする．

[2]http://taku910.github.io/mecab/

6 駄洒落検出実験

2・**3** 章にて述べたコーパスデータ及び **4** 章の子音・母音類似度行列を素性に取り入れ，**5** 章に述べた手法により構築した検出器を用いて，ブログ記事及び動画コメントデータからの駄洒落検出を試みた．開発データ及びテストデータのサイズを表 1 に示す．

7 おわりに

本稿では，異なる媒体より作成された 2 種のコーパス中の文に対して行った駄洒落検出実験の概要について述べた．発表では，検出実験の結果及び考察，ならびに駄洒落以外の多様な言語的ユーモアの包括的な認識に対する提案手法の貢献について説明する．

謝辞

本研究は科研費 (基盤研究 (C)17K00294) の助成を受けたものである．

参考文献

[1] 谷津元樹, 荒木健治: 子音の音韻類似性及び SVM を用いた駄洒落検出手法, 知能と情報（日本知能情報ファジィ学会誌）, 28 (5), pp. 875-886, 2016.

[2] 荒木健治，佐山公一，内田ゆず，谷津元樹: 駄洒落データベースの構築及び分析, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B702-3, pp.13-24, 2017.

[3] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., and Momouchi, Y.: “YACIS: A five-billion-word corpus of Japanese blogs fully annotated with syntactic and affective information”, In Proceedings of *The AISB/IACAP World Congress*, pp. 40-49, 2012.

[4] Needleman, Saul B. and Wunsch, Christian D: “A general method applicable to the search for similarities in the amino acid sequence of two proteins”, *Journal of Molecular Biology*. 48 (3), pp. 443 – 53, 1970.

[5] Mohammad, S. M., Kiritchenko, S: “Using hashtags to capture fine emotion categories from tweets”, *Computational Intelligence*, 31 (2), pp. 301-326, 2015.