ニコニココメントデータからの駄洒落検出

——【審議する】(´･ω･)(´･ω･)(･ω･`)(･ω･`)【AIをめざして】

発表者：谷津　元樹（青山学院大学，@tuu_yaa）

モチベーション

対話エイジェントやロボットに**「笑い」**という心理現象を備えさせたい

言語的ユーモア：駄洒落を含む**短文ユーモア**の認識/理解は発展途上

自然発話文に近い**短文の検出対象**→ニコニコ動画®における**動画コメント**

駄洒落の類型

並置型：

①Perfect型

種の表現　　変形表現

(霧)1 の 中 で [桐]1 を [切り倒し]2 て も [キリ]3 が ない

②Imperfect型

(きちんと) 整理 され た [キッチン] 2

音変化/挿入/削除がみられるもの

③重畳型

[すいま 千羽鶴] 3　*<文脈上あるいは既知の語>*

種の表現がテキスト中に明示されない

素材 (手法)

駄洒落データベース[1] (67,000件) を訓練データ/検証データとする

教師あり機械学習手法 (SVM, 各種 Deep Learning 手法 etc.) を
利用するための**共通フレームワークを構築，公開**[2]

動画コメントからの検出例

オレンジ字=種表現，黄緑字=変形表現，☆=今後の改良により検出予定

sm11284310.jsonl:その裁量は最良 **(並置Perfect型)**
sm16092919.jsonl:人生 井戸 井戸 **(重畳型)**☆
sm17602682.jsonl: こーしゅかの発音はこーっすか？ **(並置Imperfect型)**
sm19000864.jsonl: IKEAに行けや！ **(並置Imperfect型)**
sm19009608.jsonl: ←こんなマジニいたらマジに恐い（渾身のダジャレ） **(並置Perfect型)**
sm19027346.jsonl:「色についていろいろ教えました」審議入り **(並置Perfect型)**
sm20191681.jsonl: 遭難したんですかそうなんです・・・審議中 **(並置Perfect型)**
sm20199396.jsonl: 見事に弾道がスキマニア・クグロフwww **(重畳型)**☆
sm24090089.jsonl: 遊園地ならぬ終焉地ってか **(並置Imperfect型)**
sm25601593.jsonl: シメっぽい曲でメシが食えるのかw **(並置Imperfect型)**
sm34056964.jsonl: 内装は無いそうです(ﾎﾞｿｯ **(並置Perfect型)**

1) http://arakilab.media.eng.hokudai.ac.jp/~araki/dajare.htm
2) https://gitlab.com/m-yatsu/djr_wpsm

検出性能評価

併置型（Perf. & Imperf. 計）	43,211件 (93.8%)
重畳型	1,859件　(6.2%)

20分割交差検証法

- 使用データ (訓練/検証/テストデータ)
  - 正例：駄洒落データベース[1] (67,000 件)
    より　45,970 件を使用
  - 負例：YACIS ブログコーパス[2]
    より　45,970 文を無作為抽出
- 教師あり学習手法：Support Vector Machine (ガウスカーネル)
- 素性
  - Bag-of-Words 素性群
  - 子音音韻類似度 (従来研究[3]に基づく)

精度	再現率	F1値	正解率
90.7%	90.9%	90.8%	90.9%

課題１：検出性能の改善

課題２：重畳型駄洒落の検出（種表現＝テキスト外情報の獲得）

1) http://arakilab.media.eng.hokudai.ac.jp/~araki/dajare.htm
2) M Ptaszynski et al. "YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information" In Proc of AISB/IACAP Workshop of LaCATODA 2012, pp. 40-49, July 2012.
3) 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 28(5), 875-886, 2016.

今後の展開：フレームワークの拡張および実装ソフトの開発（応用的展開）

- 深層学習ライブラリの高速な実装を利用し，実時間での検出を目指す
- ニコニコデータセットは動画的情報もテキスト的情報をも含んでいる

→ テキスト検出処理に端を発して，**動画との同時処理へ**

応用的展開例１：コメント内リアルタイム駄洒落検出

- **コメントの流れる動画視聴時に駄洒落を検出，対象コメントを強調表示**
  - 「審議」結果のフィードバック
  - 現行の『ニコる』機能との連携
- ニコニコ生放送®などへの応用も視野に
  - リアルタイムな検出能力が必要
  - 生放送主とのインタラクションを促進
  - 『棒読みちゃん』等視聴者コメント読み上げソフトの拡張
    - 駄洒落が検出された際に笑う／沈黙する

応用的展開例２：動画内情報からの種表現獲得・駄洒落検出

- 展開例１の発展形
- 静止画／動画キャプション生成技術との融合
- **駄洒落に鋭敏に反応できるマルチモーダルな AI**

科研費(基盤研究(C)17K00294)の助成を受けています.