Génération automatique de jeux de mots à base de prénoms

Mathieu Dehouck, Marine Delaborde

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Génération automatique de jeux de mots à base de prénoms

Mathieu Dehouck[1] Marine Delaborde[1, 2]
(1) Lattice (UMR 8094), CNRS / ENS / Université Sorbonne Nouvelle
(2) LT2D (EA 7518), CY Cergy Paris Université
mathieu.dehouck@sorbonne-nouvelle.fr, marine.delaborde@cyu.fr

RÉSUMÉ

Nous présentons un système automatique de génération de blagues au format « Monsieur et Madame ». Ces blagues seront ensuite rendues accessibles sur un site internet où les visiteurs seront invités à les noter. Le tout servira ensuite à créer un corpus pour des études ultérieures.

ABSTRACT

Automatic generation of first name based wordplays

We present a system to automatically generate "Mister and Misses" type of jokes. These jokes will then be displayed on a website where visitors will have to opportunity to give feedback in order to generate a corpus.

MOTS-CLÉS : Génération automatique d'humour, jeux de mots, corpus.

KEYWORDS: Computational humor generation, puns, wordplay, corpus.

1 La détection et la génération automatique de traits d'humour

L'humour n'est pas toujours évident à déceler pour un humain et c'est donc aussi le cas pour une machine. Il existe pourtant des systèmes de classification automatique consacrés à la détection de traits d'humour, notamment avec les challenges SemEval 2017 et 2021 (Potash *et al.*, 2017; Meaney *et al.*, 2021) sur des tweets en anglais ou encore les challenges HAHA sur des tweets en espagnol (Castro *et al.*, 2018; Chiruzzo *et al.*, 2019, 2021). La génération automatique de traits d'humour est quant à elle, encore relativement peu étudiée. Amin & Burghardt (2020) ont réalisé un état-de-l'art dénombrant 12 systèmes publiés entre 1994 et 2020.

2 Des jeux de mots avec des noms propres

Nous souhaitons tester la capacité des modèles de langues à générer et/ou résoudre des blagues basées sur la prononciation [1]. En effet, les grands modèles de langues pré-entraînés sur de larges corpus sont entraînés à prédire des mots manquants (mots suivants ou mots masqués) mais ne sont à priori pas sensibles à la prononciation, qui est l'un des ressorts principaux de ce type de blagues. De plus, le passage de la prononciation de la séquence Prénom-Nom à celle de l'énoncé attendue présente parfois des changements de sons (voisement, altération des voyelles), rendant la blague d'autant plus difficile

1. Des blagues du type : « La famille *Unetelle* a trois enfants. Comment s'appellent-ils ? »

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 5 : démonstrations, pages 1–2.

à résoudre. Pour ce faire, il nous faut un corpus de référence. L'on se propose donc dans un premier temps de générer des séquences Prénom(s)-Nom qui correspondent à des énoncés valides en français.

On utilise le lexique Morphalou (ATILF, 2023) comme source de prononciation (XSAMPA) et d'information morphologique ; les prénoms quant à eux, ont été extraits du Wiktionnaire. On crée ensuite des tries (arbres de préfixes) annotés morphologiquement, associant à leur prononciation les formes orthographiques. Ainsi l'on peut trouver les mots dont la prononciation est compatible avec celle d'un prénom. Enfin, on extrait des séquences de parties-du-discours avec de l'information morphologique de treebanks Universal Dependencies (Zeman *et al.*, 2022) pour servir de colonnes vertébrales syntaxiques aux énoncés.

Les blagues ainsi générées sont rendues accessibles sur un site internet [2] où les visiteurs sont invités à les noter. Leur retour permettra d'ajouter des traits aux blagues produites en sélectionnant des critères comme « contenu explicite », « incompréhensible » ou un degré d'humour.

Références

AMIN M. & BURGHARDT M. (2020). A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, p. 29–41, Online : International Committee on Computational Linguistics.

ATILF (2023). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

CASTRO S., CHIRUZZO L. & ROSÁ A. (2018). Overview of the haha task : Humor analysis based on human annotation at ibereval 2018. In *IberEval@ SEPLN*, p. 187–194.

CHIRUZZO L., CASTRO S., ETCHEVERRY M., GARAT D., PRADA J. J. & ROSÁ A. (2019). Overview of haha at iberlef 2019 : Humor analysis based on human annotation. In *IberLEF@ SEPLN*, p. 132–144.

CHIRUZZO L., CASTRO S., GÓNGORA S., ROSÁ A., MEANEY J. & MIHALCEA R. (2021). Overview of haha at iberlef 2021 : Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, **67**, 257–268.

MEANEY J. A., WILSON S., CHIRUZZO L., LOPEZ A. & MAGDY W. (2021). SemEval 2021 task 7 : HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, p. 105–119, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.semeval-1.9.

POTASH P., ROMANOV A. & RUMSHISKY A. (2017). SemEval-2017 task 6 : #HashtagWars : Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 49–57, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/S17-2004.

ZEMAN D., NIVRE J. & AL. (2022). Universal dependencies 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

2. https://apps.lattice.cnrs.fr/aligator/