Using machine learning to classify and interpret wordplay

Aygyul Epimakhova
DATA LAB du master BIHAR de l'ESTIA,
ESTIA BERRI - Technopole Izarbel
90 Allée Fauste d'Elhuyar, BIDART, 64210, France

Abstract

In this work, we study the first task "Classify and explain instances of wordplay" set as part of the workshop project "Joker". Pilot Task 1 includes both classification and interpretation components. We use the most common methods to convert text into features. This study is based on the ML methods for elaborating an automated process of classifying and predicting missing features for test data. We use the bag-of-words model and the statistical measure of word frequency - inverse document frequency to convert text to features. Also, we apply polynomial naive Bayesian classifier and Logistic Regression to classify and predict text (with and without preprocessing). The result of the work is tables of accuracy for English and French wordplays. Examples of mostly unsuccessful and isolated relatively successful interpretations are presented. Prediction accuracy for isolated cases is less than 1%. Accuracy for the manipulation type is also not high, about 50-60%. Accuracy for other features is quite high, above 93%.

Keywords 1

Wordplay, Pun, Classification, Machine Learning, Bag-of-words, Word frequency - Inverse document frequency, Polynomial naive Bayes, Logistic regression, Accuracy, Prediction.

1. Introduction

Translation is the basis for intercultural exchange and it relies heavily on technology. However, the translation of humor and puns, which are widely represented in the culture, remains a serious problem. Humor relies on numerous cultural references, double meanings, which creates additional difficulties, including for AI-based translation systems. One of the main sources of humor is pun, which is based on the creative application or modification of the rules governing the formation of words, as well as their choice and application [1].

Preserving the wordplay can be critical to conveying meaning fully. For example, consider a pun from Lewis Carroll's Alice's Adventures in Wonderland: "'That's the reason they're called lessons', the Gryphon remarked: 'because they lessen from day to day.'" It uses the homophony of lesson and lessen for humorous effect. The French translator Henri Parisot used the words cours/cours to convey this technique: "C'est pour cette raison qu'on les appelle des cours : parce qu'ils deviennent chaque jour un peu plus courts." [1]. But in the DeepL translation, the pair leçons/diminuent is used, which makes the sentence meaningless: "'C'est pour cela qu'on les appelle des leçons', fit remarquer le Gryphon : 'parce qu'elles diminuent de jour en jour'. "

The JOKER workshop aims to bring together translators, linguists, and computer scientists to work on a creative language assessment system with the following tasks:

- Pilot task 1 is to classify individual words containing a pun according to a given typology and provide lexico-semantic interpretations.
- Pilot task 2 is to translate individual words containing a play on words.

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: aygyul.epimakhova@gmail.com (A. 3)

- Pilot task 3 is to translate entire phrases that include or contain puns.

These tasks are focused on English and French [1].

In this work, we study the first task “Classify and explain instances of wordplay” set as part of the workshop project “Joker”. This task includes both classification and interpretation components. Classification results are evaluated for accuracy, and interpretation results are evaluated in a semi-manual way [2].

The training data contains 2078 wordplays in English and 2550 in French and is presented in CSV file format with the following fields [2]:

- ID: a unique wordplay identifier
- WORDPLAY: wordplay text
- LOCATION: ambiguous words
- INTERPRETATION: explanation of the wordplay
- HORIZONTAL/VERTICAL: co-presence of source and target of the wordplay
  - In **horizontal** wordplay, both the source and the target of the wordplay
  - In **vertical** wordplay, source and target are collapsed in a single
- MANIPULATION_TYPE:
  - Identity
  - Similarity
  - Permutation
  - Abbreviation
- MANIPULATION_LEVEL: some kind of phonological manipulation:
  - Sound
  - Writing
  - Other
- CULTURAL_REFERENCE: True/False
- CONVENTIONAL_FORM: True/False

The test data contains 3255 puns in English and 4291 in French and is represented by two fields in CSV file format [2]:

- ID: a unique wordplay identifier
- WORDPLAY: wordplay text

This study is based on the ML methods for elaborating an automated process of classifying and predicting missing features for test data. Document analysis was performed fully automatically. The scripts were implemented in Python. The results were submitted in CSV file format and in Excel file format.

2. Implementation

The training data was taken from [3] and analyzed for compliance with the indicated values in [2]. Inadequate data have been excluded for the sake of purity of the experiment.

According to the two most common methods used to convert text to features [4] are:

1. Bag-of-words model which corresponds to the frequency of words (BoW) [5]
2. Statistical measure of word frequency - inverse document frequency TF-IDF, showing how important the word is in the document [6]

With the scikit-learn library, we implement them using the CountVectorizer and TfidfVectorizer classes.

In the polynomial naive Bayes algorithm [7], features follow a polynomial distribution. One of the most common uses of classifiers based on this machine learning algorithm is text classification using bag-of-word approaches or tf-idf statistical measures. That is why polynomial naive Bayes was chosen to classify the training and test data [8].

By themselves, logistic regressions are purely binary classifiers, i.e., they cannot handle target vectors with more than two classes. However, two clever extensions of logistic regression do just that [4]:

1. In one-vs-rest (OVR) logistic regression, a separate model is trained for each class to predict whether an observation is in that class or not, making it a binary classification task. Such a classifier proceeds from the fact that each classification task is independent. The OVR method is specified in the multi_class argument by default.
2. Alternatively, in polynomial logistic regression [9], the logistic function is replaced by the softmax function, a multivariable logistic function. One of the practical advantages of MLR is that its predicted probabilities using the predict_proba method are more reliable. To switch to the MNL method, we set the multi_class argument to multinomial.

3. Results

3.1. Classification

To determine the accuracy of predicted values for all wordplay properties, one-third of the training data were allocated to training and the rest served as test data with target values (Supervised learning). Accuracy is calculated from predicted values and target values and shows the proportion of accurately predicted values.

For comparison, the accuracy was calculated for three different combinations:

1. Bag of words + Polynomial naive Bayes (BoW NB)
2. TF-IDF + Polynomial Naive Bayes (TF-IDF NB)
3. TF-IDF + Logistic regression (TF-IDF LR)

Table 1
Accuracy for English wordplays

	BoW NB	TF-IDF NB	TF-IDF LR
LOCATION	0.008	0.001	0.002
INTERPRETATION	0.006	0.0	0.0
HORIZONTAL/VERTICAL	0.993	0.995	0.995
MANIPULATION_TYPE	0.490	0.480	0.527
MANIPULATION_LEVEL	0.995	0.995	0.995
CULTURAL_REFERENCE	0.946	0.946	0.946
CONVENTIONAL_FORM	0.939	0.931	0.952

Table 1 shows that the accuracy for INTERPRETATION is absolutely zero for TF-IDF NB and TF-IDF LR and LOCATION is slightly higher for BoW NB.

Table 2
Accuracy for French wordplays:

	TF NB	TF-IDF NB	TF-IDF LR
LOCATION	0.003	0.004	0.004
INTERPRETATION	0.005	0.005	0.005

HORIZONTAL/VERTICAL	0.912	0.914	0.907
MANIPULATION_TYPE	0.363	0.371	0.629
MANIPULATION_LEVEL	0.985	0.981	0.985
CULTURAL_REFERENCE	0.964	0.964	0.964
CONVENTIONAL_FORM	0.987	0.972	0.982

Interestingly, for French wordplays, the results are approximately the same for all combinations.

Table 1 and Table 2 show that high prediction accuracy is available for HORIZONTAL/VERTICAL, MANIPULATION_LEVEL, CULTURAL_REFERENCE and CONVENTIONAL_FORM

3.2. Interpretation

After determining the accuracy for all parameters, the data was trained on 2078 wordplays in English and 2550 in French and a prediction was made for the test data: 3255 puns in English and 4291 in French.

The test data contains only two fields: Id and Wordplay. The other fields need to be predicted. Examples of predicted values are shown below.

The above combinations were also used to predict test data, but only the Bag of words + Polynomial naive Bayes (BoW NB) results for English wordplays are presented here.

In the Table 3 there is an example of poor English interpretation.

Table 3
Bad English interpretation

WORDPLAY	TARGET_WORD	DISAMBIGUATION
Cliff hanger	bat	an ex axis and a why axis / an X axix and a Y axis

In the Table 4 there are some examples of the most interesting results with different values for MANIPULATION_TYPE and CULTURAL_REFERENCE.

Table 4
Interesting but isolated English interpretation

WORDPLAY	TARGET_WORD	DISAMBIGUATION	HORIZONTAL/VERTICAL	MANIPULATION_TYPE	CULTURAL_REFERENCE
Professor Grubbly-Plank	Grubbly-Plank	grubble + plank	vertical	Similarity	FALSE
'Don't you know my name ?''asked Tom swiftly.	swiftly	Tom swiftly / Tom Swifty (a kind of pun)	vertical	Identity	FALSE
How much does a hipster weigh? an Instagram. #instagramposts #instagramreels #pun #hipster #LOL #GenZ https://t.co/Gvt90HO0L	Instagram	Instagram+gram (weight measurement)	vertical	Abbreviation	FALSE

B					
There's a new TV series about a gang of Chinese zombie chefs. It's called "The Wok-ing Dead." #pun https://t.co/oIm0eT3FPC	The Wok-ing Dead	Wok + The Walking Dead	vertical	Similarity	TRUE
'I have been reading Voltaire,''Tom admitted candidly.	candidly	Candid (naïve) / Candid (character created by Voltaire)	vertical	Similarity	TRUE

MANIPULATION_LEVEL is always Sound. For this reason, this field is not shown in the Table 4.

4. Conclusion

In this work, we wanted to show how ML methods can independently cope with the task in the translation of humor and puns. Document analysis was performed fully automatically. The output test data is presented in the required CSV file format and in Excel file format with the following fields:

- RUN_ID: Run ID (as registered at the CLEF website)
- MANUAL: 0
- ID: a unique wordplay identifier from the input file
- WORDPLAY: wordplay text
- TARGET_WORD: word(s)
- DISAMBIGUATION: explanation of the wordplay
- HORIZONTAL/VERTICAL: horizontal/vertical
- MANIPULATION_TYPE: Identity/Similarity/Permutation/Abbreviation
- MANIPULATION_LEVEL: Sound/Writing/Other.
- CULTURAL_REFERENCE: True/False
- CONVENTIONAL_FORM: True/False

Several observations and brief conclusions can be made:

1. Prediction accuracy for LOCATION and INTERPRETATION is very low, less than 1% and for English INTERPRETATION accuracy with TF-IDF is zero. Wordplays are poorly predicted by the presented classification methods. Furthermore, most AI-based translation tools require a quality and quantity of training data (e.g., parallel corpora) that has historically been lacking of humour and wordplay [1]. It would be interesting to try to apply deep learning with library TensorFlow in future works.
2. The accuracy for MANIPULATION_TYPE is also not high. Interesting that the maximal accuracy for English wordplays is 53%, and for French is about 63%. We think that if there was more data to training, the result might have been better.
3. The accuracy for other features is quite high, above 93%. However:
    - MANIPULATION_LEVEL only has the predicted value of **Sound**. All participants successfully predicted all classes for MANIPULATION LEVEL. However, this success might be explained by the nature of our data as in the test set the only class was SOUND [10]
    - Signs of HORIZONTAL/VERTICAL in the bulk of **Vertical**.

- CULTURAL_REFERENCE and CONVENTIONAL_FORM - mostly **False**.

We hope that the results shown in this paper will be useful to researchers in this interesting field of humor translation.

5. `References

[1] CLEF Workshop JOKER: Automatic Wordplay and Humour Translation Liana Ermakova, Tristan Miller, Orlane Puchalski, Fabio Regattin, Élise Mathurin, Sílvia Araújo, Anne-Gwenn Bosser, Claudine Borg, Monika Bokiniec, Gaelle Le Corre, Benoît Jeanjean, Radia Hannachi, Gor˙g Mallia, Gordan Matas, and Mohamed Saki, 2022

[2] Ermakova, L., Miller, T., Puchalski, O., Regattin, F., Mathurin, É., Araújo, S., Bosser, A.-G., Borg, C., Bokiniec, M., Corre, G. L., Jeanjean, B., Hannachi, R., Mallia, Ġ., Matas, G., & Saki, M. (2022). CLEF Workshop JOKER: Automatic Wordplay and Humour Translation. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, & V. Setty (Eds.), Advances in Information Retrieval (Vol. 13186, pp. 355–363). Springer International Publishing. https://doi.org/10.1007/978-3-030-99739-7_45

[3] STID Avignon NextCloud https://guacamole.univ-avignon.fr/nextcloud/index.php/apps/files/?dir=/JOKER

[4] Chris Albon “Machine Learning with Python Cookbook”, 2019, p.303

[5] Harris Z. S., “Distributional structure”, Word, vol. 10, no 2-3, p. 146-162, 1954.

[6] Ramos J (2003) Using TF-IDF to determine word relevance in document queries. In: Proc. of the first int. conf. on machine learning

[7] Minsky, M. (1963). Steps towards artificial intelligence. Computers and Thoughts, pages 406–450.

[8] How the naive bayes classifier works in machine learning Rahul Saxena, 2017 URL: https://dataaspirant.com/naive-bayes-classifier-machine-learning/

[9] Hosmer, D.W. and Lemeshow, S. Applied logistic regression. US, Wiley-Interscience,2000

[10] Ermakova, L., Miller, T., Regattin, F., Bosser, A.-G., Mathurin, É., Corre, G. L., Araújo, S., Boccou, J., Digue, A., Damoy, A., & Jeanjean, B. (2022). Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation workshop. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, & N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)