Computational analysis and translation of wordplay

Tristan Miller

21 November 2023

Austrian Research Institute for Artificial Intelligence

Introduction

- Humour is pervasive in human dialogues and literature
- Natural language processing systems therefore must deal with it

Introduction

- Humour is pervasive in human dialogues and literature
- Natural language processing systems therefore must deal with it
- Overview of this talk:
  1. Automatically scoring/ranking jokes by humorousness
  2. Relating linguistic features of puns to humorousness
  3. Machine and machine-assisted translation of wordplay

Part I: A Bayesian approach to humorousness prediction

Introduction

- Humour is a complex phenomenon that remains challenging for computational processing
- An important first step is to recognize its presence in a text
- Previous computational approaches have treated humour detection as a binary classification task
- However, the level of appreciation (**humorousness** or **funniness**) of humour can vary

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Why model the level of appreciation?

- Conversational agents should respond with a commensurate degree of mirth

Why model the level of appreciation?

- Conversational agents should respond with a commensurate degree of mirth
- Writing assistants should identify weak jokes (and suggest better ones)

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Why model the level of appreciation?

- Conversational agents should respond with a commensurate degree of mirth
- Writing assistants should identify weak jokes (and suggest better ones)
- Machine translation systems should preserve the level of humour in the target language

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Sourcing manual humorousness annotations

- Numerical/categorical scoring

- Preference judgments

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Sourcing manual humorousness annotations

- Numerical/categorical scoring
  - ✔ Conceptually simple to process
  - ✘ Annotators can assign scores inconsistently over time
  - ✘ Annotators can interpret scores differently
  - ✘ Ordinal data cannot be treated as interval data
- Preference judgments

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Sourcing manual humorousness annotations

- Numerical/categorical scoring
  - ✔ Conceptually simple to process
  - ✘ Annotators can assign scores inconsistently over time
  - ✘ Annotators can interpret scores differently
  - ✘ Ordinal data cannot be treated as interval data
- Preference judgments
  - ✔ Allows a total sorting
  - ✔ Reduces annotation time
  - ✔ Pairwise labels can be used to infer scores
  - ✘ Number of pairs to annotate scales with $\mathcal{O}(n^2)$

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Sourcing manual humorousness annotations

- Numerical/categorical scoring
  - ✔ Conceptually simple to process
  - ✘ Annotators can assign scores inconsistently over time
  - ✘ Annotators can interpret scores differently
  - ✘ Ordinal data cannot be treated as interval data
- Preference judgments
  - ✔ Allows a total sorting
  - ✔ Reduces annotation time
  - ✔ Pairwise labels can be used to infer scores
  - ✘ Number of pairs to annotate scales with $\mathcal{O}(n^2)$
- Our approach: learn from sparse pairwise annotations

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Our contributions

1. Four novel tasks for quantifying humorousness

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Our contributions

1. Four novel tasks for quantifying humorousness
2. An annotated data set of pairwise comparisons of humorousness

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Our contributions

1. Four novel tasks for quantifying humorousness
2. An annotated data set of pairwise comparisons of humorousness
3. A Bayesian approach (GPPL) for scoring humorousness given sparse pairwise annotations

Our contributions

1. Four novel tasks for quantifying humorousness
2. An annotated data set of pairwise comparisons of humorousness
3. A Bayesian approach (GPPL) for scoring humorousness given sparse pairwise annotations
4. An empirical investigation showing that
    4.1 word embeddings and linguistic features can be used to predict humorousness

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Our contributions

1. Four novel tasks for quantifying humorousness
2. An annotated data set of pairwise comparisons of humorousness
3. A Bayesian approach (GPPL) for scoring humorousness given sparse pairwise annotations
4. An empirical investigation showing that
    4.1 word embeddings and linguistic features can be used to predict humorousness
    4.2 GPPL outperforms standard approaches (best–worst scaling, or BWS) on sparse data

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Background and related work

Learning from comparison data

- Preference learning models (e.g., Bradley–Terry, Thurstone–Mosteller) infer **rankings** from pairwise comparisons
- Best–worst scaling (BWS): annotator chooses the best and worst instances from a set
- Models such as MaxDiff infer **numerical scores** from BWS data

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Gaussian process preference learning (GPPL)

- GPPL: A Thurstone–Mosteller model that accounts for features of instances when inferring scores
  - ✔ Can make predictions on unseen instances
  - ✔ Copes well with sparse data
  - ✔ Copes well with noisy data
  - ✘ Does not scale well ($\mathcal{O}(n^3)$)

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Gaussian process preference learning (GPPL)

- GPPL: A Thurstone–Mosteller model that accounts for features of instances when inferring scores
  - ✔ Can make predictions on unseen instances
  - ✔ Copes well with sparse data
  - ✔ Copes well with noisy data
  - ✘ Does not scale well ($\mathcal{O}(n^3)$)
- We use a **scalable** method for GPPL that permits arbitrarily large numbers of instances/pairs
  - Uses stochastic variational inference to limit computational complexity
  - Outperforms SVM and BiLSTM regression models on ranking the convincingness of arguments

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Experimental data

Raw data

- SemEval-2017 pun recognition challenge data (Miller, Hempelmann, et al. 2017)
  - ✔ Purely verbal humour
  - ✔ From professional or curated sources
  - ✔ Has seen wide use outside SemEval
  - ✔ Jokes have been pre-classified by type
- 4030 one-liners: 3398 humorous and 632 not

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Annotation process

- 28 210 unique, randomly generated pairs
- Each of the 4030 texts appears in exactly 14 pairs
- Each pair was presented to five crowdsourced annotators
- Annotators were asked to indicate which text (if either) is funnier

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Gold-standard scores

- We applied BWS on the entire annotated data set to get gold-standard funniness scores
- Mean interannotator agreement: $\alpha = 0.80$
- Using fewer (1, 2, 3, 4) annotations per pair still results in good correlation ($\rho = 0.81, 0.92, 0.97, 0.99$) with the gold-standard scores

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Experiment 1:
Convergence to gold standard

Task definition

- Hypothesis:
  - Given a sufficient number of pairwise labels, the model converges close to the gold standard

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Task definition

- Hypothesis:
  - Given a sufficient number of pairwise labels, the model converges close to the gold standard
- Method:
  - Train the model on all available annotations but **without** using any feature data
  - Use the model to estimate scores for all instances
  - Rank the instances according to these scores

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Task definition

- Hypothesis:
  - Given a sufficient number of pairwise labels, the model converges close to the gold standard
- Method:
  - Train the model on all available annotations but **without** using any feature data
  - Use the model to estimate scores for all instances
  - Rank the instances according to these scores
- Evaluation:
  - Compare this ranking to BWS using Spearman's ρ

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Results

- Spearman's $\rho = 0.917$
- Unlike GPPL, BWS assigns many instances the same score, which may affect the ranking correlation
- To test this, we computed new rankings without ties, resulting in $\rho = 0.951$

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Experiment 2:
Predictive ability with full data

Task definition

- Hypothesis:
  - The model is able to generalize to unseen instances using a combination of embeddings and linguistic features

Task definition

- Hypothesis:
  - The model is able to generalize to unseen instances using a combination of embeddings and linguistic features
- Method:
  - Train the model on 60% of the instances (using both annotations and features)
  - Predict the scores on a separate 20% of the instances
  - Rank the instances according to these scores

Task definition

- Hypothesis:
    - The model is able to generalize to unseen instances using a combination of embeddings and linguistic features
- Method:
    - Train the model on 60% of the instances (using both annotations and features)
    - Predict the scores on a separate 20% of the instances
    - Rank the instances according to these scores
- Evaluation:
    - Compare this ranking to BWS using Spearman's ρ

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Features

We tested various combinations of the following features:

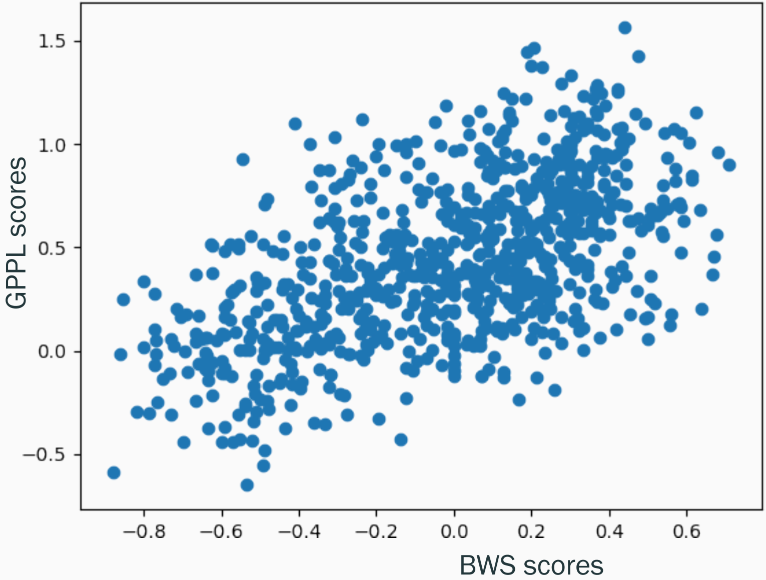
- 300-dimensional average word embeddings, using word2vec trained on Google News
- average token frequency (from a 2017 Wikipedia dump)
- polysemy (from WordNet 3.0)
- average bigram frequency (from Google Books Ngrams)

Results

features				ρ
word2vec				0.531
word2vec	frequency	polysemy		0.552
word2vec	frequency		bigrams	**0.561**
word2vec		polysemy	bigrams	0.537
word2vec	frequency	polysemy	bigrams	0.542

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

BWS vs. GPPL scores



Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Experiment 3:
Predictive ability with sparse data

Task definition

- Hypothesis:
    - With a sparser set of pairwise training labels, the model can exploit feature data to produce more accurate predictions than BWS

Task definition

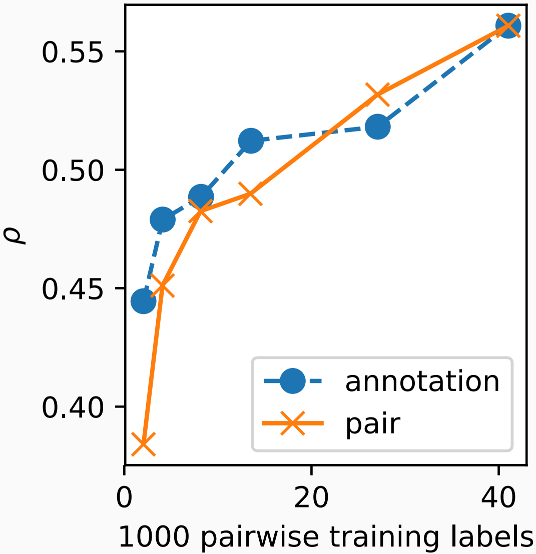
- Hypothesis:
  - With a sparser set of pairwise training labels, the model can exploit feature data to produce more accurate predictions than BWS
- Method:
  - Train the model as in Experiment 2 but on a subsample of the training data (subsampling by **annotation** or by **pair**)
  - Predict the scores on a separate 20% of the instances
  - Rank the instances according to these scores

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Task definition

- Hypothesis:
  - With a sparser set of pairwise training labels, the model can exploit feature data to produce more accurate predictions than BWS
- Method:
  - Train the model as in Experiment 2 but on a subsample of the training data (subsampling by **annotation** or by **pair**)
  - Predict the scores on a separate 20% of the instances
  - Rank the instances according to these scores
- Evaluation:
  - Compare this ranking to BWS using Spearman's ρ

Results



- smaller training sets do not reduce performance by much
- **annotation** strategy better when data is sparse, possibly due to better coverage over the feature space

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Experiment 4:
Best annotation strategy

Task definition

- Hypothesis:
  - To mitigate interannotator disagreement, obtaining the same number of annotations per pair is less effective than randomly choosing pairs to be annotated

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Task definition

- Hypothesis:
  - To mitigate interannotator disagreement, obtaining the same number of annotations per pair is less effective than randomly choosing pairs to be annotated
- Method:
  - Train the model as in Experiment 2 but on a subsample of the training data (subsampling by **annotation** or by **pair**)
  - Predict the scores on the training instances
  - Rank the instances according to these scores

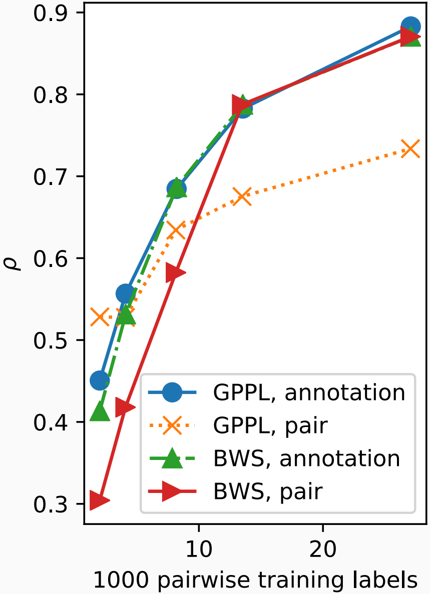
Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Task definition

- Hypothesis:
  - To mitigate interannotator disagreement, obtaining the same number of annotations per pair is less effective than randomly choosing pairs to be annotated
- Method:
  - Train the model as in Experiment 2 but on a subsample of the training data (subsampling by **annotation** or by **pair**)
  - Predict the scores on the training instances
  - Rank the instances according to these scores
- Evaluation:
  - Compare this ranking to BWS using Spearman's ρ

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Results



- GPPL exceeds BWS on very sparse data; BWS converges to GPPL as data increases
- **annotation** strategy is generally preferable, which may inform future crowdsourcing efforts

Simpson et al. (2019); Miller, Do Dinh, et al. (2020)

Part I: A Bayesian approach to humorousness prediction

Conclusion

Contributions

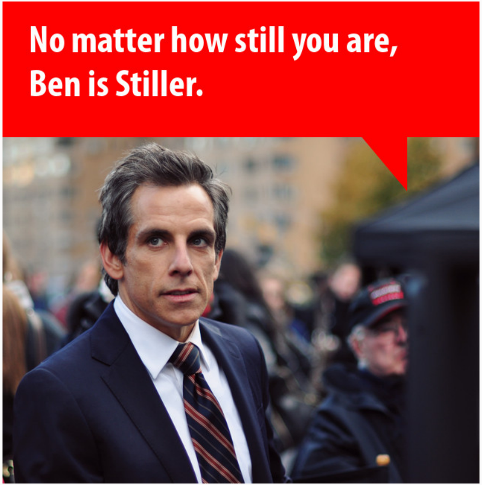
- New tasks for evaluating the degree of humorousness of short texts
- A new data set of crowdsourced preference judgments[1]
- A GPPL-based approach for estimating humorousness given word embeddings and shallow linguistic features
- Experiments showing that GPPL generalizes well, and outperforms BWS when labels are sparse

[1] `https://github.com/UKPLab/acl2019-GPPL-humour-metaphor`

Part II: What’s in a pun? Relating humorousness to linguistic features

Punning

A **punning joke** is a form of language play where a word (the **pun**) evokes the meaning of a similar-sounding word (the **target**)

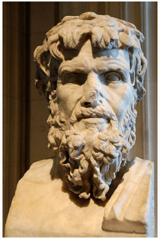


Linguistic analysis of puns

- Most linguistic analyses are taxonomic or phonological:
  - types of articulatory features involved
  - number of phonetic segments affected
  - location of changes in the lexical/syllabic structure
- Native speakers have implicit knowledge of these transformational rules; computers do not

Palmann and Miller (in preparation)

Semantics of puns



Cratylistic syllogism (Cratylus, c. 425 BCE)

Meaning motivates sound, so the meaning of similar-sounding words must be similar.

Arbitrariness of the sign (Saussure, 1916)

The relation between words and their meanings is arbitrary.

Phonology of puns

- To be recognized as wordplay, the pun and its target must be similar-sounding
- But does the degree of similarity determine how funny the pun is?
- Some studies posit such a correlation; others reject it
- Recent empirical evidence from humorous but non-punning wordplay suggests phonetic distance is indeed a factor
- Open question: Does this finding apply to puns as well?

Palmann and Miller (in preparation)

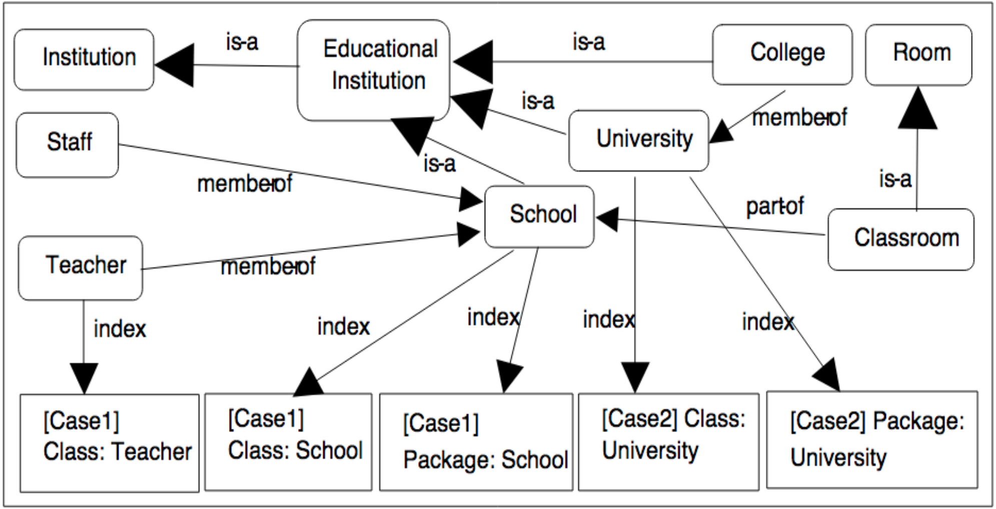
Pun typology

	homophonic	**heterophonic**
homographic	A political prisoner is one who stands behind her *convictions*.	A lumberjack's world revolves on its *axes*.
heterographic	She fell through the window but felt no *pane*.	The sign at the nudist camp read, "*Clothed* until April."

Sound similarity

- Any pair of words can be characterized by their (perceived) similarity in terms of sound or pronunciation.
- Studying pairs with a phonologically constrained relationship can help us model that relationship.
- Conversely, a model that quantifies perceived sound differences between words can assess the probability of a given relationship.
- In particular, a model of sound similarity could help detect or generate puns.

Semantic similarity



Part II: What’s in a pun? Relating humorousness to linguistic features

Research questions and methods

Research questions

1. What is the relationship between the **phonological distance** between pun and target word and the perceived funniness of a punning joke?
2. What is the relationship between the **semantic distance** between pun and target word and the perceived funniness of a punning joke?

Palmann and Miller (in preparation)

Data set

- Dataset: 2772 punning jokes (1185 **heterographic**, 1587 homographic)

Funniness ratings	Pairwise judgements	BWS & GPPL	Ranks
Phonology	IPA transcriptions	abydos.distance	Levenshtein, Covington, ALINE, Phonetic edit distance
Semantics	WordNet sense keys	nltk.wordnet word2vec	Wu-Palmer, Leacock-Chodorow, Resnik, Jiang-Conrath, Lin, Path similarity

Phonological distance measures

- Compare pun and target word based on phonetic features
- abydos.distance package (Little, 2018)

- Levenshtein distance
- Covington's distance
- ALINE distance
- Phonetic edit distance

Semantic similarity measures

- Path similarity
- Leacock-Chodorow similarity — network-based
- Wu-Palmer similarity

- Resnik similarity
- Lin similarity — information-based
- Jiang-Conrath similarity — information- & network-based

- Word2Vec similarity

Analyses and hypotheses

- Pun type: comparison of funniness ratings for homo- vs. heterographic puns

  homographic puns – higher funniness ratings

- Phonology: correlation of phonological distance & funniness ratings

  lower phonological distance – higher funniness ratings

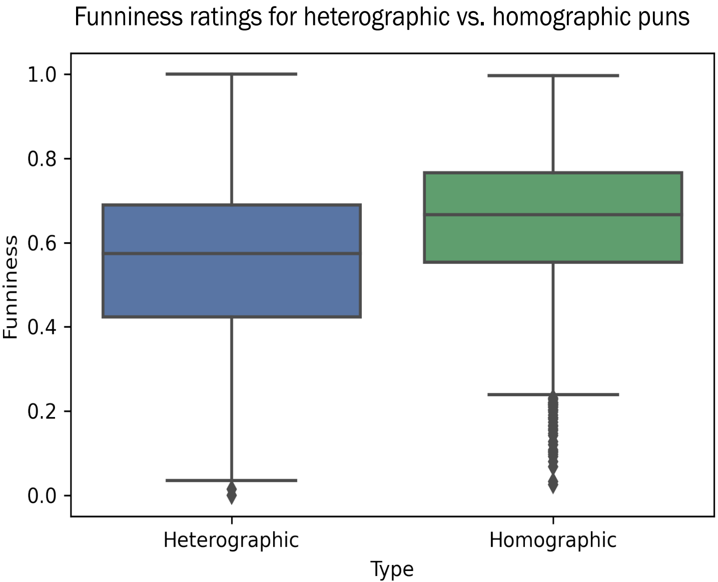
- Semantics: correlation of semantic similarity & funniness ratings

  semantic similarity in middle range – higher funniness ratings

Part II: What's in a pun? Relating humorousness to linguistic features

Results

Results: Pun type



Analyses and hypotheses

- Pun type: comparison of funniness ratings for homo- vs. heterographic puns

  homographic puns – higher funniness ratings

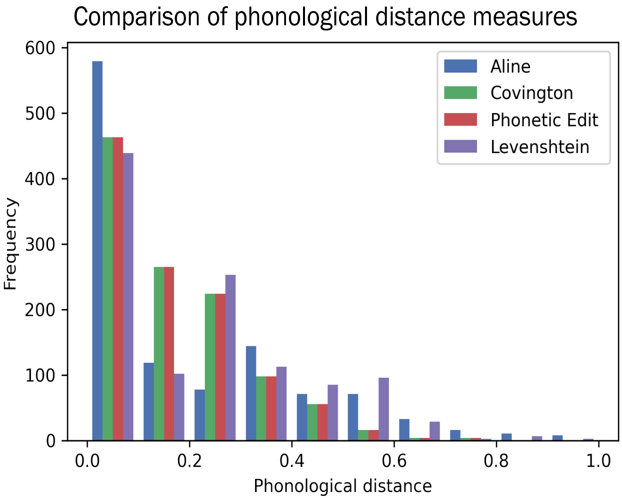
- Phonology: correlation of phonological distance & funniness ratings

  lower phonological distance – higher funniness ratings

- Semantics: correlation of semantic similarity & funniness ratings

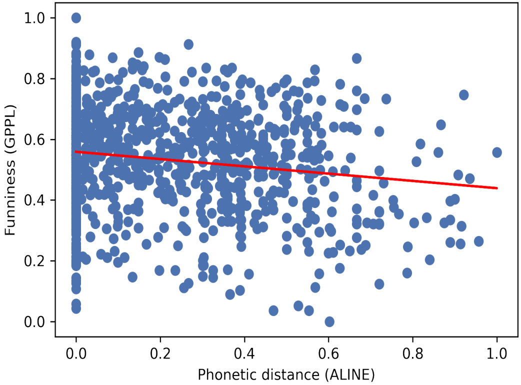
  semantic similarity in middle range – higher funniness ratings

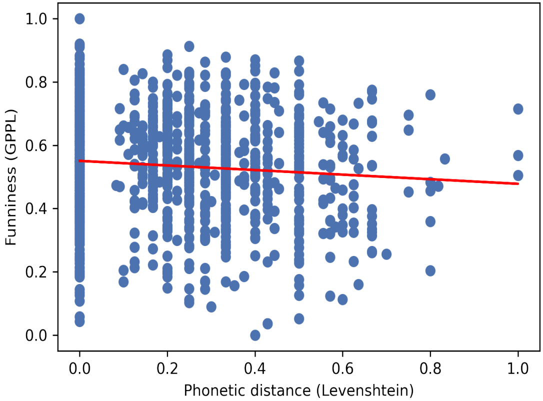
Results: Phonological distance



→ significant correlation of funniness ratings with ALINE and Levenshtein distance

Results: Phonological distance





Analyses and hypotheses

- Pun type: comparison of funniness ratings for homo- vs. heterographic puns

  homographic puns – higher funniness ratings

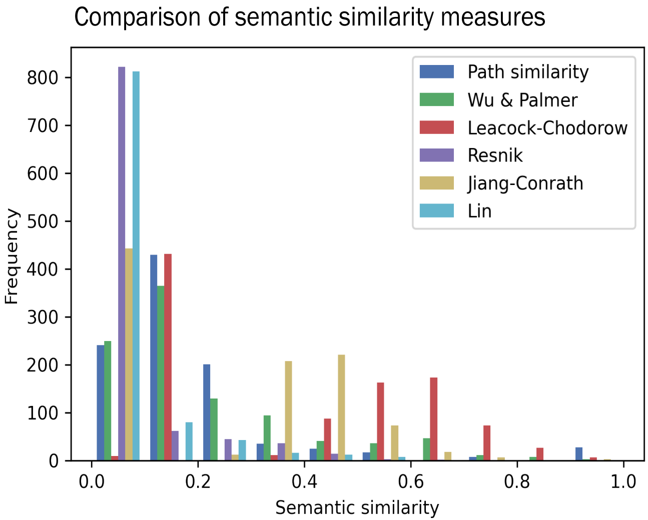
- Phonology: correlation of phonological distance & funniness ratings

  lower phonological distance – higher funniness ratings

- Semantics: correlation of semantic similarity & funniness ratings

  semantic similarity in middle range – higher funniness ratings

Results: Semantic similarity



→ no significant correlation with funniness ratings

Analyses and hypotheses

- Pun type: comparison of funniness ratings for homo- vs. heterographic puns

  homographic puns – higher funniness ratings

- Phonology: correlation of phonological distance & funniness ratings

  lower phonological distance – higher funniness ratings

- Semantics: correlation of semantic similarity & funniness ratings

  semantic similarity in middle range – higher funniness ratings

Part II: What's in a pun? Relating humorousness to linguistic features

Discussion

Discussion: Phonological distance

- Homographic puns rated as funnier than heterographic puns → homography ≈ homophony
- Lower phonological distance -- higher funniness ratings → in line with the literature (e.g. Lagerquist, 1980)
- Negative correlation only significant for Levenshtein and ALINE distance – why?

Discussion: Semantic distance

- No significance – why?
- Generally low semantic similarity values
- Maybe not semantic *similarity*, but other aspects (e.g. contextual fit)
- Semantic similarity calculations not accurate enough

Conclusion

- Lower phonological distance – higher funniness
- Facilitation of target recovery → necessary for incongruity resolution

- Many other influencing factors on funniness ratings → punning is a multi-layered phenomenon

Part III: Towards the machine(-assisted) translation of wordplay

Wordplay and translation

- Wordplay is tricky to translate, and so is widely researched in translation studies
- Translation is now a highly technologized profession
- Little/no prior work on using computers for wordplay translation
- Most language technology, including machine translation (MT), is not geared towards literary texts
- Existing digital tools ignore or eliminate linguistic anomalies and ambiguities

Miller (2019); Kolb & Miller (2022)

Punning

- Punning is a particularly common form of wordplay
- Puns employ sophisticated semantic and pragmatic mechanisms
- Puns are often held to be “untranslatable”, particularly by MT
- Can language technology nonetheless play some role in pun translation?

Miller (2019); Kolb & Miller (2022)

PunCAT

- PunCAT is our tool for computer-mediated translation of puns
- Evaluation in user study with puns from published texts
- Research questions:
    - Does PunCAT support, improve, or constrain the translation process?
    - If so, in what ways?
    - What are the tool’s benefits as perceived/described by the participants?

Part III: Towards the machine(-assisted) translation of wordplay

Background

Manual translation of puns

- **Functional equivalence:** Aim for target-language solutions that prioritize the **intention** over the literal meaning of the text
- In the case of puns, this intention is to amuse the reader in the context of the discourse
- Implications: For puns, it's OK to…
  - …substitute a different pun
  - …substitute a different form of humour
  - …omit the pun/humour altogether, as long as you compensate
- Translation strategies that preserve wordplay are preferable, but challenging to pull off

Miller (2019); Kolb & Miller (2022)

Machine(-in-the-loop) translation for literature

- Current MT can’t yet produce publication-quality output for conventional language, let alone humour and wordplay

Miller (2019); Kolb & Miller (2022)

Machine(-in-the-loop) translation for literature

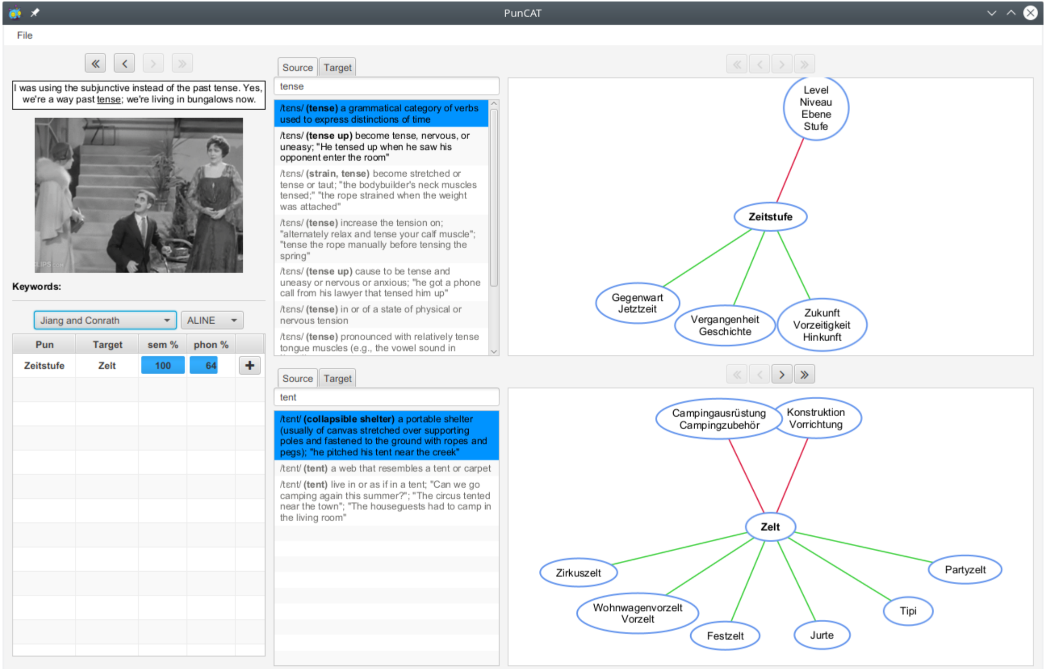
- Current MT can't yet produce publication-quality output for conventional language, let alone humour and wordplay
- AI can still play an important role in literary translation
- Rather than model the entire end-to-end translation task, put the machine in the loop:
    1. Study how human translators approach the problem
    2. Provide them with tools that support rather than replace these approaches
- Apply language technology to those subtasks it performs best
- Leave tasks that depend heavily on real-world knowledge to the human

Miller (2019); Kolb & Miller (2022)

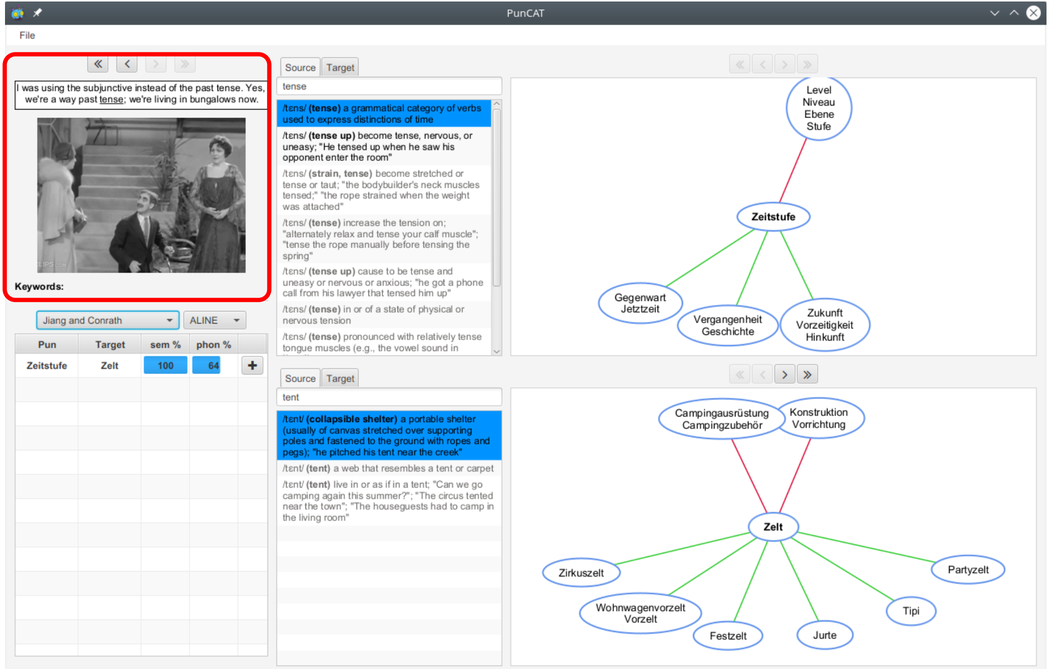
Part III: Towards the machine(-assisted) translation of wordplay

Experiment

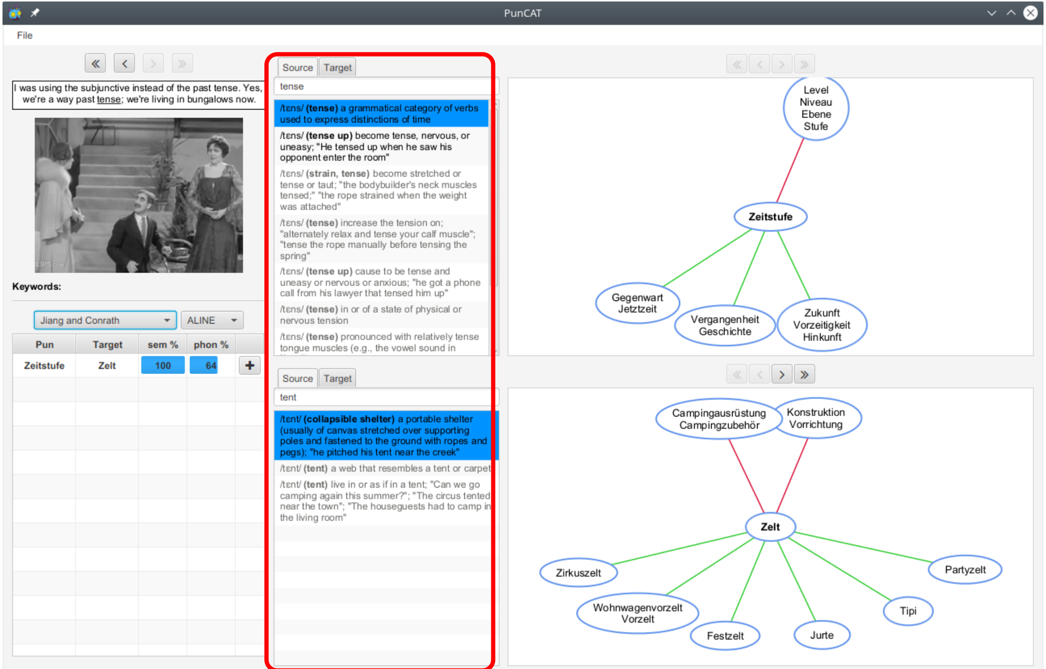
PunCAT user interface



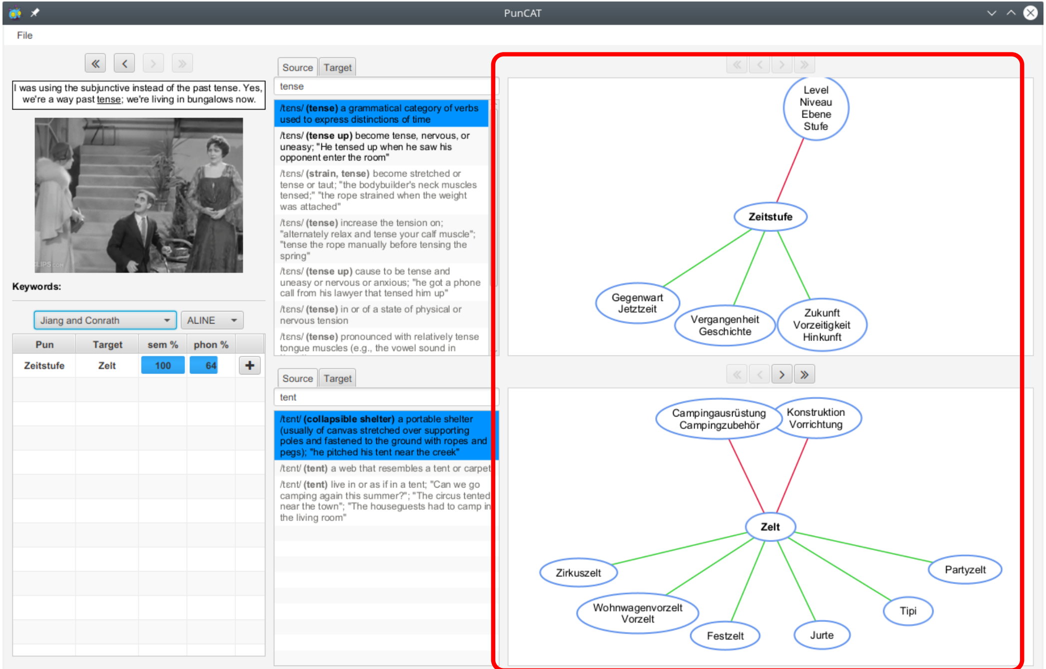
PunCAT user interface



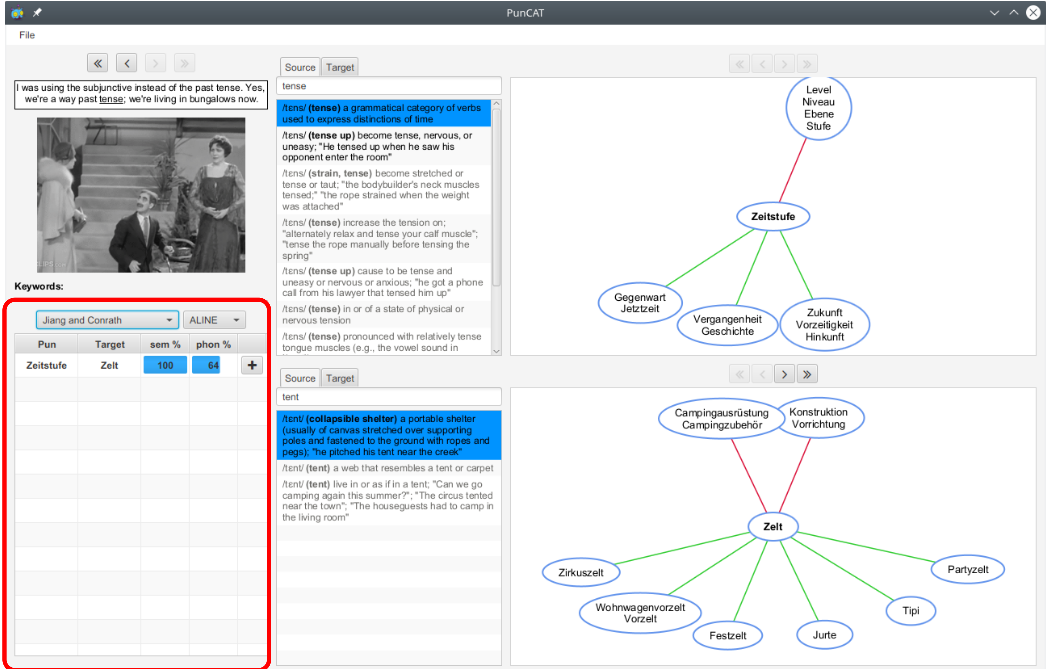
PunCAT user interface



PunCAT user interface

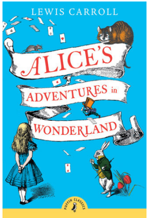


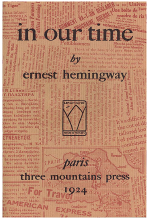
PunCAT user interface

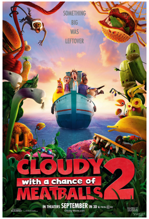


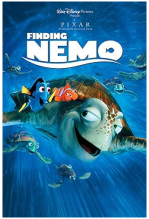
Source data

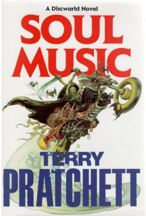
- Six puns from six published works
- All have published translations











Miller (2019); Kolb & Miller (2022)

Experimental setup

- Participants
  - 9 Master’s in Translation students at the University of Vienna
- Equipment:
  - Media lab workstation
  - Hard copy of source texts
  - Note paper

Miller (2019); Kolb & Miller (2022)

Experimental setup

- Participants
  - 9 Master's in Translation students at the University of Vienna
- Equipment:
  - Media lab workstation
  - Hard copy of source texts
  - Note paper
- Structure:
  - Two 45-minute sessions
  - Three puns to translate per session
  - PunCAT used in Session 2 only

Miller (2019); Kolb & Miller (2022)

Part III: Towards the machine(-assisted) translation of wordplay

Results and analysis

Results: Translation strategy, translation quality, and user satisfaction

- 62 target texts produced: 32 with PunCAT and 30 without
- Slightly more of the PunCAT target texts used puns (25 vs. 21)
- No significant difference in translation quality between texts produced with and without PunCAT
- Participants generally appreciated PunCAT's support with brainstorming and felt it reduced the level of stress

Miller (2019); Kolb & Miller (2022)

Part III: Towards the machine(-assisted) translation of wordplay

Conclusion

Conclusions

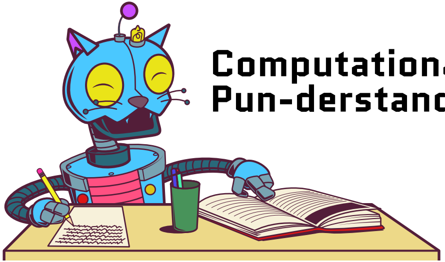
- PunCAT provides users with a specialized environment intended to structure the pun translation process without unduly constraining it
- We find good evidence that PunCAT can effectively support the translation process in terms of
  - facilitating brainstorming
  - stimulating creative thinking
  - providing inspiration
  - broadening the translator's pool of solution candidates
- But working styles vary, and PunCAT may be more suitable for some than others

Miller (2019); Kolb & Miller (2022)

Future work: Automatic translation of wordplay

- JOKER: Workshop and shared task series at CLEF 2022, 2023, 2024
- Goals:
  - bring together translators and computer scientists
  - develop evaluation frameworks for wordplay translation (data and metrics)
  - foster future work on automatic translation and automatic evaluation of translations

Ermakova, Miller, Regattin, et al. (2022); Ermakova, Miller, Bosser, et al. (2023); Ermakova, Gwenn-Bosser, et al. (2023)



Computational Pun-derstanding

https://punderstanding.ofai.at/

JOKER

https://www.joker-project.com/

References i

- Ermakova, L., A. Gwenn-Bosser, A. Jatowt, & T. Miller (2023). **"The JOKER Corpus: English–French Parallel Data for Multilingual Wordplay Recognition".** In: *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. To appear. New York, NY: Association for Computing Machinery. DOI: 10.1145/3539618.3591885.
- Ermakova, L., T. Miller, A.-G. Bosser, et al. (Apr. 2023). **"Science for Fun: The CLEF 2023 JOKER Track on Automatic Wordplay Analysis".** In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, Proceedings, Part III*. Ed. by J. Kamps, L. Goeuriot, F. Crestani, et al. Vol. 13982. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 546–556. DOI: 10.1007/978-3-031-28241-6_63.
- Ermakova, L., T. Miller, F. Regattin, et al. (2022). **"Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop".** In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)*. Ed. by A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, et al. Vol. 13390. Lecture Notes in Computer Science. Cham: Springer, pp. 447–469. DOI: 10.1007/978-3-031-13643-6_27.

- Kolb, W. & T. Miller (2022). **“Human–Computer Interaction in Pun Translation”.** In: *Using Technologies for Creative-Text Translation*. Ed. by J. L. Hadley, K. Taivalkoski-Shilov, C. S. C. Teixeira, & A. Toral. Routledge, pp. 66–88. DOI: 10.4324/9781003094159-4.
- Miller, T. (2019). **“The Punster’s Amanuensis: The Proper Place of Humans and Machines in the Translation of Wordplay”.** In: *Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology*. Shoumen: Incoma, pp. 57–64. DOI: 10.26615/issn.2683-0078.2019_007.
- Miller, T., E.-L. Do Dinh, E. Simpson, & I. Gurevych (Mar. 2020). **“Predicting the Humorousness of Tweets Using Gaussian Process Preference Learning”.** In: *Procesamiento del Lenguaje Natural* 64, pp. 37–44. DOI: 10.26342/2020-64-4.
- Miller, T., C. F. Hempelmann, & I. Gurevych (2017). **“SemEval-2017 Task 7: Detection and Interpretation of English Puns”.** In: *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 58–68. DOI: 10.18653/v1/S17-2005.

Simpson, E., E.-L. Do Dinh, T. Miller, & I. Gurevych (July 2019). **"Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning".** In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 5716–5728. DOI: 10.18653/v1/P19-1572.