Intent Classification on the ATIS dataset: comparing classical and advanced NLP methods

Seong Woo AHN
CentraleSupélec
seongwoo.ahn@student-cs.fr

Ladislas LETOURNEUR
CentraleSupélec
ladislas.letourneur@student-cs.fr

Abstract

This report presents a benchmarking study on intent classification using the ATIS dataset. The study aims to compare the performance of traditional machine learning models, such as Naive Bayes and Support Vector Machines (SVM), against advanced NLP architectures, including DistilBERT, which is a state-of-the-art transformer-based model. Our results demonstrate that advanced NLP architectures, particularly DistilBERT, significantly outperform traditional machine learning models in terms of accuracy, precision, and recall, even when the data is highly imbalanced. These findings indicate that transformer-based models are highly effective in solving the intent classification problem and have significant implications for the design and development of natural language processing systems. These results are particularly relevant for intent classification tasks in specific domains, such as airline reservations.

1 Introduction

Dialog act classification is a key component of goal-oriented dialog systems, which are designed to help users achieve specific objectives through natural language interaction. The task of dialog act classification involves identifying the purpose or intention behind each turn in the dialog, such as making a request, providing information, or expressing gratitude. This information is critical for the system to understand the user's goals and respond appropriately, making it a key aspect of natural language understanding in dialog systems [36; 26; 17; 44; 23; 39; 30; 34].

However, beyond its practical importance, dialog act classification also raises important issues related to fairness [49; 47; 32; 43], multimodality [42], and robustness [51; 3; 33; 1; 46; 15; 55; 2]. In this paper, we take a practical approach to evaluate and benchmark various machine learning and advanced NLP methods for dialog act classification, focusing on their applicability and performance on the ATIS dataset.

2 Choice of the dataset

Although there are numerous datasets available for dialog act classification [24; 4; 8; 40; 19; 11; 35; 5; 7; 22; 21; 10; 6; 9; 31], this study focuses specifically on the Airline Travel Information System (ATIS) dataset, which is a widely recognized benchmark in natural language processing (NLP) for intent classification. The ATIS dataset comprises over 4,000 spoken English language queries collected from real-world airline reservation systems, which are classified into different intent classes such as flight booking, flight query, flight schedule, and ground service. First introduced in 1990, the dataset has since become a standard reference for evaluating the performance of intent classification models.

Performing intent classification on the ATIS dataset is relevant because it simulates real-world scenarios where users interact with airline reservation systems to make bookings or inquiries. Accurately classifying user intents from these queries can help improve the overall user experience by providing more personalized and efficient responses. The ATIS dataset is also useful for evaluating the effectiveness of different intent classification algorithms and comparing their performance.

In this report, we will explore the performance of different intent classification models on the ATIS dataset. Specifically, we will focus on comparing the performance of traditional machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), and Naive Bayes with state-of-the-art deep learning models such as DistilBERT [29]. The results of our experiments will provide insights into the effectiveness of different intent classification techniques on the ATIS dataset and their potential applications in real-world scenarios.

3 Experiments Protocol

The main objective of our experiments was to assess how classical ML methods in NLP compared to state-of-the-art models for this particular dataset in intent classification. The approach that was taken was to incrementally evaluate more and more complex methods with out-of-the-box parameters on typical classification metrics (accuracy, precision, recall, f1-score). Details on the code implementation can be found in this github repository[1].

3.1 Preliminary data analysis

The first step was to analyze the ATIS dataset. The data was collected using Kaggle[2]) where `train` and `test` csv files were directly given.

Intent distribution One of the major aspects that we had to look into was the distribution of data in regards to the target variable (that is the intent). In this dataset, a total of 8 intent categories were identified:

- flight: e.g. *what flights are available from pittsburgh to baltimore on thursday morning*
- flight time: e.g. *what is the arrival time in san francisco for the 755 am flight leaving washington*
- airfaire: e.g. *cheapest airfare from tacoma to orlando*
- aircraft: e.g. *what kind of aircraft is used on a flight from cleveland to dallas*
- ground service: e.g. *what kind of ground transportation is available in denver*
- airline: e.g. *which airline serves denver pittsburgh and atlanta*
- abbreviation: e.g. *what is fare code h*
- quantity: e.g. *please tell me how many nonstop flights there are from boston to atlanta*

A plot chart was visualized to see how intents were distributed across the dataset:

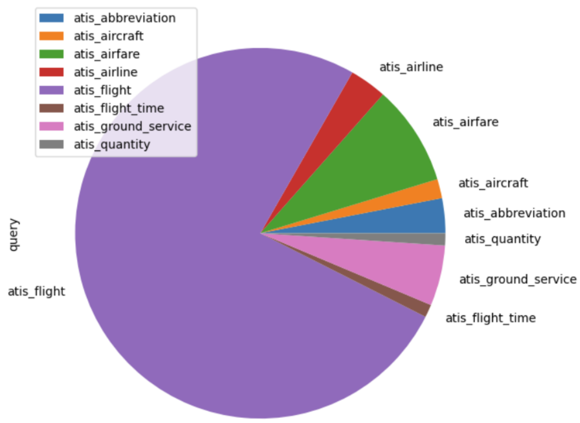

Figure 1: Distribution of intents across the ATIS dataset

Looking at the data, it can clearly be seen that it is heavily skewed, as the *flight* is the predominant label. The proportion of other intents is comparatively very small. The heavily imbalanced nature of the intent distribution was a major issue that we kept in mind throughout the project.

Query length analysis We observe in 2 that the length of the query follows a gaussian distribution. One solution to the variance in query length is to fix the length of the query. To do that we fix the length (here 100), then we truncate all query that are longer and we pad all the queries that are shorter. The result is a dataset in which all the queries have the same length.

[1] https://github.com/chrisahn99/nlp_project_intent
[2] https://www.kaggle.com/datasets/hassanamin/atis-airlinetravelinformationsystem

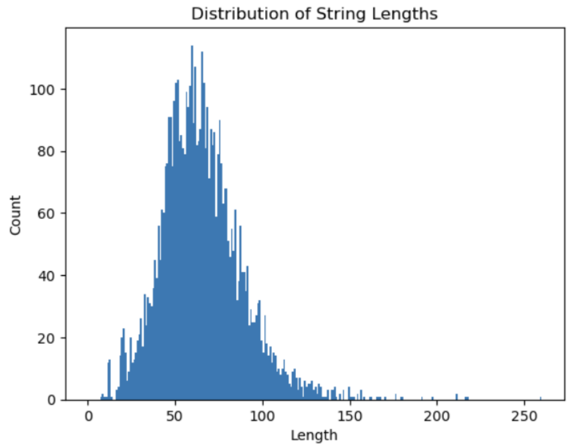

Figure 2: Distribution of the query lengths

One approach we considered was to compare the performance of the models trained on the unchanged queries, and another with these truncated queries. However, since we obtained good results on the unchanged dataset with models that we discuss about further down, we decided to leave this comparison for future works.

3.2 Classical ML algorithms

The first approach was to train test traditional machine learning techniques on the ATIS dataset. Manual and well-known pre-processing methods were also injected in the ML pipeline.

Pre-processing The pre-processing can be decomposed into the following steps:

1. punctuation removal
2. tokenization (using the NLTK `TreebankWordTokenizer`)
3. stopwords removal (using the NLTK corpus stopwords)
4. lemmatization (using the NLTK `WordNetLemmatizer`

These pre-processing steps were integrated within a pipeline function called `preprocess_atis`.

Implementation Once pre-processing was implemented, the queries were vectorized using TF-IDF vectorization (using the scikit-learn `TfidfVectorizer`, and three types of ML classifiers were trained on top of these vectorized entries:

- Logistic Regression
- SVM classifier
- Naive Bayes classifier

These classifiers were trained using the scikit-learn library on `atis_intents_train.csv` which has 762 data points, and metrics (loaded through the `metrics.classification_report` function) were evaluated on `atis_intents_test.csv` which has 4498 data points.

3.3 Advanced NLP methods

Once results were obtained for the traditional approaches, it was time to assess how Deep Learning methods performed. Numerous deep learning architectures have been tested in literature for intent classification, such as LSTMs [12], attention-based CNNs [13], and adversial multi-task learning [18].

For our implementation, we decided to take a widely-used library, HuggingFace[3], and test a very popular network, DistilBERT. DistilBERT is a transformer-based language model that has been pre-trained on a large corpus of text data. Its architecture consists of multiple transformer blocks that enable it to understand the contextual relationships between words and phrases. During training, the model learns to generate contextualized embeddings that capture the meaning of the input text. In comparison to its predecessor, BERT [20], DistilBERT is computationally lighter and faster, while maintaining a similar level of performance.

Pre-processing For the pre-processing, we decided to take automated processing pipelines from the HuggingFace `transformers` module. As such:

- `AutoTokenizer` loaded from a pretrained model (`distilbert-base-uncased` was used for the tokenization.
- `DataCollatorWithPadding` was used for sentence padding.

implementation For implementation, the HugginFace PyTorch API was used, and training was done on top of a pretrained DistilBERT model (`distilbert-base-uncased`) on 8 labels. In terms of training parameters, we used:

[3] https://huggingface.co/

- `learning_rate`: $2e^{-}5$
- `weight_decay`: 0.01
- number of training epochs: 2

Once again, train and test were done on the same datasets as the ones mentioned in the classical ML approach.

4 Results

4.1 Performance table

Here are the results that were obtained for each of the models on `atis_intents_test.csv`.

Table 1: Performance of models

Model	accuracy	precision	recall	f1
Log. Reg.	0.96	0.80	0.76	0.78
SVM	0.96	0.69	0.69	0.68
Naive Bayes	0.89	0.67	0.48	0.52
DistilBERT	0.99	0.93	1.00	0.95

4.2 Confusion matrices

The macro performance metrics given for the models does not show their performance for each of the individual intent classes. That is why confusion matrices were visualized to see which of the intents the models were having difficulties with.

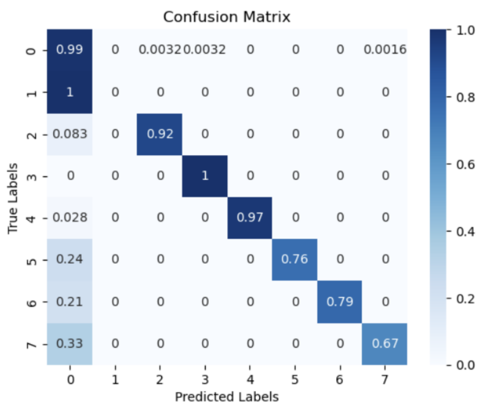

Figure 3: Confusion matrix on ATIS dataset (test) for logistic regression

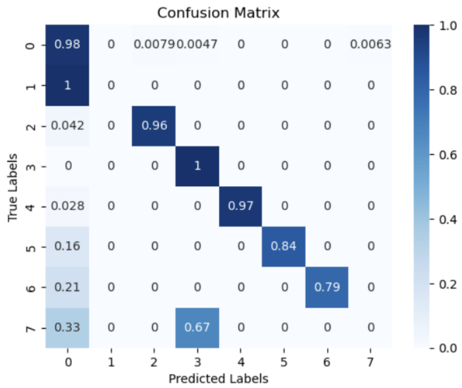

Figure 4: Confusion matrix on ATIS dataset (test) for SVM classifier

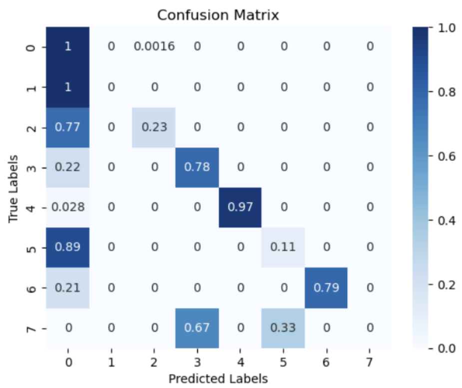

Figure 5: Confusion matrix on ATIS dataset (test) for Naive Bayes classifier

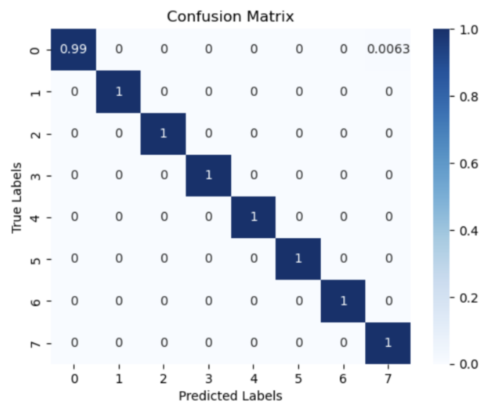

Figure 6: Confusion matrix on ATIS dataset (test) for DistilBERT

5 Discussion

5.1 Imbalanced data

What is clearly visible by comparing the classification results (2) with the confusion matrix (3,4,5,6) is that for classical ML approaches, the models have very good macro performances in terms of accuracy, but are very poor for certain classes. Often times, these classes correspond to intents that are very poorly represented (e.g. flight time - class 1, quantity - class 7), in the dataset, whereas for the predominant intents (e.g. flight - class 0) the performance is very good across all classification metrics.

This highlights the issue of having a imbalanced dataset and therefore a need to address this issue. Data augmentation through synthetic data generation methods that are adapted for NLP contexts could be integrated to enhance these problems.

5.2 Advanced NLP methods

The results shown by the DistilBERT model however really prove the high performance of state-of-the-art architectures, notably transformer architectures. Despite the imbalanced nature of the dataset, the fine-tuning of a pretrained DistilBERT model achieves near perfect classification results. It should be noted that advanced NLP models have really come far in natural language understanding (NLU) tasks, and that now libraries such as HugginFace allow for quick, easy and efficient training and implementations of such powerful models.

We can further compare the results obtained in this benchmark with existing benchmarks in literature. Comparing with the benchmark provided in [27], we see that our DistilBERT model outperforms state-of-the-art models from three years ago.

Table 2: DL model benchmark

Model	accuracy
RNN-LSTM [16]	0.93
Atten.-BiRNN [14]	0.91
Slot-Gated [25]	0.94
Joint BERT [27]	0.97
Joint BERT + CRF [27]	0.98
DistilBERT (our work)	0.99

6 Future Works

In conclusion, dialog act classification is an important task in natural language understanding that has significant implications for the development of goal-oriented dialog systems. This study has demonstrated the effectiveness of various models for classifying dialog acts, including traditional machine learning approaches and advanced NLP architectures such as transformer-based models like DistilBERT. However, there is still much work to be done in this area, particularly in terms of ensuring fairness, handling multimodal inputs, and ensuring robustness in the face of unexpected or noisy data.

Furthermore, as dialog systems become increasingly sophisticated and rely more heavily on automatic generation of natural language responses, it will be important to develop new automatic metrics to evaluate the performance of these systems in a more fine-grained and nuanced way. In particular, there is a need for automatic metrics [41; 38; 28; 52; 45; 53; 50; 54; 48; 37] that can accurately capture the quality and appropriateness of responses in different dialog contexts, as conditioned by the predicted dialog act.

Overall, the findings of this study highlight the importance of ongoing research and development in dialog act classification and its application in the design and evaluation of goal-oriented dialog systems. By continuing to explore new methods and metrics for dialog act classification, we can help to ensure that these systems are more effective, robust, and equitable for users in a wide range of domains and contexts.

References

[1] Marine Picot, Nathan Noiry, Pablo Piantanida, and Pierre Colombo. . Adversarial attack detection under realistic constraints.

[2] Marine Picot, Federica Granese, Guillaume Staerman, Marco Romanelli, Francisco Messina, Pablo Piantanida, and Pierre Colombo. . A halfspace-mass depth-based method for adversarial attack detection. *Transactions on Machine Learning Research*.

[3] Marine Picot, Guillaume Staerman, Federica Granese, Nathan Noiry, Francisco Messina, Pablo Piantanida, and Pierre Colombo. . A simple unsupervised data depth-based method to detect adversarial images.

[4] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

[5] Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: natural dialogue for speech recognition.

[6] Daniel Salber and Joëlle Coutaz. 1993. A wizard of oz platform for the study of multimodal systems. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 95–96.

[7] Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues.

[8] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

[10] Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

[11] R. Passonneau and E. Sachar. 2014. Loqui human-human dialogue corpus (transcriptions and annotations).

[12] Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth annual conference of the international speech communication association*.

[13] Zhiwei Zhao and Youzheng Wu. 2016. Attention-based convolutional neural networks for sentence classification. In *Interspeech*, volume 8, pages 705–709.

[14] Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

[15] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

[16] Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, pages 715–719.

[17] Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.

[18] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

[19] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[21] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

[22] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

[23] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

[24] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[25] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

[26] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

[27] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

[28] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

[29] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[30] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410.

[31] Alexandre Garcia, Pierre Colombo, Florence d'Alché Buc, Slim Essid, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5539–5548, Hong Kong, China. Association for Computational Linguistics.

[32] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

[33] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

[34] Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 4295–4307. Curran Associates, Inc.

[35] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

[36] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134, Online. Association for Computational Linguistics.

[37] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796.

[38] Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[39] Pierre Colombo, Chloé Clavel, Chouchang Yack, and Giovanna Varni. 2021. Beam search with bidirectional strategies for neural response generation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 139–146, Trento, Italy. Association for Computational Linguistics.

[40] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Code-switched inspired losses for spoken dialog representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8320–8337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[41] Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.

[42] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Improving multimodal fu-

sion via mutual dependency maximisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 231–245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[43] Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, Ph. D. thesis, Institut polytechnique de Paris.

[44] Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550, Online. Association for Computational Linguistics.

[45] Guillaume Staerman, Pavlo Mozharovskyi, Pierre Colombo, Stéphan Clémençon, and Florence d'Alché Buc. 2021. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv preprint arXiv:2103.12711*.

[46] Pierre Colombo, Eduardo Dadalto, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Beyond mahalanobis distance for textual ood detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 17744–17759. Curran Associates, Inc.

[47] Georg Pichler, Pierre Jean A. Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17691–17715. PMLR.

[48] Pierre Jean A. Colombo, Chloé Clavel, and Pablo Piantanida. 2022. Infolm: A new metric to evaluate summarization amp; data2text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10554–10562.

[49] Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Learning disentangled textual representations via statistical measures of similarity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2614–2630, Dublin, Ireland. Association for Computational Linguistics.

[50] Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

[51] Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2022. Rainproof: An umbrella to shield text generators from out-of-distribution data. *arXiv preprint arXiv:2212.09171*.

[52] Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. 2022. What are the best systems? new perspectives on nlp benchmarking. In *Advances in Neural Information Processing Systems*, volume 35, pages 26915–26932. Curran Associates, Inc.

[53] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

[54] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

[55] Maxime Darrin, Guillaume Staerman, Eduardo Dadalto Câmara Gomes, Jackie CK Cheung, Pablo Piantanida, and Pierre Colombo. 2023. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*.