

Vector Representations of Idioms in Conversational Systems

Tosin Adewumi*, Foteini Liwicki and Marcus Liwicki

ML Group,
EISLAB,
Luleå University of Technology, Sweden
firstname.lastname@ltu.se

Abstract

We demonstrate, in this study, that an open-domain conversational system trained on idioms or figurative language generates more fitting responses to prompts containing idioms. Idioms are part of everyday speech in many languages, across many cultures, but they pose a great challenge for many [Natural Language Processing \(NLP\)](#) systems that involve tasks such as [Information Retrieval \(IR\)](#) and [Machine Translation \(MT\)](#), besides conversational AI. We utilize the [Potential Idiomatic Expression \(PIE\)](#)-English idioms corpus for the two tasks that we investigate: classification and conversation generation. We achieve [state-of-the-art \(SoTA\)](#) result of 98% macro F1 score on the classification task by using the [SoTA T5](#) model. We experiment with three instances of the [SoTA](#) dialogue model, [Dialogue Generative Pre-trained Transformer \(DialogPT\)](#), for conversation generation. Their performances are evaluated using the automatic metric perplexity and human evaluation. The results show that the model trained on the idiom corpus generates more fitting responses to prompts containing idioms 71.9% of the time, compared to a similar model not trained on the idioms corpus. We contribute the model checkpoint/demo and code on the [HuggingFace](#) hub for public access.

Keywords: conversational systems, idioms, dialog systems, vector representation

1. Introduction

Open-domain conversational systems struggle to generate fitting responses to prompts containing idioms or figures of speech. Performance of such systems drop considerably when given context with idioms ([Jhamtani et al., 2021](#)). This challenge is not limited to open-domain conversational systems alone. [Natural Language Processing \(NLP\)](#) systems involving tasks such as [Word Sense Disambiguation \(WSD\)](#), [Information Retrieval \(IR\)](#), and [Machine Translation \(MT\)](#) also face challenges with regards to idioms ([Korkontzelos et al., 2013](#); [Adewumi et al., 2021](#)). The research question we address in this study is "does an open-domain conversational system that is idiom-aware generate more fitting responses to prompts containing idioms?" In order to investigate this question, we compare three instances of the same [SoTA](#) model: [Dialogue Generative Pre-trained Transformer \(DialogPT\)](#) by [Zhang et al. \(2020\)](#), whereby two are exposed by training to a dedicated idioms dataset and one is not. We choose the [Potential Idiomatic Expression \(PIE\)](#)-English idioms corpus by [Adewumi et al. \(2021\)](#) for this purpose. We evaluate the models using the automatic metric, perplexity, and human evaluation in two similar, but different, sets of experiments.

Two separate [NLP](#) tasks are carried out in this study. The first involves idiom identification or classification and the second involves conversation generation. Idiom identification can be essential for other [NLP](#) systems. There are usually two methods to idiom detection: type-based (depends on the expression) and token-based (depends on the context of usage) ([Peng et al., 2015a](#); [Li and Sporleder, 2009](#); [Sporleder et al., 2010a](#)). In this work, we focus on token-based, the lat-

ter.

The key contributions of this work are (1) the demonstration that an open-domain conversational system that is idiom-aware generates diverse and more fitting responses to prompts containing idioms than one that is not and (2) we obtain [SoTA](#) result in the classification task over the [PIE](#)-English idioms corpus by using the [SoTA Text-to-Text Transfer Transformer \(T5\)](#) base model, compared to the baseline result obtained by ([Adewumi et al., 2021](#)). The IdiomWOZ model checkpoint and code are hosted on the [HuggingFace](#) hub¹. Its model card is available in the appendix. The remaining parts of this paper are as follows. The Material and Methods section points out the datasets and models used. It also describes the details of the experiments carried out and the metrics of evaluation. The Results section gives results of the experiments for the two tasks, the error analysis and evaluator feedback. The Related Work section briefly discusses past efforts that are connected to this study. The Limitation section describes some of the limitation of this work and the Conclusion section summarizes this work.

2. Materials and Methods

All the experiments were performed on a shared DGX-1 machine with 8 x 32 Nvidia V100 GPUs. The operating system on the server is Ubuntu 18. It has 80 CPU cores.

2.1. Datasets

[Multi-Domain Wizard-of-Oz \(MultiWOZ\) dataset](#) MultiWOZ is a large, multi-domain, multi-topic,

¹huggingface.co/tosin/dialogpt_mwoz_idioms

and multi-task conversational dataset, originally designed for task-oriented dialogues (Budzianowski et al., 2018). It is a labelled collection of human-human written conversations and consists of more than 10,000 dialogues distributed between 70% multi-domain and 30% single domain dialogues. The data-acquisition pipeline involved crowd-sourcing without the hiring of professional annotators. Budzianowski et al. (2018) considered different dialogue scenarios that includes requests for basic information about attractions through to booking a hotel room, restaurant, train, or taxi between cities. Additional domains covered are hospital and police. It has been a standard benchmark for different dialogue problems. It was used in neural context-to-response generation experiments by Budzianowski et al. (2018) and adapted for open-domain conversational systems by Adewumi et al. (2022b; Adewumi et al. (2022a)). There are several versions of the dataset, with each new one bringing improvements (Eric et al., 2020).

PIE-English idioms corpus For both the classification and conversation generation tasks investigated in this work, the PIE-English idioms corpus is used. Hence, we discuss some of the characteristics of the corpus. It is based on example sentences from two base corpora: the British National Corpus (BNC) and UK Web Pages (UKWaC) (Ferraresi et al., 2008). About 76.94% of the samples are metaphors, making it the largest class in the dataset. Table 1 shows some examples from the dataset and the short data statement that captures other key characteristics of the dataset, as given by Adewumi et al. (2021), is given afterwards. Before training in both tasks, the corpus is split in the ratio 80:10:10 for training, dev and test sets, respectively. Three runs per experiment are conducted in order to calculate the average accuracies, F1 scores, perplexities and standard deviation. All cases are lowered and any *html* tags removed, if any, as a pre-processing step. Special characters and numbers are also removed. Data shuffling is performed before training.

Short data statement for the PIE-English idioms corpus.

This is the Potential Idiomatic Expression (PIE)-English idioms corpus for training and evaluating models in idiom identification.

The licence for using this dataset comes under CC-BY 4.0.

Total samples: 20,174

There are 1,197 total cases of idioms and 10 classes.

Total samples of euphemism (2,384), literal (1,140), metaphor (14,666), personification (448), simile (1,232), parallelism (64), paradox (112), hyperbole (48), oxymoron (48), and irony (32).

In order to use the PIE-English idioms corpus for the second task of conversation generation as intended, we

make the assumption that the corpus is suitable as a conversational dataset of dialogue turns. This assumption is valid because the sentences of the turns discuss the same cases of idioms despite being drawn from different examples from the base corpora.

2.2. Classification

Two SoTA pretrained models are compared in the classification task. These are the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and T5 (Raffel et al., 2020) models from the HuggingFace hub (Wolf et al., 2020). Both models involve their base versions. We use batch size of 64 and 16 for BERT and T5, respectively. The total training epochs for both are 6. Although the choice of hyperparameters can have significant impact on the performance of embeddings or models (Adewumi et al., 2022c; Adewumi et al., 2020), we do not carry out extensive hyperparameter exploration.

2.3. Conversation Generation

Three instances of the SoTA DialoGPT model are compared in the conversation generation task. The first instance (IdiomWOZ) is created from the model checkpoint by Adewumi et al. (2022b), which has been trained on the MultiWOZ dataset and is available on the HuggingFace hub². This is achieved by finetuning on the PIE-English idioms corpus. The second instance (IdiomOnly) is created from the original DialoGPT-medium model by Zhang et al. (2020) by finetuning on the same idioms corpus. The model checkpoint by Adewumi et al. (2022b) for the first instance is also based on the medium version of DialoGPT. The third instance (MultiWOZ) is the model checkpoint by Adewumi et al. (2022b), which was trained on the MultiWOZ dataset. For all the three instances, we set the decoding algorithm as top-k (k=100) and top-p (p=0.7). Other hyperparameters are maximum decoding length of 200 tokens, temperature of 0.8, and maximum ngram repeat limit of 3.

All three model checkpoints are then used to generate three transcripts of conversations in a first set of experiments. Ninety-four random numbers were generated and used to select the same prompts from the two test sets: the MultiWOZ and the PIE-English idioms corpus. The prompts are fed to the three models. For each dataset, thirty-two of the prompts are for generation and fifteen prompts with their test set responses (for credibility) are selected. In the second set of experiments, which has the objective of finding out fitting and diverse responses to idiom prompts, sixty-two random numbers were generated. Thirty-two are from the idioms test set and used as prompts for both the IdiomWOZ and MultiWOZ, while the remaining thirty are credibility conversations from the MultiWOZ test set. The credibility conversations are to determine the suitability of the evaluators, hence the responses to these

²huggingface.co/tosin/dialogpt_mwoz

Table 1: Samples from the PIE-English idioms corpus.

No	Samples	Class
1	Carry the day	Metaphor
2	Does the will of the Kuwaiti parliament transcend the will of the Emir and does parliament carry the day?	Metaphor
3	Time flies	Personification
4	Eighty-four!’ she giggled.’ How time flies	Personification
5	As clear as a bell	Simile
6	It sounds as clear as a bell	Simile
7	Go belly up	Euphemism
8	If several clubs do go belly up, as Adam Pearson predicts.	Euphemism
9	The back of beyond	Hyperbole
10	There’d be no one about at all in the back of beyond.	Hyperbole
11	”Why couldn’t you just stay in the back of beyond?” she said.	Hyperbole

prompts are the corresponding responses from the test sets. They are distributed at regular intervals within the transcripts.

2.3.1. Evaluation

Automatic metrics, such as BLEU or ROUGE (Lin, 2004; Papineni et al., 2002), which are common for Natural Language Generation (NLG) tasks like MT (Gehrmann et al., 2021; Vaswani et al., 2017), are sometimes viewed as inadequate for conversational systems (Jurafsky and Martin, 2020; Liu et al., 2016). This is because they do not correlate well with human assessment (Reiter, 2010). In this work, we use another common metric, perplexity, which is also used by Adwardana et al. (2020). Smaller perplexity values show that a model fits the data better as it measures how well a probability model predicts a sample, thereby corresponding to the effective size of the vocabulary (Aggarwal and Zhai, 2012).

For human evaluation, evaluators (or annotators) were recruited on Slack³. They are second/L2 (but dominant) speakers of English and are unbiased respondents who did not take part in the training of the models. For the evaluation of the transcripts, *Instruction 1* and *Instruction 2* below are the instructions for transcripts from the first and second set of experiments, respectively. Three valid evaluated transcripts from three annotators are accepted per set of experiments. The first set of transcripts are evaluated for human-likeness while the second set are based on two characteristics: more fitting and more diverse responses.

Instruction 1: Here are 94 different conversations by 2 speakers. Please, write Human-like (H) or Non-human-like (N) or Uncertain (U), based on your own understanding of what is human-like. Sometimes the speakers use idioms. If you wish, you may use a dictionary.

Instruction 2: Person 2 & Person 3 respond to Person 1. Please, write which (2 or 3) is

the a) more fitting response & b) more diverse response (showing variety in language use).

2.3.2. Credibility Unanimous Score (CUS)

In order to measure Inter-Annotator Agreement (IAA) of the conversation transcripts, we use CUS, which was introduced by Adewumi et al. (2022a). It is more intuitive, easier to calculate (based on percentages) and appears less sensitive to changes in the number of categories being evaluated, when compared to Fleiss Kappa (k). Fleiss Kappa (k) is known to be restrictive in its interpretation, depending on the number of categories (Landis and Koch, 1977), as Kappa is lower when the categories are more (Sim and Wright, 2005). According to Adewumi et al. (2022a), the assumption behind CUS is that if homogeneous samples may be used for checking the credibility of the annotators, then they may be used for establishing their agreement over the transcript. The agreement is based on unanimous votes on the homogeneous samples that are introduced. These samples may be viewed as a significant subset of the entire transcript, particularly when there’s a minimum of 30 samples, thereby fulfilling the central limit theorem. The probability of obtaining high CUS rises when the benchmark score for annotator credibility is high.

3. Results

3.1. Classification

Table 2 shows that the T5 model outperforms the BERT model. It also outperforms the best model from Adewumi et al. (2021), another BERT implementation. Results from Adewumi et al. (2021) do not provide standard deviation values and they report results only on the training and dev sets split of 85:15 ratio. From the results, it appears that the PIE-English idioms corpus is not overly challenging, at least for the T5 model, because of the high scores obtained. This may be due to the fact that the length of each sample is one sentence, or at most 2 sentences, in the corpus. The results

³slack.com

Table 2: Average accuracy & F1 results (sd - standard deviation) *data split ratio 85:15 for training:dev sets.

Model	Accuracy		weighted F1		macro F1	
	dev (sd)	test (sd)	dev (sd)	test (sd)	dev (sd)	test (sd)
BERT	0.96 (0)	0.96 (0)	0.96 (0)	0.96 (0)	0.75 (0.04)	0.73 (0.01)
T5	0.99 (0)	0.98 (0)	0.98 (0)	0.98 (0)	0.97 (0)	0.98 (0)
BERT* (Adewumi et al., 2021)	0.93	-	0.95	-	-	-

True Label	Metaphor	1,466	0	0	0	0	0	0	0	0	
	Euphemism	6	230	1	0	1	0	0	0	0	
	Simile	0	0	123	0	0	0	0	0	0	
	Personification	1	0	0	44	0	0	0	0	0	
	Literal	10	0	1	0	97	0	0	0	0	
	Oxymoron	0	0	0	0	0	5	0	0	0	
	Parallelism	0	0	0	0	0	0	7	0	0	
	Paradox	0	0	0	0	0	0	0	11	0	
	Hyperbole	0	0	0	0	0	0	0	0	5	
	Irony	0	0	0	0	0	0	0	0	0	3
		Metaphor	Euphemism	Simile	Personification	Literal	Oxymoron	Parallelism	Paradox	Hyperbole	Irony

Predicted Label

Figure 1: Confusion matrix for T5 model on the PIE-English idioms corpus test set.

are statistically significant as the p-value ($p < 0.0001$) of the two-sample t-test for the difference of two means (of the macro F1 scores) is smaller than alpha (0.05).

Error Analysis Figure 1 shows the confusion matrix of the results of the T5 model for the predictions against the true labels for the test set of the idioms corpus. The model performs substantially well even for classes that have few samples in the training set, such as *hyperbole* and *irony*. It, however, struggles mostly in correctly classifying the *literals*. It misclassified about 9.3% of them as *metaphor*, presumably because it is the largest class in the dataset.

3.2. Conversation generation

We can observe from Table 3 that the MultiWOZ model from Adewumi et al. (2022b) has the lowest average perplexity, when compared with the other two new models. This is likely because the MultiWOZ data the model was trained on is larger than the idioms corpus.

The p-value ($p < 0.0001$) of the two-sample t-test for the difference of two means (for the IdiomWOZ and IdiomOnly) is smaller than alpha (0.05), hence the results are also statistically significant. Despite the average perplexity for the IdiomOnly model being lower than the IdiomWOZ, we chose to generate responses and conduct human evaluation on the latter. This is because one of its runs had a lower perplexity, which may be deduced from the standard deviation. Besides, perplexity alone may not be sufficient to tell how good a model is (Roller et al., 2021; Hashimoto et al., 2019).

Human evaluation results are presented in Tables 4 and 5. The former is based on transcripts of 64 single-turn conversations for the first set of experiments and the latter is based on 32 single-turn conversations for the second set of experiments, after the removal of the 30 credibility conversations from each. We observe, from Table 4, that both the MultiWOZ and IdiomWOZ models have more human-like single-turn con-

Table 3: Average perplexity results. (sd - standard deviation)

Model	Perplexity	
	dev (sd)	test (sd)
IdiomWOZ	201.10 (34.82)	200.68 (34.83)
IdiomOnly	189.92 (1.83)	185.62 (2.05)
MultiWOZ (Adewumi et al., 2022b)	6.41 (-)	6.21 (-)

Table 4: Human evaluation results of 3 annotators on 3 classes for 64 single-turn conversations.

Model	Scale (majority votes)				CUS
	H (%)	U (%)	N (%)	3-way (%)	%
IdiomWOZ	39.1	10.9	37.5	12.5	80
IdiomOnly	15.6	12.5	60.9	10.9	80
MultiWOZ	62.5	1.6	32.8	3.1	80
unanimous votes - 3/3					
IdiomWOZ	20.3	0	12.5	-	80
IdiomOnly	6.3	0	31.3	-	80
MultiWOZ	45.3	0	23.4	-	80
idioms only maj. votes (32 samples)					
IdiomWOZ	30	23.3	33.3	13.3	80
IdiomOnly	26.7	0.2	36.7	16.7	80
MultiWOZ	26.7	3.3	66.7	3.3	80

Table 5: Human evaluation results of 3 annotators on 3 classes for 32 single-turn conversations.

Model	Scale (majority voting)		CUS
	More fitting (%)	More diverse (%)	%
IdiomWOZ	71.9	28.1	80
MultiWOZ	28.1	71.9	80
unanimous votes - 3/3			
IdiomWOZ	46.9	9.4	80
MultiWOZ	9.4	15.6	80

versations than other categories, under majority votes. The MultiWOZ model has the most human-like conversations. However, considering idioms only prompts, the IdiomWOZ model has the most human-like conversations. From Table 5, IdiomWOZ has more fitting conversations than the MultiWOZ, though the opposite is the case with regards to more diverse conversations. This could be due to the evaluators’ interpretation of what is diverse. For all the evaluations, we observe that there is CUS of 80%. The CUS is the same across sections in each table since the same transcript is involved for each section.

4. Discussion & Evaluator Feedback

Tables 6 and 7 show some single-turn conversations from the second transcript. Person 1 is the prompt from the PIE-English idioms test set. Person 2 is the IdiomWOZ model and it responds more with idioms. Person 3 is the MultiWOZ model.

The IdiomWOZ model, whose responses are deemed as more fitting, appears to confirm previous study, which shows that metaphors have more emotional impact than their literal equivalent (Mohammad et al., 2016). Idioms usually make utterances more colorful

(or rich) and diverse. Hence, simply reducing idioms to their literal form before feeding NLP models as practised by Jhamtani et al. (2021) may not adequately address the challenge since it implies the models still are incapable of understanding the idioms and because some idioms have more than one literal expression.

Feedback from some of the evaluators suggest the use of idioms complicated the evaluation task a bit but it was sometimes useful in identifying which of the two conversations was a more fitting or diverse option. They found it relatively easier evaluating the human-likeness characteristic in the first set of experiments but had to resort to using a dictionary for the idioms in certain cases. Some of the conversations were marked non-human-like because there was lack of connection between the prompt and the response. In the second set of transcripts, some conversations are evaluated as more fitting when the responses answer the prompts directly while some are evaluated as more diverse when the response is not a direct answer but elicits further discussion.

Table 6: Excerpt from the second transcript of single-turn conversations. Person 2 is the IdiomWOZ

Person 1	Person 2
He knew his father’s grievance ; and put off the evil day as long as he possibly could	he’s dead meat
achieving results that fly in the face of realism	I’m not here I’m just a fly on the wall
does that mean Jane is off the hook ?	She had the nerve to call time on her three-year spell in charge
he’s still sulking ; sounding like a broken record	the whole caboodle
we need new ways to do this ; to chew the fat ; to check ourselves out	the Government’s approach to the environment is geared more towards addressing the environment rather than helping the environment
the 24-year-old striker is self-critical enough to insist he is not yet firing on all cylinders	You’re the one up to your neck in it
remind her to take it off the hook	the bookie took it on the nod
A lengthy break is in order ; so make sure you keep your nose clean	they are under a cloud because of fears about US President’s wife Hillary Clinton
Utd have shown they are in a league of their own	It’s a long shot him signing for Leeds. They’re a hard nut to crack
from now until the year dot	he was the apple of her eye
it’s a good idea to go the distance	it ’ll just collapse and give up the ghost somewhere in South Africa

Table 7: Excerpt from the second transcript of single-turn conversations. Person 3 is the MultiWOZ

Person 1	Person 3
He knew his father’s grievance ; and put off the evil day as long as he possibly could	That’s great; thanks.
achieving results that fly in the face of realism	Yes; realism. I like it.
does that mean Jane is off the hook ?	I’m looking for a place to go in the centre.
he’s still sulking ; sounding like a broken record	Well then I suppose I need to find something else to do for you.
we need new ways to do this ; to chew the fat ; to check ourselves out	To each their own. Let me know if you need any more info.
the 24-year-old striker is self-critical enough to insist he is not yet firing on all cylinders	Well if you’ll excuse me... let me ask my wife. Does she think we should try for a little earlier; say 20:30?
remind her to take it off the hook	I’m sorry; I forgot to hang up. Good bye.
A lengthy break is in order ; so make sure you keep your nose clean	I am looking for a restaurant in the centre of town that serves unusual food.
Utd have shown they are in a league of their own	I’m sorry; but I’m not sure what you’re looking for. Could you please restate your requirements?
from now until the year dot	I’m sorry; but I don’t understand your request. Would you like to try a different restaurant?
it’s a good idea to go the distance	I can help with that! What are your departure and destination sites?

5. Related Work

Jhamtani et al. (2021) observed that performance dipped when some deep models were evaluated on two open-domain dialogue datasets: DailyDialog and PersonaChat, with regards to figurative language (Li et al., 2017; Zhang et al., 2018). They compared Generative Pre-trained Transformer (GPT)-2 to four other models over the datasets and noticed the drop in performance among most models. In their work, however, they proposed transforming figurative language (including idioms) to their literal form before feeding the models.

Idiom detection usually takes two approaches: type-based and token-based (Peng et al., 2015a; Cook et al., 2007; Li and Sporleder, 2009; Sporleder et al., 2010a). The type-based approach attempts to determine if an expression is an idiom, perhaps through automatic compilation of an idiom list from a corpus (Sporleder et al., 2010b). The token-based approach relies on context for distinguishing idioms (Korkontzelos et al., 2013; Sporleder et al., 2010a). Non-contextual word embeddings (such as word2vec) are used for identifying metaphors (Mao et al., 2018). However,

such an approach may underperform (Mao et al., 2018). Peng et al. (2015b) use word2vec to obtain vectors from text8 corpus. Their algorithm, based on the assumption that literal vectors are distinguished from idiom vectors by the larger inner product they produce, uses inner product of context word vectors with vector representing target expression. Bizzoni et al. (2017) use word2vec and an ANN with 1 hidden layer for detecting metaphors. Diab and Bhutada (2009) perform binary classification using support vector machine (SVM), producing literal and idiomatic expressions on a subset of the VNC-Token.

6. Limitation

The PIE-English idioms corpus that is used in this work, though relatively large and probably the first with as much as 10 classes, does not represent all the possible classes of idioms available. This may affect the performance of the ML models. Furthermore, there may be risks of offensive language or biased outputs from the model checkpoints, since data used for pretraining the deep models are from public online sources that are known to contain such risks (Zhang et al., 2020).

7. Conclusions

We answer the research question "does an open-domain conversational system that is idiom-aware generate more fitting responses to prompts containing idioms?" The answer is yes. It is, therefore, important to train open-domain conversational systems on idioms data, so as to achieve diversity and more fitting responses in ML models. This is especially since idioms are part of everyday speech in many cultures (Alm-Arvius, 2003). We achieve SoTA classification results on the PIE-English idioms corpus, using the T5 model. Future efforts may be directed at exploring more datasets of figurative language or idioms and more diverse SoTA models for training.

Appendix

Model card for DialoGPT(medium)-MultiWOZ, trained on the [PIE](#)-English idioms corpus

Details	
Model rationale	This is a fine-tuned model of DialoGPT (medium)-MultiWOZ v2.2 on the Potential Idiomatic Expression (PIE) -English idioms corpus. It is intended to be used as an idiom-aware conversational system.
Limitation	The dataset it's trained on is limited in scope, as it covers only 10 classes including metaphor, simile, euphemism, parallelism, personification, oxymoron, paradox, hyperbole, irony and literal.
Result	The perplexity achieved on the test and validation sets are 160.9 and 160.46, respectively.
Others	
Model link	https://huggingface.co/tosin/dialogpt_mwoz_idioms
More	Information about the original pre-trained model can be found at https://huggingface.co/tosin/dialogpt_mwoz
Licence	CC-BY 4.0.
Generation example from an interactive environment:	
Role	Response
User	remind her to take it off the hook.
Bot	the bookie took it on the nod
User	A lengthy break is in order ; so make sure you keep your nose clean
Bot	they are under a cloud because of fears about US President's wife Hillary Clinton
User	Some families breed like rabbits and send us all their male heirs
Bot	he was out for the count

Table 8:

A. Bibliographical References

- Adewumi, T. P., Liwicki, F., and Liwicki, M. (2020). Exploring swedish & english fasttext embeddings for ner with the transformer. *arXiv preprint arXiv:2007.16007*.
- Adewumi, T. P., Vadoodi, R., Tripathy, A., Nikolaidou, K., Liwicki, F., and Liwicki, M. (2021). Potential idiomatic expression (pie)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*.
- Adewumi, T., Adeyemi, M., Anuoluwapo, A., Peters, B., Buzaaba, H., Samuel, O., Rufai, A. M., Ajibade, B., Gwadabe, T., Traore, M. M. K., Ajayi, T., Muhammad, S., Baruwa, A., Owoicho, P., Ogunremi, T., Ngigi, P., Ahia, O., Nasir, R., Liwicki, F., and Liwicki, M. (2022a). Ìtākùròsò: Exploiting cross-lingual transferability for natural language generation of dialogues in low-resource, african languages.
- Adewumi, T., Brännvall, R., Abid, N., Pahlavan, M., Sabry, S. S., Liwicki, F., and Liwicki, M. (2022b). Småprat: Dialogpt for natural language generation of swedish dialogue by transfer learning. In *5th Northern Lights Deep Learning Workshop, Tromsø, Norway*, volume 3. Septentrio Academic Publishing.
- Adewumi, T., Liwicki, F., and Liwicki, M. (2022c). Word2vec: Optimal hyperparameters and their impact on natural language processing downstream tasks. *Open Computer Science*, 12(1):134–141.
- Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.
- Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.
- Alm-Arvius, C. (2003). *Figures of speech*. Studentlitteratur.
- Bizzoni, Y., Chatzikiyakidis, S., and Ghanimifard, M. (2017). “deep” learning : Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (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*, pages 5016–5026, Brussels, Belgium, October-November. Association for Computational Linguistics.
- Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Diab, M. and Bhutada, P. (2009). Verb noun construction mwe token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, pages 17–22.
- Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May. European Language Resources Association.
- Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.
- Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kale, M., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo Rubungo, A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Sobrevilla Cabezedo, M. A., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., and Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, August. Association for Computational Linguistics.
- Hashimoto, T. B., Zhang, H., and Liang, P. (2019). Unifying human and statistical evaluation for natural language 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 1689–1701, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Jhamtani, H., Gangal, V., Hovy, E., and Berg-Kirkpatrick, T. (2021). Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic, November. Association for Computa-

- tional Linguistics.
- Jurafsky, D. and Martin, J. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Limited.
- Korkontzelos, I., Zesch, T., Zanzotto, F. M., and Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Li, L. and Sporleder, C. (2009). Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323.
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.
- Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.
- Mao, R., Lin, C., and Guerin, F. (2018). Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia, July. Association for Computational Linguistics.
- Mohammad, S., Shutova, E., and Turney, P. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany, August. Association for Computational Linguistics.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, page 311–318, USA. Association for Computational Linguistics.
- Peng, J., Feldman, A., and Jazmati, H. (2015a). Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 507–511.
- Peng, J., Feldman, A., and Jazmati, H. (2015b). Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 507–511, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Reiter, E. (2010). 20 natural language generation. *The handbook of computational linguistics and natural language processing*, page 574.
- Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April. Association for Computational Linguistics.
- Sim, J. and Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.
- Sporleder, C., Li, L., Gorinski, P., and Koch, X. (2010a). Idioms in context: The idix corpus. In *LREC*. Citeseer.
- Sporleder, C., Li, L., Gorinski, P., and Koch, X. (2010b). Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.
- Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 2204–2213, Melbourne, Australia, July. Association for Computational Linguistics.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.