Sci

MDPI

Article

Vector Representations of Idioms in Conversational Systems

Tosin Adewumi *, Foteini Liwicki and Marcus Liwicki

ML Group, EISLAB, Luleå University of Technology, 971 87 Luleå, Sweden
* Correspondence: tosin.adewumi@ltu.se

Abstract: In this study, we demonstrate that an open-domain conversational system trained on idioms or figurative language generates more fitting responses to prompts containing idioms. Idioms are a part of everyday speech in many languages and across many cultures, but they pose a great challenge for many natural language processing (NLP) systems that involve tasks such as information retrieval (IR), machine translation (MT), and conversational artificial intelligence (AI). We utilized the Potential Idiomatic Expression (PIE)-English idiom corpus for the two tasks that we investigated: classification and conversation generation. We achieved a state-of-the-art (SoTA) result of a 98% macro F1 score on the classification task by using the SoTA T5 model. We experimented with three instances of the SoTA dialogue model—the Dialogue Generative Pre-trained Transformer (DialoGPT)—for conversation generation. Their performances were evaluated by using the automatic metric, perplexity, and a human evaluation. The results showed that the model trained on the idiom corpus generated more fitting responses to prompts containing idioms 71.9% of the time in comparison with a similar model that was not trained on the idiom corpus. We have contributed the model checkpoint/demo/code to the HuggingFace hub for public access.

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

check for updates

Citation: Adewumi, T.; Liwicki, F.; Liwicki, M. Vector Representations of Idioms in Conversational Systems. *Sci* **2022**, *4*, 37. https://doi.org/10.3390/sci4040037

Academic Editor: Anastasios Doulamis

Received: 26 June 2022
Accepted: 20 September 2022
Published: 29 September 2022

1. Introduction

Open-domain conversational systems struggle to generate fitting responses to prompts containing idioms or figures of speech. The performance of such systems drops considerably when given a context with idioms [1]. 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 [2,3]. The research question that we address in this study is: "Does an open-domain conversational system that is idiom-aware generate more fitting responses to prompts containing idioms?". We consider the English language. In order to investigate this question, we compared three instances of the same state-of-the-art (SoTA) model, the Dialogue Generative Pre-trained Transformer (DialoGPT) by [4], two instances of which were exposed in their training to a dedicated idiom dataset and one of which was not. We chose the Potential Idiomatic Expression (PIE)-English idiom corpus by [3] for this purpose because of the wide range of idiom classes that it has. We evaluated the models using an automatic metric, perplexity, and a human evaluation in two similar, but different, sets of experiments.

Two separate NLP tasks were carried out in this study. The first involved idiom classification, and the second involved conversation generation. Idiom detection can be essential for other NLP systems. There are usually two methods for idiom detection: type-based (which depend on the expression) and token-based methods (which depend on the context of usage) [5–7]. In this work, we focus on token-based methods.

The key contributions of this work are (1) the demonstration that an open-domain conversational system that is idiom-aware generates more fitting responses to prompts containing idioms than one that is not and (2) the fact that we obtained an SoTA result in the classification task using the PIE-English idiom corpus by using the SoTA Text-to-Text Transfer Transformer (T5) as a base model in comparison with the baseline result obtained by [3]. The IdiomWOZ model checkpoint is hosted on the HuggingFace hub (huggingface.co/tosin/dialogpt_mwoz_idioms accessed date: 1 September 2022). Its model card is available in Appendix A. The remainder of this paper is structured as follows. Section 3 points out the datasets and models used. It also describes the details of the experiments carried out and the metrics of the evaluation. Section 4 gives the results of the experiments on the two tasks, the error analysis, and the evaluator feedback. Section 2 briefly discusses past efforts that are connected to this study. Section 6 describes some of the limitations of this work, and Section 7 summarizes this work.

2. Related Work

Jhamtani et al. [1] observed that performance dipped when some deep models were evaluated on two open-domain dialogue datasets—DailyDialog and PersonaChat—with regards to figurative language [8,9]. They compared the Generative Pre-trained Transformer (GPT)-2 with four other models on the datasets and noticed the drop in performance among most of these models. In their work, however, they proposed transforming figurative language (including idioms) into its literal form before feeding the models. Idiom detection usually takes place with one of two approaches: type-based and token-based [5–7,10]. A type-based approach attempts to determine if an expression is an idiom, perhaps through the automatic compilation of an idiom list from a corpus [7]. A token-based approach relies on the context to distinguish idioms [2,7].

Non-contextual word embeddings (such as word2vec) are used to identify metaphors [11]. However, such an approach may underperform [11]. Peng et al. [5] used word2vec to obtain vectors from the text8 corpus. Their algorithm, which was based on the assumption that literal vectors are distinguished from idiom vectors by the larger inner product that they produce, used the inner products of context word vectors with a vector representing the target expression. Bizzoni et al. [12] used word2vec and an artificial neural network (ANN) with one hidden layer to detect metaphors. Diab et al. [13] performed binary classification using a support vector machine (SVM), which produced literal and idiomatic expressions on a subset of the verb–noun construction (VNC) tokens. Using BERT, which is based on contextual embeddings, Minghuan and Jing [14] suggested that the pretrained model is capable of accurately separating the literal and idiomatic usages of a PIE and encoding the meaning of a PIE to some extent. Aside from BERT, other contextual representation models that may be used in idiom detection include RoBERTa [15], XLNet [16], ALBERT [17], and ELECTRA [18]. Most of them learn bidirectional contexts.

3. Materials and Methods

All of the experiments were performed on a shared DGX-1 machine with 8 × 32 Nvidia V100 GPUs. LEADTEK Research Inc. Taiwan. The operating system on the server was Ubuntu 18. It had 80 CPU cores. Textual preprocessing was applied before training in all of the experiments [19]. This included the removal of URLs, excess spaces, emails, IP addresses, numbers, and special characters. Three runs per experiment were conducted in order to calculate the average accuracies, F1 scores, perplexities, and standard deviation (sd) [20]. Data shuffling was performed before training [21].

3.1. Multi-Domain Wizard-of-Oz (MultiWOZ) Dataset

The MultiWOZ dataset is a large, multi-domain, multi-topic, and multi-task conversational dataset that was originally designed for task-oriented dialogues [22]. It is a labeled collection of human–human written conversations and consists of more than 10,000 dialogues distributed between 70% multi-domain dialogues and 30% single-domain dialogues. The data-acquisition pipeline involved crowd-sourcing without the hiring of professional annotators. Ref. [22] considered different dialogue scenarios that ranged from requests for basic information about attractions to booking a hotel room, restaurant, train, or taxi between cities. Additional domains covered were hospitals and police. This has been a standard benchmark for different dialogue problems. It was used in neural context-to-response generation experiments by [22] and adapted for open-domain conversational systems by [23,24]. There are several versions of the dataset, with each new one bringing improvements [25].

3.2. The Dataset Used

The Potential Idiomatic Expression (PIE)-English idiom corpus was used for both the classification and conversation generation tasks investigated in this work. 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) [26]. About 76.94% of the samples are metaphors, making this the largest class in the dataset. Table 1 shows some examples from the dataset, and a short data statement that captures other key characteristics of the dataset, as given by [3], is given below. Before training in both tasks, the corpus is split into the ratio of 80:10:10 for the training, dev (validation set for evaluation), and test sets, respectively. In order to use the PIE-English idiom corpus for the second task of conversation generation as intended, we made the assumption that the corpus was suitable as a conversational dataset of dialogue turns. This assumption was valid because the sentences of the turns discussed the same cases of idioms despite being drawn from different examples from the base corpora.

> ***Short data statement for the PIE-English idiom corpus.***
> This is the Potential Idiomatic Expression (PIE)-English idiom 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 1197 total cases of idioms and 10 classes.
> Total samples of euphemism (2384), literal (1140), metaphor (14,666), personification (448), simile (1232), parallelism (64), paradox (112), hyperbole (48), oxymoron (48), and irony (32).

Table 1. Samples from the PIE-English idiom corpus [3].

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

3.3. Classification Task

3.3.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT [27] was used in the classification task of this study. It is an encoder stack from the Transformer architecture Vaswani et al. [28], where the base version has 12 layer blocks, a hidden size of 768, 12 self-attention heads, and 110 M parameters. It is pretrained with a deeply bidirectional method, where 15% of the words in the input are masked so that it predicts only the masked words, in what is called a masked language model (MLM).

3.3.2. Text-to-Text Transfer Transformer (T5)

The T5 [29] is also based on the Transformer architecture by Vaswani et al. [28]. It uses a different layer normalization in which there is no additive bias applied and the activations are only rescaled. Autoregressive self-attention is used in the decoder for it to attend to past outputs. The T5-Base model has about twice the number of parameters (220 M) as that of BERT-Base. Its has 12 layers each in the encoder and decoder blocks [29]. The T5 training method uses standard maximum likelihood and a cross-entropy loss.

3.3.3. Fine-Tuning Process

The two SoTA pretrained models (BERT and T5) were compared in the classification task. The models were acquired from the HuggingFace hub [30]. Both models involved their base versions and employed a linear schedule with a warmup for the learning rate (LR) adjustment. We used batch sizes of 64 and 16 for BERT and T5, respectively. The T5 batch was lower because it required more memory and did not fit on a single V100 GPU. The total number training epochs for both was 6. The T5 model took a hyperparameter called a task prefix, and we used "classification" as the prefix. Although the choice of hyperparameters can have a significant impact on the performance of embeddings or models [31,32], we did not carry out extensive hyperparameter exploration.

3.4. Conversation Generation

3.4.1. Dialogue Generative Pre-Trained Transformer (DialoGPT)-Medium

DialoGPT was trained on Reddit conversations of 147 M exchanges [4]. It is an autoregressive LM that is based on the GPT-2 [33]. In single-turn conversations, it achieved the SoTA in automatic and human evaluations. The medium model has 345 M parameters and 24 transformer layers. This was the model version that was adapted in this study. An advantage of this model is its easy adaptability to new dialogue datasets with few samples. Recent improvements to the DialoGPT model jointly trained a grounded generator and document retriever [34].

3.4.2. Fine-Tuning Process

Three instances of the SoTA DialoGPT model were compared in the conversation generation task. The first instance (IdiomWOZ) was created from the model checkpoint by [23], and it was trained on the MultiWOZ dataset and is available on the HuggingFace hub (huggingface.co/tosin/dialogpt_mwoz accessed date: 1 September 2022). This was achieved by fine tuning on the PIE-English idiom corpus. The second instance (IdiomOnly) was created from the original DialoGPT-Medium model by [4] by fine tuning on the same idiom corpus. The model checkpoint by [23] for the first instance was also based on the medium version of DialoGPT. The third instance (MultiWOZ) was the model checkpoint by [23], which was trained on the MultiWOZ dataset. For all three instances, we set the decoding algorithm to top-k (k = 100) and top-p (p = 0.7). We used only one GPU for training. The other hyperparameters were a maximum decoding length of 200 tokens, a temperature of 0.8, and a maximum ngram repeat limit of 3. Furthermore, for the fine-tuned models, we used the AdamW optimizer, the initial learning rate (LR) was 5×10^{-5}, a linear schedule with a warmup was used as a scheduler for the LR adjustment, and the total number of training epochs was 3. Due to memory constraints for the model size, the batch size was 2 and the context size was 7.

All three model checkpoints were 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 set and the PIE-English idiom corpus. The prompts were fed to the three models. For each dataset, thirty-two of the prompts were for generation, and fifteen prompts with their test set responses (for credibility) were selected. In the second set of experiments, which had the objective of finding fitting and diverse responses to idiom prompts, sixty-two random numbers were generated. Thirty-two were from the idiom test set and were used as prompts for both the IdiomWOZ and MultiWOZ sets, while the remaining thirty were credibility conversations from the MultiWOZ test set. The credibility conversations were used to determine the suitability of the evaluators; hence, the responses to these prompts were the corresponding responses from the test sets. They were distributed at regular intervals within the transcripts.

3.4.3. Evaluation

Automatic metrics, such as BLEU or ROUGE [35,36], which are common for natural language generation (NLG) tasks such as MT [28,37], are sometimes viewed as inadequate for conversational systems [38,39]. This is because they do not correlate well with human assessments [40]. In this work, we used another common metric, perplexity, which was also used by [41]. 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 [42].

For the human evaluation, evaluators (or annotators) were recruited on Slack (slack.com). They were second/L2 (but dominant) speakers of English and were unbiased respondents who did not take part in the training of the models (each annotator was paid a small amount after completing their task). For the evaluation of the transcripts, *Instruction 1* and *Instruction 2* below were the instructions for the transcripts from the first and second sets of experiments, respectively. Three valid evaluated transcripts from three annotators were accepted per set of experiments. The first set of transcripts were evaluated for human-likeness, while the second set was based on two characteristics: more fitting and more diverse responses (transcripts: drive.google.com/file/d/1f_x2KF7JfsmY2UoJk4FQCI9wfJcrKXVm/view?usp=sharing).

> ***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).

3.4.4. Credibility Unanimous Score (CUS)

In order to measure inter-annotator agreement (IAA) of the conversation transcripts, we used the CUS, which was introduced by [24]. It is more intuitive, easier to calculate (based on percentages), and to be appears less sensitive to changes in the number of categories being evaluated when compared to the Fleiss Kappa (k). The Fleiss Kappa (k) is known to be restrictive in its interpretation, depending on the number of categories [43], as the Kappa is lower when there are more categories [44]. According to [24], the assumption behind the CUS is that if homogeneous samples may be used to check the credibility of the annotators, then they may be used to establish 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 is a minimum of 30 samples, thereby fulfilling the central limit theorem. The probability of obtaining a high CUS rises when the benchmark score for annotator credibility is high.

4. Results

4.1. Classification

Table 2 shows that the T5 model outperformed the BERT model. It also outperformed the best model from [3], another BERT implementation. The results from [3] do not provide standard deviation values, and they reported results only on the training and dev set split with a 85:15 ratio. From the results, it appears that the PIE-English idiom corpus was 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 in the corpus is one sentence or, at most, two sentences. More lengthy input texts seemed to be more challenging for models than short texts/tweets because of the longer context required. The results were 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) was smaller than the alpha (0.05).

Table 2. Average accuracy and F1 results (sd—standard deviation). *: Data split ratio of 85:15 for the 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 * [3]	0.93	-	0.95	-	-	-

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 idiom corpus. The model performed substantially well even for classes that had few samples in the training set, such as *hyperbole* and *irony*. Overfitting was mitigated through the use of a pretrained model and early stopping. However, it struggled mostly in correctly classifying the *literals*. It misclassified about 9.3% of them as *metaphor* because this was the largest class in the dataset. Imbalance in training data is known to be a problem that affects model performance [45–47]. There are oversampling methods for mitigating this challenge [45].

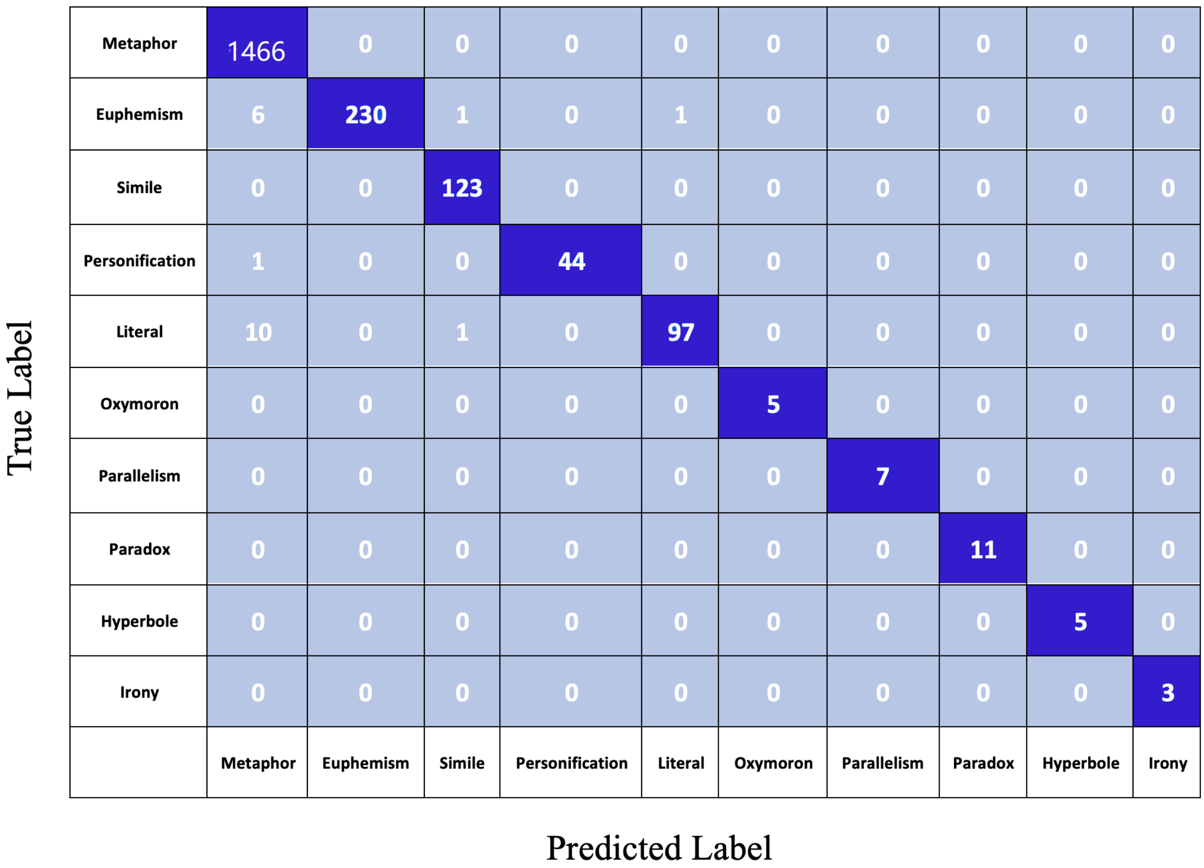

Figure 1. Confusion matrix for the T5 model on the PIE-English idiom corpus test set.

4.2. Conversation Generation

We can observe from Table 3 that the MultiWOZ model from [23] had the lowest average perplexity when compared with the other two new models. This is likely because the MultiWOZ dataset that the model was trained on were larger than the idiom corpus. The *p*-value ($p < 0.0001$) of the two-sample *t*-test for the difference of two means (for IdiomWOZ and IdiomOnly) was smaller than the alpha (0.05); hence, the results are also statistically significant. Despite the average perplexity for the IdiomOnly model being lower than that of IdiomWOZ, we chose to generate responses and conduct human evaluations on the latter. This was because one of its runs had a lower perplexity, which may have been deduced from the standard deviation. In addition, perplexity alone may not be sufficient to tell how good a model is [48,49].

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 [23]	6.41 (-)	6.21 (-)

The results of the human evaluation 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. From Table 4, one can observe that both the MultiWOZ and IdiomWOZ models had more human-like single-turn conversations than the other categories did, under majority votes. The MultiWOZ model had the most human-like conversations. However, considering idiom-only prompts, the IdiomWOZ model had the most human-like conversations. From Table 5, IdiomWOZ had more fitting conversations than MultiWOZ, though the opposite was the case with regard to more diverse conversations. This could be due to the evaluators' interpretations of what was diverse. For all of the evaluations, we observed that there was a CUS of 80%.

Table 4. Human evaluations of three annotators on three classes for 64 single-turn conversations (CUS = 80%).

Model	**Scale (Majority Votes)**			
	H (%)	**U (%)**	**N (%)**	**3-Way (%)**
IdiomWOZ	39.1	10.9	37.5	12.5
IdiomOnly	15.6	12.5	60.9	10.9
MultiWOZ	62.5	1.6	32.8	3.1
	Unanimous Votes—3/3			
IdiomWOZ	20.3	0	12.5	-
IdiomOnly	6.3	0	31.3	-
MultiWOZ	45.3	0	23.4	-
	Idiom-Only Maj. Votes (32 Samples)			
IdiomWOZ	30	23.3	33.3	13.3
IdiomOnly	26.7	20	36.7	16.7
MultiWOZ	26.7	3.3	66.7	3.3

Table 5. Human evaluation results of three annotators for 32 single-turn conversations (CUS = 80%).

Model	Scale (Majority Voting)	
	More Fitting (%)	More Diverse (%)
IdiomWOZ	71.9	28.1
MultiWOZ	28.1	71.9
	Unanimous Votes—3/3	
IdiomWOZ	46.9	9.4
MultiWOZ	9.4	15.6

5. Discussion and Evaluator Feedback

Tables 6 and 7 show some single-turn conversations from the second transcript. Person 1 was the prompt from the PIE-English idiom test set. Person 2 was the IdiomWOZ model, and it responds more with idioms that were generated. Person 3 was the MultiWOZ model. The IdiomWOZ model, whose responses were deemed to be more fitting, appeared to confirm the results of a previous study, which showed that metaphors have more emotional impact than their literal equivalents [50]. Idioms usually make utterances more colorful (or rich) and diverse. Hence, simply reducing idioms to their literal form before feeding NLP models, as practiced by [1], may not adequately address the challenge, since it implies that the models are still incapable of "understanding" the idioms and because some idioms have more than one literal expression.

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

Feedback from some of the evaluators suggested that the use of idioms slightly complicated the evaluation task, but it was sometimes useful in identifying which of the two conversations was a more fitting or diverse option. They found it relatively easier to evaluate the characteristic of human-likeness 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 as non-human-like because there was a lack of connection between the prompt and the response. In the second set of transcripts, some conversations were evaluated as more fitting when the responses answered the prompts directly, while some were evaluated as more diverse when the response was not a direct answer but elicited further discussion.

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

6. Limitations

The results in this study are based on models and data for the English language. It is uncertain if the results can be generalized across other languages. The PIE-English idiom corpus that was used in this work, though relatively large and probably the first with as many as 10 classes, does not represent all of 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 the data used to pretrain the deep models are from public online sources that are known to contain such risks [4].

7. Conclusions

We addressed 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. Therefore, it is important to train open-domain conversational systems on idiom data so as to achieve diversity and more fitting responses in ML models. This is especially important because idioms are part of everyday speech in many cultures [51]. Other practical implications of this work include the improved user experience with (open-domain) conversational systems, better nuanced translations from MT systems, and improved WSD, all resulting from idiom-aware ML models. We also achieved SoTA classification results on the PIE-English idiom corpus by using the T5 model. Future efforts may be directed at exploring more datasets of figurative language or idioms, the use of explainable artificial intelligence (XAI) to explain the predictions of the models, and more diverse SoTA models for training.

Author Contributions: Conceptualization, T.A. and F.L.; methodology, T.A.; investigation, T.A.; writing—original draft preparation, T.A.; writing—review and editing, F.L. and M.L.; supervision, F.L. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The PIE-English idiom corpus is available online (github.com/tosingithub/idesk) accessed date: 1 September 2022.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

NLP	Natural Language Processing
NER	Named Entity Recognition
SA	Sentiment Analysis
ML	Machine Learning
BoW	Bag-of-Words
CBoW	Continuous Bag-of-Words
SLTC	Swedish Language Technology Conference
ANN	Artificial Neural Network
NN	Neural Network
LSTM	Long Short-Term Memory Network
biLSTM	Bidirectional Long Short-Term Memory Network
SoTA	State-of-the-Art
NLG	Natural Language Generation
NLU	Natural Language Understanding
MWE	Multi-Word Expression
SW	Simple Wiki
MT	Machine Translation
BW	Billion Word
PIE	Potential Idiomatic Expression
IAA	Inter-Annotator Agreement
RTE	Recognizing Textual Entailment
IR	Information Retrieval
QA	Question Answering
BNC	British National Corpus
UKWaC	UK Web Pages
AI	Artificial Intelligence
GDC	Gothenburg Dialogue Corpus
dialogpt DialoGPT	Dialogue Generative Pre-trained Transformer
GPT	Generative Pre-trained Transformer
MultiWOZ	Multi-Domain Wizard-of-Oz
T5	Text-to-Text Transfer Transformer
BART	Bidirectional and Auto-Regressive Transformer
XLM-R	Cross-Lingual Model-RoBERTa
M2M	Many-to-Many Multilingual Translation Model
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
ELMo	Embeddings from Language Models
PII	Personally Identifiable Information
QG	Question Generation
TC	Text Classification
PCL	Patronizing and Condescending Language
GUS	Genial Understander System
GMB	Groningen Meaning Bank
WSD	Word Sense Disambiguation
CC-BY4	Creative Commons Attribution 4.0
CI	Confidence Interval
BLEU	Bilingual Evaluation Understudy
GDPR	General Data Protection Regulation
SVM	Support Vector Machine
VS	Vector Space

VSM	Vector Space Model
NLTK	Natural Language Toolkit
tf-idf	Term Frequency–Inverse Document Frequency
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
LSI	Latent Semantic Indexing
PLSI	Probabilistic Latent Semantic Indexing
LDA	Latent Dirichlet Allocation
LM	Language Model
biLM	Bidirectional Language Model
PoS	Part of Speech
NNLM	Neural Network Language Model
BPE	Byte-Pair Encoding
OOV	Out-of-Vocabulary
IMDB	Internet Movie Database
LR	Learning Rate
CUS	Credibility Unanimous Score
IE	Information Extraction
RL	Reinforcement Learning
MDL	Minimal Dependency Length
MLM	Masked Language Model
RQ	Research Questions
VNC	Verb–Noun Constructions
XAI	Explainable Artificial Intelligence

Appendix A

Table A1. Model card for DialoGPT(medium)-MultiWOZ, which was trained on the PIE-English idiom corpus.

	Details
Model rationale	This is a fine-tuned model of DialoGPT (medium)-MultiWOZ v2.2 on the Potential Idiomatic Expression (PIE)-English idiom corpus. It is intended to be used as an idiom-aware conversational system.
Limitation	The dataset that it is trained on is limited in scope, as it covers only 10 classes, namely, metaphor, simile, euphemism, parallelism, personification, oxymoron, paradox, hyperbole, irony, and literal.
Result	The perplexities achieved on the test and validation sets are 160.9 and 160.46, respectively.
	Others
Model link	https://huggingface.co/tosin/dialogpt_mwoz_idioms (accessed date: 1 September 2022)
More	Information about the original pretrained model can be found at https://huggingface.co/tosin/dialogpt_mwoz accessed date: 1 September 2022
Lisence	CC-BY 4.0.
	Example of generation 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

References

1. Jhamtani, H.; Gangal, V.; Hovy, E.; Berg-Kirkpatrick, T. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, 7–11 November 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7476–7485. [CrossRef]
2. Korkontzelos, I.; Zesch, T.; Zanzotto, F.M.; Biemann, C. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, GA, USA, 14–15 June 2013*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; Volume 2, pp. 39–47.
3. Adewumi, T.; Vadoodi, R.; Tripathy, A.; Nikolaidou, K.; Liwicki, F.; Liwicki, M. Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms. In Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, 21–23 June 2022; European Language Resources Association (ELRA): Paris, France, 2022.
4. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. 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, Online, 5–10 July 2020; pp. 270–278. [CrossRef]
5. Peng, J.; Feldman, A.; Jazmati, H. Classifying idiomatic and literal expressions using vector space representations. In Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 5–11 September 2015; pp. 507–511.
6. Li, L.; Sporleder, C. Classifier combination for contextual idiom detection without labelled data. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 315–323.
7. Sporleder, C.; Li, L.; Gorinski, P.; Koch, X. Idioms in Context: The IDIX Corpus. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010.
8. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, 27 November–1 December 2017; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; pp. 986–995.
9. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. 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), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2204–2213. [CrossRef]
10. Cook, P.; Fazly, A.; Stevenson, S. 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, Prague, Czech Republic, 25–27 June 2007; pp. 41–48.
11. Mao, R.; Lin, C.; Guerin, F. 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), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1222–1231. [CrossRef]
12. Bizzoni, Y.; Chatzikyriakidis, S.; Ghanimifard, M. "Deep" Learning : Detecting Metaphoricity in Adjective-Noun Pairs. In Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark, 8 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 43–52.[CrossRef]
13. Diab, M.; Bhutada, P. Verb noun construction MWE token classification. In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009), Singapore, 6 August 2009; pp. 17–22.
14. Tan, M.; Jiang, J. Does BERT Understand Idioms? A Probing-Based Empirical Study of BERT Encodings of Idioms. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 9–10 September 2021; INCOMA Ltd.: Athens, Greece, 2021; pp. 1397–1407.
15. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
16. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
17. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
18. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
19. Obaid, H.S.; Dheyab, S.A.; Sabry, S.S. The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In Proceedings of the 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), Jaipur, India, 13–15 March 2019; pp. 279–283.
20. Javed, S.; Adewumi, T.P.; Liwicki, F.S.; Liwicki, M. Understanding the Role of Objectivity in Machine Learning and Research Evaluation. *Philosophies* **2021**, *6*, 22. [CrossRef]
21. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.

22. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gašić, M. 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, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 5016–5026. [CrossRef]
23. Adewumi, T.; Brännvall, R.; Abid, N.; Pahlavan, M.; Sabry, S.S.; Liwicki, F.; Liwicki, M. Småprat: DialoGPT for Natural Language Generation of Swedish Dialogue by Transfer Learning. In Proceedings of the 5th Northern Lights Deep Learning Workshop, Tromsø, Norway, 10–12 January 2022; Septentrio Academic Publishing: Tromsø, Norway, 2022; Volume 3. [CrossRef]
24. Adewumi, T.; Adeyemi, M.; Anuoluwapo, A.; Peters, B.; Buzaaba, H.; Samuel, O.; Rufai, A.M.; Ajibade, B.; Gwadabe, T.; Traore, M.M.K.; et al. Ìtàkúròso: Exploiting Cross-Lingual Transferability for Natural Language Generation of Dialogues in Low-Resource, African Languages. *arXiv* **2022**, arXiv:2204.08083.
25. Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A.; Ku, P.; Hakkani-Tur, D. 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, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 422–428.
26. Ferraresi, A.; Zanchetta, E.; Baroni, M.; Bernardini, S. 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, Marrakech, Morocco, 1 June 2008; pp. 47–54.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
29. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
30. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45. [CrossRef]
31. Adewumi, T.; Liwicki, F.; Liwicki, M. Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks. *Open Comput. Sci.* **2022**, *12*, 134–141. [CrossRef]
32. Adewumi, T.P.; Liwicki, F.; Liwicki, M. Exploring Swedish & English fastText embeddings for NER with the Transformer. *arXiv* **2020**, arXiv:2007.16007.
33. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
34. Zhang, Y.; Sun, S.; Gao, X.; Fang, Y.; Brockett, C.; Galley, M.; Gao, J.; Dolan, B. Joint Retrieval and Generation Training for Grounded Text Generation. *arXiv* **2021**, arXiv:2105.06597.
35. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 1 September 2003; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.
36. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the ACL'02, 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
37. Gehrmann, S.; Adewumi, T.; Aggarwal, K.; Ammanamanchi, P.S.; Aremu, A.; Bosselut, A.; Chandu, K.R.; Clinciu, M.A.; Das, D.; Dhole, K.; et al. 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), Online, 1 September 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 96–120. [CrossRef]
38. Jurafsky, D.; Martin, J. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Dorling Kindersley Pvt, Limited: London, UK, 2020.
39. Liu, C.W.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv* **2016**, arXiv:1603.08023.
40. Reiter, E. 20 Natural Language Generation. In *The Handbook of Computational Linguistics and Natural Language Processing*; Wiley: Hoboken, NJ, USA, 2010; p. 574.
41. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a human-like open-domain chatbot. *arXiv* **2020**, arXiv:2001.09977.
42. Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In *Mining Text Data*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 163–222.
43. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
44. Sim, J.; Wright, C.C. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Phys. Ther.* **2005**, *85*, 257–268. [CrossRef] [PubMed]

45. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
46. Adewumi, T.; Alkhaled, L.; Mokayed, H.; Liwicki, F.; Liwicki, M. ML_LTU at SemEval-2022 Task 4: T5 Towards Identifying Patronizing and Condescending Language. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Online, 14–15 July 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 473–478. [CrossRef]
47. Sabry, S.S.; Adewumi, T.; Abid, N.; Kovacs, G.; Liwicki, F.; Liwicki, M. HaT5: Hate Language Identification using Text-to-Text Transfer Transformer. *arXiv* **2022**, arXiv:2202.05690.
48. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.L.; et al. 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, Online, 19–23 April 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 300–325. [CrossRef]
49. Hashimoto, T.B.; Zhang, H.; Liang, P. 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), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 1689–1701. [CrossRef]
50. Mohammad, S.; Shutova, E.; Turney, P. Metaphor as a Medium for Emotion: An Empirical Study. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, Berlin, Germany, 11–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 23–33. [CrossRef]
51. Alm-Arvius, C. *Figures of Speech*; Studentlitteratur: Lund, Sweden, 2003.