Generating Conceptual Metaphoric Paraphrases

Olof Gotting

Department of Linguistics
Bachelor’s Thesis 15 ECTS credits
Computational Linguistics – Bachelor’s Course, LIN621
Bachelor’s Program in Philosophy and Linguistics 180 ECTS credits
Spring semester 2021
Supervisor: Robert Östling
Swedish title: Generering av konceptuella metaforparafraser

Generating Conceptual Metaphoric Paraphrases

Olof Gotting

Abstract

Metaphoric Paraphrase generation is a relatively new and unexplored Natural Language Generation task. The aim of the task is to develop computational systems that paraphrase literal sentences into cogent metaphoric ones. Challenges in the field include representation of common sense knowledge and ensuring meaning retention when dealing with phrases that are dissimilar in their literal sense. This thesis will deal with the specific task of paraphrasing literal adjective phrases into metaphoric noun phrases, taking into consideration the preceding context of the adjective phrase. Two different systems were developed as part of this study. The systems are identical, apart from the fact that one is endowed with a knowledge representation based on Conceptual Metaphor Theory. The paraphrases generated by the systems, along with paraphrases written by a native speaker of English, were scored on the parameters of meaning retention and creativity by a crowd-sourced panel. Both systems were able to generate cogent metaphoric paraphrases, although fairly unreliably compared to the human. The system endowed with Conceptual Metaphor Theory knowledge got a lower average meaning retention score and a higher average creativity score than the system without Conceptual Metaphor Theory knowledge representation. In addition to that it was found that less similarity in sentence embeddings of literal sentences and metaphoric paraphrases of them correlates with a higher level of perceived meaning retention and a lower perceived creativity of the metaphoric paraphrase. It was also found that less difference in GPT-2 log probability between literal sentences and metaphoric paraphrases of them correlates with humans evaluating the paraphrases as less creative.

Keywords

Metaphor, Natural Language Processing, Natural Language Generation, Machine Learning, Conceptual Metaphor Theory, Computational Creativity

Sammanfattning

Generering av metaforiska parafraser är ett relativt nytt och outforskat område inom språkteknologi. Det handlar om utveckling av mjukvara som kan parafrasera meningar med bokstavlig betydelse till målande metaforiska meningar. Några av utmaningarna inom området är att representera sunt förnuft-kunskap och att garantera betydelsebevarande mellan fraser som har olika bokstavlig betydelse. Den här artikeln kommer behandla den specifika uppgiften att parafrasera bokstavliga adjektivfraser till metaforiska nominalfraser, samtidigt som adjektivfrasens föregående kontext hålls i åtanke. Två system utvecklades för denna studie. Systemen är identiska, förutom att den ena är utrustad med kunskapsrepresentation baserad på konceptuell metaforteori. Parafraserna som genererades av de olika systemen och parafraser skrivna av en person med engelska som modersmål poängsattes på parametrarna meningsbevarande och kreativitet av en panel. Båda systemen lyckades generera metaforiska parafraser, dock inte lika tillförlitligt som en människa. Parafraserna som genererades av systemet utrustat med kunskapsrepresentation baserad på konceptuell metaforteori fick lägre medelpoäng på betydelsebevarande och högre medelpoäng på kreativitet än systemet utan. Därtill upptäcktes det att mindre likhet i meningsinbäddningar mellan bokstavliga meningar och metaforiska parafraser av dem korrelerar med en högre nivå av upplevt betydelsebevarande och en lägre nivå av upplevd kreativitet hos den metaforiska parafrasen. Det upptäcktes också att mindre skillnad i GPT-2-log-sannolikhet mellan bokstavliga meningar och metaforiska parafraser av dem korrelerar med att människor skattar parafraserna som mindre kreativa.

Nyckelord

Metaforer, datorlingvistik, språkteknologi, språkgenerering, maskininlärning, konceptuell metaforteori, kreativa datorer

Contents

1 Introduction

> A metaphor is a word or expression that is used to talk about an entity or quality other than that referred to by its core, or most basic meaning. This non-core use expresses a perceived relationship with the core meaning of the word, and in many cases between two semantic fields (Deignan 2005, p. 34).

The metaphor is a fascinating property of human language. Our ability to communicate about certain things by using words for other things highlights one of the many wondrous aspects of our cognition. It is indicative of our extraordinary ability to find patterns and our inclination towards symbolism. Commonsensical pattern finding seems to be a prerequisite for understanding and using metaphor, and more and more successful attempts at endowing computers with the ability have been made (Veale 2013; Bosselut et al. 2019; Bisk et al. 2020; Chakrabarty, Zhang, et al. 2021). Practically, a computational system adept at metaphoric paraphrasing could be used as an assistant in creative writing. Or, further down the road, in conjunction with advanced general natural language generation models, to write entire poems, song lyrics, screenplays and novels. Optimistically, even to leverage our aforementioned pattern finding ability to elucidate difficult abstract concepts to humans by illustrating them in terms of concepts more familiar to us.

Many studies have been done on metaphor detection (Veale, Shutova, et al. 2016; Jang et al. 2016; Bisk et al. 2020; Mu et al. 2019; Stowe, Moeller, et al. 2019), but the studies on metaphoric paraphrase generation are few (Terai and Nakagawa 2009; Yu and Wan 2019; Stowe, Ribeiro, et al. 2020; Chakrabarty, Zhang, et al. 2021). Those have mostly focused on the paraphrasing of verbs to increase their metaphoricity, while this study will focus on paraphrasing literal adjective phrases into metaphoric noun phrases. The systems developed in this project leverage several pre trained neural networks for semantic analysis and common-sense knowledge representation to generate cogent metaphoric paraphrases.

2 Background

In this section the reader will be presented with background to the theoretical subjects this study touches on. Then the purpose of the study and the research questions will be stated.

2.1 Reasoning about Metaphors

2.1.1 The Conceptual Metaphor

The metaphor is no small part of our everyday language use and we use it to achieve artistic or rhetoric effect as well as to effectively communicate. Our mental lexicons are full of metaphors and they pervade all domains of linguistic interaction and have done so for a long time. As Aristotle (ca. 335 B.C.E./2017) wrote in Poetics:

> “But the greatest thing by far is to be a master of metaphor. It is the one thing that cannot be learnt from others; and it is also a sign of genius, since a good metaphor implies an intuitive perception of the similarity in dissimilars.”

Throughout history, what Deignan (2005) calls the “decorative approach” seems to have been the leading theory of metaphor. Like the term suggests, in that approach metaphor is viewed just as having an ornamental function in language. Literal language is viewed as standard and metaphors are used for artistic effect or when no literal words can be found or are insufficient. While this theory wasn’t an academically fleshed out one, it was throughout most of the 20th century taken for granted as commonsensical (Deignan 2005). So while the concept of the metaphor and the interest in it goes back at least to classical antiquity, the seminal work by Lakoff and Johnson (1980) renewed the field and switched its focus from the decorative approach towards the metaphor as a cognitive phenomenon. They argue that metaphors fundamentally are conceptual, meaning that the success of metaphors to convey meaning hinges on the fact that they are linguistic realizations of concepts within human cognition. This theory is known as Conceptual Metaphor Theory (CMT). In CMT a distinction is made between the *target domain* and the *source domain*. The target domain is the thing to be described, and the source domain, which is usually less abstract, is what describes it. In the classic example of “Argument is war”, “argument” is the target domain, explained by the source domain of “war”.

Advances in computers and an increase in available linguistic data have advanced the field of *Corpus linguistics*, which deals with analysis of enormous amounts of computer readable language data. This has allowed researchers to find corroborative evidence for the central theses of CMT and advance the theory as a whole (see Deignan 2005, pp. 96–99 for examples).

2.1.2 Knowledge Representation

Because knowledge seems to be a prerequisite for intelligence, Knowledge Representation is a central notion in Artificial Intelligence research. With knowledge representation the objective is to equip computational systems with information and heuristics for using that information in deciding strategies to accomplish their goals. An example of a specific type of knowledge representation is that of an *ontology*. This category of knowledge representation, which takes its name from the philosophical studies of what exists, is made up of representative models for reasoning about the world. An ontology is a mapping of a population of intelligent agents’ concepts and the relations between them. Or, as succinctly defined by Gruber (2007) - “... An explicit specification of a conceptualization”. In an ontology, the knowledge being represented is typically of sets of individual concepts, their attributes, and relations between individuals or sets (Gruber 2007). In this study then, the knowledge to be represented is in the domain of metaphor. More precisely, as CMT is the metaphor theory espoused here, knowledge representation will be used to allow for reasoning about how English speakers talk, as well as think about certain things by using terms for other things.

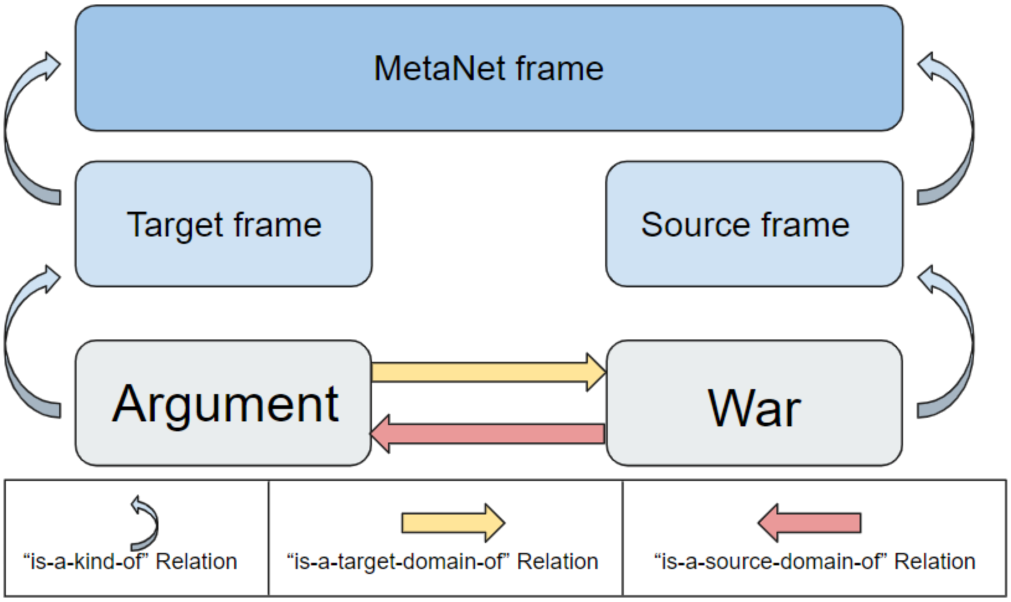

Figure 1: *The MetaNet based knowledge representational ontology.*

2.1.3 MetaNet

MetaNet (Dodge et al. 2015) is a repository of formalized frames and metaphors. The project contains a dataset of metaphors and the domain frames within them in accordance with CMT. The metaphors, the frames and the relations between them have been manually added to the repository by linguists trained in CMT. For this project, metaphors, frames and frame-frame relations were scraped from the MetaNet wiki [1]. This data, which is a form of knowledge representational ontology, was used as the main source of knowledge representation in the project. The set of individuals in the ontology contains all the frames, each of which has the property of being either a target frame or a source frame, and the relations between individuals are of them either being a source, or a target, of another frame. See Figure 1 for an illustration of the ontology.

2.2 Computational Creativity

Defined by Colton and Wiggins (2012) as

> "The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative",

Computational Creativity is a subfield of Artificial Intelligence which studies computers as something approximating creative agents in their own right, rather than merely tools to be used in the creative endeavors of us humans. It's a discipline where you ask if and how computational systems can become co-creators or even autonomous creators (Veale and Pérez y Pérez 2020, p. 551).

[1] https://metaphor.icsi.berkeley.edu/pub/en/

The task of metaphoric paraphrase generation, which is the subject of this paper, fits right in the field of computational creativity. An advanced system for understanding and generating metaphors could help us in creative endeavors, be they literary or rhetorical or otherwise. The most obvious one is being a tool for writers, helping them achieve more expressive and cogent writing. Endowing computers with more advanced knowledge of metaphor could also be useful for linguistic human-machine interaction, as used in products such as consumer-grade virtual home assistants, chat bots, and automated phone systems. If CMT is accurate, an even more sophisticated understanding of metaphor could help us in understanding how we think and how our minds are influenced by exposure to language.

2.3 Related Research

2.3.1 Metaphor Masking

Stowe, Ribeiro, et al. (2020) details two systems they created for generating metaphoric paraphrases. In it they focus on paraphrasing verbs, using a lexical replacement baseline and a sequence to sequence model. In an evaluation survey, neither of the developed systems come close to being as good at metaphoric paraphrasing as a literary expert. Their baseline utilizes WordNet (Fellbaum 2010), which is a lexical database of semantic relations between words, to pick out all hyponyms of the original verb as candidates for replacing the verb. The output embedding of each candidate, and the mean output embedding of the rest of the sentence is collected. Cosine similarity of each word embedding and the mean context embedding is then calculated, and the candidate most similar is chosen as the winner.

The other system by Stowe, Ribeiro, et al. (2020), dubbed Metaphor Masking , was trained by masking literal verbs in sentences and through parallel training data teaching the system to replace the masked verb with a metaphoric verb using a transformer architecture (Vaswani et al. 2017) based on the OpenNMT-py (Klein et al. 2017) sequence to sequence (seq2seq) learning paradigm. Seq2seq neural networks are called that because they convert a sequence into another sequence. They work by encoding a sequence (like a sentence in natural language) to a vector space and then decoding it into whatever it is programmed to produce as its output. Models with the transformer architecture can be contrasted with those of another seq2seq paradigm - Long short-term memory (LSTM). The LSTM architecture is based on recurrent neural networks through which information is passed in loops to allow for keeping some context of previous parts of the current sequence. Transformers, on the other hand, don't have any recurrent connections, but instead rely on what is known as self-attention layers (Jurafsky and Martin 2020). Simplified, self-attention allows for the Transformer system to find the relevance of a given element of a sequence to other elements in the context. The training data used by Stowe, Ribeiro, et al. (2020) contained sentences with a verb annotated as either metaphoric or literal. There were 35,415 sentences in total, of which 11,593 contained a metaphoric verb. The best paraphrase of the Metaphor Masking system, according to the annotators, was "She *was saddened* by his refusal of her invitation" into "She *besieged* by his refusal of her invitation".

2.3.2 SCOPE

Chakrabarty, Muresan, et al. (2020) developed an approach, called SCOPE, for generating similes. Similes are related to metaphor, but while a metaphor is saying something *is* some other thing, a simile says something *is like* some other thing. They frame the task as a style-transfer problem, meaning that they aim for a solution that paraphrases sentences to convey the same meaning. The style should change, while the semantic value is kept intact. Their system, using a simile corpus of data collected from social media, a common sense knowledge base and a seq2seq model with a transformers architecture, managed to generate similes better than two literary experts 37% of the time. An example of what SCOPE generated: “It was obscene, but she was drawn to it, *fascinated*” was paraphrased with a simile into “It was obscene, but she was drawn to it, *like a moth to a flame*”.

2.3.3 MERMAID

Chakrabarty, Zhang, et al. (2021) identified three challenges that pervaded the few earlier metaphor generation projects:

> “1) the lack of training data that consists of pairs of literal utterances and their equivalent metaphorical version in order to train a supervised model; 2) ensuring that amongst the seemingly endless variety of metaphoric expressions the generated metaphor can fairly consistently capture the same general meaning as the literal one, with a wide variety of lexical variation; and 3) computationally overcome the innate tendency of generative language models to produce literal text over metaphorical one.”

Their solution to challenge 1 was to develop an approach for collecting metaphoric sentences from poetry to create a parallel corpus of literal sentences. The creation of the parallel corpus was done with the BERT (Devlin et al. 2019) transformer model, which they fine-tuned on a metaphor detection corpus (Steen 2010). Using this method they found 518,865 sentences in Jacobs (2018) that could confidently be assumed to contain a metaphoric verb. A BERT (Devlin et al. 2019) model was then used to find the top verb candidates for literal replacement of the metaphoric verb.

Their approach is based on an underlying theory of metaphors being symbols, and to make sure meaning is preserved in these candidates, and thus provide a solution to challenge 2, they use a knowledge representational model (Bosselut et al. 2019) to find matching symbolism between literal and metaphoric verbs.

To make sure the system wouldn’t tend to generate literal text, when metaphoric text is what is aimed for, and thus solve challenge 3, they modified the decoder of BART (Lewis et al. 2019). BART is a Transformer that combines functionality of BERT-like encoders and GPT-like decoders. Their modification of the decoder enabled their system favor more metaphorical verbs.

This system by Chakrabarty, Zhang, et al. (2021), named MERMAID, performed much better than the Metaphor Masking paradigm by Stowe, Ribeiro, et al. (2020) in a human evaluation survey, where syntax, semantics, creativity and metaphoricity were scored for paraphrases generated by several systems. MERMAID was also used to enhance poems from a poetry subreddit. The poems enhanced by the system were preferred by human evaluators 68% of the time.

“The scream *filled* the night” was paraphrased by MERMAID into “The scream *pierced* the night”. This paraphrase outscored two creative writing experts in the human evaluation survey:

2.4 Purpose and Research Questions

The purpose of this research is to investigate if a logistic regression model trained on the three feature variables *Sentence vector similarity*, *Difference in sentiment* and *difference in sentences’ GPT-2 log probability* can be used to reliably paraphrase adjective phrases into cogent metaphoric noun phrases. In addition to that, what effect the feature variables have on the perceived meaning retention capability and creativity of the paraphrases will be studied. It will also be evaluated how knowledge representation in the form of MetaNet (Dodge et al. 2015) affects the cogency of the paraphrases. The aim is to, through design and comparative evaluation of systems with the paraphrasing task mentioned above elucidate what works and what doesn’t.

Question 1: What is a viable design of a system for paraphrasing adjective phrases into cogent metaphoric noun phrases?

Question 2: How do the three feature variables correlate with the human evaluation scores of the paraphrases' meaning retention and creativity?

Question 3: How does the implementation of CMT knowledge representation affect the cogency of the generated metaphoric paraphrases of literal sentences, as evaluated by humans?

3 Data

This section will describe what data was collected for this project, as well as how it was collected. It will also be made clear what motivated the data selection.

3.1 Synonyms, Hyponyms and Related Phrases

A large repository of synonyms, hyponyms and related phrases for the Metanet frames were collected in order to allow for more diverse paraphrases, by making the selection of possible candidates larger. For each metaphor frame on the Metanet wiki, if there were related phrases in the section "Relevant Lexical Units" on their wiki entry, they were collected. Hyponyms for all Metanet frames and their related phrases were collected from WordNet (Fellbaum 2010), if there were any. Hyponyms, rather than hypernyms, were actively sought out when collecting the data because they are more likely to carry strong connotations (Carter et al. 2001). It was also hypothesized that hypernyms are more likely to be explanatory, rather than metaphoric. With now having a collection of metaphor frames, related phrases and hyponyms, synonyms to these were obtained through the online lexicon Synonyms.com (2021). The final set contained 1519 unique phrases, each mapped to its semantically corresponding Metanet frames. To put this in the context of the ontology of Figure 1: The ontology has now been extended with a large amount of particulars. These particulars are the synonyms, hyponyms and related phrases and have a "is-a-synonym-hyponym-or-related-phrase-of" Relation with at least one Target frame or Source frame in the ontology.

3.2 Gigaword

English Gigaword (Napoles et al. 2012) is an archive of part of speech tagged news articles in English. This data was used to collect sentences to be paraphrased. The reason Gigaword was used to supply the sentences to be paraphrased was to avoid the bias there could have been if they were chosen by someone working on the project. Sentences of the form "[noun phrase] is/are/was/were [adjective phrase]", for example "Our culture is coarse and vulgar", were collected and categorized by how many adjectives were in the adjective phrase. 13230 sentences with one adjective were collected, along with 765 containing two adjectives, and 60 containing three adjectives. From each of these three sets 33 sentences were selected uniformly at random and put in a final set of 99 sentences.

3.3 Paraphrases Written by a Human

A native speaker of American English was recruited to paraphrase the 99 randomly selected sentences. They were instructed to paraphrase the adjective phrase of each sentence into a metaphoric indefinite noun phrase starting with ’a’ or ’an’, and including at most one adjective and no proper noun. All 99 human-written metaphoric paraphrases were used to train a logistic regression classifier. 66 of them were also used in the survey for comparison with the sentences generated by the programs being developed.

4 Method

The methods used to develop the paraphrasing systems and the survey for human evaluation of the metaphoric paraphrases generated by the systems will be explained here, but before that a quick overview of how the programs work will be given.

4.1 Brief Overview of the Paraphrasing Systems

The programs take a sentence such as “Our culture is coarse and vulgar.” as input. They check for possible noun phrases to replace the adjective phrase (“coarse and vulgar” in this case). This creates new sentences (among others “Our culture is a parasite.” and “Our culture is a thing.”), each of which is checked and given a probability. The 10 sentences deemed most probable to be a good metaphoric paraphrase are then endowed with an adjective (which might result in “Our culture is a great parasite.” etc.) and checked for probability of being a good metaphoric paraphrase again. Now, the sentence deemed most probable, with or without adjective, is selected as the winning candidate. In this case one system generated “Our culture is an abomination.”, and the other “Our culture is a strange thing.”

4.2 Computational Tools for Semantic Reasoning

The selection of computational tools for semantic reasoning was motivated by the hypothesis that the systems for metaphoric paraphrasing would need to be able to do three kinds of semantic analysis:

1. **Meaning analysis.** For a paraphrase to be good it should convey the same meaning as the phrase it is rewording. So any paraphrasing system should be able to reason about meaning. This was done using Sentence-BERT (Reimers and Gurevych 2019), detailed in 4.2.1.

2. **Sentiment analysis.** Paraphrasing an adjective phrase into a metaphoric noun phrase, it needs to be made sure that the sentiment stays the same in the metaphor. A system dealing with metaphors should have the information that the adjective phrase “happy, comfortable, and pretty” is closer in sentiment to “angel” than it is to “demon”. This was done using VADER (Hutto and Gilbert 2014), detailed in 4.2.2.

3. **Surprisal analysis.** While a metaphoric paraphrase should retain meaning, it should not retain the strictly literal meaning. For it to be cogent and creative it should to some degree be unique or unexpected. Therefore a system for metaphoric paraphrasing should be able to judge how surprising a word or phrase is, given its context. This was done using GPT-2 (Radford et al. 2019), detailed in 4.2.3.

4.2.1 Sentence-BERT

Sentence-BERT (Reimers and Gurevych 2019) is a pre-trained language model that was engineered to be proficient at deriving sentence embeddings for use in semantic analysis. Embeddings represent the semantic properties of a piece of text with numeric vectors. Two sentences (or two words) similar in meaning, will have similar embeddings. Calculating the cosine distance between two embeddings will give a value that represents the semantic similarity (Widdows 2004, p. 157). The lower the cosine distance, the more semantically similar the linguistic units are.

4.2.2 VADER

With sentiment analysis the goal is to ascertain which emotions a given piece of natural language evokes, and representing that information. VADER (Hutto and Gilbert 2014) is a rule based model for general sentiment analysis. It can score words or sentences according to how positive or negative their sentiment is. The "compound" score will be used in this study, where the input is given a score between -1 (most negative) and 1 (most positive). For example "very bad" is scored as -0.5849 "decent" as 0 and "very good" as 0.4927. For this project the VADER package from the open source programming suite NLTK (Loper and Bird 2002) was used to analyze sentiment in order to generate more cogent paraphrases.

4.2.3 GPT-2

GPT-2 (Radford et al. 2019) is an artificial intelligence trained on an enormous amount of data to be able to synthesize the next item in an arbitrary information sequence. It has a wide range of possible uses, the most well-known of which is generating text fairly close to something that could've been written by an actual person. In this project, however, GPT-2 will be used to measure the amount of surprisal, or unexpectedness, sentences contain. The less probable a sentence is, the more surprising.

4.3 Logistic Regression Model

Logistic regression is regarded as the baseline supervised machine learning algorithm of natural language processing (Jurafsky and Martin 2020). For this project, a binary logistic regression classifier was trained on a set containing 99 sentences marked as "bad paraphrases" and 99 sentences marked as "good paraphrases" [1]. The "bad paraphrases" were sentences in which the adjective phrase of the original sentence had been replaced with a randomly generated noun phrase. The noun phrase was selected uniformly at random from the set of 1519 frames mentioned in 3.1 and had a 50% probability of containing an adjective, selected uniformly at random from a set containing the 500 most common adjectives in the Brown Corpus (Francis and Kucera 1979). The "good metaphoric paraphrases" were paraphrases of the original sentence written by a native speaker of American English. The purpose of endowing the programs with a logistic regression model is to enable it to make judgments about the probability that a generated sentence is a good metaphoric paraphrase. These judgments are based on the trained model's coefficients for the three feature values mentioned earlier - cosine distance of the Sentence-BERT embeddings, difference in sentiment, and difference in GPT-2 log probability.

4.4 Paraphrase Generation

Two systems for paraphrasing sentences of the form "[noun phrase] is/are/was/were [adjective phrase]", here called *original sentences*, into metaphoric sentences of the form "[target frame] is/are/was/were [source frame]", *metaphoric paraphrases*, were created. "Our culture is coarse and vulgar." is an example of an *original sentence* of which "Our culture is rough experience" is a *metaphoric paraphrase*. This section will go over both systems, starting with the one endowed with MetaNet based knowledge representation.

[1] The logistic regression model used L2 regularization with regularization strength $\lambda = 1$.

4.4.1 The Conceptual Metaphor Paraphraser

This is the most advanced system developed for this project, and uses all the collected knowledge representational data. Given an original sentence from Gigaword (Napoles et al. 2012) of the above form to paraphrase, the program first finds the set of possible source frames that metaphorically illustrate the original sentence's target frame. The set of possible source frames consists of the frames that have a "is-a-source-domain-of" relation to the original sentence's noun phrase. Each of the them then replace the adjective phrase of the original sentence, one at a time. This creates new sentences — *paraphrase candidates*. Each paraphrase candidate is then evaluated on three feature variables:

1. The cosine distance of the Sentence-BERT embeddings of the original sentence (u) and the paraphrase candidate (u').

$$1 - \frac{u \cdot u'}{||u||_2 \cdot ||u'||_2} \tag{1}$$

2. The difference in sentiment of the original sentence (v) and the paraphrase candidate (v'), according to VADER.

$$|v - v'| \tag{2}$$

3. The difference in log probability of the original sentence (s) and the paraphrase candidate (s'), as determined by GPT-2, divided by the number of characters in the original sentence.

$$\frac{logP(s) - \ logP(s')}{|s|} \tag{3}$$

The 10 most probable paraphrase candidates, according to the logistic regression model, are saved. Then, from a set of 500 adjectives, those most similar to the original sentence in sentiment, according to their 'compound' score in VADER (Hutto and Gilbert 2014), are saved as possible adjectives to be included in the final paraphrase. Adjectives already in the original sentence are excluded. [2] These adjectives are now added to the beginning of the source frame of each of the top 10 paraphrase candidates. Then the three parameter evaluation process is repeated, this time with the paraphrase candidates containing adjectives. Finally, the most probable paraphrase, with or without adjective, according to the logistic regression model is selected as the winning candidate.

[2] Narrowing down the adjective selection was done in order to allow for faster execution of the programs, since the VADER and GPT-2 models require a lot of computing power. The maximum allowed difference in VADER 'compound' score was set at .2, which is an arbitrary number.

4.4.2 The Non-conceptual Metaphor Paraphraser

This system is identical to the Conceptual Metaphor Paraphraser, except that the MetaNet knowledge representation has been taken out. So the set of possible source frame candidates for each sentence here contains all 1519 noun phrases, whereas it in the previously explained system only contained those source frames that were conceptual metaphors (or synonyms or related phrases of such) of the original sentence's noun phrase. This version of the program, then, while set to generate metaphoric paraphrases, is not forced to generate them according to CMT.

4.5 Human Evaluation

A panel of 10 annotators was assembled to evaluate the metaphoric paraphrases of 66 literal sentences. The annotators were all competent English users with at least 30 ECTS credits or equivalent in linguistics. Each annotator was tasked with scoring the paraphrases generated by the Conceptual Metaphor Paraphraser, the Non-Conceptual Metaphor Paraphraser and one paraphrase written by a human. Each participant did this for 20 sentences selected uniformly at random from the set of 66 literal sentences. The participants were presented with two questions for each paraphrase:

1. How well does the paraphrase retain the meaning of the original sentence?
   Scored by the participants on a Likert scale as:
   1. Not at all, **2.** Somewhat well, **3.** Fairly well, **4.** Very well.

2. How creative is the paraphrase?
   Scored by the participants on a Likert scale as:
   1. Not at all, **2.** Somewhat creative, **3.** Fairly creative, **4.** Very creative.

In an effort to avoid confusion about what is meant by creativity, the participants were presented with this definition of the notion:

> In this survey, *creative* is defined as "characterized by **expressiveness** and **originality**".
> So, for a metaphoric sentence to be creative, it should carry strong connotations and communicate its meaning vividly, as well as contain some degree of uniqueness or unexpectedness. However, it should not be regarded as creative if it's so unique or unexpected that its meaning becomes unclear.
> A very creative sentence would fit well in a work of fiction, like a novel or a work of poetry, and not fit at all in a legal document or an instruction manual.

The order of the paraphrases was randomized and participants were not informed that some paraphrases had been generated by a program and that some had been written by an actual person.

5 Results

A total of 600 annotations were made by the survey participants. Each paraphrase was scored at least 3 times. Inter-annotator agreements computed using Krippendorff's alpha for meaning retention and creativity were 0.51 and 0.24 respectively.

5.1 Results by Paraphrasing Method

	Meaning retention score mean	Creativity score mean	Paraphrases with a meaning retention score mean of ≥ 2 (Creativity score mean for those)
Conceptual	1.68	1.86	38% (2.18)
Non-Conceptual	2.16	1.75	60% (1.73)
Human-written	2.98	2.45	91% (2.53)

Table 1: *Results of the human evaluation survey*

See Appendix A for the mean meaning retention and creativity scores of every paraphrase.

5.2 Difference Between Systems

Annotating a paraphrase with a score of 2 on meaning retention indicated that the annotator judged that the paraphrase retained the meaning of the original sentence "somewhat well". Thus the paraphrases with a mean meaning retention score of less than 2 were regarded as total failure of paraphrasing. For the Conceptual Metaphor Paraphraser, these failed instances constituted 62% of the paraphrases. For the Non-Conceptual Metaphor Paraphraser, 40% were failures of paraphrasing. A two-tailed Student's t-test run on the meaning retention score means for the systems revealed that the difference was significant ($p < .001$). Both the systems developed for this project performed poorly on the evaluation survey, compared to the human-written paraphrases. In the instances where meaning retention was successful, however, the Conceptual Metaphor Paraphraser yielded the higher mean score for creativity. But this statistic is not significant (*p=.22*).

5.3 Correlation Between Feature Values and Human Evaluation Scores

A linear regression analysis was run on the survey results of the Conceptual Metaphor Paraphraser, the Non-Conceptual Metaphor Paraphraser and the human-written paraphrases combined in order to allow for evaluation of if any of three feature variables - cosine distance of the Sentence-BERT embeddings, difference in sentiment, and difference in GPT-2 log probability, correlated with the paraphrases' meaning retention or creativity scores.

For meaning retention, the Sentence-BERT cosine distance variable was the only statistically significant one ($p < .001$). Its correlation coefficient had a value of 6.06, indicating that for every .1 increase in cosine distance between the Sentence-BERT embeddings of a literal sentence and a system generated or human-written metaphoric paraphrase, the meaning retention score is increased by .61.

For creativity, two variables were statistically significant — the Sentence-BERT cosine distance variable ($p < .001$)) and the difference in GPT-2 log probability ($p < .05$). The correlation coefficient of the Sentence-BERT cosine distance variable was 5.52, indicating that for every .1 increase in cosine distance between the Sentence-BERT embeddings of a literal sentence and a system generated or human-written metaphoric paraphrase, the creativity score is increased by .55. The coefficient of the GPT-2 sentence probability parameter was -.51. So a smaller difference in sentence probability correlates with a lower creativity score.

6 Discussion

This section will contain discussion and evaluation of the data, the methodology, and the results.

6.1 Discussion by Research Questions

Question 1: What is a viable design of a system for paraphrasing adjective phrases into cogent metaphoric noun phrases?

The three feature variables used here can be used with a logistic regression model to generate cogent metaphoric paraphrases, though fairly unreliably. Across the two systems developed for this study, they managed to generate metaphoric paraphrases that retained the meaning of the original sentences at least somewhat well for about half of the instances. That can safely be assumed to be better than chance, but far from as reliably as paraphrases written by a native speaker of English. To draw conclusions about whether this is due to the three feature variables, or something else in the data or methodology will require further discussion, which is to follow in 6.2 and 6.3.

Question 2: How do the three feature variables correlate with the human evaluation scores of the paraphrases' meaning retention and creativity?

The only significant coefficient on meaning retention was surprising — a higher Sentence-BERT cosine distance correlates with a higher meaning retention score. This will be discussed in 6.4.3. The same correlation was found in the creativity scores, where it was not surprising. Just as it was not surprising that sentences with a lower surprisal value were regarded as less creative.

Question 3: How does the implementation of CMT knowledge representation affect the cogency of the generated metaphoric paraphrases of literal sentences, as evaluated by humans?

CMT knowledge representation, as implemented with the ontology of this study, weakens the meaning retention capability of the developed metaphoric paraphrasing system. Further discussion about whether this can be used to argue that CMT knowledge representation is not useful in metaphoric paraphrase generation will follow in 6.2.

6.2 Discussion of Data

Ideally, the dataset would have been much larger and the knowledge representation more advanced. Many MetaNet (Dodge et al. 2015) frames are umbrella concepts, rather than commonly used terms. For example, the Conceptual Metaphor Paraphraser paraphrases "The effect is beautiful and unexpected." into "The effect is an amazing object transfer.", and "The relationship is good, fair and balanced." into "The relationship is a happy bounded region in space.". There is obviously a shortcoming in the data here, as no hyponyms or frame semantic knowledge representations of many of these specific umbrella concepts were collected. There is a sophisticated ontology built into MetaNet, with a lot of additional data and semantic relations between frames, but it was outside the scope of this research project to collect it all. Leveraging the full MetaNet knowledge representation in conjunction with FrameNet (Baker et al. 1998) would allow for a massive knowledge representational ontology, which could help with metaphoric paraphrase generation. Even so, keeping in mind the success of the MERMAID (Chakrabarty, Zhang, et al.

2021) metaphoric paraphrasing system, which used automatic knowledge graph construction (Bosselut et al. 2019), there is evidence of that approach being satisfactory.

English Gigaword (Napoles et al. 2012) was chosen as the source of original sentences to paraphrase because of the need for a large part of speech tagged corpus. The sentences were also shown to be well suited for metaphoric paraphrasing, because the amount of metaphoric adjectives in them was low. This was evaluated as positive, because if there had been a lot of metaphoric adjectives, paraphrasing would have been from metaphor to metaphor, rather than literal to metaphor, which is what was aimed for. Out of the 66 original sentences, just two contained only metaphoric adjectives: "Our marriage was strong." and "The market is healthy.". Ideally, though, these two sentences would have been replaced while collecting data.

The person who wrote the paraphrases was without academic credentials in creative writing or literature, which they would ideally have had. However, this was not viewed as a major flaw of the project, because the purpose of the human written paraphrases was to compare them with those generated by the programs. The comparison made clear that the systems couldn't match a layperson in metaphoric paraphrase generation, so it can safely be assumed that they wouldn't be able to match an expert either.

6.3 Discussion of Methodology

Because the Conceptual Metaphor Paraphraser and the Non-Conceptual Metaphor Paraphraser were run on a regular desktop CPU, rather than a GPU server, some compromises had to be made. Ideally, all possible combinations of adjectives and source frames would have been analyzed and ranked in finding the best paraphrase, but because of the limited computing power, that was not possible.

The participants of the human evaluation survey were asked two questions about each paraphrase: "How well does the paraphrase retain the meaning of the original sentence?" and "how creative is the paraphrase?". Earlier research (Stowe, Ribeiro, et al. 2020; Chakrabarty, Zhang, et al. 2021) asked human evaluation participants to score the "fluency" - the syntactic and grammatical viability of the generated paraphrases. For this study, that feature variable was regarded as unnecessary. The syntax of the original sentence and its paraphrase guaranteed syntactic viability.

Following Stowe, Ribeiro, et al. (2020), one factor that would have been have been interesting to evaluate in the survey is the question of the relevance of the source frame to the target frame in the metaphoric paraphrase. This feature variable was however discarded for the final version of the survey. That was due to the fact that it would have increased the time needed to complete the survey by presenting the participants with the required explanations and definitions of Metaphors generally and CMT specifically, as well as giving them extra scoring work to do. Since the participants were all volunteers, the aim was to design the survey so that it could be finished in 25 minutes.

That was also the reason for not including a baseline system that made random paraphrases. The output from a system for randomizing metaphoric paraphrases was used to train the logistic regression model from section 4.3, but those paraphrases were not included for scoring in the final survey.

6.4 Discussion of Results

6.4.1 Comparison with other studies

Chakrabarty, Zhang, et al. (2021) asked participants of the human evaluation task to rate paraphrases generated by their system and four other baselines, including the Lexical Replacement and Metaphor Masking systems of Stowe, Ribeiro, et al. (2020), on four factors, including meaning retention and creativity. Their annotators scored on a five point scale. Here those results will be converted by multiplication of .8 to fit with the four point scale used in this study before discussion. With scores converted, the MERMAID system got a mean score of 2.68 on meaning retention, and 2.8 on creativity, which was very close to the scores of one of the experts they had contracted to write paraphrases. This system was far more successful than the Conceptual Metaphor Paraphraser and the Non-Conceptual Metaphor Paraphraser, and its paraphrases actually scored higher on creativity than those written by this study's layman contractor.

Chakrabarty, Zhang, et al. (2021) evaluation of Stowe, Ribeiro, et al. (2020) Lexical replacement system got an average score for meaning retention of 2.07, which is better than the Conceptual Metaphor Paraphraser, but worse than the score for the Non-Conceptual Metaphor Paraphraser. Its creativity score of 1.73 is similar enough to those of the systems developed here that nothing conclusive can be said about it.

Conversion of the scores from Chakrabarty, Zhang, et al. (2021) evaluation on a 5-point scale, to that of the 4-point scale used in this study on the Metaphor Masking system of Stowe, Ribeiro, et al. (2020), reveals that both of the systems developed for this study outperformed the Metaphor Masking system on meaning retention and creativity.

It should be kept in mind, however, that the projects by Chakrabarty, Zhang, et al. (2021) and Stowe, Ribeiro, et al. (2020) focused on paraphrasing verbs, while paraphrasing of adjective phrases into metaphoric noun phrases was the task here, so a straight comparison of human evaluation scores can't safely be assumed to be more than an approximation of the different systems' general paraphrasing ability. The seq2seq based MERMAID model should in practice be able to be retrained for paraphrasing adjective phrases into metaphoric noun phrases, which would be a good direction for future studies. A more general metaphoric paraphrasing system would be of much more practical use than the part of speech limited ones that have been developed so far.

6.4.2 Discussion of Inter-annotator Agreement

The Krippendorff's alpha coefficient of meaning retention was 0.51, which indicates a slightly higher level of inter-annotator agreement than in the studies of Chakrabarty, Muresan, et al. (2020) and Chakrabarty, Zhang, et al. (2021). The fact that all survey participants in this study had some knowledge of linguistics might have been a contributing factor to the relatively high agreement.

The Krippendorff's alpha coefficient of creativity, however, was 0.23 in this study, which is significantly lower. Chakrabarty, Muresan, et al. (2020) and Chakrabarty, Zhang, et al. (2021) did not explicitly define the notion of creativity as was done in this study, which indicates that the attempt to do so here was futile or even counter-productive.

6.4.3 Discussion of Correlation Between Feature Values and Human Evaluation Scores

Analysis of the linear regression results made clear that increase in perceived creativity correlated with a lower GPT-2 probability score, which was expected. Less probable means more unexpected, which in turn means more creative. That a higher sentence embedding cosine distance was correlated with increased perceived creativity, was not surprising either. Having two phrases express the same thing while being very different in literal meaning is per definition a creative use of metaphor.

What was surprising, however, was that a higher cosine distance between the sentence embeddings correlates with a higher meaning retention score. The point of the Sentence-BERT model (Reimers and Gurevych 2019) is to evaluate how semantically similar sentences are, so the higher the cosine distance, the less similarity should be between them in terms of meaning. An explanation of this peculiarity was found by calculating the average of the cosine distance of the paraphrases generated by the Conceptual Metaphor Paraphraser, the Non-Conceptual Metaphor Paraphraser and the human written paraphrases.

The mean cosine distance between the sentence embeddings of the original sentence and the human written paraphrases actually was the highest, with the Conceptual Metaphor Paraphraser having a slightly lower mean and the Non-Conceptual Metaphor Paraphraser a lower still. Further investigation revealed that this might have been caused by nominalizations that the developed systems, in particular the Non-Conceptual Metaphor Paraphraser, tended to generate. Examples of nominalizing paraphrases that it generated are "The face is black." into "The face is a black.", "Their host is blind and deaf." into "Their host is a blindness." and "Our marriage was strong." into "Our marriage was a strength.". The cosine distance between the sentence embeddings of these sentences were all very low. While there is obviously semantic similarity in these examples, they just aren't good paraphrases. This highlights a possible shortcoming in using Sentence-BERT for automatic evaluation of metaphoric paraphrases. The surprising correlation of a higher cosine distance between the sentence embeddings and a higher meaning retention score was found for each paraphrasing method, when analyzed individually by linear regression.

7 Conclusions

Question 1: What is a viable design of a system for paraphrasing adjective phrases into cogent metaphoric noun phrases?

The reliability of the systems developed for this project at generating metaphoric paraphrases is far from that of a human. The less restrained system, called the Non-Conceptual Metaphor Paraphraser, did however manage to generate somewhat meaning retaining metaphoric paraphrases for more than half of the instances. Keeping in mind recent research in the area and their performance compared to that of the Conceptual Metaphor Paraphraser and the Non-Conceptual Metaphor Paraphraser, it can safely be inferred that a basic machine learning algorithm such as a logistic regression model is somewhat functional, but not the optimal way to go about the task at hand.

Question 2: How do the three feature variables correlate with the human evaluation scores of the paraphrases' meaning retention and creativity?

The statistically significant correlations this study has found are the following:

1. Higher cosine distance between sentence-BERT embeddings of literal sentences and metaphoric paraphrases of them correlates with a higher level of perceived meaning retention and a lower perceived creativity of the metaphoric paraphrase.
2. Less difference in GPT-2 log probability between literal sentences and metaphoric paraphrases of them correlates with humans evaluating the paraphrases as less creative.

Question 3: How does the implementation of CMT knowledge representation affect the cogency of the generated metaphoric paraphrases of literal sentences, as evaluated by humans?

In the systems for metaphoric paraphrase generation developed as part of this project, implementation of a shallow CMT-based knowledge representational ontology weakened the system's ability to generate cogent metaphoric paraphrases, as evaluated by humans. From this it can not be concluded that CMT knowledge representation has no place in future projects with a similar task. A very recent study by Chakrabarty, Zhang, et al. 2021 utilized symbolic knowledge representation to develop a model for metaphoric paraphrasing that yielded results close to that of a human expert. Since symbolism is a central notion in CMT this indicates that the problem with the systems of this study was not CMT knowledge representation as such, but the specific implementation. Future research could look into using more of the knowledge representational data available on MetaNet, and a larger repository of synonyms, hyponyms and related terms of the frames.

7.1 Summary

This study showed that using neural networks for semantic analysis in conjunction with a logistic regression model and CMT knowledge representation is a working, but unreliable way to generate metaphoric paraphrases. The results also show some indication that Sentence-BERT (Reimers and Gurevych 2019) is unreliable in representing semantic similarity between literal sentences and metaphoric paraphrases of them. A sentence and its metaphoric paraphrase being less similar in their Sentence-BERT embeddings does however correlate with humans evaluating the paraphrase as more creative. So does a sentence and its metaphoric paraphrase having higher difference in probability according to GPT-2 (Radford et al.

2019). For this project the source frames and target frames of MetaNet (Dodge et al. 2015) were collected, along with information about which of the other frames each frame was a target or source of, and synonyms and hyponyms of all frames. The failure of the systems to reliably generate cogent metaphoric paraphrases and their tendency to generate nonsensical paraphrases, indicate that the body of knowledge representational information just explained was unsatisfactory for the task.

7.2 Future Research

A future direction to study is development of more general systems that can do metaphoric paraphrasing on any sentence. Following the success Chakrabarty, Zhang, et al. (2021) had with automatic knowledge base construction, this is an area that should be explored further. Using the entire knowledge representation on offer by MetaNet (Dodge et al. 2015) is also an interesting future direction for study, especially if one is trying to achieve a more general metaphoric paraphrasing system. The knowledge contained in MetaNet could be very useful for a creative writers' assistant, because it contains information about how we usually think, talk, and write about things. In writing a work of fiction, this information could prove useful in assisting the writer in finding a more commonplace metaphor, or a more fantastical one. An issue in the field is the lack of datasets containing literal sentences and metaphoric counterparts of them. Creation of such datasets, be it by manual or automatic means would greatly help the NLP research area of Metaphoric paraphrase generation.

References

Aristotle (ca. 335 B.C.E./2017). *Poetics*. Trans. by Ingram Bywater. DIGIREADS.COM. ISBN: 9781420956450. URL: https://books.google.se/books?id=llxdswEACAAJ.

Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). The Berkeley FrameNet Project. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 86–90.

Bisk, Yonatan, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. (2020). Piqa: Reasoning about physical commonsense in natural language. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 7432–7439. DOI: https://doi.org/10.1609/aaai.v34i05.6239.

Bosselut, Antoine, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi (2019). *COMET: Commonsense Transformers for Automatic Knowledge Graph Construction*. arXiv: 1906.05317 [cs.CL].

Carter, Ronald, Angela Goddard, Maggie Bowring, Danuta Reah, and Keith Sanger (2001). *Working with texts: a core introduction to language analysis*. Psychology Press.

Chakrabarty, Tuhin, Smaranda Muresan, and Nanyun Peng (2020). Generating similes effortlessly like a Pro: A Style Transfer Approach for Simile Generation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Online: Association for Computational Linguistics, pp. 6455–6469. DOI: 10.18653/v1/2020.emnlp-main.524.

Chakrabarty, Tuhin, Xurui Zhang, Smaranda Muresan, and Nanyun Peng (2021). MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding. In: arXiv: 2103.06779.

Colton, Simon and Geraint A Wiggins (2012). Computational Creativity: The Final Frontier? In: *ECAI*.

Deignan, Alice (2005). *Metaphor and corpus linguistics*. Vol. 6. John Benjamins Publishing.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dodge, Ellen, Jisup Hong, and Elise Stickles (2015). MetaNet: Deep semantic automatic metaphor analysis. In: *Proceedings of the Third Workshop on Metaphor in NLP*. Denver, Colorado: Association for Computational Linguistics, pp. 40–49. DOI: 10.3115/v1/W15-1405.

Fellbaum, Christiane (2010). WordNet. In: *Theory and Applications of Ontology: Computer Applications*. Ed. by Roberto Poli, Michael Healy, and Achilles Kameas. Dordrecht: Springer Netherlands, pp. 231–243. ISBN: 978-90-481-8847-5. DOI: 10.1007/978-90-481-8847-5_10.

Francis, W Nelson and Henry Kucera (1979). Brown corpus manual. In: *Letters to the Editor* 5.2, p. 7.

Gruber, Tom (2007). Ontologies. In: *Encyclopedia of Database Systems*. Ed. by Ling Liu, Tom, and M. Tamer Özsu.

Hutto, Clayton and Eric Gilbert (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 1. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Jacobs, Arthur (2018). The Gutenberg English Poetry Corpus: Exemplary Quantitative Narrative Analyses. In: *Frontiers in Digital Humanities* 5, p. 5. DOI: 10.3389/fdigh.2018.00005.

Jang, Hyeju, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rose (2016). Metaphor detection with topic transition, emotion and cognition in context. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 216–225.

Jurafsky, Daniel and James H Martin (2020). Speech and language processing (draft). In: *Chapter 5: Logistic Regression (Draft of December 30, 2020)*. URL: https://web.stanford.edu/~jurafsky/slp3/5.pdf.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, pp. 67–72. URL: https://www.aclweb.org/anthology/P17-4012.

Lakoff, George and Mark Johnson (1980). *Metaphors we live by*. University of Chicago Press.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: arXiv: 1910.13461 [cs.CL].

Loper, Edward and Steven Bird (2002). NLTK: The Natural Language Toolkit. In: arXiv: cs/0205028.

Mu, Jesse, Helen Yannakoudakis, and Ekaterina Shutova (2019). Learning outside the box: Discourse-level features improve metaphor identification. In: arXiv: 1904.02246 [cs.CL].

Napoles, Courtney, Matthew R Gormley, and Benjamin Van Durme (2012). Annotated gigaword. In: *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pp. 95–100.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). Language models are unsupervised multitask learners. In: *OpenAI blog* 1.8, p. 9.

Reimers, Nils and Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: arXiv: 1908.10084.

Steen, Gerard (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. Vol. 14. John Benjamins Publishing.

Stowe, Kevin, Sarah Moeller, Laura Michaelis, and Martha Palmer (2019). Linguistic Analysis Improves Neural Metaphor Detection. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, pp. 362–371. DOI: 10.18653/v1/K19-1034.

Stowe, Kevin, Leonardo Ribeiro, and Iryna Gurevych (2020). Metaphoric Paraphrase Generation. In: arXiv: 2002.12854.

Synonyms.com (2021). https://www.synonyms.com. (Accessed on 03/12/2021).

Terai, Asuka and Masanori Nakagawa (2009). A neural network model of metaphor generation with dynamic interaction. In: *International Conference on Artificial Neural Networks*. Springer, pp. 779–788.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In: arXiv: 1706.03762.

Veale, Tony (2013). Less Rhyme, More Reason: Knowledge-based Poetry Generation with Feeling, Insight and Wit. In: *ICCC*, pp. 152–159. URL: http://www.computationalcreativity.net/iccc2013/download/iccc2013-veale-2.pdf.

Veale, Tony and Rafael Pérez y Pérez (Nov. 1, 2020). Leaps and Bounds: An Introduction to the Field of Computational Creativity. In: *New Generation Computing* 38.4, pp. 551–563. ISSN: 1882-7055. DOI: 10.1007/s00354-020-00116-w.

Veale, Tony, Ekaterina Shutova, and Beata Beigman Klebanov (2016). *Metaphor: a computational perspective*. Synthesis lectures on human language technologies 31. OCLC: 944304883. San Rafael: Morgan & Claypool. ISBN: 978-1-62705-850-6.

Widdows, Dominic (2004). *Geometry and Meaning*. Vol. 172. CSLI lecture notes series. CSLI Publications. ISBN: 978-1-57586-448-8.

Yu, Zhiwei and Xiaojun Wan (2019). How to Avoid Sentences Spelling Boring? Towards a Neural Approach to Unsupervised Metaphor 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, Minnesota: Association for Computational Linguistics, pp. 861–871. DOI: 10.18653/v1/N19-1092.

Appendix A Human Evaluation Survey Results

MR is the meaning retention score mean, **Cre** is the creativity score mean.

Original	**Human**	MR	Cre	**Conceptual**	MR	Cre	**Non-Conc**	MR	Cre
The effect is beautiful and unexpected.	The effect is a pleasing surprise.	3.8	2.5	The effect is an amazing object transfer.	1.2	1.0	The effect is an amazing quality.	2.8	2.0
The decision is ill-conceived and lamentable.	The decision is a regrettable outcome.	2.9	1.6	The decision is an ill acceptation.	2.1	2.2	The decision is a loss.	2.5	2.1
The land is poor.	The land is an impoverished parcel.	2.7	2.8	The land is an abomination.	1.5	2.0	The land is a poverty.	2.3	1.3
Our culture is coarse and vulgar.	Our culture is a rough experience.	2.0	2.0	Our culture is an abomination.	2.0	2.0	Our culture is a strange thing.	1.0	1.0
The decision was mutual and amicable.	The decision was an agreeable affair .	3.0	2.0	The decision was an acceptance.	2.0	1.0	The decision was an understanding.	3.0	1.0
The group is standard and essential.	The group is a bland diet.	2.0	3.0	The group is a necessary body.	4.0	2.0	The group is a nonstandard.	1.0	1.0
His information was accurate.	His information was a sure thing.	2.7	1.7	His information was a correct contents.	1.3	1.0	His information was a right.	1.7	1.0
Our marriage was strong.	Our marriage was a rock.	3.8	3.0	Our marriage was a happy knit.	2.5	3.2	Our marriage was a strength.	2.8	1.8
This activity is deliberate, illegal and unwarranted.	This activity is an illicit conduct.	4.0	2.5	This activity is a wrong thing.	3.0	2.0	This activity is an evil.	2.2	2.5
The rule is unfair.	The rule is an injustice.	4.0	1.0	The rule is a position.	1.0	1.0	The rule is a terrible thing.	1.0	2.0
His rival is telegenic, loquacious and charismatic.	His rival is an excellent showman.	3.5	2.2	His rival is a big go.	1.8	1.8	His rival is a close ally.	1.0	1.2
The effect is agreeable.	The effect is a warm handshake.	2.3	3.0	The effect is an object transfer.	1.0	1.0	The effect is a good thing.	3.3	1.0

The vote was free and fair.	The vote was an upright procedure.	3.5	2.8	The vote was a go.	1.8	1.5	The vote was a free hand.	1.8	1.5
Her life is casual, expensive and spare.	Her life is a carefree luxury.	3.0	3.0	Her life is a simple way.	2.0	1.0	Her life is a simple living.	2.5	1.5
The relationship is good, fair and balanced.	The relationship is a state of parity.	3.0	2.3	The relationship is a happy bounded region in space.	1.7	2.3	The relationship is a perfect deal.	3.3	2.0
This conclusion is irrefutable and undeniable.	The conclusion is an indubitable fact.	4.0	2.0	This conclusion is a false acceptance.	1.0	1.5	This conclusion is a complete disputation.	1.8	2.2
The community was rural, marginalized and illiterate.	The community was an agrarian backwater.	3.6	3.6	The community was a local household.	1.2	1.8	The community was a different population.	1.6	1.8
The work is complex, painstaking and slow.	The work is a gruelling venture.	4.0	3.0	The work is an expensive occupation.	2.0	2.5	The work is a long effort.	3.5	2.0
Our performance was depressing and humiliating.	Our performance was a total embarrassment.	4.0	1.5	Our performance was a sad sign.	2.0	2.5	Our performance was a sad loss.	3.0	2.5
The market was small and insignificant.	The market was an insubstantial pittance.	2.5	3.5	The market was a no star.	2.5	3.5	The market was a low quantity.	2.0	1.0
The choice is obvious.	The choice is a safe candidate.	3.5	3.5	The choice is a real question.	1.0	2.0	The choice is a decision.	1.0	2.0
A beatification is inopportune and premature.	A beatification is a rash inconvenience.	2.8	1.8	A beatification is a necessary sign.	1.0	1.0	A beatification is a thing.	1.2	1.2
The group was clear, detailed and stable.	The group was a unified whole.	4.0	3.0	The group was a pretty complex.	1.0	1.0	The group was a good organization.	4.0	1.0

The living was hard, brutal and dangerous.	The living was a savage reality.	3.7	3.0	The living was a journey.	1.7	2.0	The living was a life.	1.0	1.7
The population was impoverished and demoralized.	The population was a stray dog.	1.8	2.5	The population was a poor heart.	2.2	2.8	The population was a loss.	1.5	1.5
A deal is close.	A deal is a bird in the hand.	1.0	1.5	A deal is a holding.	1.0	1.0	A deal is a conclusion.	1.0	1.5
The change is large, rapid and inexorable.	The change is a substantial unavoidability.	2.0	2.0	The change is a slow motion.	1.0	1.0	The change is a real thing.	3.0	2.0
A job is tough.	A job is a hardship .	3.0	1.8	A job is a construction.	1.2	1.2	A job is a challenge.	3.2	1.5
The voting was fair and balanced.	The voting is a just preceding.	2.7	2.0	The voting was a blessing.	1.3	2.7	The voting was a winner.	3.0	1.7
The decision was obvious.	The decision is a walk in the park.	3.0	3.0	The decision was an acceptance.	1.5	1.5	The decision was a taking.	2.5	1.0
Its spine was loose and soft.	Its spine was a floppy arc.	3.7	3.3	Its spine was a container for emotions.	1.0	2.0	Its spine was a strain.	1.0	1.0
The election was transparent, democratic and clean.	The election was an equitable process.	3.7	1.3	The election was a clear go.	2.0	1.3	The election was a winner.	1.7	2.7
The conflict is inevitable.	The conflict is an inescapable consequence.	4.0	2.2	The conflict is a war.	1.0	1.4	The conflict is a real thing.	2.4	2.2
The information was sparse and inadequate.	The information was a deficiency.	2.3	2.3	The information was a poor contents.	1.7	1.3	The information was a poor resource.	3.3	2.3
That opposition is marginal.	That opposition is a footnote.	2.0	3.5	That opposition is a minor question.	3.5	2.0	That opposition is a thing.	1.5	2.0

The offense is ineffective and dull.	The offense is a fruitless undertaking.	3.3	2.3	The offense is a boring game.	2.0	2.0	The offense is a loss.	1.7	1.7
The work is dangerous, demeaning and exhausting.	The work is a perilous journey.	2.5	3.5	The work is a nasty occupation.	3.0	2.0	The work is a strain.	2.0	1.0
The service was erratic.	The service was a disarray.	3.0	1.8	The service was an unusual unit.	1.5	1.7	The service was a challenge.	2.0	2.0
The discussion was collegial and constructive.	The discussion was a productive collaboration.	2.0	1.0	The discussion was a classic one-on-one physical fighting.	1.0	2.0	The discussion was a civility.	2.0	1.0
Her government was corrupt, inept and undemocratic.	Her government was an dishonourable oligarchy .	4.0	3.0	Her government was a mediocrity.	1.0	1.5	Her government was a corruptness.	3.0	2.0
The work was methodical and productive.	The work was a well thought out success.	3.0	1.5	The work was a smooth occupation.	2.2	2.5	The work was a consistent effort.	2.5	1.0
His subject was varied, humorous and brilliant.	His subject is a diverse whimsicality.	1.0	1.0	His subject was a wonderful eyes and ears.	1.0	2.0	His subject was a wonderful thing.	3.0	2.0
The face is black.	The face is a nightly hue.	3.8	3.8	The face is a container for emotions.	1.0	2.0	The face is a black.	1.2	2.0
The vote was legitimate, free and democratic.	The vote was justifiable endeavour.	3.0	2.2	The vote was a victory.	2.0	1.5	The vote was a free hand.	1.5	2.2
|>———->->–>| The house is tiny, cramped and worn.	The house is a paltry dishevelment.	2.0	3.0	The house is a home.	1.0	1.0	The house is a thing.	1.0	1.0
The experience is easy, seamless and integrated.	The experience is a pleasurable whole.	2.7	2.7	The experience is a perfect commute.	2.7	4.0	The experience is a clean functionality.	2.7	2.3

The group was pooped and proud.	The group was a ragged mutt.	2.0	3.0	The group was a happy face.	1.0	1.0	The group was a triumph.	1.5	1.0
This land was barren, dangerous and useless.	This land was a wasteland.	3.8	2.8	This land was a ruin.	3.2	2.2	This land was a dead world.	3.2	3.2
The country is safe, stable and peaceful.	The country is a prosperous story.	2.5	2.5	The country is a good person.	2.0	2.0	The country is a nation.	2.0	1.0
The election was free and fair.	The election was a democratic victory.	4.0	1.5	The election was a victory.	1.0	1.0	The election was a right.	2.0	3.0
Its execution was illegal.	Its execution was a travesty.	1.7	2.0	Its execution was a sign.	1.0	2.0	Its execution was a crime.	4.0	1.3
The information is relevant.	The information is an apropos reference .	2.5	3.0	The information is historical contents.	1.0	1.5	The information is a reference.	3.0	2.0
The market is healthy.	The market is a robust performer.	1.5	2.2	The market is a good man.	2.0	2.2	The market is a good thing.	3.5	1.8
The think tank was controversial.	The think tank was a disputed source.	3.3	3.3	The think tank was a family.	1.0	2.3	The think tank was a controversy.	3.0	1.0
His knowledge is profound and deep.	His knowledge was a vast well.	3.8	4.0	His knowledge is a vast level.	2.0	1.5	His knowledge is a richness.	2.8	2.8
The election was clean and fair.	The election was an amicability.	1.5	1.5	The election was a go.	1.5	1.5	The election was a winner.	1.0	1.5
The subject is young, female and white.	The subject is a caucasian woman.	3.7	1.3	The subject is an all eyes and ears.	1.0	2.0	The subject is a people.	1.7	1.7
The information was proprietary.	The information was a exclusive property.	3.4	2.0	The information was a sensitive contents.	1.8	1.9	The information was a monopoly.	2.0	2.2
The market is fragile and frightened.	The market is a tense tinderbox.	3.0	4.0	The market is a scared man.	3.0	2.5	The market is a strain.	2.0	2.0

Their effect is regional, spotty and short-lived.	Their effect is a inconsequential dalliance.	1.7	2.0	Their effect is a cultural object transfer.	1.0	1.0	Their effect is a narrow influence.	2.3	2.0
The election was fair and square.	The election was a great accomplishment.	1.3	2.0	The election was a victory.	2.0	2.0	The election was a winner.	1.3	1.7
The election is open, free and fair.	The election was a nonpartisan landslide.	1.3	1.3	The election is a race.	1.0	1.3	The election is a free hand.	1.7	1.3
My heart is secure and content.	My heart is a stable landmark.	2.7	3.0	My heart is a warmth.	1.3	1.7	My heart is a joy.	1.3	1.0
The effect is poisonous.	The effect is a great detriment.	2.0	2.0	The effect is an object transfer.	1.0	1.0	The effect is a nasty thing.	2.0	1.0
Their host is blind and deaf.	Their host is an imperceptive mole.	2.7	4.0	Their host is a machine.	1.0	2.3	Their host is a blindness.	1.3	1.7
The economy was different.	The economy was a different picture.	3.0	2.9	The economy was a big shift.	2.1	2.0	The economy was a change.	2.1	1.4

Stockholm University
SE-106 91 Stockholm, Sweden
Telephone +46 (0)8 16 20 00
https://www.su.se/