The Delta of Thought: Channeling Rivers of Commonsense Knowledge in the Sea of Metaphorical Interpretations

Antonio Lieto[1] , **Gian Luca Pozzato**[2] , **Stefano Zoia**[2]

[1]University of Salerno, Cognition Interaction and Intelligent Technologies Lab, DISPC, Italy
[2]University of Turin, Department of Computer Science, Italy
alieto@unisa.it, {name.surname}@unito.it

Abstract

We propose a system called **MET**$^{\mathsf{CL}}$ (**M**etaphor **E**laboration in **T**ypicality-Based **C**ompositional **L**ogic) able to generate and identify metaphors by using the $\mathbf{T}^{\mathsf{CL}}$ reasoning framework, specialized in human-like commonsense concept combination. We show that **MET**$^{\mathsf{CL}}$ is able to improve both state-of-the-art Large Language Models (e.g DeepSeek-R1, GPT-4o, Qwen2.5-Max) and symbolic ones in the task of metaphor identification. Additionally, we show how the metaphors generated by **MET**$^{\mathsf{CL}}$ are generally well accepted by human subjects. The obtained results are encouraging and pave the way to research in automatic metaphor generation and comprehension based on the assumption that metaphors interpretation can be partially regarded as a categorization problem relying on generative commonsense concept combination.

1 Introduction

Metaphors are known to be pervasive in human language and thought. Given their mutability and the creativity underlying their production, their automatic processing is a challenging Artificial Intelligence (AI) problem. The study of metaphors, on the other hand, enjoys a cross-fertilization where AI serves as a test bed for hypotheses about the functioning of biological minds, while the advancements in Cognitive Science pave the way for the emerging of new ideas and applications in AI [Lieto, 2021]. In this work, we first review the main theories of metaphor processing in cognitive science as well as the main tasks and models developed in the AI literature. We then introduce the system **MET**$^{\mathsf{CL}}$, which exploits the $\mathbf{T}^{\mathsf{CL}}$ logic in its reasoning engine. $\mathbf{T}^{\mathsf{CL}}$ is able to deal with the problem of commonsense compositionality and, in the present work, it has been applied to generate a high level semantic representation of a metaphor's meaning by using its, human-like, concept combination strategy. Last, we present the twofold evaluation conducted to test the system and consisting in (i) an automatic evaluation of the classification capabilities of metaphors of our system compared to other state-of-the-art neural models and symbolic resources and in (ii) a human evaluation of the generated metaphors.

1.1 Theories of Metaphors

The most successful metaphor theories in Cognitive Science are known as *Analogy*, *Categorization*, *Conceptual Mapping* [Holyoak and Stamenković, 2018; Carston, 2012; Wilson, 2011]. Each approach takes on different perspectives, but some fundamental assumptions are common. Metaphors always involve two elements: the first, called *target* (or *topic*), identifies what is talked about, the second (*source* or *vehicle*) is the concept used to characterize the target.

The Analogy model states that metaphors arise from similarities between concepts belonging to different domains. Analogical reasoning allows one to identify similarities based on relations among entities, rather than solely on the entities themselves. In computational implementations of such theory [Gentner and Forbus, 2011], the reference domain is represented as a graph where each node is a concept and the edges are the relationships between them. The *structure mapping* extension [Gentner, 1983] postulates that metaphor comprehension can be implemented as an *alignment* process that finds "maximal consistent subgraphs within the source and target that yield a one-to-one (isomorphic) mapping between one another".

The Categorization approach sees metaphors as category statements. The underlying idea is that the source can assume either a concrete, literal, meaning (e.g. "furnace" = an enclosed structure in which material can be heated) or a more abstract and categorical one in metaphorical expressions (e.g. "furnace" = a hot place, as in "the streets were a furnace"). The target, on the other hand, is a prototypical member of that category. The main computational process to extract an abstract meaning from the source (and consequently the meaning of the metaphor) is *conceptual combination* between source and target [Holyoak and Stamenković, 2018]. For example, in a metaphorical expression like "My lawyer is a shark", the metaphorical meaning of the expression is intended as a combination of its constituents "shark" (source) and "lawyer" (target), intended, in such metaphorical declination, as a prototype of the figurative meaning of the source "shark". From a computational perspective, this type of combination is less demanding than structure mapping, since it operates on the representation of the source and target concepts alone (not on their relations, if any).

Conceptual Mapping (also called Conceptual Metaphor Theory) affirms that in metaphorical expressions a source do-

main (generally concrete) is mapped on a target domain (generally more abstract). Such conceptual mappings (or "conceptual metaphors") include, for example, LIFE IS A JOURNEY and POVERTY IS A DISEASE. In this view, the conceptual mappings are stated in the form of category statements (like in the categorization approach), though they are typically interpreted as mappings (following the analogy approach). The comprehension of a metaphor, according to this theory, is carried out as a form of constrained analogical reasoning: the links between the two domains are retrieved from knowledge structures instead of being calculated from scratch.

Due to the huge variability of the metaphors phenomenon, there is *de facto* no general theory able to account for all the aspects of metaphorical processing. As we will describe in the following sections, our work naturally resonates and falls within the Categorization paradigm, as we propose to deal with metaphorical meanings through concept combination. As we will show, such an account proves to be a relevant complement to the current state of the art systems and resources developed for automatic metaphor processing.

1.2 Metaphor Processing: tasks and models

Many efforts have been dedicated to the development of systems able to detect, understand or generate metaphors. One aspect of metaphorical expressions that makes them particularly tricky to elaborate automatically is their formal variability. As Holyoak and Stamenković [2018] put it, simpler syntactic forms include nominal metaphors ("The stock is a rollercoaster", where the focus is on a noun), predicate metaphors ("The flower purred in the sunshine", based on a verb), and attributive metaphors ("The weary mountain", based on an adjective) and, overall, there is neither a fixed syntactic structure for metaphors nor a unique pattern for the parts of speech it can involve.

Metaphor processing includes three main computational tasks [Ge *et al.*, 2023]: metaphor identification, metaphor interpretation and metaphor generation. Different approaches have been developed to perform each task. Because of the high variability in metaphor, the majority of metaphor processing techniques presented in the literature are dedicated to a small subset of cases. In particular, nominal metaphors and predicate metaphors are often the target of these works, given their high frequency in language, simple syntactic form and high data availability. In the next section we review some of the most prominent works on metaphor processing and of the developed datasets.

Metaphor identification task This task consists in the recognition of an input as a metaphorical or a literal expression. The input can be a whole sentence, a token or a word couple. Performing the task requires somehow pointing out what features define a metaphor, focusing on the ontological differences between literal and metaphorical expressions. As noted by Tsvetkov *et al.* [2014], due to their formal variability, distinguishing between metaphorical and literal expressions can sometimes be hard even for humans, and manual annotation of data may depend on a subjective component. Many hypotheses have been tested, relying either on statistical learning or linguistic insights.

Wan *et al.* [2020] adopted an interesting strategy using modality norms together with word embeddings as the input for a neural network system. Modality norms express for each word a measure of strength for six sensorimotor modalities: auditory, gustatory, haptic, visual, olfactory and interoceptive. The hypothesis behind their application of modality norms is that "metaphor manifests a concept mismatch (modality shift in particular) between source and target". The resulting system outperformed several deep learning baselines, corroborating this hypothesis. Another notable attempt to draw knowledge from different modalities was made by Shutova *et al.* [2016], who developed a metaphor identification method based on both text and images.

Metaphor interpretation task Metaphor interpretation consists in the extraction of a metaphor's meaning. Based on their output type, metaphor interpretation systems are classified into three categories [Ge *et al.*, 2023]:

1. Property extraction systems: output the common features of source and target.
2. Word-level paraphrasing systems: replace each metaphorical word with a correspondent literal term.
3. Explanation pairing systems: provide a full explanation, similar to a dictionary definition.

The task can be understood as generating a representation of the ground of a metaphor, given the source and the target (or the whole metaphorical sentence). This is almost literally the task definition used by Song *et al.* [2021], who tackled the problem as a graph completion. Their system's goal is to generate triplets of the form (source, attribute, target), where the source and the target are given in input and the attribute is automatically extracted from a set of candidates to represent a common property of source and target concepts.

Rai *et al.* [2019] proposed an emotion-driven approach to metaphor interpretation arguing that metaphors are better understood when charged with the extraction of affective labels attached to the words, rather than solely identifying similarity between source and target. Mao *et al.* [2022] tested a metaphor interpretation system as preprocessing on a sentiment analysis task, finding an improvement in the sentiment analysis classifier performances.

Recently, Large Language Models (LLMs) have been leveraged for the metaphor interpretation task. In particular, Ichien *et al.* [2023] applied GPT-4 to novel literary metaphors and evaluated the output against the interpretations provided by a group of college students. Human judges, blind to the involvement of an AI model, rated the automatically produced interpretations as superior to the ones provided by humans.

Metaphor generation task This task concerns the automatic production of metaphors. Metaphor generation systems are classified as [Ge *et al.*, 2023]:

1. Verb substitution systems: replace a literal verb of the sentence with a metaphorical one.
2. Metaphorical expression surface realization systems, or MESRs: can produce one or more words to complete a metaphor. For example: given source and target, some MESRs can generate a list of properties that link them.

3. Sentence generation systems: given the target, can generate a whole metaphorical sentence

Due to the relatively simple definition of the task, the approach to verb substitution tends to be always the same, although the applied technologies can change: first, the model is prompted to perform the verb substitution, and then it is checked that the result is metaphorical. As an example, Chakrabarty *et al.* [2021] fine-tuned a sequence-to-sequence model and tested it as a tool to enhance human-written poems, by endowing it with metaphor generation utilities.

Zheng *et al.* [2020] presented a distance-based MESR system designed to be incorporated into a chatbot. Their first step is to identify a set of target candidates and a set of source candidates, based on concept frequency in human-computer conversations and on a concreteness score. As a second step, they located targets and sources in a word embedding vector space. Then, given a source-target pair, they look for a word that can act as a "connection" (ground) between the two based on the distance between the vectors representing the words. Finally, their system can output the realized metaphor, using the "connecting" word as an explanation of the metaphor.

Veale and Hao [2007] used explicit similes (such as "the streets were as hot as a furnace") as a case base for figurative expressions, managing to acquire a large knowledge base on the most salient properties of each source concept. They also developed a web tool named Aristotle2, which can perform both metaphor comprehension and metaphor generation based on this knowledge base.

As for the interpretation task, LLMs were recently applied to metaphor generation. Contributions such as the work from Ding *et al.* [2023] show both the potential utility and the risks of these models, using GPT-3 to enhance cross-domain analogical reasoning. A more structured tool was presented by Kim *et al.* [2023]. They developed a system that helps in creating extended metaphors, which are particularly relevant and frequently used in science writing.

The system **MET**$^{\mathsf{CL}}$ introduced in this work (see Section 2) aims at providing a significant contribution in both metaphor identification and generation tasks.

1.3 MetaNet and related resources

Another relevant related work in the context of metaphor processing is represented by MetaNet [Dodge *et al.*, 2015]: a metaphor research project developed by a wide network of researchers from several universities in the USA. They developed a structured repository of metaphors based on the Conceptual Metaphor Theory, called MetaNet Metaphor Wiki, that is - to date - the widest available resource of conceptual metaphors able to cover all the different typologies of these linguistic and semantic expressions. In MetaNet a conceptual metaphor (such as POVERTY IS A DISEASE) is represented as a node in a graph, linked with frames explicitly representing its source and target concepts (e.g., POVERTY IS A DISEASE has as *source frame* Disease and as *target frame* Poverty). Other edges of the graph express lexical and semantic relations both between metaphors and between frames. In turn, MetaNet's frames are represented together with a list of *lexical units* that evoke them (e.g., the Disease frame is evoked by "disease", "illness" and "sickness" lexical units, while the Poverty frame is evoked by "poverty", "impoverishment" and "indigence") and can point to related FrameNet frames. Moreover, each conceptual metaphor is provided with a list of metaphorical expressions as examples of lexical realizations (e.g., POVERTY IS A DISEASE is exemplified by "The epidemic of poverty is spreading in America").

A formal ontology representing MetaNet's data, called Amnestic Forgery [Gangemi *et al.*, 2018], was later developed by a different research group, systematizing its content and allowing SPARQL queries on it. Amnestic Forgery was designed as an extension of the linked data hub Framester [Gangemi *et al.*, 2016], which enables access to several linguistic resources, particularly focusing on frame semantics. Currently, Framester contains Amnestic Forgery, that reflects the contents of MetaNet Wiki with neglectable differences. Interestingly enough, the authors of Amnestic Forgery explicitly refer to conceptual blending [Fauconnier and Turner, 2002] as the key cognitive mechanism that enables metaphor interpretation. Focusing on adjective-noun modification, they show that literal phrases (e.g. "business relation") can be interpreted as establishing a new referential frame, which is obtained by simple, conservative frame composition (Business+PersonalRelationship). Differently, in metaphorical expressions (e.g. "frosty relation") a non-conservative composition is performed, and the new frame that emerges inherits only part of the *core* frame (PersonalRelationship), while some roles are substituted by the *modifying* one (Temperature). As we will describe below, the way the authors suggest to apply conceptual blending perfectly resonates both with the Description Logic of typicality and with the HEAD-MODIFIER heuristic that constitute two fundamental principles of $\mathbf{T}^{\mathsf{CL}}$ logic and of its implementation in **MET**$^{\mathsf{CL}}$.

2 The MET$^{\mathsf{CL}}$ System

MET$^{\mathsf{CL}}$ is a metaphor generation and classification system exploiting the Description Logic (DL) of concept combination $\mathbf{T}^{\mathsf{CL}}$ as reasoning engine. In this section, we first provide a high-level overview of $\mathbf{T}^{\mathsf{CL}}$ and then we describe how such logic framework has been applied and integrated in **MET**$^{\mathsf{CL}}$.

2.1 Overview of T$^{\mathsf{CL}}$ for knowledge generation

The logic $\mathbf{T}^{\mathsf{CL}}$ is a compositional reasoning framework developed by Lieto and Pozzato [2020] and employed in a number of applications [Chiodino *et al.*, 2020; Lieto *et al.*, 2021]. It is able to account for the generation of novel chunks of knowledge following a process of human-like concept combination (including conceptual blending) by explicitly relying on a formalization of the prototype theory [Rosch and Mervis, 1975]. In particular, this framework has been shown to be able to account for the phenomenon of the composition of prototypical representations. This aspect is relevant since, according to a well-known argument, prototypes are not compositional [Fodor, 1981; Osherson and Smith, 1981; Murphy, 2016]. For instance, consider a concept like *pet fish*: it results from the composition of the concepts *pet* and *fish*. However, the prototype of *pet fish* cannot result from the composition of the prototypes of a pet and a fish: e.g. a typical pet is furry and warm, a typical fish is grayish, but a typical pet fish is neither

furry and warm nor grayish (typically, it is red). $\mathbf{T}^{\mathsf{CL}}$, on the other hand, proved to be able to generate knowledge by using this type of commonsense compositionality.

The logic $\mathbf{T}^{\mathsf{CL}}$ is based on three main ingredients. The first one relies on the Description Logic of typicality $\mathcal{ALC} + \mathbf{T_R}$ introduced in [Giordano *et al.*, 2015] which allows one to describe the *prototype* of a concept. In this logic, *typical* properties can be directly specified by means of a *typicality* operator $\mathbf{T}$, and a TBox can contain inclusions of the form $\mathbf{T}(C) \sqsubseteq D$ to represent that "typical Cs are also Ds". As a difference with standard DLs, in the logic $\mathcal{ALC} + \mathbf{T_R}$ one can consistently express exceptions and reason about defeasible inheritance as well. The semantics of $\mathbf{T}$ is characterized by the properties of *rational logic*, recognized as the core properties of nonmonotonic reasoning. As a second ingredient, the logic $\mathbf{T}^{\mathsf{CL}}$ exploits a distributed semantics similar to the one of probabilistic DLs known as DISPONTE [Riguzzi *et al.*, 2015], imposing to label inclusions $\mathbf{T}(C) \sqsubseteq D$ with a real number between 0.5 and 1, representing its degree of belief/probability and assuming that each axiom is independent from each others. As an example, we can formalize that we believe that a typical athlete is fit with degree 0.9, whereas we believe that, normally, athletes are young, but with degree 0.75, with the inclusions $0.9 :: \mathbf{T}(Athlete) \sqsubseteq Fit$ and $0.75 :: \mathbf{T}(Athlete) \sqsubseteq Young$, respectively. Degrees of belief in typicality inclusions allow one to define a probability distribution over *scenarios*: roughly speaking, a scenario is obtained by choosing, for each typicality inclusion, whether it is considered as true or false. As a third ingredient, the logic $\mathbf{T}^{\mathsf{CL}}$ employs a heuristics inspired by cognitive semantics [Hampton, 1987] for the identification of a dominance effect between the concepts to be combined: for every combination, it is distinguished a HEAD, representing the stronger element of the combination, and a MODIFIER.

The basic idea of the logic $\mathbf{T}^{\mathsf{CL}}$ is as follows: given a Knowledge Base (KB) and two concepts C_H (HEAD) and C_M (MODIFIER) occurring in it, only *some* scenarios are considered in order to define a revised knowledge base, enriched by typical properties of the combined concept $C \sqsubseteq C_H \sqcap C_M$ obtained by considering blocks of scenarios with the same probability, in decreasing order starting from the highest one. Here all the inconsistent scenarios are discarded. Given a consistent scenario w so selected, the ultimate output is a KB whose set of typicality properties is enriched by all $\mathbf{T}(C_H \sqcap C_M) \sqsubseteq D$ that are entailed from w in the logic $\mathbf{T}^{\mathsf{CL}}$.

In the context of the $\mathbf{MET}^{\mathsf{CL}}$ system, the logic $\mathbf{T}^{\mathsf{CL}}$ has been used to generate, starting from natural language sentences, metaphorical expression via conceptual combination. The distinction concerning HEAD and MODIFIER, intrinsic in $\mathbf{T}^{\mathsf{CL}}$, has been used and mapped within the one between source and target in the context of metaphor literature.

2.2 Pipeline implementation

The system $\mathbf{MET}^{\mathsf{CL}}$ consists in a pipeline of 3 modules (Figure 1). Module 1 handles dataset building and preprocessing. Module 2 is used for the generation of the prototypical representation of concepts. The module generates a text file for each concept involved in some metaphor (either as the source or as the target), containing the prototype of that concept.

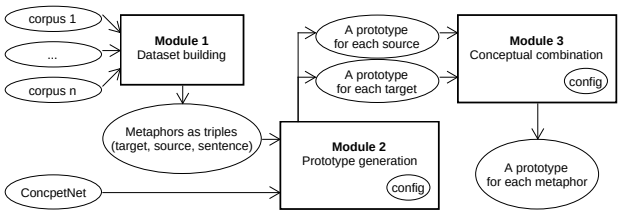


Figure 1: The pipeline architecture of $\mathbf{MET}^{\mathsf{CL}}$.

Module 3 performs the conceptual combination of the two concepts (source and target) involved in a metaphor, applying the $\mathbf{T}^{\mathsf{CL}}$ logic. The output is the prototypical representation of the metaphorical concept.

We implemented two variants of the pipeline, allowing to apply it both to metaphorical expressions and to collections of conceptual metaphors like MetaNet's taxonomy. The metaphorical expressions can be any sentences, but the system needs source and target concepts to be explicitly annotated. On the other hand, conceptual metaphors in MetaNet are linked with rich data, that we leverage to achieve the representations of source and target concept.

As a running example, we will use the metaphorical expression "The epidemic of poverty is spreading in America" and its corresponding conceptual metaphor in MetaNet, POVERTY IS A DISEASE.

Dataset building The implementation of the first module is straightforward for metaphorical expressions. It simply filters and orders sentences in a structured TSV file containing triples of the form (source, target, sentence), e.g. ("epidemic", "poverty", "The epidemic of poverty is spreading in America"). For conceptual metaphors, instead, Module 1 builds an extended representation. To each conceptual metaphor m is associated a list of candidate source and a list of candidate target concepts, including:

1. source/target frame of m;
2. candidate source/target derived from m's name, assuming that it has the form [TARGET] [BE] [SOURCE];
3. source/target frame's *relevant FrameNet frames*;
4. frames subsuming the source/target frame;
5. all above rules applied to the metaphors subsuming m.

Such list is kept ordered by distance (in terms of number of relations in MetaNet's graph) from m. For example, for POVERTY IS A DISEASE we get a candidate source list that looks like ["disease", "affliction", "harm"] and a candidate target list that looks like ["poverty", "status", "problem"].

Prototype generation The second module generates two prototypes representing the source and the target concepts. Our application to metaphorical expressions represents concepts based on meaningful features extracted from ConceptNet 5 relations ([Speer *et al.*, 2017], https://conceptnet.io) and on ConceptNet Numberbatch similarity score [Speer and Lowry-Duda, 2017]. For the example expression "The epidemic of poverty is spreading in America", we build a representation for the concepts "epidemic" and "poverty" by ex-

ploring relevant relations from ConceptNet. In the case of conceptual metaphors, we try to build the prototype for each element in the list produced by Module 1, until a well formed prototype can be computed.

Concept combination Once we have a prototypical representation for both source and target concepts of a metaphor, we apply $\mathbf{T}^{\mathsf{CL}}$ to combine them, by assigning the HEAD role to the target concept, and the MODIFIER role to the source. This assignment also resonates with Gangemi *et al.* [2018] account of conceptual blending as a cognitive tool to metaphorical interpretations. For example, for the metaphor POVERTY IS A DISEASE, we combine "disease" (source) and "poverty" (target). Each of these concepts will have properties with an associated probability extracted from ConceptNet (e.g. "infectious" 0.909, "illness" 0.904 for "disease", and "financial condition" 0.865, "undesirable" 0.833 for "poverty"). With $\mathbf{T}^{\mathsf{CL}}$, we obtain a concept "poverty-disease", inheriting some properties from the source and other from the target (like in the *pet fish*) and constituting a high level representation of the semantics of the metaphor.

3 Evaluation Methodology

In order to evaluate $\mathbf{MET}^{\mathsf{CL}}$ and its capability of metaphor classification and generation we conducted two experiments:

1) An automatic evaluation of the coverage of the MetaNet taxonomy of metaphors tested both on a novel metaphor dataset produced (the NN-450, described below) and on the examples of metaphors included in the MetaNet resource itself. The evaluation has been carried out by using the classification capabilities of a number of state-of-the-art Large Language Models (LLMs), including GPT-4o [Hurst *et al.*, 2024], DeepSeek-R1 [Guo *et al.*, 2025] and Llama 3.2 [Meta, 2024], in both a zero shot and few-shot classification mode. The evaluation included as well, as a baseline, an assessment of a frame-based classification approach, comparing the performance of such classical symbolic approachs to LLMs. We also tested the classification capabilities of $\mathbf{MET}^{\mathsf{CL}}$ on both datasets (based on the number of metaphors the system was able to generate and, as a consequence, to classify later on) and obtained classification results showing that $\mathbf{MET}^{\mathsf{CL}}$ output enables an improvement (a positive performance delta) for the current state of the art models both in terms of number of classified metaphors and, overall, on the precision and recall of their provided output.

2) A human evaluation of the metaphors generated as compound concepts by $\mathbf{MET}^{\mathsf{CL}}$. Compared to the above tasks, this one can be seen as an instance of the metaphor generation task based on MESRs (described above). In this case - we pose under scrutiny the quality of the properties associated by $\mathbf{MET}^{\mathsf{CL}}$ to a particular metaphor, and constituting themselves a high level semantic representation of the metaphor's meaning. Such properties are ranked by human judges, asked to evaluate if they convey the appropriate meaning.

3.1 NN-450 metaphor dataset

The NN-450 dataset was obtained by merging 3, already existing, datasets: the Gordon Metaphor corpus by Gordon *et al.* [2015], which provides manually validated metaphor annotations; the VU Amsterdam Metaphor Corpus [Steen *et al.*, 2010]; and the relatively small Metaphor Detection Dataset developed by Mensa *et al.* [2018]. The result is an enlarged unified corpus of metaphorical sentences of which we used only the nominal metaphors subset (NN-450).

3.2 Automatic Evaluation Rationale

As a first experiment, we tested to what extent MetaNet's taxonomy of metaphors – manually developed by linguists and considered a reference point in the literature – was actually able to cover (i.e. to classify) the metaphors coming from another dataset (i.e. the NN-450) as well as the ones included in the set of 831 MetaNet's examples. The rationale of this evaluation was to assess if and to what extent the metaphors generated by $\mathbf{MET}^{\mathsf{CL}}$ could improve the current state of the art systems in case of missing coverage. To test such a coverage capability of the MetaNet resource we tested several LLMs in two different classification settings: zero-shot and few-shot learning. Given a metaphorical sentence annotated for source and target concepts (e.g. the sentence "The epidemic of poverty is spreading in America", with source "epidemic" and target "poverty"), the task is to identify, if present, the corresponding conceptual metaphor (in the example, POVERTY IS A DISEASE) in MetaNet.

In the zero-shot setting we compared state of the art LLMs with a frame-based baseline grounded in the major available symbolic resource: i.e. MetaNet itself. The baseline approach tries to identify the source and target frames of a metaphor according to frame semantics: if the source (or target) concept is listed between the lexical units of a frame, we assume that such frame is evoked. To maximize the coverage, we extend the lexical units of MetaNet's frames relying on FrameNet, ConceptNet and WordNet [Ruppenhofer *et al.*, 2016; Speer *et al.*, 2017; Fellbaum, 1998]. Then we look for a conceptual metaphor having as *source frame* a frame evoked by the source concept and as *target frame* a frame evoked by the target concept. If such a metaphor exists, it is returned in output. Similarly, for what concerns the LLMs, we checked to what extent they were able to classify the examples provided in the two datasets according to the MetaNet taxonomy. In this case, the evaluation task can be seen as a multiclass classification where an expression is classified into one of MetaNet's conceptual metaphors (or in the synthetic OTHER class if no one of the existing is applicable). We used a zero-shot classification pipeline using a number of different models: the largest of which are DeepSeek-R1 and GPT-4o (see Table 1).

The same experimental rationale was followed also for the few-shot setting. The only difference, in this case, was that such approach was usable only for LLMs. The few-shot learning assessment was used in order to assess wether the use of example-based learning strategies could improve the classification results of LLMs and, if this was the case, if the delta provided by $\mathbf{MET}^{\mathsf{CL}}$ was still significant.

In these classification scenarios, by MetaNet's construction [Dodge *et al.*, 2015], if an expression cannot be referred to a corresponding conceptual metaphor, we conclude that the correct conceptual metaphor is absent from MetaNet, indicating that MetaNet is, in some way, incomplete. We also

compared this datum with the ability of **MET**$^{\text{CL}}$ to generate a representation of any metaphor for which a conceptual representation of source and target was provided.

3.3 Human Evaluation Rationale

Our second experiment consists in a human evaluation of the metaphorical concepts generated via concept combination. We tested a convenience sampling of 70 people, mainly researchers and students, asking them to rate, on a scale from 1 to 10, if and to what extent the properties selected by **MET**$^{\text{CL}}$ were able to convey the intended metaphorical meaning. Overall 630 metaphors, coming from the two datasets described above, were evaluated.

Considering the two versions of the pipeline described in Section 2.2, we conducted two rounds of the evaluation. In the first, the generation involved a conceptual metaphor, using MetaNet to identify the source and target concepts to combine. In the second, we focused on metaphorical expressions (i.e. sentences conveying a metaphor), where the source and target to combine were explicitly annotated based on the linguistic datum. The main difference between these two versions lies in the fact that conceptual metaphors belong to a higher level of abstraction, whereas the combinations obtained starting directly from the words expressed in a sentence are more grounded in the lexicon.

Survey structure The survey consisted in a 10-point scale evaluation of the quality of the generated compound concepts where, ideally, each compound concept represents the meaning of a metaphor, based on its source and target. Each evaluation question ended with the list of the typical properties for the compound concept and their probability score. An example of evaluation question is:

> Please consider the following metaphorical sentence: POVERTY IS A DISEASE and rate on a scale between 1 (worst) and 10 (best) if, overall, the following features associated to the metaphorical concept make sense to understand its metaphorical meaning.
>
> - infectious: 0.909
> - illness: 0.904
> - financial condition: 0.865
> - ...

Before the evaluation (completely anonymous), each participant was asked if they were a native English speaker or not (since all the generated metaphorical concepts came from the two English-based datasets). If the participant answered not to be a native English speaker, an additional question asked a self assessment of written English skills on a 5-point scale.

After the evaluation, the last two questions asked to rate on a 10-point scale how much and how many of the metaphors presented in the survey were hard to grasp or difficult to understand. The metaphors generated by **MET**$^{\text{CL}}$ that each participant had to evaluate were randomly selected.

4 Results, Discussion and Conclusion

For what concerns the first type of automatic evaluation conducted, we report that **MET**$^{\text{CL}}$ was able to generate metaphors via conceptual combination for 441 out of the 448 sentences in NN-450 (98.44%) and for 459 out of 831 sentences (55.23%) of MetaNet examples. The lower percentage in the MetaNet case is partly a consequence of the fact that 261 examples do not have an explicit realization of either the source or the target frame (that on the other hand represents an explicit prerequisite for the automatic generation of metaphors in **MET**$^{\text{CL}}$).

The results of the different classification systems (LLMs and frame-based classification) in the zero-shot setting - and concerning the coverage of the MetaNet taxonomy of metaphors - are shown in Table 1, with a model on each row. For the two datasets NN-450 and MetaNet, the table shows the percentage of sentences that each model (M) was able to classify in MetaNet's taxonomy and the delta that can be obtained by integrating M with **MET**$^{\text{CL}}$. For MetaNet it is also reported the the percentage of correctly classified examples and the precision (number of correctly classified examples / number of classified examples) of each model. This was possible since, in the MetaNet dataset, each sentence is provided as an example of a conceptual metaphor: this allows us to use such binding as a gold standard to compute the precision and recall of the classifier. Finally, the table shows the classification results obtained by each system extending the original MetaNet taxonomy with the metaphors generated by **MET**$^{\text{CL}}$ (EXTENDED MetaNet in the table). Despite, formally, this multiclass classification problem is harder than the previous one (since there are more classes), the results show - for the LLMs - an improvement in precision and recall in all the cases and, overall, an improvement also in the obtained classification. This shows that integrating LLMs with the results of **MET**$^{\text{CL}}$ improves both the number of classifications and their correctness, thus suggesting that the classes generated by **MET**$^{\text{CL}}$ are salient for the sentences to be classified.

Comparing the different models on the classification task, we can see that the frame-based approach is able to classify only a small portion of the datasets. The LLMs, on the other hand, tend to classify almost every sentence into a conceptual metaphor. Surprisingly, smaller models (like RoBERTa large and ELECTRA small) achieve higher coverage compared to larger ones (DeepSeek-R1 and GPT-4o). However, such classifications suffer of poor precision and recall.

Table 2 shows the results of the same experiments in a few-shot setting. Using the LLMs for text generation, we generated a prompt that asked to classify a sentence into a set of classes, providing some examples. We can see that the behavior of the LLMs is completely different from the Zero-shot setting, with the smaller models failing to effectively classify even a single sentence. This is a consequence of the text-generation approach, which is open ended and does not guarantee that the result will be one of the possible classes. For the models that do provide some classifications, we can again see a considerable improvement in precision and recall when run on the extended version of the dataset.

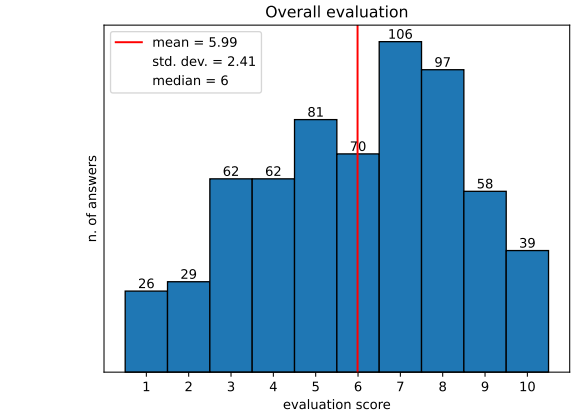
For what concerns the human evaluation of the generated metaphors, Figure 2a shows the overall results: the generated combinations aiming at representing the meaning of a metaphor were generally accepted, with a mean evaluation of 5.99/10 and a median of 6, with a standard deviation of 2.41. Figures 2b and 2c show the evaluations for metaphors generated using the conceptual metaphor approach (mean 5.87/10, median 6, std. dev. 2.35) and the metaphorical expressions approach (mean 6.14/10, median 6, std. dev. 2.47). These results

Table 1: Zero-shot classification

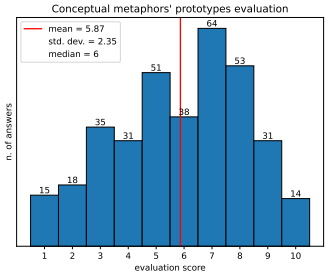
		NN-450 Dataset		MetaNet				EXTENDED MetaNet		
Model (M)	# params	Classifications	$\Delta cl. = \textbf{MET}^{\textbf{CL}} \setminus M$	Classifictions	$\Delta cl. = \textbf{MET}^{\textbf{CL}} \setminus M$	Recall	Precision	Classifications	Recall (delta)	Precision (delta)
GPT-4o	unknown	79.02%	**+20.31%**	89.65%	**+5.17%**	33.69%	37.58%	90.25%	49.94% (**+16.25%**)	55.33% (**+17.75%**)
DeepSeek-R1	671B	79.02%	**+20.31%**	96.51%	**+1.93%**	49.10%	50.87%	93.86%	49.58% (**+0.48%**)	52.82% (**+1.95%**)
Qwen2.5-Max	unknown	31.92%	**+67.86%**	84.84%	**+6,50%**	36.10%	42.55%	68.47%	38.15% (**+2,05%**)	55.71% (**+13,16%**)
BLOOMZ 560m	560M	99.33%	**+0.67%**	99.40%	**+0.12%**	9.75%	9.81%	99.76%	37.55% (**+27.8%**)	37.64% (**+27.83%**)
BART large	407M	97.99%	**+2.01%**	98.19%	**+0.60%**	12.88%	13.11%	98.80%	37.06% (**+24.18%**)	37.52% (**+24.41%**)
RoBERTa large	355M	98.44%	**+1.56%**	98.32%	**+0.84%**	11.31%	11.51%	99.28%	35.26% (**+23.95%**)	35.52% (**+24.01%**)
ELECTRA small	14M	99.78%	**+0.22%**	100.00%	0%	12.52%	12.52%	100.00%	47.77% (**+35.25%**)	47.77% (**+35.25%**)
frame-based (baseline)	symbolic	16.89%	**+81.47%**	16.00%	**+39.95%**	10.23%	63.91%	16.00%	10.23%	63.91%

Table 2: Few-shot classification

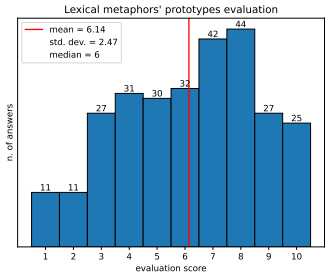
		NN-450 Dataset		MetaNet				EXTENDED MetaNet		
Model (M)	# params	Classifications	$\Delta cl. = \textbf{MET}^{\textbf{CL}} \setminus M$	Classifications	$\Delta cl. = \textbf{MET}^{\textbf{CL}} \setminus M$	Recall	Precision	Classifications	Recall (delta)	Precision (delta)
GPT-4o	unknown	84.15%	**+15.85%**	88.21%	**+5.90%**	34.06%	38.61%	87.36%	42.96% (**+8.9%**)	49.17% (**+10.56%**)
DeepSeek-R1	671B	70.98%	**+28.79%**	95.79%	**+1.81%**	46.57%	48.62%	97.95%	47.17% (**+0.6%**)	48.16% (-0.46%)
Qwen2.5-Max	unknown	51.56%	**+47.54%**	85.80%	**+7.46%**	29.46%	34.92%	73.04%	31.53% (**+2,07%**)	43.16% (**+8,24%**)
BLOOMZ 7b1	7.1B	78.79%	**+21.21%**	88.57%	**+6.02%**	5.05%	5.71%	89.05%	13.96% (**+8.91%**)	15.68% (**+9.97%**)
Falcon 7b Instruct	7B	79.69%	**+20.09%**	70.28%	**+14.92%**	1.44%	2.05%	84.36%	5.29% (**+3.85%**)	6.28% (**+4.23%**)
LLaMa 3.2 3B Instruct	3.21B	20.31%	**+78.35%**	29.72%	**+38.75%**	3.85%	12.96%	35.02%	9.27% (**+5.42%**)	26.46% (**+13.5%**)
Mistral 3b instruct	3B	0%	**+98.44%**	0%	**+55.23%**	NA	NA	0.12%	0%	0%
ELECTRA large	335M	0%	**+98.44%**	0%	**+55.23%**	NA	NA	0%	NA	NA
RoBERTa base	125M	0%	**+98.44%**	0%	**+55.23%**	NA	NA	0%	NA	NA



(a) Overall.



(b) Conceptual metaphors.



(c) Metaphorical expressions.

Figure 2: The overall results of the human evaluation (a), and the ones assigned to the combinations representing conceptual metaphors (b) and metaphorical expressions (c).

suggest that metaphor systems grounded directly in the lexicon, instead of relying on the top-down knowledge instilled in resources like MetaNet, are more able to capture the intended metaphorical meaning. Overall, to the best of our knowledge, the reported user study represents the first attempt to actually evaluate in a more systematic way (i.e. not with few annotators but with a controlled user experiment involving a relevant number of human judges) the generative capability of AI systems for metaphors. Our results are not comparable with other works obtained in the metaphor generation tasks since, in those cases, either the tasks under focus are diverse or, in situations similar to ours and aiming at evaluating the quality of the generated representations of metaphors, e.g. [Zheng *et al.*, 2020], the metrics used are qualitative (i.e. Likert scales). Overall, the main finding reported in this work is that, at least for a subset of metaphorical expressions, the generative strategy proposed in $\textbf{MET}^{\textbf{CL}}$ represents a valid complement able to improve the classification capabilities of both LLMs and more traditional frame-based approaches. From a practical perspective, this result paves the way to research in metaphor elaboration that hybridize neural and symbolic systems. From a theoretical point of view, on the other hand, this experimental datum suggests that, in absence of a comprehensive theory of metaphorical processing able to account for all aspects of this phenomenon, the categorization approach plays certainly a role to be further investigated. In addition, the reported data show how top-down approaches based on metaphor classifications (resulting in resources like MetaNet) fail to capture the whole spectrum of conceptual metaphors and, as such, need to be integrated with systems like $\textbf{MET}^{\textbf{CL}}$. We are now extending both the automatic and human evaluation in order to enrich the robustness of our findings. In addition, since experimental evidences in cognitive science have suggested that metaphors are the most common form of linguistic instantiation of analogical abilities [Gentner and Clement, 1988; Tourangeau and Rips, 1991], we are extending our investigation towards analogical reasoning[1].

[1]Data and code described in this paper are available at https://github.com/StefanoZoia/METCL

References

Robyn Carston. Metaphor and the literal/non-literal distinction. In Keith Allan and Kasia M. Jaszczolt, editors, *The Cambridge Handbook of Pragmatics*, pages 469–492. Cambridge University Press, 1 edition, January 2012.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online, 2021. Association for Computational Linguistics.

Eleonora Chiodino, Davide Di Luccio, Antonio Lieto, Alberto Messina, Gian Luca Pozzato, and Davide Rubinetti. A knowledge-based system for the dynamic generation and classification of novel contents in multimedia broadcasting. In *ECAI 2020*, pages 680–687. IOS Press, 2020.

Zijian Ding, Arvind Srinivasan, Stephen Macneil, and Joel Chan. Fluid Transformers and Creative Analogies: Exploring Large Language Models' Capacity for Augmenting Cross-Domain Analogical Creativity. In *Creativity and Cognition*, pages 489–505, Virtual Event USA, June 2023. ACM.

Ellen Dodge, Jisup Hong, and Elise Stickles. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado, 2015. Association for Computational Linguistics.

Gilles Fauconnier and Mark Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, New York, 2002.

Christiane Fellbaum. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686, 1998.

Jerry A Fodor. *The present status of the innateness controversy*. unknown, 1981.

Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. Framester: A Wide Coverage Linguistic Linked Data Hub. In Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali, editors, *Knowledge Engineering and Knowledge Management*, volume 10024, pages 239–254. Springer International Publishing, Cham, 2016. Series Title: Lecture Notes in Computer Science.

Aldo Gangemi, Mehwish Alam, Valentina Presutti, and others. Amnestic forgery: An ontology of conceptual metaphors. *Frontiers in Artificial Intelligence and Applications*, 306:159–172, 2018. Publisher: IOS Press.

Mengshi Ge, Rui Mao, and Erik Cambria. A survey on computational metaphor processing techniques: from identification, interpretation, generation to application. *Artificial Intelligence Review*, August 2023.

Dedre Gentner and Catherine Clement. Evidence for relational selectivity in the interpretation of analogy and metaphor. *Psychology of learning and motivation*, 22:307–358, 1988.

Dedre Gentner and Kenneth D. Forbus. Computational models of analogy. *WIREs Cognitive Science*, 2(3):266–276, May 2011.

Dedre Gentner. Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science*, 7(2):155–170, April 1983.

Laura Giordano, Valentina Gliozzi, Nicola Olivetti, and Gian Luca Pozzato. Semantic characterization of rational closure: From propositional logic to description logics. *Artificial Intelligence*, 226:1–33, 2015.

Jonathan Gordon, Jerry Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. A corpus of rich metaphor annotation. In Ekaterina Shutova, Beata Beigman Klebanov, and Patricia Lichtenstein, editors, *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66, Denver, Colorado, June 2015. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

James A Hampton. Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, 15(1):55–71, 1987.

Keith J. Holyoak and Dušan Stamenković. Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, 144(6):641–671, June 2018.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors, 2023. Publisher: arXiv Version Number: 1.

Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 115–135, Pittsburgh PA USA, July 2023. ACM.

Antonio Lieto and Gian Luca Pozzato. A description logic framework for commonsense conceptual combination integrating typicality, probabilities and cognitive heuristics. *Journal of Experimental and Theoretical Artificial Intelligence*, 32(5):769–804, 2020.

Antonio Lieto, Gian Luca Pozzato, Stefano Zoia, Viviana Patti, and Rossana Damiano. A commonsense reasoning framework for explanatory emotion attribution, generation and re-classification. *Knowledge-Based Systems*, page 107166, 2021.

Antonio Lieto. *Cognitive design for artificial minds*. Routledge, 2021.

Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. MetaPro: A computational metaphor processing model for text preprocessing. *Information Fusion*, 86-87:30–43, October 2022.

Enrico Mensa, Aureliano Porporato, and Daniele P. Radicioni. Grasping metaphors: Lexical semantics in metaphor analysis. In Aldo Gangemi, Anna Lisa Gentile, Andrea Giovanni Nuzzolese, Sebastian Rudolph, Maria Maleshkova, Heiko Paulheim, Jeff Z. Pan, and Mehwish Alam, editors, *The Semantic Web: ESWC 2018 Satellite Events*, pages 192–195, Cham, 2018. Springer International Publishing.

AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024, 2024.

Gregory L. Murphy. Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, 23(4):1035–1042, August 2016.

Daniel N. Osherson and Edward E. Smith. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1):35–58, January 1981.

Sunny Rai, Shampa Chakraverty, Devendra K. Tayal, Divyanshu Sharma, and Ayush Garg. Understanding Metaphors Using Emotions. *New Generation Computing*, 37(1):5–27, January 2019.

Fabrizio Riguzzi, Elena Bellodi, Evelina Lamma, and Riccardo Zese. Reasoning with probabilistic ontologies. In Qiang Yang and Michael J. Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4310–4316. AAAI Press, 2015.

Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute, 2016.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California, 2016. Association for Computational Linguistics.

Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. A Knowledge Graph Embedding Approach for Metaphor Processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:406–420, 2021.

Robert Speer and Joanna Lowry-Duda. ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada, 2017. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), February 2017.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, and T. Pasma. *A method for linguistic metaphor identification. From MIP to MIPVU.* Number 14 in Converging Evidence in Language and Communication Research. John Benjamins, 2010.

Roger Tourangeau and Lance Rips. Interpreting and evaluating metaphors. *Journal of Memory and Language*, 30(4):452–472, 1991.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Tony Veale and Yanfen Hao. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *AAAI*, volume 2007, pages 1471–1476, 2007.

Mingyu Wan, Baixi Xing, Qi Su, Pengyuan Liu, and Chu-Ren Huang. Sensorimotor Enhanced Neural Network for Metaphor Detection. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 312–317, Hanoi, Vietnam, October 2020. Association for Computational Linguistics.

Deirdre Wilson. Parallels and differences in the treatment of metaphor in relevance theory and cognitive linguistics. *Intercultural Pragmatics*, 8(2), January 2011.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. "Love Is as Complex as Math": Metaphor Generation System for Social Chatbot. In Jia-Fei Hong, Yangsen Zhang, and Pengyuan Liu, editors, *Chinese Lexical Semantics*, volume 11831, pages 337–347. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.