Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5810–5819
Marseille, 11–16 May 2020

Figure Me Out: A Gold Standard Dataset for Metaphor Interpretation

Omnia Zayed, John P. McCrae, Paul Buitelaar
Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland
{omnia.zayed, john.mccrae, paul.buitelaar}@insight-centre.org

Abstract

Metaphor comprehension and understanding is a complex cognitive task that requires interpreting metaphors by grasping the interaction between the meaning of their target and source concepts. This is very challenging for humans, let alone computers. Thus, automatic metaphor interpretation is understudied in part due to the lack of publicly available datasets. The creation and manual annotation of such datasets is a demanding task which requires huge cognitive effort and time. Moreover, there will always be a question of accuracy and consistency of the annotated data due to the subjective nature of the problem. This work addresses these issues by presenting an annotation scheme to interpret verb-noun metaphoric expressions in text. The proposed approach is designed with the goal of reducing the workload on annotators and maintain consistency. Our methodology employs an automatic retrieval approach which utilises external lexical resources, word embeddings and semantic similarity to generate possible interpretations of identified metaphors in order to enable quick and accurate annotation. We validate our proposed approach by annotating around 1,500 metaphors in tweets which were annotated by six native English speakers. As a result of this work, we publish as linked data the first gold standard dataset for metaphor interpretation which will facilitate research in this area.

Keywords: metaphor, interpretation, dataset, annotation, tweets, lexical resources.

1. Introduction

Metaphor is a crucial aspect of human cognition and communication. The computational processing of metaphors has gained wide attention lately by focusing on two tasks, namely, metaphor identification and metaphor interpretation. Metaphor identification is concerned with recognising the metaphoric word or expression in a given sentence while metaphor interpretation focuses on "translating" the metaphor to its literal meaning. The interpretation task is very important to fully understand the intended meaning of the metaphor, however it is much less explored compared to the identification task. One reason is that it is very exhausting for humans to comprehend the interaction between the target and the source components of the metaphoric expression. Although native speakers unconsciously grasp such interaction, asking a human annotator to translate such a cognitive process and interpret a metaphoric expression is a very demanding task. This is the reason behind the lack of publicly available datasets for metaphor interpretation, which in turn hinders the development of this topic.

There are several approaches to address metaphor interpretation among which:

1. *Lexical Substitution* (lexical paraphrasing) where the metaphoric word/phrase is replaced with its literal counterpart to clarify its semantic meaning. This task is viewed as single-word (lexical) substitution (Shutova, 2010; Shutova et al., 2012; Bollegala and Shutova, 2013);

2. *Paraphrase Generation* (inference of meaning) where the full sentence including the metaphoric expression is transformed using more literal words (Bizzoni and Lappin, 2018);

3. *Definition Generation* (interpretation or definition assignment) where a full interpretation (explanation) of the metaphoric expression is provided (Martin, 1990) in a way similar to dictionaries or lexicons.

Table 1 gives examples of the three aforementioned approaches of metaphor interpretation. The choice of the approach depends on the application. In this work, we view metaphor interpretation as a definition generation (explanation) task focusing on finding out the meaning of a given metaphoric expression and explain it in literal words. There are a variety of applications that can benefit from interpreting metaphors, including language learning and text simplification (Barbu et al., 2015; Wolska and Clausen, 2017; Bingel et al., 2018) as well as lexical resources creation and development (Krek et al., 2018).

Approach	Metaphor	Interpretation
lexical substitution (Shutova et al., 2010)	**brush aside** accusation	reject
paraphrase generation (Bizzoni and Lappin, 2018)	The crowd was a river in the street.	The crowd was large and impetuous in the street.
definition generation (Martin, 1990)	How do I **kill the process**?	to terminate computer process.

Table 1: Metaphor interpretation approaches with examples from previous studies.

Manually annotating a dataset for metaphor interpretation (either to provide a definition/explanation or to paraphrase the expression) is a very demanding task which requires effort and time from a human annotator to figure out the meaning of a given metaphor and provide a literal explana-

tion (if possible[1]) for it. Moreover, it is a highly subjective task; the meaning of an expression can vary from one annotator to the other depending on the context and the cultural background of the annotator. This will introduce a question of accuracy and consistency of the created dataset and the submitted annotations. The available datasets have important limitations in terms of size, representativeness and quality as will be discussed in Section 2.. This work attempts to address these issues by introducing an annotation scheme that employs lexical resources to assist in the creation of the interpretations. We design this scheme with the goal of reducing the cognitive load for annotators while maintaining accuracy and consistency based on our previous experience and conversations with expert annotators. Our approach employs dictionaries to automatically compile a list of possible definitions for a given metaphoric expression. These possible candidates of interpretations are generated by employing semantic similarity based on word embeddings. As a result, we produce the first gold standard dataset of metaphor interpretations.

This annotation task closely resembles word sense disambiguation (WSD) in that given a metaphoric verb the goal (of the human annotator) is to identify its closest (literal) meaning among the automatically generated list of candidates. External knowledge resources including machine readable dictionaries and lexicons have been widely used in WSD (Ide and Véronis, 1993; Agirre and Stevenson, 2007; Navigli, 2009). We will discuss the criteria of choosing the resources utilised in this work in Section 3..

Linguistic metaphors can be expressed in various syntactic structures. The majority of previous work focused on modelling verbal and adjectival metaphoric expression (Shutova, 2015). Corpus studies showed that verbs are the most frequent metaphorical expressions (Cameron, 2003; Shutova and Teufel, 2010) which encouraged the majority of systems pertained to metaphor processing to focus on the metaphorical usage of verbs. Thus, in this work, we focus on verb-direct object metaphoric expressions. We create our dataset of metaphor definitions by interpreting around 1,500 metaphoric expression identified in an existing tweets dataset (Zayed et al., 2019) and providing their literal meaning. To the best of our knowledge, there is no publicly available annotated dataset of this kind and we believe that this resource will be invaluable for the development and evaluation of computational models for metaphor interpretation.

2. Related Work

There exist only two datasets for metaphor interpretation, one prepared for lexical substitution and the other for paraphrase generation. Shutova (2010) introduced a corpus-based approach that addressed metaphor interpretation as a lexical paraphrasing task focusing on subject-verb and verb-object metaphoric expressions. In this work, each metaphoric verb is substituted by its literal counterpart (literal paraphrase/synonym). A dataset of 46 sentences covering 61 metaphoric verbs from a subset of the British National Corpus (BNC) (Burnard, 2009) is created to evaluate the approach. In order to annotate this dataset, five native speakers were asked to write down all suitable literal paraphrases for the highlighted metaphorical verbs. For example, the possible paraphrases given by the annotators for *"leak report"* are *"reveal, disseminate, publish, divulge, let out, disclose"*. There is no information available regarding the inter-annotator agreement as the final dataset is compiled by incorporating all of the annotations. This dataset is the only dataset available for single-word metaphor paraphrasing (lexical substitution) focusing on metaphoric verbs. Despite its limited size, it was used to evaluate other metaphor paraphrasing systems (Shutova et al., 2012; Bollegala and Shutova, 2013). The dataset is not directly available online but can be obtained upon request from the authors.

More recently, Bizzoni and Lappin (2018) created a dataset to judge paraphrases of metaphoric sentences. Their dataset consists of 200 metaphorical sentences, each sentence has four ranked candidate paraphrases. The candidate paraphrases were labelled on a 1-4 scale based on the degree to which they paraphrase the metaphoric sentence. The dataset covers metaphors with various syntactic structures including: noun phrases, verbs, adjectives and multi-word metaphors. The metaphoric sentences were either selected from published sources or devised manually by the authors. Also, the provided candidate paraphrases were created manually by the authors themselves. Finally, all the sentences were revised by a native speaker. The dataset is publicly available online[2].

The discussed datasets have important limitations in terms of size, representativeness and quality. Both datasets are relatively small which limits their usage for machine learning applications. Also, they are restricted to a small subset of metaphors which limits their metaphoric coverage and representativeness. Moreover, their annotation technique influences their quality as both datasets are not verified in terms of inter-annotator agreement. In this work, we avoid these limitations while creating our dataset. We considered several aspects to ensure the dataset quality including:

- data selection to ensure metaphoric coverage and representativeness.
- data compilation to ensure annotations consistency and quality
- native human annotators' training and expertise
- clear annotation scheme and guidelines

3. Data Preparation

In this section, we discuss the preparation steps behind our dataset. We first describe the criteria that we followed to select a dataset of already identified metaphors. Our main concern while choosing a dataset of metaphors is to ensure wide coverage and representativeness. We then demonstrate how we compiled the data by employing existing lex-

[1]The debate of whether a metaphor can be paraphrased or translated into its literal meaning or not is out of the scope of this paper. Stewart (1971) provides details about the different views on this issue.

[2]https://github.com/yuri-bizzoni/Metaphor-Paraphrase

ical resources with the goal to reduce the cognitive load on the annotators while maintaining accuracy and consistency.

3.1. Data Source

The first step towards creating our dataset is to have a manually annotated dataset where the metaphors are identified. Since we are interested in verb-noun metaphoric expressions, our initial consideration was to explore existing annotated datasets designed to identify verb-noun grammatical relations for metaphoricity. There exist two datasets of this kind; the first one is introduced by Shutova et al. (2016) which is an adaptation of the dataset introduced by Mohammad et al. (2016), referred to as the MOH dataset. The original MOH dataset was created by annotating different senses of verbs in the example sentences in WordNet (Fellbaum, 1998) for metaphoricity. Shutova et al. (2016) extracted the subject-verb and verb-direct object grammar relations from the MOH dataset and created a subset of 647 instances out of which 316 instances are metaphorical and 331 instances are literal. The second dataset to consider is introduced by Zayed et al. (2019) who created a dataset of around 2,500 tweets annotated to support the identification of verb-direct object expressions in which around 55% of the instances are metaphoric expressions. The dataset comprises emotional tweets of general topics as well as political tweets related to Brexit covering a wide range of verbs including light and aspectual verbs along with various associated abstract and concrete concepts (nouns). Five native annotators performed the annotation of this dataset and the inter-annotator agreement was carried out to asses the quality of the annotations by means of Fleiss' kappa (Fleiss, 1971) which averaged 0.75. The MOH dataset has ∼300 metaphoric instances while Zayed's dataset has ∼1,500. Since we are looking for wider coverage of verb-noun metaphoric expressions, we chose Zayed's dataset as our dataset of identified metaphors to be interpreted. Table 2 shows examples of instances appearing in Zayed's tweets dataset. It will be interesting to analyse the effect of the noisy user-generated text of the tweets on interpreting metaphors.

Tweet	Metaphoric Expression
its great to be happy, but its even better to **bring happiness** to others.	bring happiness
make memories you will look back and smile at.	make memories
make or **break moment** today! together we are stronger! vote remain #strongerin #euref	break moment
...cameron can not win this #euref without your support. how many will **lend their support** to...	lend their support

Table 2: Examples of instances appearing in Zayed's tweets dataset showing verb-direct object metaphoric expressions that can be used as targets for interpretation.

3.2. Data Compilation

Now that we have a set of sentences (tweets) with identified metaphors (verb-noun pairs) that needs to be interpreted, the direct approach would be to ask human annotators to write down a definition of each metaphoric expression. As discussed earlier, this task will be very demanding and highly subjective. It will require a lot of time and cognitive effort from the annotators to interpret the metaphor after understanding the interaction between its components (the tenor or the noun and the vehicle or the verb). With the aim to reduce this cognitive load and maintain consistency, we bootstrap an initial list of possible interpretations for the highlighted metaphor (target verb-noun pair) from lexical resources and provide it to the annotators.

The idea comes from the question: what would a language learner (a non-native speaker) do when encountering a new[3] metaphoric expression in a given text? One way could be to look it up in a dictionary. Since there is no specific dictionary for metaphors, sometimes the full expression could be found in a dictionary where very conventionalised metaphors are labelled as idioms[4]. For the majority of cases, where there is no direct match of the whole metaphoric expression (verb-direct object pair) in a dictionary, the user could start looking for the verb in the dictionary. Then, try to find the nearest definition that can match the metaphoric sense of the verb and at the same time represent its interaction with the accompanying noun.

To automate this idea, we have two approaches to pursue; first, to check out metaphors that are labelled as idioms in lexical resources and extract their definitions. Second, to check out the nearest definition of the verb in a dictionary that could be applied to the noun to convey a metaphoric sense. Both methods should be validated by human annotators.

3.2.1. Metaphors in Wiktionary Idioms

An idiom is a phrase or an expression consisting of a group of words that conveys a figurative meaning different from their literal one. This meaning cannot be guessed from the meanings of the individual words, thus an idiom is considered an inseparable lexical unit. On the other hand, a metaphor is an analogy where a concept (represented by a word sense) is borrowed to represent another concept by exploiting common properties between both concepts (Lakoff and Johnson, 1980). Unlike idioms, the meaning of a metaphor can be determined by understanding its individual lexical units even if the listener did not encounter it before (Crystal, 2008).

Commonly used metaphors which became conventionalised in the language found their way into lexical resources (dictionaries) under the idioms category. Although we argue against this generalisation from a linguistic point of view, it is understandable to assign conventionalised metaphors (fixed expression) to an already existing label rather than creating a new one. Wiktionary[5], which is a multilingual online lexicon (dictionary) edited and maintained by volunteers in a collaborative way, has a large set

[3]By "new" here we do not mean "novel" in the absolute sense but we mean that the language learner did not know the metaphoric expression beforehand.

[4]The difference between metaphors and idioms is out of the scope of this paper. But for the sake of clarity, we briefly discuss it in the next sub-section.

[5]https://www.wiktionary.org

of idioms under the *English Idioms Category*[6]. Wiktionary is the largest available collaboratively constructed lexicon and is an important resource for natural language processing research (Meyer and Gurevych, 2012). In this work, we used Wiktionary's API[7] to query the idioms category in order to automatically get the definition of metaphoric expressions in our dataset. Table 3 shows examples of the metaphors labelled as idioms and their retrieved definition.

Metaphor	Definition
blow someone's mind	to astonish someone, to flabbergast someone.
break a law	to violate a law.
build bridges	to establish links or friendly relations.
cast one's vote	to vote for something.
take a chance	to risk doing something; to try something risky.

Table 3: Examples of the metaphoric expressions from the tweets dataset found under Wiktionary's English Idioms Category.

Although this category contains around 8,000 idioms, only around 10% of the identified metaphors in the tweets dataset were found under this category. It means that our dataset contains only ∼140 conventionalised metaphors which are considered fixed expressions and labelled as idioms in Wiktionary. This motivated us to proceed with our second idea of finding the nearest definition of the metaphoric expression in a dictionary as will be discussed in the next section.

3.2.2. Nearest Definitions in a Dictionary

Consider the highlighted metaphoric expression in the following tweet:

> *I want him to participate in Presidential Elections so we can defeat him and* ***break his ego****[...]*

In this example, the concrete (physical) concept of a brittle object represented by the verb *"break"* is borrowed to express an abstract (emotional) concept represented by the noun *"ego"*. Although the metaphoric expression *"break ego"* is not directly found in a dictionary, there will be a sense for the verb *"break"*, in almost any dictionary, that is related to destroying emotions or a person's spirit, will or determination which is, in a sense, related to the concept of the noun *"ego"*. Table 4 shows the definition of the verb *"break"* related to emotional concepts in several dictionaries.

Our hypothesis is that measuring the semantic similarity of the noun of the metaphoric expression against each sense of the verb retrieved from a dictionary can reflect the interaction between the meaning of the components of the metaphor and, in turn, reveal the nearest definition of the metaphoric expression. To examine this hypothesis, we modelled this idea by employing a dictionary API, pre-trained word embeddings and cosine similarity.

Dictionary	Definition
Wiktionary[8]	to cause (a person or animal) to lose spirit or will; to crush the spirits of.
WordNet[9]	weaken or destroy in spirit or body.
Oxford[10]	Crush the emotional strength, spirit, or resistance of.
Oxford Learner's[11]	to destroy something or make somebody/something weaker; to become weak or be destroyed
Longman[12]	to make someone feel that they have been completely defeated and they cannot continue working or living.
Macmillan[13]	to destroy someone's confidence, determination, or happiness.

Table 4: The definition of the verb *"break"* that is related to *"destroying emotions"* in different dictionaries.

In this work, we represent a sense of a verb by its definition in a dictionary along with the accompanied contextual examples (example sentences). We used the Oxford Learner's Dictionary to retrieve the definitions of a given verb and the example sentences. The reason behind choosing Oxford Learner's Dictionary is that it offers many contextual examples for each word compared to other dictionaries that we examined including Wiktionary and WordNet. More contextual examples will help us to better model the sense of the verb. We also considered other factors while choosing the dictionary including the number and granularity of senses that a word has.

The first step, in our approach, is to retrieve the definitions and the sentence examples of each verb in our dataset of metaphors in order to represent the different senses of the verb. Given a metaphoric expression, we then use GloVe (Pennington et al., 2014) pre-trained word embeddings to calculate the cosine similarity between the sense of the verb (represented by the definition and the contextual examples) and the noun of this given metaphoric expression. Our approach can be formulated as follows:

- Given a set of metaphoric expressions of verb-noun pairs $M = \{(V, N)\}$, suppose that each verb in M has a set of senses S_v in the dictionary represented by its definition and the sentences examples.

- Each sense is represented by a sequence of words $w_{v,i,1}, w_{v,i,2}, ..., w_{v,i,l}$ where l is the number of words in the i^{th} sense of the verb v in the dictionary for each $i \in S_v$.

- The cosine similarity between the embeddings of the noun n in the metaphoric expression represented as x_n and the embeddings of the words of the verb sense combined into a single vector by mean pooling as $x'_{v,i}$ can be calculate as follows:

$$Similarity = cos(x_n, x'_{v,i}); \forall i \in S_v \quad (1)$$

[6] https://en.wiktionary.org/w/index.php?title=Category:English_idioms

[7] https://www.mediawiki.org/wiki/API:Query

[8] https://en.wiktionary.org/wiki/break

[9] last sense: http://bit.ly/2TlWxyx

[10] http://bit.ly/2x7SYUb

[11] http://bit.ly/3aoa12v

[12] http://bit.ly/2wtmtiI

[13] http://bit.ly/32Kg7HS

This gives a list of senses (definition and example sentences) ranked according to the similarity score.

– The top three definitions are then obtained as possible candidates to interpret the given metaphor (v, n) according to the highest similarity score. Initial evaluations demonstrated that selecting the top three definitions was a sufficient trade-off between reducing cognitive load and maintaining accuracy.

We used the Gensim Python library (Rehurek and Sojka, 2010) and the 300-dimensional GloVe embeddings pretrained on the Common Crawl dataset. Table 5 lists the nearest three definitions from Oxford Learner's Dictionary, ranked by the cosine similarity score, which could interpret the given metaphoric expressions based on the similarity between the noun of the metaphoric expression and the sense of its verb.

Our dataset now comprises around 1,500 tweets with highlighted metaphoric expressions and a list[14] of possible interpretations for each highlighted expression. The annotators will be asked to select one interpretation from the list or provide their own interpretation in case no applicable definition can be found.

4. Annotation Process

We set up our annotation task on Amazon Mechanical Turk (MTurk). Six native English speakers were hired to annotate the dataset whose field of study is English. It is worth mentioning that all annotators have the same nationality to rule out cultural background bias. This section describes how we set up the task on MTurk.

Task Definition. Given a tweet with a highlighted metaphoric expression, the main goal of the task is to select the most probable definition/interpretation (if exists) of the highlighted expression among the given definitions (similar to manual sense disambiguation but for the metaphoric expression). If the given list does not contain a definition that correctly interprets the metaphor, the annotator is asked to provide a simple definition that explains both the verb and the noun of the metaphor. The annotators are encouraged to consider explaining the meaning of the metaphoric expression to a child, a language learner or a person with a learning difficulty.

Guidelines. Each tweet has a highlighted metaphoric expression of a verb-direct object syntactic structure. The annotators were instructed to follow the following set of guidelines:

1. Read the whole tweet to establish a general understanding of the meaning.
2. Focusing on the highlighted expression, read the given definitions and determine which one is the most probable (nearest) definition of the highlighted metaphor. In case no applicable choice is found, select "not applicable".
3. In case of choosing "not applicable", provide a definition to interpret and explain the metaphor in few words.

These steps were represented in the task as three questions appearing to the annotators on MTurk as shown in Figure 1. A free text area was provided under each tweet to allow the annotator to write their comments, insights or any confusing issues about the tweet content. The annotators went through a training phase by taking a demo task to familiarise them with the platform and to clarify the annotation process.

Task Design. We designed the annotation task as pages of 10 tweets each. We estimated the time taken to annotate around 60 tweets to be one hour; therefore, we paid $1.80 for each page. This comes down to $12 per hour, which aligns with the minimum wage regulations of the country where the authors resided at the time of this publication.

5. Dataset Evaluation and Analysis

In this section, we provide a description of our assessment of the annotation results. We also discuss our observations and analysis of the dataset. Moreover, we will discuss the points of agreement and disagreement between the annotators along with statistical analysis of the dataset.

5.1. Evaluation

In order to evaluate the reliability of the annotation scheme, the inter-annotator agreement (IAA) was measured in terms of Fleiss' kappa (Fleiss, 1971) among the six annotators. We consider each definition in the list as a category and the annotator's definition as a category, so in total we have four categories. Fleiss' kappa is then calculated as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

where $\bar{P}$ is the mean proportion of agreement between k annotators and $\bar{P}_e$ is the mean proportion of agreement by chance.

Among the six annotators, the IAA averaged 0.272 for four categories on 1,301 annotated instances. Based on Landis and Koch (1977) scale, a fair agreement was achieved despite the subjectivity of the task.

We were interested to analyse the best obtained IAA by varying the number of annotators depending on the majority of the annotated (non-skipped) instances. We calculated the IAA between the best (top) five, four and three annotators, respectively, who tend to agree the most as shown in Table 6. From this analysis we observed that: 1) in case of the five annotators who agreed the most, the discarded annotator was the one who tend to chose the customised definition more often; 2) while in the case of the three annotators who agreed the most, the discarded two annotators were the ones who tend to choose the dictionary definition more often (as will be discussed in detail in subsection 5.2.). Having such versions of the dataset will allow the users to choose the subset that better suits their application. A higher quality dataset can be obtained from the

[14] We shuffled this list before giving it to the annotators in order to avoid the bias of selecting the first choice every time.

Metaphoric Expression	Definition	Cosine Similarity
bind country	to unite people, organizations, etc. so that they live or work together more happily or effectively	**0.620**
	to force somebody to do something by making them promise to do it or by making it their duty to do it	0.573
	to tie somebody/something with rope, string, etc. so that they/it cannot move or are held together firmly	0.422
hit economy	to have a bad effect on somebody/something	**0.524**
	to reach a particular level	0.519
	to experience something difficult or unpleasant	0.414
lend support	to give or provide help, support, etc.	**0.743**
	to give money to somebody on condition that they pay it back over a period of time and pay interest on it	0.404
	to give a particular quality to a person or a situation	0.375
meet fear	to experience something, often something unpleasant	**0.669**
	to be in the same place as somebody by chance and talk to them	0.557
	to touch something; to join	0.535
promote intolerance	to help something to happen or develop	**0.385**
	to move somebody to a higher rank or more senior job	0.116
	to move a sports team from playing with one group of teams to playing in a better group	0.085

Table 5: Examples of the nearest definitions from Oxford Learner's Dictionary that could interpret the given metaphoric expressions based on the cosine similarity between the noun of the metaphoric expression and the verb sense.

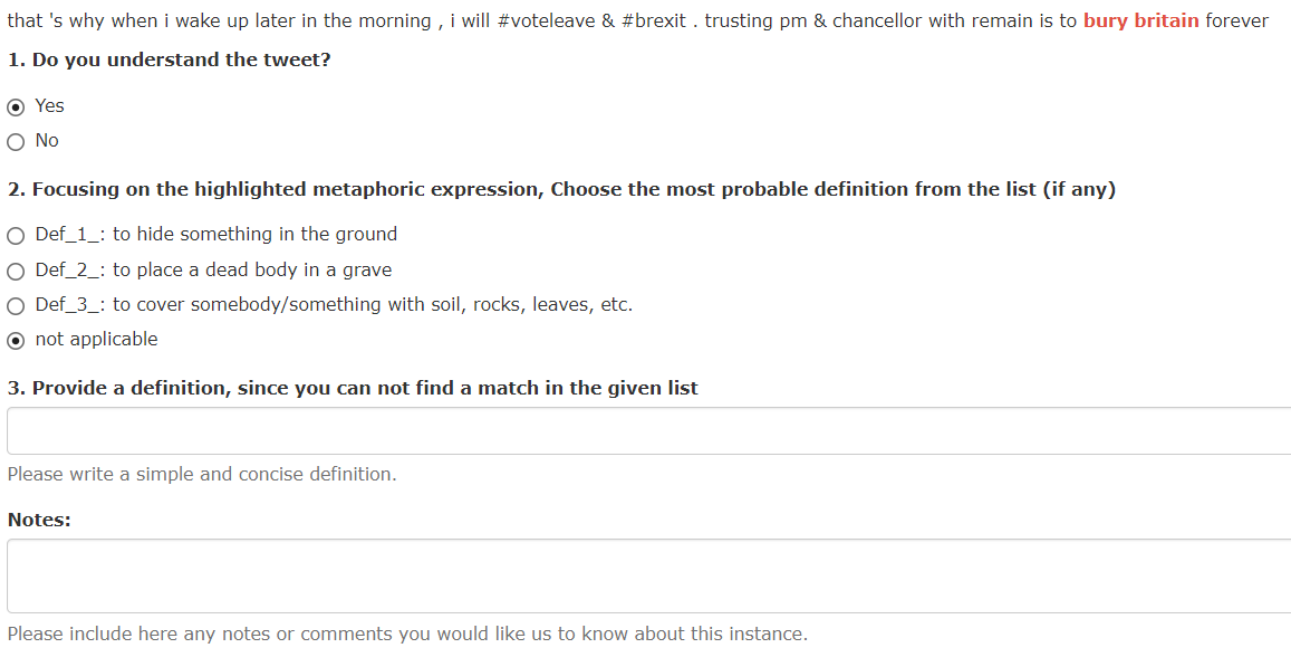

Figure 1: An example from the annotation task given to the annotators on MTurk.

instances which have majority vote over 60% with a moderate agreement strength of 0.48 in terms of Fleiss' kappa.

5.2. Analysis

Definition Choice: In 70.82% of the cases, the annotators preferred to choose a definition from the suggested ones. On the other hand, they opt to provide their own definition of the metaphoric expression either in the cases of encountering uncommon usage of the verb in a metaphoric way such as *"wash off all your sadness"*, *"open your heart"* and *"bring cheers"* or if the suggested definitions from the dictionary do not accurately reflect the metaphoricity of the expression such as *"take a stand"*, *"make a conscious effort"* and *"reduce anxiety"*. Figure 2 illustrates the percentage of choosing to provide an interpretation for each annotator. One of the annotators always preferred to write his own interpretations (definitions) of the metaphoric expressions; he provided an interpretation for 88.16% of the instances. We plan, as a future work, to validate the annotators' provided definitions by either 1) looking into ranking measures such as the "mean average precision" or "mean reciprocal rank" or 2) performing a review by an expert na-

	Annotated Instances	Fleiss' Kappa	Agreement Strength
Top three annotators	1,353	0.436	Moderate
Top four annotators	1,352	0.425	Moderate
Top five annotators	1,304	0.386	Fair
All six annotators	1,282	0.27	Fair
high quality subset with majority vote >60 (six annotators)	**676**	**0.48**	**Moderate**

Table 6: Dataset analysis based on the agreement strength per number of annotators.

tive annotator who will go through the write-in definitions and consolidate them.

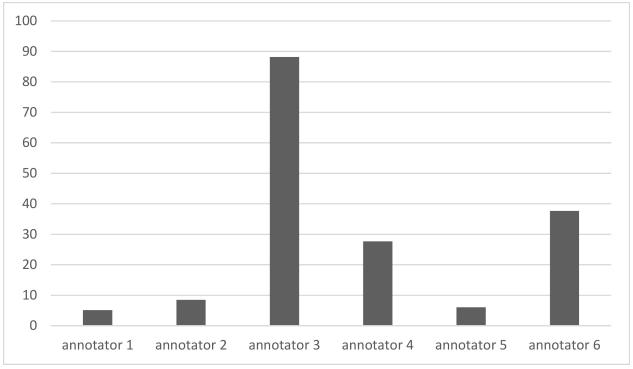

Figure 2: Percentage of providing a customised interpretation (definition) per annotator.

Points of (Dis-)agreements: We analysed the points of (dis-)agreements between the annotators. Almost half of the provided annotations have a majority vote greater than 60% which yields a moderate IAA of 0.48 in terms of Fleiss kappa. The majority of disagreements centred around whether the suggested definition in the dictionary is enough to represent the metaphoric sense of the expression or not. Tables 7 and 8 shows examples of the agreements and disagreements between the six annotators. For example, the six annotators agreed that the suitable definition for the metaphoric expression *"release pain"* is the one from the Oxford Learner's Dictionary as shown in in Table 7 whereas they opt for providing their own definition for the metaphoric expression *"brushing up my german"*. Table 9 gives more information about the statistics of the annotated dataset.

The effect of tweets: Although the context where the metaphoric expression appears is important to understand the expression, the noisy ungrammatical text of the tweets affected the annotation process. We observed that two annotators find it difficult to understand around 50 tweets, therefore, they skipped them which affected the overall agreement. The rest of the annotators did not skip them but they provided some notes about them. According to the annotators the reasons behind skipping these tweets were: 1) they do not understand the topic of the tweet at all (sarcasm, science fiction or games); 2) there is not enough information about the noun to give a definition; 3) the tweet is not grammatically correct to convey a meaning.

Annotators' Experience: Some of the annotators raised the issue of using metaphors while defining a metaphor. The annotators had to make sure not to use metaphors when writing their own definitions, which they found difficult. For instance, one annotator encountered the metaphoric expression *"stand a chance"* and she wanted to write *"to take/have an opportunity"* which is another metaphor; therefore she had to think of another definition using literal words. The majority of annotators agreed that sometimes using a metaphor is the easiest way to express what the author wants to say and here lies the difficulty of the metaphor interpretation task itself. It is worth mentioning also that the genre of the tweets affected the annotators' experience. Some annotators found many of the metaphoric expressions in the political tweets very straightforward and obvious, but when it came to emotional or motivational metaphors they found them slightly harder to define in simple terms.

6. Dataset Publication as Linked Data

We believe that this resource can be used to enrich Wiktionary (or any lexical resource) by including a metaphor category similar to the idioms one. Therefore, in order to provide access to the data and promote reusability, we will provide the dataset as a linked open dataset. As the original annotators chose the definitions from the provided suggestions obtained from the Oxford Learner's Dictionary, which is not possible to republish due to copy-rights, we instead provide the links by reference to the website. In particular, we refer to the sense *IDs* as links and publish the annotations in the Resource Description Framework (RDF) as linked data as shown in Figure 3.

In this case, we provide a direct link to the definition and a hash of the definition, which can be used to verify the definition has not changed. A script is provided with the download that fetches the definitions, verifies that they match the required hash and produces the results as comma-separated values. The customised definitions by the annotators will be provided as well.

7. Conclusions

In this paper, we presented our work on creating the first gold standard dataset for metaphor interpretation along the more complex "definition generation" approach which provides full explanation of a given metaphoric expression. We demonstrated our methodology on preparing the dataset which combines an automatic retrieval approach with manual annotation to ensure wide coverage, accuracy and consistency. We were able to employ lexical resources, word embeddings and semantic similarity to assist in the annotation process with the aim to reduce the cognitive load on the annotators and to address the subjectivity of interpreting metaphoric expressions. As a result, we annotated around 1,500 metaphoric verb-direct object expressions in tweets. Our methodology and annotation scheme can be generalised to annotate metaphors of any syntactic structure in any text genre/type. We believe that this dataset will be invaluable for the development and evaluation of approaches for metaphor interpretation.

We will release the full set of ∼1,500 annotated instances, including the annotators customised definitions as linked

Metaphoric Expression	Definition	Source
repay the tremendous support	to give something to somebody or do something for them in return for something that they have done for you	Oxford
release old emotional pain	to express feelings such as anger or worry in order to get rid of them	
ruin all the fun	to damage something so badly that it loses all its value, pleasure, etc.; to spoil something	
brushing up my german	to improve on something that one used to excell at	annotator provided
defeating brexit	to defeat the opposing group, argument, party etc.	
ramp up production	to increase the rate of production of somethings	

Table 7: Examples of agreements among all annotators (100% majority vote).

Metaphoric Expression	Definition	Source
take control	to capture a place or person; to get control of something	Oxford
take a minute	to need or require a particular amount of time	
finds fear	to have a particular feeling or opinion about something	
checked out this new friend	to look at information showing or pictures of a new supporter	annotator provided
wash off all your sadness	to stop feeling a particular emotion	
brings cheers	to make someone/group of people to feel a certain emotion	

Table 8: Examples of disagreements among all annotators (less than 60% majority vote).

```
<#anno1>
  <#metaphor> "ignited a new passion"@en ;
  <#interpretation> [
    dc:source
     <https://www.oxfordlearnersdictionaries.com/definition/english/ignite_1#ignite_sng_1>;
    <#hash> "70B6783C04E770A02409174F97089E58";
    <#annotators> 2;
    <#majorityVote> 0.334;
    <#cosineSim> 0.520;
  ],
  [
    <#hash> "B501696811F1198BCFF3435E3822B571", "04BA979E7D9900B23321CE7318265E5F",
"433578FE1D3F6301616A61D732927B54", "EA9FA1DB0B8050D611715B92E7567B12";
    <#annotators> 4;
    <#majorityVote> 0.667;
    skos:definition "to cause something to happen or begin", "to make someone start
feeling a particular way", "made people more interested than ever", "to start
something/feelings"
  ]
```

Figure 3: Section of the dataset as published as linked data.

data in RDF format to promote reusability and to facilitate its incorporation to other lexicons such as Wiktionary and WordNet. Moreover, we will release the high-quality subset of the data where we only consider the instances with more than 60% majority agreement and a moderate inter-annotator agreement of 0.48 in terms of Fleiss' kappa. As a future work, we plan to consolidate the annotators' provided definitions by looking into ranking measures such as the "mean average precision" or "mean reciprocal rank". A native speaker will go through the provided definitions and set a reference one in order to apply these methods.

8. Acknowledgements

This work was supported by: 1) Science Foundation Ireland under grant number 12/RC/2289_2 (Insight). 2) the Prêt-à-LLOD project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 825182.

9. References

Agirre, E. and Stevenson, M. (2007). Knowledge sources for WSD. In E. Agirre et al., editors, *Word sense disambiguation: Algorithms and applications*, chapter 8, pages 217–251. Springer, NY, USA.

Aspect	Value
total # of tweets	1394
# of unique (lemmatised) verb-direct objects	1394
average tweet length	22.36
# of skipped (non-understandable) tweets by all annotators	5
maximum # of skipped tweets by one annotator	50
minimum # of skipped tweets by one annotator	1
total # of annotated instances by six annotators	1,301
total # of annotated instances by three annotators	1,353
maximum # of instances with annotator's provided definition	1,147
% of instances annotated with agreement majority vote greater than 60%	52.02%
% of instances annotated with agreement majority vote less than 60%	47.9%
% of customised definitions by all annotators	29.2%

Table 9: Statistics of the annotated dataset.

Barbu, E., Martín-Valdivia, M., Martínez-Cámara, E., and López, L. (2015). Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42:5076–5086, July.

Bingel, J., Paetzold, G., and Søgaard, A. (2018). Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, NM, USA, August. Association for Computational Linguistics.

Bizzoni, Y. and Lappin, S. (2018). Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the first Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana, June. Association for Computational Linguistics.

Bollegala, D. and Shutova, E. (2013). Metaphor interpretation using paraphrases extracted from the web. *PLoS ONE*, 8(9):1–10, September.

Burnard, L. (2009). About the British National Corpus. `http://www.natcorp.ox.ac.uk/corpus/index.xml`.

Cameron, L. (2003). *Metaphor in educational discourse*. Advances in Applied Linguistics. Continuum, London, UK.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Blackwell Publishing.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ide, N. and Véronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time. In *Proceedings of the Workshop on Knowledge Bases and Knowledge Structures*, pages 257–266, Tokyo, Japan.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C., and Wissik, T. (2018). European lexicographic infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress*, Ljubljana, Slovenia, july.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago, USA.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Martin, J. H. (1990). *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA, USA.

Meyer, C. M. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger et al., editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford: Oxford University Press, November.

Mohammad, S. M., Shutova, E., and Turney, P. D. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, *Sem '16, pages 23–33, Berlin, Germany.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2), February.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1532–1543, Doha, Qatar, October.

Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Shutova, E. and Teufel, S. (2010). Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC '10, pages 255–261, Malta, May.

Shutova, E., de Cruys, T. V., and Korhonen, A. (2012). Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of the 24th International Conference on Computational Linguistics*, COLING '12, pages 1121–1130, Mumbai, India, December. Association for Computational Linguistics.

Shutova, E., Kiela, D., and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, pages 160–170, San Diego, CA, USA, June.

Shutova, E. (2010). Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '10, pages 1029–1037, Los Angeles, CA, USA, June. The Association for Computa-

tional Linguistics.

Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623, December.

Stewart, D. (1971). Metaphor and paraphrase. *Philosophy and Rhetoric*, 4(2):111–123.

Wolska, M. and Clausen, Y. (2017). Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zayed, O., McCrae, J. P., and Buitelaar, P. (2019). Crowdsourcing a high-quality dataset for metaphor identification in tweets. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*, LDK '19, Leipzig, Germany, May.