Adaptation of Word-Level Benchmark Datasets for Relation-Level Metaphor Identification

Omnia Zayed, John P. McCrae, Paul Buitelaar
Insight SFI Research 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 processing and understanding has attracted the attention of many researchers recently with an increasing number of computational approaches. A common factor among these approaches is utilising existing benchmark datasets for evaluation and comparisons. The availability, quality and size of the annotated data are among the main difficulties facing the growing research area of metaphor processing. The majority of current approaches pertaining to metaphor processing concentrate on word-level processing due to data availability. On the other hand, approaches that process metaphors on the relation-level ignore the context where the metaphoric expression. This is due to the nature and format of the available data. Word-level annotation is poorly grounded theoretically and is harder to use in downstream tasks such as metaphor interpretation. The conversion from word-level to relation-level annotation is non-trivial. In this work, we attempt to fill this research gap by adapting three benchmark datasets, namely the VU Amsterdam metaphor corpus, the TroFi dataset and the TSV dataset, to suit relation-level metaphor identification. We publish the adapted datasets to facilitate future research in relation-level metaphor processing.

1 Introduction

Metaphor is a ubiquitous figurative device that represents the interaction between cognition and language (Cameron and Low, 1999). A metaphor contains an implied analogy where a concept (represented by a word sense) is borrowed to represent another concept by exploiting common or single properties of both concepts. Generally, a metaphor has two main components, the tenor and the vehicle; the relation between them is called the ground. The tenor represents the topic of the metaphor while the vehicle is the term used metaphorically and the ground gives the metaphor its meaning (End, 1986). Perceiving these components is essential to fully comprehend the metaphor. In this work, we adopt the conceptual metaphor theory (CMT) by Lakoff and Johnson (1980) to view metaphor where there is an underlying mapping between a source domain (the vehicle) and a target domain (the tenor). For example, a concept such as *"fragile object"* (source domain/vehicle) can be borrowed to express another such as *"emotions"* (target domain/tenor). This conceptual metaphor *"Emotions are Fragile Objects"* can be expressed in our everyday language in terms of linguistic metaphors such as *"shattered my emotions"*,*"break his soul"*,*"crushed her happiness"*, *"fragile emotions"* and *"brittle feelings"*.

Due to their nebulous nature, metaphors are quite challenging to comprehend and process by humans, let alone computational models. This intrigued many researchers to develop various automatic techniques to process metaphor in text. Metaphor processing has many potential applications, either as part of natural language processing (NLP) tasks such as machine translation (Koglin and Cunha, 2019), text simplification (Wolska and Clausen, 2017; Clausen and Nastase, 2019) and sentiment analysis (Rentoumi et al., 2012) or in more general discourse analysis use cases such as in analysing political discourse (Charteris-Black, 2011), financial reporting (Ho and Cheng, 2016) and health communication (Semino et al., 2018).

The computational processing of metaphors can be divided into two tasks, namely metaphor identification and its interpretation. While the former is concerned with recognising the metaphoric word or expression in a given sentence, the latter focuses on discerning the meaning of the metaphor. Metaphor identification is studied more extensively than metaphor interpretation, in part due to the availability of datasets. Identifying metaphors in

Proceedings of the Second Workshop on Figurative Language Processing, pages 154–164

https://doi.org/10.18653/v1/P17

text can be done on either the sentence, grammatical relation or word levels. Sentence-level approaches classify the whole sentence that contains the metaphoric word/expression without explicit annotation of the source and target domain words. Relation-level metaphor identification focuses on certain grammatical relations by looking at pairs of words where both the source and target domain words are classified as a metaphoric expression. It is also referred to as phrase-level metaphor identification due to the way a sentence is divided into sub-phrases with various syntactic structures (we use these two terms indistinguishably in the context of this paper). The most commonly studied grammatical relations are verb-noun and adjective-noun relations where the metaphoricity of the verb or the adjective (source domain/vehicle) is discerned given its association with the noun (target domain/tenor). Finally, word-level metaphor identification approaches treat the task as either sequence labelling or single-word classification. In both methods, only the source domain words (vehicle) are labelled either as metaphoric or literal given the context. Many approaches are designed to identify metaphors of different syntactic types on the word-level but the most frequently studied ones are verbs.

In this paper, we are interested in relation-level metaphor identification focusing on the data availability for this level of processing. The next section explains, in detail, the difference between word-level and relation-level metaphor analysis highlighting the research gap that we aim to tackle.

2 Word-Level vs. Relation-Level Metaphor Analysis

Although the main focus of both the relation-level and word-level metaphor identification is discerning the metaphoricity of the vehicle (source domain words), relation-level approaches attend to the tenor (target domain words) associated with the vehicle under study during processing the metaphor which, in turn, gives the model a narrower focus in a way that mimics human comprehension of metaphors. Thus, processing metaphors on the word-level could be seen as a more general approach where the tenor of the metaphor is not explicitly highlighted as well as the relation between the source and the target domains. On the other hand, relation-level metaphor identification explicitly analyses the tenor and the relation between the source and the target domains. Figure 1 illustrates the difference between the levels of metaphor identification.

Stowe and Palmer (2018) highlighted the importance of integrating syntax and semantics to process metaphors in text. Through a corpus-based analysis focusing on verb metaphors, the authors showed that the type of syntactic construction (dependency/grammar relation) a verb occurs in influences its metaphoricity.

Relation-level metaphor processing requires an extra step to identify the grammatical relations (i.e. dependencies) that highlight both the tenor and the vehicle. Thus, it might be seen that processing metaphors on the word-level is more straight forward and raises the question: why do we need relation-level metaphor identification? Relation-level metaphor identification can be used to support metaphor interpretation and cross-domain mappings. Metaphor interpretation focuses on explaining or inferring the meaning of a given metaphorical expression. As explained earlier, the tenor (target domain words) is the topic of the metaphor that gives the metaphor its meaning. Therefore, relation-level identification is an important initial step that facilitates inferring the meaning of a given expression. Cross-domain mappings focuses on identifying the relation between the source and target domain concepts in a way that mimics the human formulation of metaphors. This mapping is produced by studying a set of multiple metaphorical expressions that describe one concept in terms of another. Hence, identifying metaphors on the relation-level is employed to support such mappings in order to create knowledge-bases of metaphoric language.

The levels of processing metaphors should be taken into consideration when designing and developing a computational model to identify metaphors and hence choosing the annotated dataset accordingly for evaluation and comparison. Shutova (2015), Parde and Nielsen (2018) and Zayed et al. (2019) provided extensive details about existing datasets for metaphor identification in English text. The authors highlighted the level of annotation for each dataset among other properties. The widely used benchmark datasets are TroFi (Birke and Sarkar, 2006), VU Amsterdam metaphor corpus (VUAMC) (Steen et al., 2010) and MOH (Mohammad et al., 2016) for word-level metaphor identification, whereas TSV (Tsvetkov et al., 2014), the adaptation of MOH by Shutova et al. (2016), and

Example: *The diligent reporter said that the new rules sparked a heated debate in the media on the weekend.*

Word-level Metaphor Identification:

implicit relations and no identification of the tenor (target domain words)

*The diligent reporter said that the new rules **sparked** a **heated** debate **in** the*

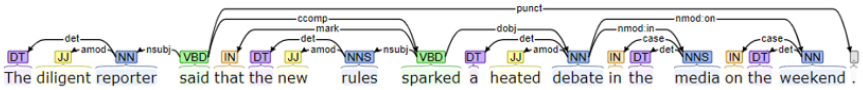
0 0 0 0 0 0 0 0 1 0 1 0 1 0

*media **on** the weekend.*

0 1 0 0

Relation-level (phrase-level) Metaphor Identification:

explicit relations and explicit identification of the tenor (target domain words)



grammar relation	expression	
amod	**diligent** reporter	0
nsubj	reporter **said**	0
nsubj	rules **sparked**	1
amod	**heated** debate	1
dobj	**sparked** debate	1
case	**in** media	1
case	**on** weekend	1

Figure 1: An illustration of the difference between word-level and relation-level metaphor identification. Stanford CoreNLP is used to generate the dependencies.

Zayed's Tweets (Zayed et al., 2019) datasets are utilised for relation-level metaphor identification.

Approaches addressing the task on the word-level are not fairly comparable to relation-level approaches since each task deals with metaphor identification differently. Therefore, given the distinction of the tasks definition, the tradition of previous work in this area is to compare the word-level metaphor identification approaches against each other on either the TroFi, VUAMC or MOH datasets. On the other hand, relation-level approaches are compared against each other on either the TSV, Shutova's adaptation of MOH or Zayed's Tweets datasets. Although, the VUAMC is the most well-known and widely used corpus for metaphor identification, it is not possible to apply it to relation-level metaphor identification without further annotation effort. This is also the case for the TroFi dataset which is one of the earliest balanced datasets annotated to identify metaphoric verbs on the word-level. On the other hand, the TSV dataset is the only available annotated dataset for relation-level metaphor identification that addresses adjective-noun grammatical relations. However, the main issue with this dataset is the absence of full sentences in the training set leaving a relatively small test set that has full sentences which limits its usage for state-of-the-art approaches that rely on using the full context.

One limitation of word-level annotation is the implicit level of analysis discussed earlier. Direct mapping from word-level to relation-level annotation is not straight forward and requires extra annotation effort. Consider the following examples that contain verb metaphors:

(1) The speech stirred the emotions.

(2) "history will judge you at this moment."

(3) Citizens see hope in the new regulations.

Identifying metaphoric verbs on the word-level will result in recognising the verbs *"stirred"*, *"judge"* and *"see"* as metaphoric in examples (1), (2) and (3), respectively. In example (1), both the subject and the object are responsible for the metaphoricity of the verb; while in example (2), the subject gave the verb its metaphoricity and in example (3) the object did. This is done implicitly in word-level annotation/identification. On the other hand, if we consider relation-level processing, the tenor associated with the verb has to be explicitly highlighted. Thus, annotating the above examples on the relation-level focusing on verb-direct object relations (i.e. dobj) will result in identifying the expressions *"stirred the emotions"* and *"see hope"*

as metaphoric in examples (1) and (3), respectively and ignoring example (2) since *"history will judge"* is a subject-verb (i.e. *nsubj*) relation. Therefore, adapting existing datasets annotated on the word-level is required to arrive at explicit analysis of the tenor and the relation between the source and the target domains.

In this work, we take a step towards filling this research gap by introducing an adapted version of benchmark datasets to better suit relation-level (phrase-level) metaphor identification. We adapt the VUAMC and the TroFi dataset to identify verb-noun metaphoric expressions. Moreover, we extend the relation-level metaphor identification TSV dataset by providing context for the adjective-noun relations in its training set. We publish the adapted version of the datasets according to the licensing type of each of them to facilitate research on metaphor processing.

3 Related Work

This work is inspired by Tsvetkov et al. (2014) and Shutova et al. (2016) who attempted to adapt existing word-level metaphor identification datasets to suit their relation-level (phrase-level) identification approaches. Shutova et al. (2010) was the first to create an annotated dataset for relation-level metaphor identification. The Robust Accurate Statistical Parsing (RASP) parser (Briscoe et al., 2006) was utilised to extract verb-subject and verb-direct object grammar relations from the British National Corpus (BNC) (Burnard, 2007). The dataset comprises around 62 verb-noun pairs of metaphoric expressions, where the verb is used metaphorically given the complement noun (tenor).

The TroFi dataset, which was designed to classify particular literal and metaphoric verbs on the word-level, was adapted by Tsvetkov et al. (2014) in order to extract metaphoric expressions on the relation-level. The authors parsed the original dataset using the Turbo dependency parser (Martins et al., 2010) to extract subject-verb-object (SVO) grammar relations. The final dataset consists of 953 metaphorical and 656 literal instances. In the same work, Tsvetkov et al. also prepared a relation-level metaphor identification dataset, referred to as the TSV dataset, focusing on adjective-noun grammar relations. We will further describe this dataset in Section 4.

More recently, Shutova et al. (2016) adapted the benchmark MOH dataset, which was initially created to extract metaphoric verbs on the word-level, to suit relation-level metaphor identification of verb-noun relations. Verb-direct object and verbs-subject dependencies were extracted and filtered yielding a dataset of 647 verb–noun pairs, out of which 316 instances are metaphorical and 331 instances are literal.

To the best of our knowledge, there is no attempt to adapt the benchmark VU Amsterdam metaphor corpus, referred to as VUAMC, to suit relation-level metaphor identification. This has discouraged other researchers focusing on relation-level approaches to employ this dataset such as the work done by Rei et al. (2017), Bulat et al. (2017), Shutova et al. (2016) and Tsvetkov et al. (2014) who did not evaluate or compare their approaches using this dataset. In this paper, we introduce the first adapted version of the VUAMC. Furthermore, we adapt the TroFi and the TSV datasets to better suit relation-level metaphor processing.

4 Datasets

As mentioned in Section 2, the widely used benchmark datasets for word-level metaphor identification are TroFi, VUAMC and MOH datasets, while TSV, Shutova's adaptation of MOH and Zayed's Tweets datasets are commonly used for relation-level metaphor identification. Table 1, adapted from (Zayed et al., 2019), revisits the properties of each dataset . In this work, we focus on the word-level VUAMC, and the TroFi dataset in addition to the relation-level TSV dataset as the largest and extensively used datasets for metaphor identification. In this section, we discuss each dataset in detail.

The VU Amsterdam Metaphor Corpus (VUAMC)[1], introduced by Steen et al. (2010), has become one of the most well-known metaphor corpus existing nowadays. It is the largest corpus annotated for metaphors and has been used extensively to train, evaluate and compare models that identify metaphors on the word-level. The corpus consists of 117 randomly selected texts from the BNC Baby version which comprises various text genres, namely academic, conversation, fiction and news. The corpus is annotated for metaphors on the word-level, regardless of the word's syntactic type, through a collaborative annotation scheme. The employed annotation scheme is referred to as the metaphor identification procedure (MIPVU) by

[1] Also referred to, in literature, as the VUA dataset or the VUA metaphor corpus.

Level of analysis	Dataset	Syntactic structure	Text type	Size	% Metaphors
word-level	TroFi Example Base (Birke and Sarkar, 2006)	verb	50 selected verbs (News)	3,727 sentences	57.5%
	VUAMC (Steen et al., 2010)	all POS	known-corpus (The BNC)	~16,000 sentences (~200,000 words)	12.5%
	MOH (Mohammad et al., 2016)	verb	selected examples (WordNet)	1,639 sentences	25%
relation-level (phrase-level)	TSV (Tsvetkov et al., 2014)	adjective–noun	selected examples (Web)	~2,000 adj-noun pairs	50%
	adaptation of MOH (Shutova et al., 2016)	verb-direct object; subject-verb	selected examples (WordNet)	647 sentences	48.8%
	Zayed's Tweets (Zayed et al., 2019)	verb-direct object	Tweets (general and political topics)	~2,500 tweets	54.8%

Table 1: Statistics of the widely used benchmark datasets for linguistic metaphor identification.

which a strong inter-annotator agreement of 0.84 is obtained, in terms of Fleiss' kappa (Fleiss, 1971), among four annotators. The dataset is published in an XML format; Figure 2 shows an example of the corpus where the metaphoric words are tagged as *function="mrw"*.

```
<w lemma="such" type="DT0">Such </w>
<w lemma="language" type="NN1">language </w>
<w lemma="focus" type="VVD-VVN">
  <seg function="mrw" subtype="PP" type="met"
  vici:morph="n">focused</seg>
</w>
<w lemma="attention" type="NN1">attention </w>
<w lemma="on" type="PRP">
  <seg function="mrw" type="met" vici:morph="n">on</seg>
</w>
<w lemma="the" type="AT0">the </w>
<w lemma="individual" type="NN2">individuals </w>
<w lemma="or" type="CJC">or </w>
<w lemma="group" type="NN2">groups </w>
<w lemma="who" type="PNQ">who </w>
<w lemma="be" type="VBD">were </w>
<c type="PUQ">'</c>
<w lemma="break" type="VVG">
  <seg function="mrw" type="met" vici:morph="n">breaking
  </seg></w>
<w lemma="the" type="AT0">the </w>
<w lemma="law" type="NN1">law</w>
```

Figure 2: An example form the VU Amsterdam metaphor corpus (VUAMC) showing the data annotation format and the metaphoric words labelled with the metaphor-related word tag *(function="mrw")*.

The NAACL 2018 Metaphor Shared Task (Leong et al., 2018) employed the VUAMC in order to develop, train and test systems to identify metaphors on the word-level. The shared task consisted of two tracks, which are 1) *All Part-Of-Speech (POS)* to identify nouns, verbs, adverbs and adjectives that are labelled as metaphorical; 2) *Verbs* track which is concerned only with identifying metaphorical verbs. All forms of the verbs: *"be, do,* and *have"* are excluded for both tracks. The corpus is then divided into training and test sets according to the focus of each track. A script is provided to parse the original VUAMC.xml file[2] which contains the corpus, since the corpus is not directly downloadable due to licensing restrictions. In this paper, we utilise the dataset from the *Verbs* track from this shared task. Table 2 shows the statistics of the dataset as highlighted in (Leong et al., 2018).

Data	Training			Test		
	#texts	#tokens	%M	#texts	#tokens	%M
Academic	12	4,903	31 %	4	1,259	51%
Conversation	18	4,181	15%	6	2,001	15%
Fiction	11	4,647	25%	3	1,385	20%
News	49	3,509	42 %	14	1,228	46%

Table 2: Statistics of the training and test data in the "Verbs" track in the NAACL metaphor shared task. %M is the percentage of metaphors.

The main limitation of the VUAMC, and any dataset that stems from it, is that it only suits the identification of metaphors on the word-level. Thus, it is not possible to apply the VUAMC in its current state to relation-level metaphor identification and there are no larger dataset designated to support relation-level metaphor identification since the size of Shutova's adaptation of MOH and Zayed's Tweets datasets is relatively small for training state-of-the-art neural models.

The TroFi Dataset is one of the earliest metaphor identification datasets introduced by Birke and Sarkar (2006, 2007). The dataset focuses on the metaphoric usage of 50 selected verbs and comprises 3,727 English sentences extracted from the 1987-1989 Wall Street Journal (WSJ) corpus. The metaphoricity of the selected verbs on the word-level is identified by manual annotation. The inter-annotator agreement was calculated on a random sample of 200 annotated

[2]The VUAMC was available online at: http://ota.ahds.ac.uk/headers/2541.xml but the website was unresponsive at the time of this publication.

sentences scoring 0.77 in terms of Cohen's kappa (Cohen, 1960) among two annotators. The dataset had been used to evaluate the performance of many word-level metaphor identification systems. In order to use this dataset for relation-level metaphor identification of verb-noun relations, further annotation is required to highlight the complementing noun (tenor) of each metaphoric verb (vehicle).

The TSV Dataset (Tsvetkov et al., 2014) was created to support relation-level metaphor identification approaches that focus on adjective-noun grammatical relations. The dataset comprises ∼2,000 adjective-noun pairs which were selected manually from collections of metaphors on the Web. It is divided into 1,768 pairs as a train set and 200 pairs as a test set. As mentioned earlier, only the test set contains the full sentences which was obtained from the English TenTen Web corpus (Jakubíček et al., 2013) by utilising SketchEngine[3] (Kilgarriff et al., 2014). The annotation scheme depended on the intuition of the human annotators to define the metaphoric expressions. An inter-annotator agreement of 0.76, in terms of Fleiss' kappa, was obtained among five annotators on the test set. The main limitation of this dataset is the absence of the full sentences in the training set which forces the models employing it to either ignore the context that surrounds the adjective-noun pairs or to use the small test set in a cross-validation experimental setting which makes the model prone to overfitting.

5 Dataset Adaptation Methodology

In this section, we discuss the methodology of adapting the VUAMC, TroFi and TSV datasets to better suit relation-level (phrase-level) metaphor processing.

5.1 VUAMC and TroFi dataset Adaptation

As discussed earlier, relation-level metaphor identification focuses on a specific grammatical relation that represents the source and target domains of the metaphor. The datasets that are initially annotated for word-level processing have the source domain words (vehicle) labelled as a metaphor regardless of its tenor since it is word-by-word classification. Therefore, in order to adapt them to suit relation-level processing, the associated target domain words (tenor) need to be identified.

[3] http://www.sketchengine.eu

Our approach towards adapting the datasets annotated on the word-level is as follows:

1. select the benchmark dataset which is originally annotated on the word-level;
2. extract particular grammatical relations focusing on the vehicle as the head of the relation (e.g. the verb in a *dobj* or adjective in *amod* relation);
3. retrieve the gold labels from the original dataset based on the metaphoricity of the vehicle;
4. verify the correctness of the retrieved relations and the assigned gold label.

In this work, we employ the Stanford dependency parser (Chen and Manning, 2014) to identify grammar relations. The recurrent neural network (RNN) parser, pre-trained on the WSJ corpus, is used from within the Stanford CoreNLP toolkit (Manning et al., 2014).

For the VUAMC adaptation, as discussed in Section 4, we utilise the training and test splits provided by the NAACL metaphor shared task in the *Verbs* track. We focus on this track since we are interested in verb-noun relations. The verbs dataset consists of 17,240 annotated verbs in the training set and 5,874 annotated verbs in the test set. First, we retrieved the original sentences of these verbs from the VUAMC since the shared task released their *ids* and the corresponding gold labels. This yielded around 10,570 sentences in both sets. Then, we parsed these sentences using the Stanford parser and extracted the verb-direct object (i.e. *dobj*) relations, discarding the instances with pronominal or clausal objects[4]. The extracted relations are then filtered to exclude parsing-related errors. Manual inspection is done to ensure that, in a given *dobj* relation, the verb is metaphoric due to the associated object (more details will be given in Section 6). The final adapted dataset comprises 4,420 sentences in the training set and 1,398 in the test set.

For the TroFi dataset adaptation, we utilise the 3,737 manually annotated English sentences from Birke and Sarkar (2006)[5]. Each sentence contains either literal or metaphorical use for one of 50 English verbs. These sentences were parsed to extract dependency information. Then, we filtered

[4] This is done automatically using regular expressions to select the grammatical relations with certain POS tags.

[5] http://natlang.cs.sfu.ca/software/trofi.html

the extracted relations to only select the *dobj* relations that include verbs from the 50 verbs list and to eliminate mis-parsing cases. This resulted in a dataset of 1,535 sentences.

Table 4 shows the statistics of the adapted VUAMC and TroFi dataset after applying the quality assessment in Section 6. Examples of the annotated sentences from the adapted VUAMC and TroFi dataset are listed in Table 5 as they appear in the adapted relation-level version.

5.2 TSV Dataset Adaptation

Our main goal when adapting the TSV relation-level dataset is to provide a context for the balanced training set of 1,768 metaphoric and non-metaphoric adjective-noun pairs. Table 3 gives examples for the adjective-noun expressions appearing in the original TSV training set[6]. This will allow the computational models to benefit from the contextual knowledge that surrounds the expression. The method used to achieve this goal is to query the Twitter Search API[7] using the adjective-noun pairs and retrieve tweets as the context around these expressions. Among the main motivations behind selecting the user-generated text (tweets) to expand this dataset are: 1) to encourage and facilitate the study of metaphors in social media contexts; 2) the availability of Twitter data as well as the ease of use of the Twitter API.

Metaphor	Non-metaphor
blind faith	blind patient
deep sorrow	deep cut
empty life	empty house
fishy offer	frozen food
heated criticism	heated oven
raw idea	raw vegetables
shallow character	shallow water
warm smile	warm day

Table 3: Examples of the annotated adjective-noun expressions in the TSV training dataset.

For each expression in the training set, a tweet is retrieved given that its length is more than 10 words and it does not contain more than four hashtags or mentions to ensure that the retrieved context has enough information. Then, the tweets are preprocessed to remove URLs and duplicate tweets. This yielded an adapted training set of 1,764 tweets that contains metaphoric and non-metaphoric expressions of adjective-noun relations. The next step is to ensure the quality of the retrieved content in terms of keeping the metaphoricity of the original expression. This is done manually as will be discussed in the next section. Table 4 provides the statistics of the adapted TSV training dataset after expanding it with full sentences (tweets). Examples of the annotated tweets from the adapted TSV training dataset are given in Table 5.

[6] https://github.com/ytsvetko/metaphor

[7] https://developer.twitter.com/en/docs/api-reference-index

6 Quality Assessment and Enhancement

In order to assess the quality of the adapted datasets, we suggested a preliminary quality assessment scheme and tested it through an initial experiment on a randomly sampled subset from each dataset. We then employed this scheme to ensure the quality of the whole datasets.

6.1 Initial Quality Assessment Experiment

In this pilot experiment, we randomly sampled 100 sentences from each dataset. We then asked two native English speakers with background in (computational) linguistics to manually identify the quality of the retrieved sample. Since the datasets were previously annotated, our main concerns for evaluation are as follows:

For each instance in the VUAMC and the Trofi dataset:

1. to check that the *dobj* dependency is syntactically valid;
2. to ensure that the verb is metaphoric due to the associated object;
3. check if the expression is really a metaphor.

For each instance in the TSV dataset:

1. to ensure that the tweet is in understandable English;
2. to check that the *amod* dependency is syntactically valid;
3. to ensure that the provided context (scrapped tweets) preserves the metaphoricity of the expression.

For the VUAMC, the annotators agreed that in 81.1% of the metaphoric cases, the the metaphoricity of the verb is due to the complement direct-object. However, the annotators raised some issues regarding the original annotation of the VUAMC using the MIPVU procedure. Their main concerns

	VUAMC* (NAACL metaphor shared task data)		TroFi Dataset	TSV Dataset
	training set	test set		training set
targeted grammar relation	verb-direct object		verb-direct object	adjective-noun
# sentences	4,420	1,398	1,535	1,764
# metaphoric instances	1,675	586	908	881
# non-metaphoric instances	2,745	812	627	883
% metaphors	37.96%	41.92%	59.15%	49.94%

Table 4: Statistics of the adapted VUAMC, TroFi and TSV benchmark datasets. *The training and test sets from the "Verbs" track in the NAACL metaphor shared task.

	ID	Text	Expression	Label
VUAMC	fpb-...._1150_5	I want you to break the news gently to Gran.	break the news	1
	crs-...._35_12	The Community Health Team had major responsibility for assessing children and recommending provision.	recommending provision	0
TroFi	wsj13:9766_16	And even when that loophole was closed, in 1980, the Japanese decided to absorb the tariff rather than boost prices.	absorb the tariff	1
	wsj67:11208_14	Because they 're so accurate, cruise missiles can use conventional bombs to destroy targets that only a few years ago required nuclear warheads.	destroy targets	0
TSV (Training)	1248238...	@sacrebleu141 @FasslerCynthia But it's exactly what the left wants. Trains the people into blind obedience	blind obedience	1
	1248271...	Still have nightmares about waiting tables many years later. Hands down the hardest, most stressful job I've ever had.	stressful job	0

Table 5: Examples from the adapted VUAMC, TroFi and TSV benchmark datasets showing the targeted expression and the provided label (1:metaphor; 0:non-metaphor).

were: 1) the quality of the original annotation which is done on the word-level without explicitly highlighting the tenor or the ground of the metaphor; 2) the consistency of the annotations across the corpus which relied on the annotators' intuition of the basic meaning of the given word and the definition of metaphor. This informal discussion with our expert annotators confirmed our initial concerns that the VUAMC is not really sufficient for metaphor processing in its current state.

The annotators highlighted that the TSV and TroFi datasets have more reliable annotations that align well with the linguistic definition of metaphor than the VUAMC. We attribute that to the following reasons: 1) the TSV dataset was originally annotated on the relation-level with explicit labelling of the tenor; 2) the TroFi dataset comprises carefully selected examples of metaphoric and literal usages for 50 particular verbs. For the TroFi dataset, the annotators agreed that the all the verbs in the random set were used metaphorically due to the associated direct-object without raising any concerns regarding the original annotation of the dataset.

The manual inspection of the random subset of the TSV dataset revealed that, surprisingly, the provided context for the adjective-noun expressions preserved the meaning and the metaphoric sense of all the queried expressions. We suspected that some ambiguous cases might led to ambiguous contexts. For example, the expression *"filthy man"*, which is marked as a metaphor in the dataset, could be used literally to describe the hygienic state of a person; however, the retrieved tweet preserved the metaphoric sense of this expression that describes the morality of a person. This might be due to the following reasons: 1) the conventionality and frequency of adjective-noun metaphoric expressions; 2) the nature of the user-generated (conversational) text of the tweets allows the usage of figurative and metaphoric expressions more frequently than their literal counterparts; 3) the nature of the expressions in the TSV dataset itself in terms of abstractness and concreteness. Further corpus studies are required to investigate this finding.

6.2 Data Filtering and Quality Enhancement

Based on the conclusions of the initial quality assessment, an expert annotator[8] is asked to review the three adapted datasets for quality enhancement following the same scheme. Table 6 includes de-

[8]by "expert" we mean having a computational linguistic background and extensive experience in metaphor processing.

tailed statistics of this quality assessment.

To enhance the **TSV** dataset and ensure its quality, if any of the aforementioned problems is detected the annotator provided another tweet by manually searching Twitter. This is done in a similar way to that adopted by Tsvetkov et al. (2014) while preparing the TSV test set. The annotator noticed that sometimes the tweets contain code-mixed text in English and other language written in Latin letters. These instances are replaced by understandable ones. For the **TroFi** dataset and the **VUAMC**, the annotator corrected the detected parsing errors if possible otherwise the erroneous instances are discarded. Moreover, if the expression is metaphoric due to the associated subject (not the direct object), the expression is corrected and labelled as having an *nsubj* dependency. These expressions are not excluded from the data. Finally, when the annotator disagrees about the metaphoricity of a given instance, it has to be checked first in the original VUAMC dataset and if no annotation error is detected then the instance is flagged to have an annotation disagreement with what the annotator believed to be a metaphor. Aligning with the other two annotators of the pilot experiment, quality and consistency issues are raised about the VUAMC annotation. For example, the verb *"commit"* is labelled five times as a metaphor with the nouns *"acts, bag, government,* and *offence(s)"* and three times as literal with the nouns *"rape,* and *offence(s)"* in very similar contexts. As shown in Table 6, the annotator flagged around 5% of the data for annotation doubt or inconsistency. The majority of the inconsistent annotations revolves around the verbs *"receive, form, create, use, make, recognise, feel, enjoy,* and *reduce"*.

Dataset		Total % accepted by annotator
TSV	the Tweet is in understandable English?	70%
	the relation is syntactically valid?	82.75%
	did the context (tweet) kept the metaphoric sense of the expression?	99.36%
TroFi	the relation is syntactically valid?	98.52%
	the verb is metaphoric or literal due to the associated object?	100%
VUAMC	the relation is syntactically valid?	98.42%
	the verb is metaphoric or literal due to the associated object?	98.5%
	annotation disagreement or inconsistency	5.45%

Table 6: Statistics of the quality assessment of the three adapted datasets showing the total percentage of instances accepted by the annotator.

7 Conclusion and Future Work

In this paper, we took a step towards filling the gap of the availability of large benchmark datasets for relation-level metaphor processing in English text by utilising existing word-level datasets. We employed a semi-automatic approach to adapt the VUAMC to better suit identifying metaphors on the relation-level without the need for extensive manual annotation. We also adapted the TroFi dataset, one of the earliest word-level datasets for metaphor identification of verbs, to support verb-noun metaphor identification. Furthermore, we extended the TSV dataset which was originally annotated on the relation-level focusing on adjective-noun relations by assigning context to its expressions from Twitter. This will encourage research in this area to work towards understanding metaphors in social media. As a result of this work, we publish an adapted version of these benchmark datasets which will facilitate research on relation-level metaphor identification focusing on verb-direct object and adjective-noun relations.

This paper also provides an extensive review of the different levels of metaphor processing and the importance of relation-level metaphor identification. We question the reliability of word-level metaphor processing and annotation in general highlighting the reasons behind that. We provided a brief data analysis in this regard that we are planing to extend as a continuation of this work.

In future work, we will expand the adapted VUAMC to include verb-subject (i.e. *nsubj*) and adjective-noun (i.e. *amod*) relations. Moreover, we plan to consolidate these adapted datasets in one repository categorised by data source and text genre. We also plan to invest extra annotation effort to ensure the consistency of the annotated instances across the different datasets using weakly supervised approaches.

Acknowledgments

This work was supported by Science Foundation Ireland under grant number SFI/12/RC/2289_2 (Insight).

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '06, pages 329–336, Trento, Italy.

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, NY, USA.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, COLING-ACL '06, pages 77–80, Sydney, Australia.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 523–528, Valencia, Spain.

Lou Burnard. 2007. Reference guide for the British National Corpus (XML edition).

Lynne Cameron and Graham Low. 1999. *Researching and Applying Metaphor*. Cambridge Applied Linguistics. Cambridge University Press, Cambridge, UK.

Jonathan Charteris-Black. 2011. Metaphor in Political Discourse. In *Politicians and Rhetoric: The Persuasive Power of Metaphor*, pages 28–51. Palgrave Macmillan UK, London.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 740–750, Doha, Qatar.

Yulia Clausen and Vivi Nastase. 2019. Metaphors in text simplification: To change or not to change, that is the question. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 423–434, Florence, Italy.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Laurel J End. 1986. Grounds for metaphor comprehension. *Knowledge and language*, pages 327–345.

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

Janet Ho and Winnie Cheng. 2016. Metaphors in financial analysis reports: How are emotions expressed? *English for Specific Purposes*, 43:37 – 48.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. In *Proceedings of the 7th International Corpus Linguistics Conference*, CL '13, pages 125–127, Lancaster, UK.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, pages 7–36.

Arlene Koglin and Rossana Cunha. 2019. Investigating the post-editing effort associated with machine-translated metaphors: a process-driven analysis. *The Journal of Specialised Translation*, 31(01):38–59.

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

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, LA, USA.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '14, pages 55–60, Baltimore, MD, USA.

André Martins, Noah Smith, Eric Xing, Pedro Aguiar, and Mário Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 34–44, Cambridge, MA, USA.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 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.

Natalie Parde and Rodney Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC '18, pages 1535–1540, Miyazaki, Japan.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 1537–1546, Copenhagen, Denmark.

Vassiliki Rentoumi, George A. Vouros, Vangelis Karkaletsis, and Amalia Moser. 2012. Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective. *ACM Transactions on Speech and Language Processing*, 9(3):1–31.

Elena Semino, Zsofia Demjen, Andrew Hardie, Sheila Alison Payne, and Paul Edward Rayson. 2018. *Metaphor, Cancer and the End of Life: A Corpus-based Study*. Routledge, London, UK.

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

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 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.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1002–1010, Beijing, China.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.

Kevin Stowe and Martha Palmer. 2018. Leveraging syntactic constructions for metaphor identification. In *Proceedings of the Workshop on Figurative Language Processing*, pages 17–26, New Orleans, LA, USA.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 248–258, Baltimore, MD, USA.

Magdalena Wolska and Yulia Clausen. 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. Association for Computational Linguistics.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*, LDK '19, Leipzig, Germany.