Linguistic Analysis Improves Neural Metaphor Detection

Kevin Stowe, Sarah Moeller, Laura Michaelis, Martha Palmer
University of Colorado, Boulder, CO 80309
[kest1439, samo9533, laura.michaelis, mpalmer]@colorado.edu

Abstract

In the field of metaphor detection, deep learning systems are the ubiquitous and achieve strong performance on many tasks. However, due to the complicated procedures for manually identifying metaphors, the datasets available are relatively small and fraught with complications. We show that using syntactic features and lexical resources can automatically provide additional high-quality training data for metaphoric language, and this data can cover gaps and inconsistencies in metaphor annotation, improving state-of-the-art word-level metaphor identification. This novel application of automatically improving training data improves classification across numerous tasks, and reconfirms the necessity of high-quality data for deep learning frameworks.

1 Introduction

Humans use metaphors to conceptualize abstract and often difficult concepts by employing knowledge of more concrete domains. They are prevalent in speech and text, and allow us to communicate more effectively and more imaginatively. The fact that they are commonplace and easily understood by humans makes appropriate interpretation of them essential for high quality natural language processing applications.

The primary linguistic and cognitive theory of metaphor is conceptual metaphor theory (Lakoff and Johnson, 1980; Lakoff, 1993), which theorizes that metaphors are primarily a mental activity, and the language is merely a side effect of these "conceptual" metaphors. From this, it is posited that metaphors are agnostic with regard to syntactic structure: a conceptual mapping can be expressed through whatever syntax the speaker desires. This is apparent from evidence that many metaphoric predications have the same syntactic properties as their literal counterparts. Goldberg (1995) observes that metaphorical ditransitive sentences like "It gave me a headache" do not differ syntactically from literal ditransitive sentences like "She gave me the account." Accordingly, to find syntactic hallmarks of metaphorical meaning we do not look generally for particular syntactic constructions, but rather for mismatches of various kinds between specific verbs' ordinary syntactic behavior and their behavior under metaphoric interpretation.

Perhaps the primary source of verbal syntactic variability is the set of argument-structure constructions identified by Goldberg (1995). One such construction is Caused Motion (CM), illustrated by the sentence "They pushed it down the hall". CM can augment the array of semantic roles supplied by the verb, as in "They laughed me out of the room". Augmentation often entails a metaphoric construal: here the verb "laugh", otherwise a single-argument verb, is paired with both a theme argument (the direct object) and a PP location argument, and the resulting predication expresses metaphorical rather than literal motion (Hwang, 2014).

Despite this connection between verbal syntax and metaphoric properties, most computational approaches to metaphor eschew syntax for more semantic features. While these have proven effective, metaphor detection remains a difficult task. This could be due to many factors, but a primary reason is the lack of adequate training data. Annotation of metaphor has proven to be extremely difficult, as is evident by the variety of schemes used to attempt to achieve consistent annotation.[1] This has led to a lack of "big data" for training models, as well as inconsistencies and gaps in the data that is available.

In this work, we show that syntactic properties

[1]For a review of systems, see Veale et al. (2016)

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 362–371
Hong Kong, China, November 3-4, 2019.

can be used to improve training data, which is beneficial to metaphor processing systems. Deep learning models require sufficient quality data, which is lacking for many metaphorical expressions. We automatically fill gaps in metaphor training data by exploiting syntax in two ways: first, we use the syntactically-motivated lexical resource VerbNet to identify additional data through metaphoric and literal sense identification, and second, we use syntactic properties of certain lexemes, which allow us to identify relevant sentences via dependency parses. These methods yield training data that improves performance for metaphor classification across a variety of tasks.

2 Related Work

While most computational metaphor processing methods rely heavily on lexical semantics, many previous approaches also employ syntactic structures to varying degrees. Most prior work involving argument structure is based on the idea of selectional preferences: certain verbs prefer certain arguments when used literally, and others when used metaphorically. This idea is captured by determining what kinds of arguments fill syntactic and semantic roles for specific verbs.

The CorMet system (Mason, 2004) employs this paradigm, and is similar to ours in their collection of key verbs and analysis of syntactic arguments and semantic roles. They automatically collect documents for particular domains based on key words, and identify selectional preferences based on the WordNet hierarchy for verbs in these particular domains. For example, they find that *assault* typically takes direct objects of the type *fortification* in the *MILITARY* domain. This allows them to make inferences about when selectional preferences are adhered to, and they can then identify mappings between different domains. While their task is fundamentally different, their usage of syntactic frames to identify relevant arguments is very similar to our work. However, rather than identify preferences, we are using syntactic frames to identify whether the verbs are possibly used metaphorically. Our methods require less adherence to semantic properties, which they retrieve from WordNet. Our methods are also inherently somewhat more noisy: while there is evidence that syntactic frames can be indicative of metaphoric properties, these properties are rarely observed deterministically.

Gedigian et al. (2006) use FrameNet and PropBank annotation to collect data, focusing on the FrameNet frames *MOTION* and *CURE*. They use PropBank argument annotations as features, resulting in metaphoric classification accuracy on these domains of over 95%, although this is only slightly above the most frequent class baseline (92%). They collect data from lexical resources and then annotate it for metaphoricity, which is similar to our approach of analyzing the resources and word senses for metaphors.

Shutova et al. (2013) also employs selectional preferences based on argument structure, identifying verb-subject and verb-direct object pairs in corpora. They begin with a seed set of metaphoric pairs, similar to our methods of collecting instances based on syntactic information. They use these seed pairs to identify new metaphors, similar to our usage of syntactic patterns to identify training data. Their methods are based on the selectional preferences of verbs, and thus are less concerned with the variety of syntactic patterns metaphors can participate in. We will identify much more complex syntactic patterns, and we then use the data for training metaphor systems rather than identifying selectional preferences.

Stowe et al. (2018) use syntactic structures directly for feature-based machine learning methods. They highlight the distribution of various syntactic patterns in corpora, and extract features based on dependency parses to improve classifier performance. While their results outperform lexical baselines, they still lag behind other metaphor detection systems, with F1 scores of 53.1 for verbs and 50.5 for nouns on the Vrije Universtat Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010). We improve on their work by employing deep learning architecture while still attempting to leverage syntactic information. As many deep learning algorithms (including the recurrent neural networks used here) natively capture long-distance dependencies, direct inclusion of syntactic features is likely not productive. By capturing additional data, we can take advantage of linguistic analysis to improve deep learning-based metaphor detection.

With regard to datasets and tasks, metaphor processing has suffered from a lack of consistent evaluation methods. The metaphor shared task provided a standard evaluation procedure that has greatly helped with system comparison (Leong

et al., 2018). They use the VUAMC, providing a train/test split that has been used to regularize the evaluation of metaphor identification systems. For both sections of the task (identifying all metaphoric words and identifying verbs), four of the top five systems use some form of long short-term memory network (LSTM). The system of Wu et al. (2018) performed best on both tasks (F1 of .651 on all parts of speech, and .672 for verbs) using a combination of a convolutional neural network (CNN) and bidirectional LSTM.

Since the shared task, a variety of other approaches have been developed using similar deep learning techniques. Most recently, the work of Gao et al. (2018) achieved state-of-the-art performance on the shared task data as well as a variety of other datasets, including the TroFi (Birke and Sarkar, 2006) and Mohammad et al. (2016) datasets, using Bi-LSTM models coupled with GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) embeddings. For the VUAMC shared task, they report F1 scores of .726 for all parts of speech and .697 for verbs. For the Mohammad et al. dataset, they report an average F1 score of .791, and for the Trofi they report an F1 score of .72, slightly lower than the current state of the art (.75 of Köper et al. (2017)).

Despite recent advances in evaluation and algorithm performance, the task still remains difficult, with the highest F1 scores nearing only .73 on the VUAMC data. This is likely due to the relatively small dataset size (app. 200,000 words), which is in part caused by difficulties in annotation. We aim to overcome some of this difficulty by automatically extracting additional training data with lexical and syntactic methods.

3 Methods

We aim to improve training data for metaphor processing by performing linguistic analysis on difficult verbs to uncover the syntactic properties that can potentially influence their metaphoricity. Our work is focused on verbs: they form the foundation of many metaphoric expressions, and evidence from construction grammar and frame semantics has shown that syntactic properties can often influence the types of metaphors that are produced (Sullivan, 2013). We show that identification of anomalous syntactic structures can provide evidence towards metaphoricity, and can be leveraged to automatically extract training data that improves classification performance.

This is done through two paths: first, we explore the lexical semantic resource VerbNet, an ontology of English verbs that contains rich syntactic and semantic information. From VerbNet we explore verb senses that can potentially be deterministically metaphoric or literal, and extract training data from existing VerbNet annotation. Second, we analyze syntactic patterns from Wikipedia data. We identify patterns that indicate metaphoric or literal senses of verbs, and then extract additional data based on these patterns.

3.1 Finding Difficult Verbs

First, we need to select verbs to analyze. Our goal is to find verbs that are likely difficult for classifiers, as well as those that are frequent enough to have a significant impact. This is accomplished through two avenues: first, we examine all the verbs in the training data, and analyze those that have the most even class balance between literal and metaphoric uses. We refer to these verbs as our most "ambiguous" verbs. For our preliminary experiments, we selected the ten most ambiguous verbs which occurred at least ten times in the training data.

Second, we employed the metaphor detection system of Gao et al. (2018). We trained the system on the provided VUAMC shared task training data and ran it on their validation set. We then analyzed which verbs were most frequently misclassified in the validation data, to determine where additional data would be most effective. We chose to use the VUAMC data for this task due to its size and status as the standard for metaphor identification. As with the most ambiguous verbs, we selected the ten verbs with the lowest F1 score that occurred at least ten times in the data for analysis. The verbs chosen through these analyses are shown in Table 1. In theory, expanding the number of verbs analyzed would yield more data and improve performance, but as an experimental baseline ten verbs is sufficient for analysis and classifier improvement.

For each of these verbs, we performed two kinds of analysis. First, we explored their metaphoric and literal usage in VerbNet. Second, we examined their syntactic properties for metaphoric and literal patterns.

3.2 VerbNet

VerbNet is a lexical resource that currently categorizes 6,791 verbs into 329 verb classes based

Most Ambiguous Verbs				Most Misclassified Verbs				
Verb	**Met Tokens**	**Lit Tokens**	**% Met**	**Verb**	**FP**	**FN**	**Correct**	**% Correct**
encourage	6	6	.5	spend	7	0	4	.363
blow	5	5	.5	include	8	1	10	.526
conduct	5	5	.5	play	3	3	8	.571
show	34	33	.49	hold	3	3	8	.571
find	60	62	.49	stop	7	1	12	.600
fall	18	19	.51	reduce	2	2	6	.600
hold	28	30	.52	get	15	21	74	.673
bring	36	33	.48	suggest	4	0	9	.692
put	57	52	.48	meet	2	1	7	.700
allow	19	21	.48	discuss	1	2	7	.700

Table 1: Difficult Verbs from the VUAMC data: on the left, the verbs with the most even split between literal and metaphoric. On the right, verbs in the validation set that were most misclassified. Restricted to verbs where count ≥ 10.

on their syntactic and semantic behavior (Kipper-Schuler, 2005).[2] These verb classes are based on the work of Levin (1993), who shows that for many verbs their semantics can be determined by the syntactic alternations they participate in, arguing that "the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large degree determined by its meaning." (pg 1)

VerbNet is primarily composed of verb "classes": these classes are a hierarchically structuring of verb senses based on their syntactic and semantic behavior. Each class contains a list of verb senses, the syntactic frames that these verbs can participate in, a first-order semantic predicate representation for the class's meaning, and the thematic roles the verb takes as arguments. These thematic roles, which are fairly coarse-grained roles such as Agent, Theme, and Patient, are often marked with selectional restrictions. For example, many classes have Agents that are marked as +ANIMATE, indicating the Agent of the verb must be an animate entity.

VerbNet has practical applications for word sense disambiguation and semantic role labelling, and numerous annotation projects have been done to tag data with the correct VerbNet senses. Our goal is to identify which particular VerbNet senses are typically metaphoric or literal, and extract sentences tagged with these VerbNet senses.

For each verb, we examined the VerbNet classes in which it appears. We looked at VerbNet annotation, the example sentences, the selectional preferences on the class's thematic roles, and the semantic predicates. From this we assessed whether the sense of the verb in each class was typically metaphoric or literal. Consider the verb "grow". It is present in two particular VerbNet classes: **GROW-26.2** and **CALIBRATABLE_COS-45.6**. The **GROW-26.2** class has an animate Agent role, and produces a concrete Product out of a concrete Material:

1. A private farmer in Poland is free to buy and sell land, hire help, and decide what to **grow**.

2. It's the kind of fruit that **grew** freely and that you could help yourself to.

We note from the semantics and annotated examples, we expect this sense of grow to typically be literal. However, in the **CALIBRATABLE_COS-45.6** class, it contains a Value role that moves along a scale by a certain Extent. These examples all appear to be metaphoric, evoking the MORE IS UP mapping[3]:

1. Exports in the first eight months **grew** only 9%.

2. Non-interest expenses **grew** 16% to $496 million.

This allows us to extract new training data using these classes. We use a repository of manually annotated VerbNet senses, containing approximately 150,000 annotated verbs (Palmer et al.,

[2] https://verbs.colorado.edu/verbnet/

[3] Examples from the VerbNet annotation data from Palmer et al. (2017)

2017). We found all annotated instances of "grow" in the **GROW-26.2** class, and considered them to be literal, and all instances of "grow" from **CALIBRATABLE_COS-45.6** were considered metaphoric. This process was completed for all of the verbs in Table 1. This analysis only includes the particular verb in question. We believe "grow" in the **GROW-26.2** class is typically literal, and "grow" in the **CALIBRATABLE_COS-45.6** class is typically metaphoric, but this does not necessarily extend to other verbs in these classes. We will discuss the possibility of expanding this analysis to include all verbs for particular classes in Section 6.

Note that we only consider the verbs in these instances: we have no knowledge of the metaphoricity of the verbs' arguments. For each verb, we extracted up to 100 annotations for each sense that we determined to be largely metaphoric or literal.

3.3 Syntactic Patterns

As a second path for finding additional data, we explore the syntactic properties of metaphoric expressions. While metaphors are traditionally seen as cognitive, and relatively unaffected by surface syntactic realizations, there is recent evidence based in construction grammar that syntactic structures can influence the source and target domain elements of metaphoric expressions (Sullivan, 2013; David and Matlock, 2018; David, 2017). We expand on this idea: we believe that not only can syntactic structures indicate source and target elements, but they can also indicate metaphoricity.

We see this in English with verbs like hemorrhage, which is almost always used metaphorically when it is used transitively[4]:

- GM was supporting this event even as they were **hemorrhaging cash**.
- For 30 straight years, American organized labor has been **hemorrhaging members**.

When used intransitively, hemorrhage is almost always literal:

- Cerebral AVMs often have no symptoms until they rupture and **hemorrhage**.
- Michael **hemorrhaged** and sustained a massive stroke to the left side of his brain.

[4]Examples from SketchEngine (Kilgarriff et al., 2014) http://www.sketchengine.eu/

This is likely due to the fact that literal use of "hemorrhage" contains an understood argument, blood, which is the most natural object of the verb. If the use is intended in a less literal way, which requires on over syntactic object, the null "blood" object needs to be overridden. While not all verbs have this direct relation between argument number and metaphoricity, we believe that the type and number of syntactic arguments of a verb can be indicative of unmarked usage, and may be utilized as a method for automatically extracting training data for metaphor classification. This analysis doesn't reflect linguistic facts: it is possible to construct sentences in which the intransitive use of "hemorrhage" is metaphoric ("after the stock market crashed, the company hemorrhaged"), as well as transitive usages that are literal ("after the surgery, the patient hemorrhaged blood"). However, we find that in the majority of cases, metaphoricity aligns with the argument structure, and these contrived examples are exceedingly rare.

For each verb in our list, we analyzed all the sentences from the VUAMC training data as well as 50 additional sentences from Wikipedia that contained the verb, and attempted to discover syntactic patterns that are indicative of metaphoricity. We examined argument structure, active vs passive voice, prepositional complements, aspect, idiomatic combinations and other surface syntactic properties. We created a short list of the most likely candidates for literal and metaphoric syntactic patterns. We then extracted up to 100 sentences from Wikipedia that matched these syntactic patterns.

A brief overview of the syntactic patterns and VerbNet class analysis is shown in Table 2; the full extraction rules, code, and data will be released upon publication.

Note that for many cases, it was difficult to determine what was literal and what was metaphoric. Highly polysemous verbs like "get" in particular are problematic: they contain many different meanings and usages that can often be annotated inconsistently, so strong metaphoric or literal patterns were impossible to identify.

As with the "hemorrhage" examples above, these patterns are not deterministic. The syntactic structures analyzed aren't always metaphoric or literal, but they are consistent enough to be useful for extracting additional training data. For each verb we attempted to extract up to 100 samples

Verb	Lit. Syn. Patterns	Met. Syn. Patterns	Lit. VerbNet Classes	Met. VerbNet Classes
encourage	NP V NP {TO} VP	NP V NP	advise-37.9	amuse-31.1
find	NP V PRO VP *find out*, *find dead*	NP V NP {TO BE} ADJ	get-13.5.1	declare-29.4
fall	NP V ADV, NP V	WH NP V *fall in*, *fall to*	escape-51.1	calibratable_cos-45.6 convert-26.6.2 long-32.2 acquiesce-95.1 die-42.4
spend	NP V NP V {ON} NP	*spend time* *spend life*	pay-68	consume-66 spend_time-104
play	NP V PP *play with*	-	meet-36.3 performance-36.7 play-114.2	trifle-105.3 use-105.1
suggest	negation	-	say-37.7	reflexive_appearance-48.1.2
meet	NP V *meet for/at/to*	-	contiguous_location-47.8	satisfy-55.7

Table 2: Example analysis of syntactic patterns and VerbNet classes.

	From VerbNet		From Syn. Patterns	
Verb	Count	% Met	Count	% Met
encourage	86	.611	200	.497
blow	-	-	99	.946
conduct	-	-	200	.503
show	-	-	-	-
find	407	.300	255	.047
fall	314	.601	600	.749
hold	913	.445	487	.560
bring	-	-	500	.601
put	-	-	-	-
allow	2	1	300	.334
spend	439	.630	341	.553
play	52	.196	343	0
stop	482	.208	-	-
reduce	-	-	-	-
suggest	307	.003	12	0
meet	455	.229	399	0
Total	3985	.442	3736	.424

Table 3: Total samples extracted from VerbNet classes and syntactic patterns, along with the percentage of extracted samples that are metaphoric.

for each VerbNet sense and each syntactic pattern. This is to prevent the extracted data becoming saturated with extremely common senses or patterns. Many senses and patterns are rare, and fewer than 100 instances were collected. A summary of the extracted data by verb is shown in Table 3.

In total, we extracted 3,985 samples from VerbNet annotation and 3,736 samples from Wikipedia based on syntactic samples for our analyzed verbs. Each sample is an entire sentence containing the verb in question, for which we can provide automatic annotation based on our VerbNet and syntactic analyses. We can treat this as distantly supervised data: we have beliefs about the metaphoric and literal labels for the verbs in each sentence extracted, but these aren't always deterministic: errors in syntactic pattern matching, anomalous examples, and other factors introduce inaccuracies in these samples.

4 Tasks

In order to show the efficacy of our extracted data, we add this data to the standard datasets and evaluate performance on a variety of metaphor processing tasks. For a relevant comparison to contemporary research, we evaluate our results using the baseline system of Gao et al. on five different tasks. As per their work, we experiment with two different models: a sequence based model (dubbed "SEQ") that performs best when all parts of speech contain metaphor tags, and a "classification" model ("CLS"), which tags individual verbs as metaphoric or not.

4.1 Sequential Model (SEQ)

The sequential model takes as input sentences from VUAMC data, each with a binary metaphor tag. They represent each word as the concatenation of a 300 dimension GloVe embeddings with an ELMo vector. These are then input to a bidirectional LSTM. These sequential models are particularly useful for encoding relations among distant words, and have proven effective on a large number of tasks for which each word in a sentence has a tag.

4.2 Classification Model (CLS)

The classification model represents each verb in the VUAMC data as its own instance, maintaining the sentential context, and these each retain their annotation as either metaphoric or not. As with the

sequential model, they are input to a bidirecitonal LSTM using GloVe and ELMo vectors. They also include an index embedding and attention layer to encode the location of the target word.[5]

These models are used over a variety of datasets: the dataset from Mohammad et al. (2016), the Trofi dataset (Birke and Sarkar, 2006), and the VUA metaphor corpus which is the basis for the metaphor shared task (Steen et al., 2010). They use a section of the Mohammad et al. dataset dubbed "MOH-X", consisting of 636 example sentences from 214 verbs taken from WordNet, annotated as metaphoric or literal. The Trofi dataset contains 3,737 sentences from 50 different verbs which were automatically clustered into metaphoric or literal clusters. The sequence tagger shows best performance when all words in the sentence are metaphoric, as is the case with the VUAMC data. The classification model performs best when only a single word is metaphoric, as in the MOH-X and Trofi datasets. While the VUAMC is the basis of our analysis, we will also examine how adding additional impacts results on the MOH-X and Trofi datasets; their setup as classification tasks more accurately mirrors the additional data, as there is only one potentially metaphoric word per sample.

4.3 Architecture

We replicate the architectures of Gao et al., using the same experimental set-up. For the classification model, we can include our extra data as-is, with metaphor annotations based on our analysis. For the sequential model, we consider only the verb analyzed as metaphoric, leaving the rest of the words tagged as literal. In order to judge performance, we run three experimental setups: one with the additional VerbNet samples, one with the additional samples generated via syntactic patterns, and one with both. We experimented with tuning hyperparameters (learning rate, dropout, and the size of the hidden layer), but found no significant improvements over their experimental setup. We did make one modification: we increased the amount of training epochs in proportion to the amount of training data added. This allows for the model to be sufficiently trained over all the data.

The VUAMC data has been split into training and test sections for the shared task, and these sections are also used by Gao et al. We will adopt this split. For the MOH-X and Trofi datasets, they run 10-fold cross validation and report the mean F1 score. Due to the variable nature of neural models and the relatively small dataset size, we include experiments to calculate the statistical significance of our methods. We split these datasets into 75% training, 25% test, mirroring the VUAMC data, and ran classification 10 times. We then calculated the means and standard deviations. We also ran bootstrap estimation for all tasks, reevaluating using random replacement over 10^6 iterations (Efron, 1979; Berg-Kirkpatrick et al., 2012). We consider improvement significant when the mean and standard deviation from both methods yield p values of less than .01.

5 Results

The results from our additions on the original tasks are shown in Figure 1, and the improvements over the baseline for each method are outlined in Table 4. For each task, we display the results of the original Gao baseline, along with the addition of VerbNet samples, syntactic pattern-based samples, and both. For each of these, we show the mean and standard deviation from running the task 10 times.

We find that adding VerbNet samples, syntactic patterns, and both datasets all always produces a significant improvement over the baseline. Adding this additional data outperforms the Gao et al. sequence tagging algorithm on the VUA shared task data for both verbs and all parts of speech. We also see improvements in the classification model, and on the Trofi dataset. It is important to note that the extreme variability in the results for these smaller datasets. We found our improvements on the Trofi dataset to be significant, while the MOH-X results were not significant. This is likely due to the size of the dataset: the MOH-X data contains only 636 samples, leading to high variance in performance. Further evaluation is necessary to determine the consistent effect of this data.

For verb sequence tagging, the VerbNet data yielded the best performance, while for all parts of speech the additional syntactic data performed best. This may be because the VerbNet data comes specifically from VerbNet annotation, relying strictly on VerbNet senses. VerbNet is grounded in syntactic alternations, but individually VerbNet senses occasionally encode metaphor

[5]Full details and code for each of these models can be found at https://github.com/gao-g/metaphor-in-context/

	Task				
Additional Data	MOH-X	Trofi	VUA CLS (Verbs)	VUA SEQ (Verbs)	VUA SEQ (All)
Baseline	.653	.658	.665	.682	.728
+VN Data	.681*	.672	.673	**.696**	.736
+SYN Data	.704*	.672	.677	.694	**.738**
+Both	.683*	**.684**	**.679**	.695	.735

Table 4: Mean F1 scores over 10 iterations, for each model and dataset added. Bootstrap sampling indicated that these improvements are all significant over the baseline, excepting the MOH-X dataset. Due to high variability, the MOH-X results (*) were not significant improvements over the baseline ($p > .01$)

.

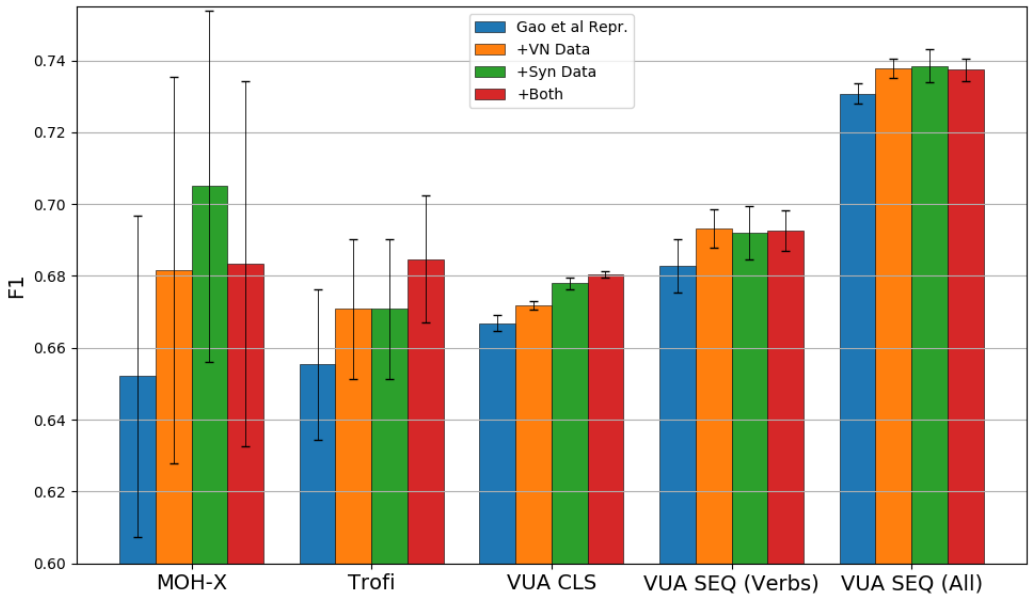

Figure 1: Results for each task. Results shown are the mean with standard deviations from running 10 iterations of each model for each task.

without direct relation to the verb's arguments. The syntactic data directly relies on patterns which include other parts of speech: arguments, prepositions, and idiomatic expressions. These extra components of the analysis may make the data more broadly applicable to all parts of speech, driving the improvement in the sequence tagging of all words.

Adding both distantly supervised datasets improved performance over adding either individually only for the classification-based tasks, where only one word per sentence has a tag (the Trofi and VUAMC verb classification tasks). Only on the Trofi dataset was the improvement from adding both datasets significantly higher than the improvement from adding the best individual dataset. For the sequence-based tagging of the VUAMC, adding both yielded negligible improvements. As adding both datasets was effective for classification tasks, we believe the difficulty in combining both datasets in the sequence models is due to excessive noise from the non-target words of the samples. We default to marking every word other than the target verb in the sentence as literal, so the additional data is understandably less informative for sequence tagging problems. It is likely that the combination of VerbNet data and syntactic pattern-based data caused additional noise: the two datasets may in places be contradictory, particularly with regard to these non-target elements.

6 Conclusions

We show that using external data found through syntactic structures and lexical resources can be used to improve deep learning methods for metaphoric classification. This is due to regular syntactic patterns of metaphoric usage, and the idea that the semantics of verbs can be dependent on the syntactic patterns that it participates in. For future improvements, there are other resources available that could be leveraged in the same way. PropBank (Palmer et al., 2005), FrameNet (Baker

et al., 1998), and WordNet (Fellbaum, 2010) all offer some syntactic and/or semantic information, and data annotated with these resources could prove another valuable source of additional samples.

We also only examine some basic syntactic patterns for a small number of verbs, and this was done manually. Improved methods for automatically detecting relevant syntactic patterns as well as further effort in manual identification of syntactic properties of metaphoric samples could increase the amount of data extracted. Further linguistic analysis of constructions that either require or prohibit metaphoric interpretations could improve both automatic metaphor processing and our broader understanding of linguistic metaphors. Additionally, we only look at specific verbs within VerbNet classes. All verbs within VerbNet classes share syntactic and semantic properties, so it is likely that we can extend our verb-level analysis to a broader class-level analysis. A straightforward extension of this work would be to analyze VerbNet classes as being metaphoric or literal, and extracting data for all verbs within a given class.

Finally, while they have proven invaluable for the standardization of metaphor processing, there are still gaps and inconsistencies in our metaphor datasets. Extracting additional training data based on syntactic patterns likely was effective in this case in part due to the idiosyncrasies of the previous datasets, which may over-annotate possible metaphors. This procedure yields a large number of conventional metaphors, which lack novelty, are very frequent, and are perhaps more amenable to being discovered via syntactic patterns. More data annotated for metaphor is essential to improve deep learning methods for metaphor processing, and while we are attempting to overcome these gaps with outside resources, further quality metaphor annotation would prove especially valuable to the field.

Acknowledgements

We gratefully acknowledge the support of the Defense Threat Reduction Agency, HDTRA1-16-1-0002/Project #1553695, eTASC - Empirical Evidence for a Theoretical Approach to Semantic Components and a grant from the Defense Advanced Research Projects Agency 15-18-CwC-FP-032 Communicating with Computers, a subcontract from UIUC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any government agency.

References

C. F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98*, pages 86–90, Montreal, QC.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. "An Empirical Investigation of Statistical Significance in NLP". In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2006. "A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language". In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Oana David. 2017. "Computing metaphor: The case of MetaNet". *Cambridge Handbook of Cognitive Linguistics*, pages 574–589.

Oana David and Teenie Matlock. 2018. Cross-linguistic automated detection of metaphors for poverty and cancer. *Language and Cognition*, 10(3):467–593.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*.

George A. Fellbaum, Christiane; Miller. 2010. WordNet. `http://wordnet.princeton.edu/`. Accessed: 2019-02-28.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. "Neural Metaphor Detection in Context". In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613. Association for Computational Linguistics.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, ScaNaLU '06, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adele Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Jena Hwang. 2014. *"Identification and Representation of Caused-Motion Constructions"*. Ph.D. thesis, University of Colorado, Boulder.

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.

Karen Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Maximilian Köper and Sabine Schulte im Walde. 2017. "Improving Verb Metaphor Detection by Propagating Abstractness to Words, Phrases and Individual Senses". In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain. Association for Computational Linguistics.

George Lakoff. 1993. The Contemporary Theory of Metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, pages 202–251. Cambridge University Press.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago and London.

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. Association for Computational Linguistics.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Zachary J. Mason. 2004. "CorMet: A computational, corpus-based conventional metaphor extraction system". *Computational Linguistics*, 30(1):23–44.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. "Metaphor as a Medium for Emotion: An Empirical Study". In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Association of Computational Linguitics*, 31(1):71–106.

Martha Palmer, James Gung, Claire Bonial, Jinho Choi, Orin Hargraves, Derek Palmer, and Kevin Stowe. 2017. The Pitfalls of Shortcuts: Tales from the word sense tagging trenches. In *Essays in Lexical Semantics and Computational Lexicography - In honor of Adam Kilgarriff*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification*. John Benjamins Publishing Company.

Kevin Stowe and Martha Palmer. 2018. Leveraging Syntactic Constructions for Metaphor Processing. In *Workshop on Figurative Language Processing*, New Orleans, Louisiana.

Karen Sullivan. 2013. *Frames and Constructions in Metaphoric Language*. John Benjamins.

T. Veale, E. Shutova, and B. Klebanov. 2016. *Metaphor: A Computational Perspective*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. "Neural Metaphor Detecting with CNN-LSTM Model". In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114. Association for Computational Linguistics.