International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 117–140
Received 25/01/2015, Accepted 25/02/2015, Final 04/03/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Cross-lingual Semantic Generalization for the Detection of Metaphor

MICHAEL MOHLER
MARC TOMLINSON
BRYAN RINK
Language Computer Corporation, Richardson, TX

ABSTRACT

In this work, we describe a supervised cross-lingual methodology for detecting novel and conventionalized metaphors that derives generalized semantic patterns from a collection of metaphor annotations. For this purpose, we model each metaphor annotation as an abstract tuple – (source, target, relation, metaphoricity) – that packages a metaphoricity judgement with a relational grounding of the source and target lexical units in text. From these annotations, we derive a set of semantic patterns using a three-step process. First, we employ several generalized representations of the target using a variety of WordNet information and representative domain terms. Then, we generalize relations using a rule-based, pseudo-semantic role labeling. Finally, we generalize the source by partitioning a semantic hierarchy (defined by the target and the relation) into metaphoric and non-metaphoric regions so as to optimally account for the evidence in the annotated data. Experiments show that by varying the generality of the source, target, and relation representations in our derived patterns, we are able to significantly extend the impact of our annotations, detecting metaphors in a variety of domains at an F-measure of between 0.88 and 0.92 for English, Spanish, Russian, and Farsi. This generalization process both enhances our ability to jointly detect novel and conventionalized metaphors and enables us to transfer the knowledge encoded in metaphoricity annotations to novel languages.

Keywords: Metaphor detection, generalization, semantic modeling, WordNet, transfer learning

1. INTRODUCTION

Metaphor is the air we breathe. It stirs the emotions, arouses the senses, and serves as a vehicle for describing and reasoning about difficult concepts – all while using language that is both clear and familiar. Metaphor is everywhere in human language, hiding in plain sight. For this reason, it is crucial for technologies that seek to model and understand human language to be capable of correctly identifying and interpreting metaphor. Indeed, metaphor has been found to confound both statistical and knowledge-based techniques for natural language processing across a wide variety of applications including textual entailment, text summarization, word sense disambiguation, semantic textual similarity, question answering, and event extraction. In this work, we propose a methodology that derives generalized semantic patterns from existing metaphor annotations in order to detect a wide variety of metaphoric language as either a stand-alone system or as a component in a larger supervised or unsupervised metaphor detection system.

Although there have been many influential theories regarding the cognitive basis of metaphor, the most prominent is Lakoff's Contemporary Theory of Metaphor [13, 11], which popularized the idea of a conceptual metaphor mapping. Within the cognitive framework of a given conceptual mapping, terms pertaining to one concept (the *source*) can be used figuratively to express some aspect of another concept (the *target*). For example, the conceptual metaphor "Life is a Journey" indicates a cognitive lens through which the target concept "life" may be more easily discussed and understood. This particular mapping allows us to speak of one being stuck in a "dead-end" job, of a crucial decision as being a "fork in the road", and of someone's life "taking a wrong turn".

Existing work on the identification of metaphor can be broadly categorized as either feature-based or example-based. Feature-based metaphor identification is based upon the

assumption that metaphoric usages in context typically have certain characteristics that serve as cues for indicating non-literal usage. Such characteristic indicators of non-literal usage include the pairing of abstract with concrete terms [28, 4, 3, 27], the violation of selectional preference [7, 19, 10, 29], dissonance between a term and its greater context [26, 1, 25], explicit linguistic cues [16], and semantic unrelatedness between terms [22, 5]. In general, such methods are successful at detecting novel metaphors, but have difficulty in detecting commonly used figurative language or "conventionalized metaphors" for which modeling selectional preference, contextual relatedness, and semantic mismatch is more complex.

Example-based methods seek to detect metaphor by comparing candidate texts to a set of known metaphors using abstraction hierarchies [18], known conceptual metaphor domain interactions [21], semantic signatures [20], models of metaphoricity priors [24], probabilistic typicality hierarchies [15], or simply large lexical stores of conventionalized Metaphor [14]. Of course, methods that compare candidates to known metaphors are able to reliably detect those that are most commonly used, but they require some additional framework to generalize from these examples to less common or even novel metaphoric utterances.

We propose an approach to metaphor identification that follows the example-based paradigm but differs from existing work in that we (1) explicitly explore a variety of methods for generalization; (2) determine the impact that such methods have on overall metaphor identification performance; and (3) apply these examples cross-lingually to maximize performance in novel languages. In particular, we generalize our annotations, which consist of the tuple (source, target, relation, metaphoricity), in three ways:

1. We generalize target lexemes using semantic categories and a domain-level groupings of terms and exploit the dichotomy between abstract and concrete nouns;

2. We generalize source/target relations in text by converting dependency relations into broader pseudo-semantic relations using a rule-based approach;
3. We generalize source lexemes by exploiting the hypernymy links in a semantic hierarchy (i.e., WordNet).

For each generalized target/relation pair, we map the associated source lexemes from our annotations onto the semantic hierarchy. In particular, we make use of a largescale, multilingual semantic knowledge base (i.e., a multilingual WordNet) that combines groupings of related objects (i.e., synsets and semantic categories) with a hierarchical tree structure (i.e., hypernymy relations). For a given target/relation pair, we then define a semantic pattern as a single node in this hierarchy with an associated metaphoricity judgement which indicates that that node, and all of its descendents in the hierarchy, are or are not metaphors unless they are under the influence of a different pattern node. In other words, the metaphoricity judgement of each node in the hierarchy is determined by the metaphoricity associated with the pattern node that is its nearest ancestor. The set of pattern nodes is selected using a dynamic programming algorithm to optimally select (and assign judgements to) a minimal set of nodes in the hierarchy so as to account for all of the annotation evidence. In effect, we are using a multilingual knowledge base to generalize from the given examples (of both literal and metaphoric usages) so as to partition the semantic space that it defines into regions of likely metaphoricity and regions of unlikely metaphoricity. By generalizing in this way, we are able to transfer knowledge of metaphor to languages with no metaphor annotations.

The remainder of this work is organized as follows. In Section 2, we survey related work in metaphor identification. Then, in Section 3, we describe the two components of our system – generalizing through the knowledge base to produce a set of semantic patterns and using the resulting patterns to detect metaphor in unseen data. Section 4 describes the provenance and the characteristics of the multilingual datasets we use to train and evaluate our system. In Section 5 we present our experiments and

discuss our results. Finally, we share the insights gained from these experiments in Section 6 and offer our recommendations for moving forward.

2. RELATED WORK

One of the earliest works in example-based metaphor processing is that of Martin [17], who sought to enable an automated Unix help client (uc) to detect and interpret conventional metaphor. This system had as its backing an abstraction hierarchy which was used to enable the interpretation components of the system to generalize. In particular, it would attempt to apply manually-coded interpretations associated with a particular type of metaphor, and then, if it failed, would move away from the original metaphor to more and more abstract representations until the metaphor could be understood. Originally, the system was backed by only twenty-two core metaphors (with interpretations), but these were further expanded to 200 as part of the Berkeley Master Metaphor List [12]. In many ways, we follow the intuitions of this work on a much larger scale, using a wide variety of metaphoricity annotations across multiple unrelated domains in multiple languages.

Krishnakumaran and Zhu [10] introduced a key observation – that metaphors can be categorized according to their relationship to some non-metaphoric unit in the text and that the characteristics that indicate metaphoricity differ by category. They proposed three types of metaphors – IS-A metaphors (Type I), verb-noun metaphors (Type II), and adj-noun metaphors (Type III). In order to detect these three types of metaphors, they made extensive use of the WordNet hypernym structure (to rule out literal IS-A relations) and a verb-adj/noun co-occurrence matrix (to rule out common pairings). In effect, they employed co-occurrence information to partition the knowledge base into regions of conventional usage versus unconventional usage which is, arguably, an approximation of (novel) metaphorical usage. More recently, Li et al. [15] built upon this idea by combining extracted figurative comparisons with a probabilistic IS-A knowledge base (ProBase) to partition the semantic space

associated with particular dependency relations in a much more principled way.

Likewise, Hovy et al. [9] focused on the semantic relation between pairs of words in a text by building a tree-kernel classifier that uses (1) a vector representation of individual words and (2) several tree representations of the word. These representations included lemmatized versions of the words, POS tags of the words, and WordNet semantic classes (i.e., lexicographer files). This work represents a clear attempt to generalize from the surface form of the words in their training data. However, by using the full parse tree instead of a relation between the source and target, their methodology was particularly vulnerable to data sparsity.

Similar to our approach in rationale, if not methodology, is Mohler et al. [20], which compared sentence-level utterances against a large collection of sentences validated as either containing metaphors or not. In particular, they sought to compare sentences within a semantic space (defined by a WordNet- and Wikipedia-based "semantic signature"). While this approach successfully addressed the semantics of metaphor, it did not consider the relationship between the source and the target within a sentence. As such, it was heavily impacted by noise associated with the wider context of the sentence.

At the forefront of large-scale, example-based metaphor detection is the work of Levin et al. [14], who have produced a resource containing common metaphors associated with a variety of target concepts in English, Spanish, Russian, and Farsi. Moving beyond the relational categories of Krishnakumaran and Zhu [10], they predicted metaphoricity between lexical pairs in a variety of dependency relations: subj-verb, obj-verb, adv-verb, adj-noun, noun-pred, noun-noun, noun-poss, and noun-prep-noun. The most common terms that co-occur with a target term (e.g., "poverty", "wealth", "taxation") were manually analyzed and annotated as being either conventionalized metaphors or literal language. However, other than normalizing for conjunctions and several semantically "light" nouns (e.g., containers, quantifiers, partitives), no generalization was carried out. Without employing generalization, it is not possible to detect

novel metaphors using a resource such as theirs.

3. METHODOLOGY

We propose a supervised approach to the identification of metaphor which seeks to generalize over an existing dataset annotated for metaphoricity. We model each annotation (without contextual information) using the abstract tuple – (source, target, relation, metaphoricity). It is our hypothesis that the vast majority of utterances in text, represented as such a tuple, are either consistently metaphoric or consistently nonmetaphoric, with the wider context of the utterance having little to no effect on this property.[1] Building on this hypothesis, we assert that a model of the prior metaphorical likelihood of all possible utterances – that is, all possible source/target pairs in all dependency relations[2] – represents a complete solution to the metaphoricity problem. We seek to approximate this level of knowledge by deriving and using semantic patterns that are capable of grouping the metaphoricity decisions of individual examples and applying them to a bounded region of this tuple space. These decisions can then be propagated to utterances in larger and more general regions without the need for humans to annotate such (potentially novel) utterances directly.

While this approach was developed as a supplement to an existing feature-based metaphor identification system [3], we have also made use of it as a stand-alone system, and we employ

---

[1] That said, we acknowledge several classes of utterances such as "Cholera is a disease of poverty"; "Men are animals"; and "The rock began to sing (in a dream)" for which this hypothesis is insufficient. However, we believe that the appropriate means for handling such cases is to determine a metaphoric or non-metaphoric prior likelihood for the utterance and to overcome this prior in anomalous cases using supplementary, context-dependent components tailored to such cases.

[2] This is defined by the set of tuples $(S^V,\ T^V,\ R^N,\ m \in [0..1])$ where V is the vocabulary size and N is the number of possible syntactic elations between two words within a sentence.

it as such throughout this work. Our approach consists of two stages – (1) generalizing our annotations into semantic patterns using a three-step process; and (2) using these semantic patterns to detect linguistic metaphors in unseen data. In this section, we describe these two components with a particular focus on our distinct techniques for individually generalizing the target, the relation, and the source. We then describe our method for combining the results yielded by these patterns to arrive at a single decision for a given input.

3.1. *Generalizing over existing annotations*

In order to effectively generalize from our metaphoricity annotations, we first individually generalize the target and the relation. For the target, we make use of a variety of semantic information associated with its representation in WordNet including (1) its associated synset, (2) its semantic category, and (3) whether it can be considered "concrete" or "abstract". In addition, we generalize the target using a manual grouping of terms related to a small set of target domains. The domains under consideration – GOVERNMENT, BUREAUCRACY, DEMOCRACY, ELECTIONS, POVERTY, WEALTH, and TAXATION – were selected to represent a variety of distinct domains with different characteristics.[3] Relations are generalized using a rule-based, pseudo-semantic role labeling process which maps from dependency chains to a small set of abstract semantic relations.

For each target/relation representation pair, we define a WordNet semantic hierarchy upon which we can map individual source lexemes, along with their metaphoricity annotations. Once these annotations have been linked to WordNet, it is possible to begin the process of deriving semantic patterns. We define a semantic pattern according to the tuple, $(g(S, x), T, R, m)$, where T corresponds to some representation of the target lexeme, R corresponds to some representation of the relation between a

[3] These domains correspond to those under consideration as part of the IARPA Metaphor program.

source and target within a text, m corresponds to a binary metaphoricity decision (metaphor or non-metaphor), and the function g (S,x) defines a group within the semantic space S (such as all nodes in a subtree) that contains the annotated source x. In the sections to follow, we will describe several techniques for representing T and R at various levels of generality along with our methodology for partitioning the semantic space S (for a given T, R) into regions within which we assert that m must be either always true or always false.

Generalizing Targets In addition to the lexical (surface form) representation of a target, T_{LEX}, we propose four alternative representations. First, we link the target lexeme to the WordNet synset that corresponds to its most frequent sense[4] – T_{SYN}. This enables us to directly propagate metaphoricity annotations to all synonyms of a given term. More importantly, this allows us to propagate annotations to synonyms cross-lingually using the cross-lingual links (within a synset) associated with our multi-lingual version of WordNet.

Next, we represent a given target using the WordNet semantic category (i.e., lexicographer file) associated with its most frequent sense – T_{SEM}. This permits us to infer the metaphoricity of the tuple (devours, NOUN.STATE, dobj) from the annotated example (devours, poverty, dobj). This is because the semantic category NOUN.STATE includes additional target lexemes "health", "silence", "guilt", and "comfort" – none of which is literally capable of "devouring".

Third, we partition the noun hierarchy of WordNet into an abstract-concrete dichotomy – T_{ABS} – which indicates whether the synset PHYSICAL ENTITY is or is not a direct or indirect hypernym of the target. While this generalizes the targets in a very coarse way, a significant number of physical interaction

[4] This was determined to be sufficient for the limited number of target lexemes that we consider, but will need to be reconsidered moving forward.

verbs (e.g., “drop”, “carry”, “throw”, “touch”, “eat”) will be metaphoric for all abstract nouns.

Finally, we represent the target according to the domain or topic with which it is associated – T_{DOM}. For our purposes, this corresponds to a manually-created list of lexical items, each related to one of the seven domains mentioned above and separated according to part-of-speech. This permits us to group together such target lexemes as “taxes”, “taxation”, and “income tax”.

3.2. *Generalizing relations*

For representing the relation, R, which links the source and the target in text, we explore two possibilities. First, we use the dependency relation (as determined by MaltParser) between the source and the target directly – R_{DEP}. This relation is then post-processed to remove conjunction relations (“conj”), so that both “government” and “bureaucracy” in the phrase “government and bureaucracy punish us” will have the same subject relation to the verb “punish”.

In order to better promote generalization, we also represent the relation, R, by transforming it into a language-independent, pseudo-semantic relation – R_{SEM} – derived using a small number of manually-crafted rules over raw dependency relations. For our purposes, we define a small set of pseudo-semantic relations – AGENCY, PATIENCY, and MODIFICATION[5] between the source and target. As an example, each of the following phrases associates the target (“wealth”) as an AGENT to some form of the source verb “enslave”: “wealth enslaves people” (nsubj), “wealth continues to enslave people” (nsubj+xcomp^{-1}), “wealth which enslaves people” (rcmod^{-1}), “causes wealth to enslave people” (infmod^{-1}), “for wealth to enslave people” (prep for+infmod^{-1}), and “chained to my enslaving wealth” (amod^{-1}). Because in each of these cases “wealth” is metaphorically “enslaving” someone, we avoid sparsity (and improve coverage) by propagating our annotations to all of these relations through

[5] We are concerned here with adjectival modifiers only.

this generalization step. We define equivalent rules to transform raw dependency relations to these three pseudosemantic relations for each of our four languages.[6] Since these pseudo-semantic relations are consistent across languages, their use enables us to propagate annotations across languages by pairing these relations with any of the language independent target representations (i.e., T_{SYN}, T_{ABS}, T_{SEM}, or T_{DOM}) and deriving cross-lingual semantic patterns for the pair's associated hierarchy.

Table 1. *Candidate LM sources – categorized according to metaphoricity – for "bureaucratic [SOURCE]" given the adjective-noun dependency relation (amod)*

Metaphoric	Candidate	LM Sources	Non-Metaphoric	Candidate	LM Sources
Maze	Monster	Fiefdoms	Process	Management	Inconsistencies
Nightmare	Sorcery	System	Control	Activity	Administration
Hell	Basement	Perdition	Adjudication	Tenure	Occupation

Generalizing Sources Throughout the Knowledge Base Once we have defined a target/ relation pair, we begin the process of generalizing our source lexemes and deriving semantic patterns. Table 1 shows a variety of source lexical items taken from our annotations that share an "amod^{-1}" relation with the target lexeme "bureaucratic"[7] Before generalizing, we must first link each of these annotated source lexemes into a semantic knowledge base, such as WordNet[8]. WordNet represents an ideal knowledge base for our purposes due to its ability to simultaneously group individual senses (i.e., synsets and semantic categories) and to define a hierarchical structure within groups using hypernymy relations. However, the problem of

[6] While these rule-based groupings do not approach the accuracy of state-of-the-art semantic role labeling (SRL) systems (at least in English), we believe that their quality is sufficient for the task of metaphor identification. The use of move advanced SRL technologies for this task remains an open problem.

[7] These are organized as "metaphoric" or "non-metaphoric" according to the predominant metaphoricity judgement in our annotations.

word sense ambiguity makes it non-trivial to link individual lexical items to particular synsets. This is even more problematic when the lexical items to be linked are being used in non-literal ways. To account for this, we perform a light disambiguation step that filters out potential senses whose annotated metaphoricity neighborhood (in the semantic hierarchy) fails to match the metaphoricity of the annotation. For example, the lexeme "Hell" can be linked to either a synset that contains the word "Perdition" or a synset with a hypernym of "Mischief". Given the existence of an annotation for "Perdition" with the same metaphoricity, the first synset is preferred. After limiting the number of potential word-sense mappings in this way[8], annotated source terms are linked to all remaining senses.

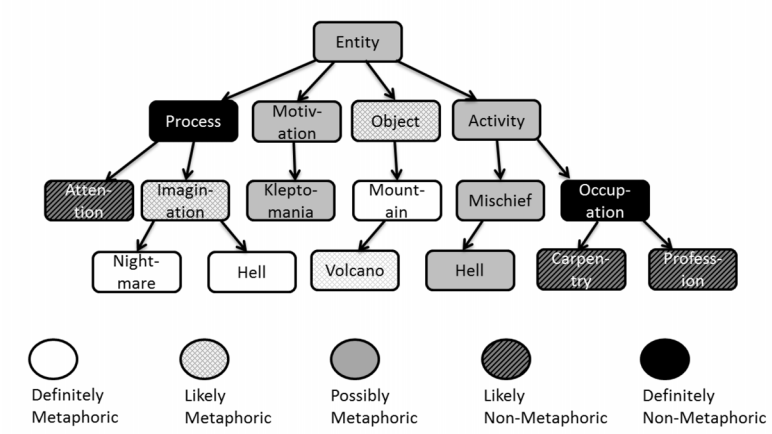

Figure 1. *Categorizing nodes in the source semantic hierarchy based upon the annotated source lexemes from Table 1.*

After each annotation has been linked to one or more nodes in the source semantic hierarchy, metaphoricity decisions are propagated to the nodes themselves. Each node in the hierarchy is categorized as one of the following:

[8] Further details omitted, due to space.

1. **definitely non-metaphoric** – if the annotations for a particular node are entirely negative;
2. **definitely metaphoric** – if the annotations for the node are at least partially positive;
3. **likely non-metaphoric** – if there are no direct annotations, but all of the node's direct or indirect ancestors or descendants are known to be non-metaphoric;
4. **likely metaphoric** – if there are no direct annotations, but all of the node's direct or indirect ancestors or descendants are known to be metaphoric;
5. **possibly metaphoric** – if there is no indication at all among the existing annotations (or if the results are mixed).

The associated categorization of each node in a simplified source hierarchy (based on the annotations from Table 1) is shown in Figure 1.

At this point, we must define our semantic patterns by selecting a set of nodes in the source hierarchy (and assign a metaphoricity judgement to each) such that the metaphoricity judgement associated with each node's closest pattern-node ancestor will determine whether it is likely to be metaphoric or not. We constrain our pattern selection process such that it is constrained to correctly categorize all "definitely metaphoric" and "definitely non-metaphoric" nodes (i.e., those representing the annotations themselves). The remaining categories ("likely metaphoric", "possibly metaphoric", and "likely nonmetaphoric") represent a continuum along which the system can increase recall at the expense of precision. For tasks that require a higher recall, the system can be further constrained to select, for instance, "likely metaphoric" nodes as metaphors as well. Any nodes in groups that have not been constrained in this way may be dominated by patterns suggesting metaphoricity or non-metaphoricity or they may be dominated by no pattern at all.

In order to encourage generalization throughout the source hierarchy, we wish to select a minimal set of patterns nodes that satisfy our constraints. These nodes can be selected efficiently using a straightforward, tree-based dynamic programming methodology in which the state space is defined by two

dimensions: (1) the current node, *N*, and (2) the current metaphoricity judgment, *m*, defined by the pattern that dominates it (i.e., the nearest pattern node on its ancestor chain). If the system has been constrained to judge the current node as metaphoric or non-metaphoric (as described above), some semantic pattern must be correctly applied to this node – either the pattern that currently dominates the node, or a new pattern defined on the node itself. Otherwise, if the current node is under no constraints, we are free to assign a metaphoricity pattern or a non-metaphoricity pattern to the node or to add no pattern and allow any ancestor pattern in effect to continue dominating the hierarchy. This process is described more formally by the following equations[9] where r(*N*) represents the required constraints on node N – i.e., "metaphor" (met), "non-metaphor" (lit), "unconstrained" (unc) – and N_c represents a direct descendant of N in the source semantic hierarchy:

$$f(N,m) = \begin{cases} \min\limits_{x \in h(m)} g(N,m,x) & \text{if } r(N) = unc \\ g(N,m,r(N)) \end{cases} \quad (1)$$

$$g(N,m,x) = \begin{cases} \sum\limits_{c} f(N_c,x) & \text{if } m = x \\ 1 + \sum\limits_{c} f(N_c,x) \end{cases} \quad (2)$$

$$h(m) = \begin{cases} \{met, lit, unc\} & \text{if } m = unc \\ \{met, lit\} \end{cases} \quad (3)$$

The count of the minimal number of patterns can be computed by summing the results of $f(N_R,$ "unc") for each root node, N_R, in

[9] In effect, $h(m)$ is the set of potential pattern decisions that can be selected based on the ancestor patterns; $g(N, m, x)$ is the number of new patterns required to change the current node *N*, from m to x, and to correctly cover all descendants; and $f(N,m)$ is the number of new patterns required to cover the current node, *N*, and all of its descendants given the ancestor pattern m.

the hierarchy. Once these counts have been computed over the state space, selecting the nodes that lead to this optimal state can be accomplished by greedily traversing the state space along locally optimal paths. Note that the result of this process is a partitioning of the entire semantic space of the source hierarchy into metaphoric, non-metaphoric, and unclear regions for a given target/relation pair. Once a set of pattern nodes has been selected that satisfies our constraints, they can be used either as a stand-alone, semantics-only metaphor detection system or in conjunction with an existing metaphor detection system.

3.3. *Detection of metaphors in unseen text*

Employing these patterns in a stand-alone metaphor detection system requires two things: (1) a strategy for selecting potential source/target pairs in text and (2) an algorithm for combining the metaphoricity decisions associated with multiple target/relation representations. For the first of these, we begin with a list of lexical items that have been manually associated with our seven target domains. These have been supplemented using word senses gathered from target domain signatures for each domain in the manner of Bracewell et al.[2]. For each target term selected, we consider all content words in the same sentence as potential sources.[10]

Once the source/target pairs have been extracted from a text, they are converted into (source, target, relation) tuples – for each target/relation representation – and compared against the patterns derived in Section 3.1. We then treat our patterns as a cascade, analyzing groups in increasing order of abstractness of the target representation – T_{LEX}, T_{SYN}, T_{SEM}, T_{DOM}, T_{ABS}. Within each group, we compare the two relation representations – R_{DEP} and R_{SEM} – which can either provide the same metaphoricity

[10] We additionally collapse hyphenated terms and collocations when selecting both the source and the target. For collocations and hyphenated terms that contain a target term, we produce a sub-collocate candidate pair – e.g., (stricken, poverty, dep) for “poverty-stricken”.

judgement, different metaphoricity judgments, or indicate that there is no clear decision. If they agree (or if only one provides a clear answer), then that metaphoricity decision is assigned to the pair. If they disagree (or cannot provide an answer), the next group is considered. If none of the groups in the cascade results in a response, the metaphoricity of the pair remains unclear and is reported as such. This represents a cascading structure from specific to more general representations of the target and relation.

4. DATASETS

In order to evaluate our methodology for generalizing over metaphor annotations, we make use of four datasets (in four languages – English, Spanish, Russian, and Farsi) developed by a team of annotators with native-level proficiency. The size and characteristics of each dataset are summarized in Table 2. The first dataset (ANN) consists of examples selected by the annotators using targeted web searches for representative (non-conventionalized) metaphors for particular pairs of source and target concepts that were of interest at the program level. While this dataset is a good source of novel metaphors in a variety of source and target domains, it includes only a single annotation (with source and target) for the full sentence and does not include any non-metaphor annotations.

The second dataset (REC) was developed to address this problem by providing a more natural source of data for training the machine learning component of our overall system with a significant number of non-metaphor annotations. Individual documents were selected automatically to be annotated thoroughly. For these documents, annotators were provided with all source/target pairs that had been selected as described in Section 3.3. This is the most 'natural' of our four datasets. The third (EVAL) was annotated in the same way by our annotators, but was provided by a third party for the purpose of evaluating our system's ability to detect metaphors in unseen data. As such, it has a slight bias towards metaphoricity.

Table 2. *Datasets used in our experiments with size and balance information. The REDUCED set corresponds to the subset of the combined dataset that can be represented using a pseudosemantic relation representation as described in Section 3.2.*

Language	Dataset	Metaphors	Non-Metaphors	Unclear	Language	Dataset	Metaphors	Non-Metaphors	Unclear
	ANN	8,667	0	125		ANN	4,308	0	353
	REC	402	25,238	340		REC	223	43,241	1,213
EN	EVAL	284	7,130	177	ES	EVAL	177	15,533	400
	SYS	8,507	29,563	6,562		SYS	7,053	28,782	8,633
	TOTAL	17,860	61,931	7,204		TOTAL	11,761	87,556	10,599
	REDUCED	5,512	7,080	2,683		REDUCED	5,158	10,946	4,745
	ANN	3,436	0	322		ANN	2,229	0	168
	REC	1,204	29,225	758		REC	367	51,884	188
RU	EVAL	341	5,538	321	FA	EVAL	186	21,081	222
	SYS	5,823	14,664	2,834		SYS	7,883	27,186	6,369
	TOTAL	10,804	49,427	4,235		TOTAL	10,665	100,151	6,947
	REDUCED	6,261	8,548	2,215		REDUCED	242	2,453	82

Our final and largest dataset (SYS) consists of a validated subset of the output from both our machine learning-based detection system and earlier versions of the standalone semantic generalization component described which is the focus of this work. Our full dataset consists of vastly more examples that have not been validated. Those that have been validated were selected using an $ad\ hoc$ active learning setting that considered a variety of characteristics – including similarity to existing annotations, target/source diversity, relation diversity, system confidence, and disagreement between the ML system and the semantic generalization components.

The datasets labeled "REDUCED" in Table 2 represent a subset of the combined dataset (all four of the above) which consists of those only instances that can be represented using a pseudo-semantic relation – i.e., metaphors (or literal utterances) with subj-pred, obj-pred, or adj-noun relations. Only this subset can be used and tested crosslingually.

For each dataset, annotators were asked to judge metaphoricity according to criteria comparable to the MIP annotation guidelines [23]. Following the insights of Dunn [6], we have instructed the annotators to employ a four-point metaphoricity scale corresponding to: (0) no metaphoricity, (1)

possible (or weak) metaphoricity, (2) likely (or conventionalized) metaphoricity, or (3) clear metaphoricity. In some cases, multiple annotators provided scores for individual instances, and so we use the average score across all annotators. Using these averages, we categorize each annotated instance as "metaphoric" (score≥1.5), "non-metaphoric" (score<0.5), or "unclear" (0.5<score≤1.5).

Table 3. *Experiments showing the performance of each target/relation representation alongside the combined "cascade" and a lexical baseline. This represents a 10-fold cross-validation over the combined dataset for a given language. For those using the R_{SEM}, the REDUCED dataset it used. The numbers reported above correspond to the F-measure for detecting metaphor. When there is a range specified (low/high), it is due to the way that we are categorizing annotations marked as "unclear". On the low end, annotated examples labeled "unclear" are ignored entirely, while on the high end, any "unclear" examples are labeled as "metaphor" by the system to improve precision. Recall is unaffected by this distinction*.

Target/Relation	ENGLISH	SPANISH	RUSSIAN	FARSI	**Target/Relation**	ENGLISH	SPANISH	RUSSIAN	FARSI
CASCADING	**0.92**	**0.90/0.91**	**0.88**	**0.86/0.87**	BASELINE	0.73	0.76	0.53	0.78/0.79
T_{LEX}, R_{DEP}	0.91	**0.90/0.91**	0.87	0.85/0.86	T_{LEX}, R_{SEM}	0.89/0.90	0.90/0.91	0.87/0.88	0.62
T_{DOM}, R_{DEP}	0.86/0.87	0.87/0.88	0.82/0.83	0.82/0.83	T_{DOM}, R_{SEM}	0.84/0.85	0.87/0.88	0.82	0.60/0.61
T_{SYN}, R_{DEP}	0.87	0.86	0.82	0.80/0.81	T_{SYN}, R_{SEM}	0.86	0.87/0.88	0.81	0.56
T_{SEM}, R_{DEP}	0.83/0.84	0.75/0.76	0.71	0.78/0.79	T_{SEM}, R_{SEM}	0.79	0.67/0.68	0.65/0.66	0.52
T_{ABS}, R_{DEP}	0.82/0.83	0.74/0.75	0.54	0.75/0.77	T_{ABS}, R_{SEM}	0.78/0.79	0.65/0.66	0.47/0.48	0.51/0.52

In deriving our semantic patterns (cf. Section 3), we have not made use of instances labeled "unclear".

5. EXPERIMENTS

In order to highlight the contributions of our approach to semantic generalization for metaphor detection, we have carried out two experiments. First, we evaluate the performance of the semantic patterns derived for individual target/relation representation pairs as well as the performance of the full system in the cascading framework described in Section 3.3. Then, we determine the extent to which annotations from one or more

languages can be applied to the task of metaphor identification in a separate language for which no annotations are available.

5.1. *Monolingual generalization experiments*

In our first experiment, we test the ability of our semantic generalization component (using each of the target/relation representations described in 3.1) to detect metaphors in a monolingual setting. We compare each target/relation representation (used in isolation) against both our cascading combination of patterns described in Section 3.3 and a fully lexical baseline. This baseline consists of the following: for each example in the test fold, we find an exact lexical match of the tuple ($S,\ T_{LEX},\ R_{DEP}$) – meaning the hierarchy was not used – and then apply the most common metaphoricity decision from our annotations. If no lexical matches are found, it is labeled as "unclear".

We have performed our experiment over the combined datasets described in Section 4 using a 10-fold cross-validation – that is, for evaluation against each fold of the data, we develop our semantic patterns over the remaining 9 folds. We report the results of these experiments in Table 3.

The lexical baseline system predictably resulted in very high precision (> 95%) with a comparatively low recall (appx. 50%). For each language, the cascading combination system performed best, highlighting the advantage associated with our overall methodology. Among the individual target/relation pairs, performance was comparable (slightly better) using the raw dependency relation compared to using the pseudosemantic relations. For the target representation, the lexical (T_{LEX}) representation resulted in the best performance followed by the manual domain-level term groupings (T_{DOM}) and the synsets (T_{SYN}). The remaining representations – WordNet semantic categories (T_{SEM}) and the abstract/concrete noun dichotomy (T_{ABS}) – resulted in lower performance due to their coarseness, especially for verbs and adjectives.

Table 4. *Experiments in applying annotations from the other three languages to a given language using cross-lingual target/relation representations*

	ENGLISH	SPANISH	RUSSIAN	FARSI
RECALL	.28	0.48	0.16	0.20
PRECISION	0.72/0.78	0.58/0.72	0.72/0.75	0.26/0.29
F-MEASURE	0.40/0.41	0.53/0.57	0.26	0.22/0.24

5.2. *Cross-lingual generalization experiments*

In our second experiment, we attempt to detect metaphors in the combined dataset of one language using patterns developed from the datasets of the remaining three languages. That is, we make use of no native-language annotations in determining metaphoricity. In particular, these experiments make use of only the "REDUCED" dataset from Section 4 for which we are able to convert each annotation's source/target relation into a cross-lingual, pseudo-semantic relation. All target representations (except for the T_{LEX}) can be applied cross-lingually. The results of these experiments are shown in Table 4.

For English, Spanish, and Russian, the precision of the resulting system at detecting metaphor is above 70%, while recall for these three languages ranges from 16-48%. We believe that this represents good out-of-the-box performance on a novel language with no annotated data. For Farsi, on the other hand, precision is much poorer, but it is difficult to draw conclusions, due to the small size of the "REDUCED" dataset in this language.

6. CONCLUSION

In this work, we have presented a novel approach to the generalization of metaphoricity annotations across a semantic hierarchy. We have clearly shown the advantage of generalizing sources throughout this hierarchy with an F-measure above 0.85 for four languages in 10-fold cross-validation experiments over a very large annotated dataset of metaphoricity. This is compared to a purely lexical baseline with F-measure between 0.53-0.79 across all languages. In a monolingual setting, the benefits

associated with generalizing the targets across the semantic hierarchy are less clear. We theorize that this can mostly be attributed to the restricted number of domains and target lexical items that are represented in our dataset and that this type of generalization would have a more significant effect when applied to novel domains. Likewise, the pseudosemantic relations we propose were not shown to outperform raw dependency relations within our monolingual datasets.

However, we have shown that the pseudo-semantic relations are able to be applied to the task of cross-lingual metaphor detection. For three of our languages, we were able to detect metaphors with over 70% precision and a recall ranging from 16-48%. This clearly shows the utility of our approach in tackling the task of metaphor detection in novel languages with a minimal development cost.

In future work, we intend to apply our approach to semantic generalization to a variety of novel domains to better explore the effect of target-level generalization. In addition, we hope to supplement WordNet with additional pseudo-semantic categories derived using the distributional similarity of terms for use in languages where there is no corresponding version of WordNet or where its quality is poor. Finally, we plan to analyze the performance of our semantic generalization component in an active (or co-active) learning framework in combination with a feature-based metaphor detection system.

ACKNOWLEDGMENTS

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0025. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government."

REFERENCES

1. Bogdanova, D. 2010. A framework for figurative language detection based on sense differentiation. In proceedings of the *ACL 2010 Student Research Workshop* (pp. 67-72), Association for Computational Linguistics.
2. Bracewell, D., Tomlinson, M. & Mohler, M. 2013. Determining the conceptual space of metaphoric expressions. In *Computational Linguistics and Intelligent Text Processing* (pp. 487-500), Springer.
3. Bracewell, D., Tomlinson, M., Mohler, M. & Rink, B. 2014 A tiered approach to the recognition of metaphor. In *Computational Linguistics and Intelligent Text Processing*.
4. Broadwell, G. A., Boz, U., Cases, I., Strzalkowski, T., Feldman, L., Taylor, S., Shaikh, S., Liu, T., Cho, K. & Webb, N. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 102-110), Springer.
5. Dunn, J. 2013. What metaphor identification systems can tell us about metaphor-in-language. *Meta4NLP 2013* (p. 1).
6. Dunn, J. 2014. Measuring metaphoricity. In proceedings of the *52nd annual meeting of the Association for Computational Linguistics* (pp. 745-751), Vol. 2, Association for Computational Linguistics Stroudsburg, PA.
7. Fass, D. 1991. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17/1, 49-90.
8. Fellbaum, C. 1998. *WordNet, An Electronic Lexical Database*, The MIT Press.
9. Hovy, D., Srivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., Sanders, W. & Hovy, E. 2013. Identifying metaphorical word use with tree kernels. *Meta4NLP 2013* (p. 52).
10. Krishnakumaran, S. & Zhu, X. 2007. Hunting elusive metaphors using lexical resources. In proceedings of the *Workshop on Computational approaches to Figurative Language* (pp. 13-20), Association for Computational Linguistics.
11. Lakoff, G. 1993. The contemporary theory of metaphor. *Metaphor and Thought*, 2, 202-251.
12. Lakoff, G. 1994. *Master Metaphor List*. University of California.
13. Lakoff, G. & Johnson, M. 1980. *Metaphors we live by*, 111. Chicago London.
14. Levin, L., Mitamura, T., Fromm, D., MacWhinney, B., Carbonell, J., Feely, W., Frederking, R., Gershman, A. & Ramirez, C. 2014.

Resources for the detection of conventionalized metaphors in four languages. In proceedings of the *Ninth Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

15. Li, H., Zhu, K.Q. & Wang, H. 2013. Data-driven metaphor recognition and explanation. *TACL 1* (pp. 379-390).
16. Li, L. & Sporleder, C. 2010. Linguistic cues for distinguishing literal and non-literal usages. In proceedings of the *23rd International Conference on Computational Linguistics*: Posters (pp. 683-691), Association for Computational Linguistics.
17. Martin, J. H. 1994. MetaBank: A knowledge-base of metaphoric language conventions. *Computational Intelligence*, 10/2, 134-149.
18. Martin, J. 1990. A computational model of metaphor interpretation. Academic Press Professional, Inc.
19. Mason, Z. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30/1, 23-44.
20. Mohler, M., Bracewell, D., Hinote, D. & Tomlinson, M. 2013. Semantic signatures for example based linguistic metaphor detection. *Meta4NLP 2013* (p. 27).
21. Nayak, S. & Mukerjee, A. 2012. A grounded cognitive model for metaphor acquisition. In *AAAI*.
22. Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N. & Frieder, O. 2013. Metaphor identification in large texts corpora. *PloS one*, 8/4, e62343.
23. Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22/1, 1-39.
24. Sardinha, T. B. 2010. A program for finding metaphor candidates in corpora. *ESPecialist*, 31/1, 49-67.
25. Schulder, M. & Hovy, E. 2014. Metaphor detection through term relevance. *ACL 2014* (p. 18).
26. Sporleder, C. & Li, L. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In proceedings of the *12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 754-762), Association for Computational Linguistics.
27. Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E. & Dyer, C. 2014. Metaphor detection with cross-lingual model transfer. In: *Proceedings of ACL*.
28. Turney, P. D., Neuman, Y., Assaf, D. & Cohen, Y. 2011. Literal and metaphorical sense identification through concrete and abstract

context. In proceedings of the *2011 Conference on the Empirical Methods in Natural Language Processing* (pp. 680-690).
29. Wilks, Y., Galescu, L., Allen, J. & Dalton, A. 2013. Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction. *Meta4NLP 2013* (p. 36)

MICHAEL MOHLER
LANGUAGE COMPUTER CORPORATION,
RICHARDSON, TX.
E-MAIL: <MICHAEL@LANGUAGECOMPUTER.COM>
HTTP://WWW.LANGUAGECOMPUTER.COM

MARC TOMLINSON
LANGUAGE COMPUTER CORPORATION,
RICHARDSON, TX.
E-MAIL: <MARC@LANGUAGECOMPUTER.COM>
HTTP://WWW.LANGUAGECOMPUTER.COM

BRYAN RINK
LANGUAGE COMPUTER CORPORATION,
RICHARDSON, TX.
E-MAIL: < BRYANG@LANGUAGECOMPUTER.COM>
HTTP://WWW.LANGUAGECOMPUTER.COM