Data-Driven Metaphor Recognition and Explanation

Hongsong Li
Microsoft Research Asia
hongsli@microsoft.com

Kenny Q. Zhu
Shanghai Jiao Tong University
kzhu@cs.sjtu.edu.cn

Haixun Wang
Google Research
haixun@google.com

Abstract

Recognizing metaphors and identifying source-target mappings is an important task as metaphorical text poses a big challenge for machine reading. To address this problem, we automatically acquire a metaphor knowledge base and an isA knowledge base from billions of web pages. Using the knowledge bases, we develop an inference mechanism to recognize and explain the metaphors in the text. To our knowledge, this is the first purely data-driven approach of probabilistic metaphor acquisition, recognition, and explanation. Our results show that it significantly outperforms other state-of-the-art methods in recognizing and explaining metaphors.

1 Introduction

A metaphor is a way of communicating. It enables us to comprehend one thing in terms of another. For example, the metaphor, *Juliet is the sun*, allows us to see Juliet much more vividly than if Shakespeare had taken a more literal approach. We utter about one metaphor for every ten to twenty-five words, or about six metaphors a minute (Geary, 2011).

Specifically, a metaphor is a mapping of concepts from a source domain to a target domain (Lakoff and Johnson, 1980). The source domain is often concrete and based on sensory experience, while the target domain is usually abstract. Two concepts are connected by this mapping because they share some common or similar properties, and as a result, the meaning of one concept can be *transferred* to another. For example, in "Juliet is the sun," *the sun* is the source concept while *Juliet* is the target concept. One interpretation of this metaphor is that both concepts share the property that their existence brings about warmth, life, and excitement. In a metaphorical sentence, at least one of the two concepts must be explicitly present. This leads to three types of metaphors:

1. *Juliet is the sun.* Here, both the source (*sun*) and the target (*Juliet*) are explicit.

2. *Please wash your claws before scratching me.* Here, the source (*claws*) is explicit, while the target (*hands*) is implicit, and the context of *wash* is in terms of the target.

3. *Your words cut deep.* Here, the target (*words*) is explicit, while the source (possibly, *knife*) is implicit, and the context of *cut* is in terms of the source.

In this paper, we focus on the *recognition* and *explanation* of metaphors. For a given sentence, we first check whether it contains a metaphoric expression (which we call metaphor recognition), and if it does, we identify the source and the target concepts of the metaphor (which we call metaphor explanation). Metaphor explanation is important for understanding metaphors. Explaining type 2 and 3 metaphors is particularly challenging, and, to the best of our knowledge, has not been attempted for nominal concepts[1] before. In our examples, knowing that *life* and *hands* are the target concepts avoids the confusion that may arise when the source concepts (*sun* and *claws*) are used literally in understanding the sentences. This, however, does not

[1]Nominal concepts are those represented by noun phrases.

mean that the source concept is a useless embellishment. In the 3rd sentence, knowing that *words* is mapped to *knife* enables the system to understand the emotion or the sentiment embedded in the text. This is the reason why metaphor recognition and explanation is important to applications such as affection mining (Smith et al., 2007).

It is worth noting that some prefer to consider the verb "cut", rather than the noun "words", to be metaphoric in the 3rd sentence above. We instead concentrate on nominal metaphors and seek to explain source-target mappings in which at least one domain is a nominal concept. This is because verbs usually have nominal arguments, as either subject or object, thus explaining the source-target mapping of the nominal argument covers most, if not all, cases where a verb is metaphoric.

In order for machines to recognize and explain metaphors, it must have extensive human knowledge. It is not difficult to see why metaphor recognition based on simple context modeling (e.g., by selectional restriction/preference (Resnik, 1993)) is insufficient. First, not all expressions that violate the restriction are metaphors. For example, *I hate to read Heidegger* violates selectional restriction, as the context (embodied by the verb *read*) prefers an object other than a person (*Heidegger*). But, *Heidegger* is not a metaphor but a metonymy, which in this case denotes *Heidegger's books*. Second, not every metaphor violates the restriction. For example, *life is a journey* is clearly a metaphor, but selectional restriction or preference is helpless when it comes to the isA context.

Existing approaches based on human-curated knowledge bases fall short of the challenge. First, the scale of a human-curated knowledge base is often very limited, which means at best it covers a small set of metaphors. Second, new metaphors are created all the time and the challenge is to recognize and understand metaphors that have never been seen before. This requires extensive knowledge. As a very simple example, even if the machine knows *Sports cars are fire engines* is a metaphor, it still needs to know what is a sports car before it can understand *My Ferrari is a fire engine* is also a metaphor. Third, existing human-curated knowledge bases (including metaphor databases and the WordNet) are not probabilistic. They cannot tell how typical an instance is of a category (e.g., a *robin* is a more typical bird than a *penguin*), or how popular an expression (e.g., *a breath of fresh air*) is used as a source concept to describe targets in another concept (e.g., *young girls*). Unfortunately, without necessary probabilistic information, not much reasoning can be performed for metaphor explanation.

In this paper, we address the above challenges. We start with a probabilistic isA knowledge base of many entities and categories harnessed from billions of web documents using a set of strict syntactic patterns known as the Hearst patterns (Hearst, 1992). We then automatically acquire a large probabilistic metaphor database with the help of both syntactic patterns and the isA knowledge base (Section 3). Finally we combine the two knowledge bases and a probabilistic reasoning mechanism for automatic metaphor recognition and explanation (Section 4).

This paper makes the following contributions:

1. To our knowledge, we are the first to introduce the metaphor explanation problem, which seeks to recover missing or implied source or target concepts in an implicit metaphor.

2. This is the first big-data driven, unsupervised approach for metaphor recognition and explanation. One of the benefits of leveraging big data is that the knowledge we obtain is less biased, has great coverage, and can be updated in a timely manner. More importantly, a data driven approach can associate with each piece of knowledge probabilities which are not available in human curated knowledge but are indispensable for inference and reasoning.

3. Our results show the effectiveness both in terms of coverage and accuracy of our approach. We manage to acquire one of the largest metaphor knowledge bases ever existed with a precision of 82%. The metaphor recognition accuracy significantly outperforms the state-of-the-art methods (Section 5).

2 Related Work

Existing work on metaphor recognition and interpretation can be divided into two categories: *context-oriented* and *knowledge-driven*. The approach proposed in this paper touches on both categories.

2.1 Context-oriented Methods

Some previous work relies on *context* to differentiate metaphorical expressions from literal ones (Wilks, 1978; Resnik, 1993). The selection restriction theory (Wilks, 1978) argues that the meaning of an expression is restricted by its context, and violations of the restriction imply a metaphor.

Resnik (1993) uses KL divergence to measure the *selectional preference strength (SPS)*, i.e., how strongly a context restricts an expression. Although he did not use this measure directly for metaphor recognition, SPS (and also a related measure called the selection association) is widely used in more recent approaches for metaphor recognition and interpretation (Mason, 2004; Shutova, 2010; Shutova et al., 2010; Baumer et al., 2010). For example, Mason (2004) learns domain-specific selectional preferences and use them to find mappings between concepts from different domains. Shutova (2010) defines metaphor interpretation as a paraphrasing task. The method discriminates between literal and figurative paraphrases by detecting selectional preference violation. The result of this work has been compared with our approach in Section 5. Shutova et al. (2010) identify concepts in a source domain of a metaphor by clustering verb phrases and filtering out verbs that have weak selectional preference strength. Baumer (2010) uses semantic role labeling techniques to calculate selectional preference on semantic relations instead of grammatic relations for metaphor recognition.

A less related but also context-based work is analogy interpretation by relation mapping (Turney, 2008). The problem is to generate mapping between source and target domains by computing pair-wise co-occurrences for different contextual patterns.

Our approach uses selectional restriction when enriching the metaphor knowledge base, and adopts context preference when explaining type 2 and 3 metaphors by focusing on the nearby verbs of a potential source or target concept.

2.2 Knowledge-driven Methods

A growing number of works use knowledge bases for metaphor understanding (Martin, 1990; Narayanan, 1997; Barnden et al., 2002; Veale and Hao, 2008). MIDAS (Martin, 1990) checks if a sentence contains an expression that can be explained by a more general metaphor in a human-curated metaphor knowledge base. ATT-Meta (Barnden et al., 2002) performs metaphor reasoning with a human-curated metaphor knowledge base and first order logic, and it focuses on affection detection (Smith et al., 2007; Agerri, 2008; Zhang, 2010). Krishnakumaran and Zhu (2007) use the isA relation in WordNet (Miller, 1995) for metaphor recognition. Gedigian et al. (2006) use FrameNet (Fillmore et al., 2003) and Probank (Kingsbury and Palmer, 2002) to train a maximum entropy classifier for metaphor recognition. TroFi (Birke and Sarkar, 2006) redefines literal and non-literal as two senses of the same verb and provide two senses with seed sentences from human-curated knowledge bases like WordNet, known metaphor and idiom sets. For a given sentence containing target verb, it compares the similarity of the sentence with two seed sets respectively. If the sentence is closer to the non-literal sense set, the verb is recognized as non-literal usage.

While the above work all relies on human curated data sets or manual labeling, Veale and Hao (2008) introduced the notion of talking points which are figurative properties of noun-based concepts. For example, the concept "Hamas" has the following talking points: *is_islamic:movement* and *governs:gaza_strip*. They automatically constructed a knowledge base called *Slip Net* from WordNet and Web corpus. Concepts that are connected on the Slip Net can "slip" to one another and are hence considered related in a metaphor. However, straightforward traversal on the Slip Net can become computationally impractical and the authors did not elaborate on the implementation details. In practice, the knowledge acquired in this paper is much larger but our algorithms are computationally more feasible.

3 Obtaining Probabilistic Knowledge

In this section, we describe how to use a large, general-purpose, probabilistic isA knowledge base Γ_H to create a probabilistic metaphor dataset Γ_m. Γ_H contains isA pairs as well as scores associated with each pair. The metaphor dataset Γ_m contains metaphors of the form: $(source,\ target)$, and a weight function P_m that maps a metaphor pair to a probabilistic score. The purpose of creating Γ_H is

to help clean and expand Γ_m, and to perform probabilistic inference for metaphor detection.

3.1 IsA Knowledge Γ_H

Γ_H, a general-purpose, probabilistic isA knowledge base, was previously constructed by Wu et al. (2012).[2] Γ_H contains isA relations in the form of (x, h_x), a pair of hyponym and hypernym, for example, (*Steve Ballmer*, *CEO of IT companies*), and each pair is associated with a set of probabilistic scores. Two of the most important scores are known as *typicality*: $P(x|h_x)$, the typicality of x of category h_x, and $P(h_x|x)$, the typicality of category h_x for instance x, which will be used in metaphor recognition and explanation. Both scores are approximated by frequencies, e.g.,

$$P(x|h_x) = \frac{\text{\# of } (x, h_x) \text{ in Hearst extraction}}{\text{\# of } h_x \text{ in Hearst extraction}}$$

In total, Γ_H contains 16 million unique isA relationships, and 2.7 million unique concepts or categories (the h_x's in (x, h_x) pairs). The importance of big data is obvious. Γ_H contains millions of categories and probabilistic scores for each category which enables inference for metaphor understanding, as we will show next.

3.2 Acquiring Metaphors Γ_m

We acquire an initial set of metaphors Γ_m from similes. A simile is a figure of speech that explicitly compares two different things using words such as "like" and "as". For example, the sentence *Life is like a journey* is a simile. Without the word "like," it becomes a metaphor: *Life is a journey.* This property makes simile an attractive first target for metaphor extraction from a large corpus. We use the following syntactic pattern for extraction:

$$\langle \text{target} \rangle \textbf{ BE/VB like [a] } \langle \text{source} \rangle \qquad (1)$$

where **BE** denotes *is/are/has been/have been*, etc., **VB** denotes verb other than **BE**, and $\langle$target$\rangle$ and $\langle$source$\rangle$ denote noun phrases or verb phrases.

Note that not every extracted pair is a metaphor. *Poetry is like an art* matches the pattern, but it is not a metaphor because poetry is really an art. We will use Γ_H to clean such pairs. Furthermore, due to the idiosyncrasies of natural languages, it is not trivial to correctly extract the $\langle$target$\rangle$ and the $\langle$source$\rangle$ from each sentence that matches the pattern. We develop a rule-based system that contains more than two dozen rules for extraction. For example, a rule of high-precision but low-recall is "$\langle$target$\rangle$ must be at the beginning of a sentence or the beginning of a clause (e.g., following the word *that*)".

Finally, from 8,552,672 sentences that match the above pattern (pattern 1), we obtain 1.2 million unique (x, y) pairs, and after filtering, we are left with close to 1 million unique metaphor pairs, which form the starting point of Γ_m.

3.3 Cleaning, Expanding, and Weighting Γ_m

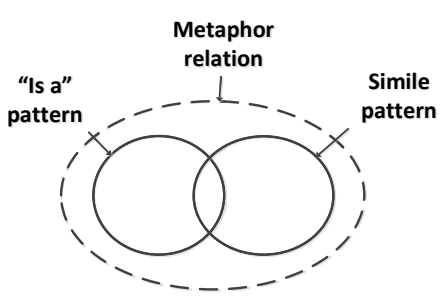
The simile pattern only allows us to extract some metaphor pairs (see Figure 1a). To expand Γ_m, we use a more flexible but also noisier pattern to extract more candidate metaphor pairs from billions of sentences in the web corpus:

$$\langle \text{target} \rangle \textbf{ BE [a] } \langle \text{source} \rangle \qquad (2)$$

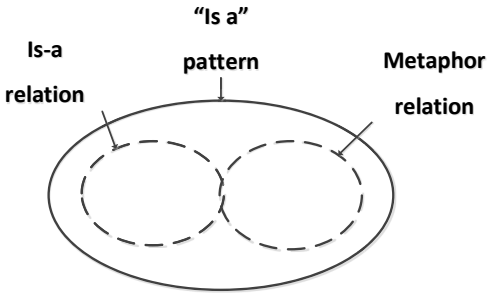
The above "is a" pattern covers metaphors such as *Life is a journey*. But many pairs thus extracted are not metaphors, for example, *Malaysia is a tropical country*. That is, pairs extracted by the "is a" pattern contains at least two types of relations: the isA relations and the metaphor relations (see Figure 1b). The problem is to distinguish one from the other. In theory, the set of all IsA relations, I, and the set of all metaphor relations, M, do not overlap, because by definition, the source concept and the target concept in a metaphor are *not* the same thing. Thus, our intuition is the following. The pairs produced by the simile pattern, called S, is a subset of M, while the pairs extracted from the Hearst pattern, called H, is also a subset of I. Since M and I hardly overlap, S and H should have little overlap, too. In practice, very few people would say something like *journeys such as life*. Figure 1b illustrates this scenario.

To verify this intuition, we randomly sampled 1,000 sentences and manually annotated them. Of these sentences, 40 contain an IsA relation, of which 27 are enclosed in a Hearst's pattern and 13 can be extracted by the "is a" pattern. Furthermore, 28 of these 1000 sentences contain a metaphor expression, and within the 28 metaphors, 15 are embedded in a

[2] Dataset can be found at `http://probase.msra.cn/`.



(a) Patterns supporting metaphor relation



(b) Relations extractable from"is a" pattern

Figure 1: Patterns vs. Relations

simile pattern. More importantly, there is no overlap between the IsA relations and metaphors (and hence the similes).

In a larger scale experiment, we crawled 1 billion sentences which match the "is a" pattern (2) from the web corpus. From these, we extracted 180 million unique (x, y) pairs. 24.8% of Γ_H can be found in "is a" pattern pairs, while 16.8% of Γ_m can be found in "is a" pattern pairs. Further more, there is almost no overlap between Γ_H and Γ_m: 1.26% of Γ_H can be found in Γ_m, and 1.31% of Γ_m can be found in Γ_H.

Our goal is to use the information collected through the syntactic patterns to enrich the metaphor relations or Γ_m. Armed with the above observations, we make two conclusions. First, the $(life, journey)$ pair we extracted from *life is a journey* is more likely a metaphor since it does not appear in the set extracted from Hearst patterns. Second, if any existing pair in Γ_m also appears in Γ_H, we can remove that pair from Γ_m.

From the 180 million unique (x, y) pairs we extracted earlier, by filtering out low frequency pairs [3] and those pairs in Γ_H, we obtain 2.6 million of fresh metaphors. This is almost 3 times larger than initial metaphor set obtained from the simile pattern.

We further expand Γ_m by adding metaphors derived from Γ_m and Γ_H. Assume $(x, y) \in \Gamma_m$, and $(x, h_x) \in \Gamma_H$, then we add (h_x, y) to Γ_m. As an example, if $(Julie, sun) \in \Gamma_m$, then we add $(person\_name, sun)$ to Γ_m, since $(Julie, person\_name) \in \Gamma_H$. This enables the metaphor detection approach we describe in Section 4. Note that we ignore transitivity in the isa relations from Γ_H as such transitivity is not always reliable. For example, car seat is a chair, and chair is furniture, but car seat is not furniture. How to handle transitivity in a data driven isA taxonomy is a challenging problem, and is beyond the scope here.

Finally, we calculate the weight of each metaphor (x, y). The weight $P_m(x, y)$ is calculated as follows:

$$P_m(x, y) = \frac{\text{occurrences of } (x, y) \text{ in isA pattern}}{\text{occurrences of isA pattern}} \tag{3}$$

The weights of derived metaphors, such as $(person\_name, sun)$, are calculated as follows:

$$P_m(h_x, y) = \sum_{(x, h_x) \in \Gamma_H} P_m(x, y) \tag{4}$$

4 Probabilistic Metaphor Understanding

In this paper, we consider two aspects of metaphor understanding, metaphor recognition and metaphor explanation. The latter is needed for type 2 and 3 metaphors where either the source or the target concept is implicit or missing. Next, we describe a probabilistic approach to accomplish these two tasks.

4.1 Type 1 Metaphors

In a type 1 metaphor, both the source and the target concepts appear explicitly. When a sentence matches "is a" pattern (pattern 2), it is a potential metaphor expression. The first noun in the pattern is the target candidate, while the second noun is the source candidate. To recognize type 1 metaphors,

[3] Specifically, we randomly sample pairs of frequency 1, 2, ..., 10 from Γ_m and check the precisions of each group. We filter out pairs with frequency less than 5 to optimize the precision.

we first obtain the candidate (source, target) pair from the sentence. Then, we check if we have any knowledge about the (source, target) pair.

Intuitively, if the pair exists in the metaphor dataset Γ_m, then it is a metaphor. If the pair exists in the is-A knowledge base Γ_H, then it is not a metaphor. But because Γ_m is far from being complete, if a pair exists in neither Γ_m nor Γ_H, there is a possibility that it is a metaphor we have never seen before. In this case, we reason as follows.

Consider a sentence such as *My Ferrari is a beast*. Assume (*Ferrari*, *beast*) $\notin \Gamma_m$, but (*sports car*, *beast*) $\in \Gamma_m$. Note that (*sports car*, *beast*) may itself be a derived metaphor which is added into Γ_m in metaphor expansion, and the original metaphor extracted from the web data is (*Lamborghinis*, *beast*). Furthermore, from Γ_H, we know *Ferrari* is a *sports car*, that is, (*Ferrari*, *sports car*) $\in \Gamma_H$, we can then infer that *Ferrari* to *beast* is very likely a metaphor mapping.

Specifically, let (x, y) be a pair we are concerned with. We want to compute the odds of (x, y) representing a metaphor vs. a normal is-A relationship:

$$\frac{P(x,y)}{1-P(x,y)} \tag{5}$$

where $P(x, y)$ is the probability that (x, y) forms a metaphor. Now, combining the knowledge we have in Γ_H, we have

$$P(x,y) = \sum_{(x,h_x)\in\Gamma_H} P(x,h_x,y) \tag{6}$$

Here, h_x is a possible superconcept, i.e., a possible interpretation, for x. For example, if $x = apple$, then two highly possible interpretations are *company* and *fruit*. In Eq.(6), we want to aggregate on all possible interpretations (all superconcepts) of x. This is possible because of the massive size of the concept space in Γ_H.

We can rewrite Eq.(6) to the following:

$$P(x,y) = \sum_{(x,h_x)\in\Gamma_H} P(y|x,h_x)P(x|h_x)P(h_x) \tag{7}$$

Here, $P(y|x, h_x)$ means when x is interpreted as an h_x, the probability of y as a target metaphorical concept for h_x. Thus, given h_x, y is independent with x, so $P(y|x, h_x)$ can be simply replaced by $P(y|h_x)$. We can then rewrite Eq.(7) to:

$$\begin{aligned} P(x,y) &= \sum_{(x,h_x)\in\Gamma_H} P(y|h_x)P(x|h_x)P(h_x) \\ &= \sum_{(x,h_x)\in\Gamma_H} P(h_x,y)P(x|h_x) \end{aligned} \tag{8}$$

It is clear $P(h_x, y)$ is simply $P_m(h_x, y)$ in Eq.(4) given by the metaphor dataset Γ_m. Furthermore, $P(x|h_x)$ is the typicality of x in the h_x category, and $P(h_x)$ is the prior of the category h_x. Both of them are available from the isA knowledge base Γ_H. Thus, we can calculate Eq.(8) using information in the two knowledge bases we have created.

If the odds in Eq.(5) is greater than a threshold δ, which is determined empirically to be $\delta = \frac{P(metaphor)}{P(isa)}$ [4], we declare (x, y) as a metaphor.

4.2 Context Preference Modeling

It is more difficult to recognize metaphors when the source concept or the target concept is not explicitly given in a sentence. In this case, we rely on the context in the sentence.

Given a sentence, we find metaphor candidates and the context. Here, *candidates* are noun phrases in the sentence which can potentially be the target or the source concept of a metaphor, while *context* denotes words that have a grammatic dependency with the candidate. The dependency can be subject-predicate, predicate-object, or modifier-header, etc. The context can be a verb, a noun phrase, or an adjective which has certain preference over the target or source candidate. For example, the word *horse* prefers verbs such as *jump*, *drink* and *eat*; the word *flower* prefers modifiers such as *red*, *yellow* and *beautiful*.

In this work, we focus on analyzing the preferences of verbs using subject-predicate or predicate-object relation between the verb and the noun phrases. We select 2,226 most frequent verbs from the web corpus. For each verb, we construct the distribution of noun phrases depend on the verb in the sentences sampled from the web corpus. The noun phrases are restricted to be those that occur in Γ_H.

[4] This is the ratio between the number of metaphors and is-a pairs in a random sample of "is a" pattern sentences.

More specifically, for any noun phrase y that appears in Γ_H, we calculate the following

$$P_r(C|y) = \frac{f_r(y,C)}{\sum_C f_r(y,C)} \quad (9)$$

where $f_r(y,C)$ means the frequency of y and context C with relation r. Note we can build preference distribution for context other than verbs since, in theory, r can be any relation (e.g. modifier-head relation).

4.3 Type 2 and Type 3 Metaphors

If a sentence contains type 2 and type 3 metaphors, either the source or the target concepts in the sentence is missing. For each noun phrase x and a context C in such a sentence, we want to know whether x is of literal or metaphoric use. It is a metaphoric use if the selectional preference of some y, which is a source or target concept of x in Γ_m, is larger than the selectional preference of any super-concept of x in Γ_H, by a factor δ. Formally, there exists a y where $(x,y) \in \Gamma_m$ or $(y,x) \in \Gamma_m$, such that

$$\frac{P(y|x,C)}{P(h|x,C)} \geq \delta, \quad \forall (x,h) \in \Gamma_H. \quad (10)$$

To compute (10), we have

$$\begin{aligned} P(y|x,C) &= \frac{P(x,y,C)}{P(x,C)} \\ &= \frac{P(x,y)P(C|x,y)}{P(x,C)} \end{aligned} \quad (11)$$

Assuming x is a target concept and y is a source concept (a Type 3 metaphor), we can obtain $P(x,y)$ by Eq.(8). [5] Furthermore, C is independent of x in a type 2 or 3 metaphor, since a metaphor is an unusual use of x (the target) within a given context. Therefore $P(C|x,y) = P(C|y)$, where $P(C|y)$ is available from Eq. (9).

Similarly, we have

$$P(h|x,C) = \frac{P(x,h)P(C|h)}{P(x,C)} \quad (12)$$

where $P(x,h)$ is obtained from Γ_H and $P(C|h)$ is from the context preference distribution.

[5] Type 2 metaphors can be handled similarly.

To explain the metaphor, or uncover the missing concept,

$$\begin{aligned} y^* &= \underset{y \wedge (y,x) \in \Gamma_m}{\arg\max} P(y|x,C) \\ &= \underset{y \wedge (y,x) \in \Gamma_m}{\arg\max} P(y,x)P(C|y) \end{aligned}$$

As a concrete example, consider sentence *My car drinks gasoline*. There are two possible targets: *car* and *gasoline*. The context for both targets is the verb *drink*. Let x = car. By Eq.(11), we first find all y's for which $(car,y) \in \Gamma_m$ or $(y,car) \in \Gamma_m$. We get terms such as *woman*, *friend*, *gun*, *horse*, etc. When we calculate $P(car,y)$ by Eq.(8), we also need to find hypernyms of *car* in Γ_H, which may include *vehicle*, *product*, *asset*, etc. For each candidate y, $P(y|car,C)$ is calculated by metaphor knowledge $P(x,y)$ and context preference $P(C|y_i)$. Table 1 shows the result. Since the selectional preference of *horse* (from Γ_m) is much larger than other literal uses of *car*, this sentence is recognized as a metaphor, and the missing source concept is *horse*.

Table 1: Log probabilities (M: Metaphor, L:Literal).

Type	y_i	log $P(y_i, car)$	log $P(C\|y_i)$	log $P(y_i \|car, C)$
L	vehicle	-6.2	-∞	-∞
L	product	-6.9	-∞	-∞
L	asset	-6.3	-∞	-∞
M	woman	-8.5	-2.8	-11.3
M	friend	-8.0	-3.0	-11.0
M	gun	-8.4	-∞	-∞
M	horse	-8.2	-2.4	-10.6
...	...	...	...	...

5 Experimental Result

We evaluate the performance of metaphor acquisition, recognition and explanation in our system and compare it with several state-of-the-art mechanisms.

5.1 Metaphor Acquisition

From the web corpus, we collected 8,552,672 sentences matching the "is like a" pattern (pattern 1) and we extracted 932,621 unique high quality simile mappings from them. These simile mappings became the core of Γ_m. Γ_H contains 16,736,068 unique isA pairs. We also collected 1,131,805,382

sentences matching the "is a" pattern (pattern 2), from which 180,446,190 unique mappings were extracted. These mappings contain both metaphors and isA relations. From there, we identified 2,663,127 pairs of metaphors unseen in the simile set. These new metaphor pairs were added to Γ_m. Random samples show that the precisions of the core metaphor dataset and the whole dataset are 93.5% and 82%, respectively. All of the above datasets, a sample of context preference, as well as the test sets mentioned in this section can be found at `http://adapt.seiee.sjtu.edu.cn/~kzhu/metaphor`.

5.2 Type 1 Metaphor Recognition

We compare our type 1 metaphor recognition with the method (known as KZ) by Krishnakumaran and Zhu (2007). For sentences containing "*x* is a *y*" pattern, KZ used WordNet to detect whether y is a hypernym of x. If not, then this sentence is considered a metaphor. Our test set is 200 random sentences that match the "*x* BE a *y*" pattern. We label a sentence in the set as a metaphor if the two nouns connected by BE do not actually have isA relation; or if they do have isA relation but the sentence expressed a strong emotion [6].

Table 2: Type 1 metaphor recognition

	Precision	Recall	F1
KZ	13%	30%	18%
Our Approach	**73%**	**66%**	**69%**

The result is summarized in Table 2. KZ does not perform as well due to the small coverage of WordNet taxonomy. Only 33 out of 200 sentences contain a concept x that exists in WordNet and has at least one hypernym. And among these, only 2 sentences contain a y which is the hypernym ancestor of x in WordNet. Clearly, the bottleneck is the scale of WordNet.

5.3 Type 2/3 Metaphor Recognition

For type 2/3 metaphor recognition, we compare our results with three other methods. The first competing method (called SA) employs the *selectional association* proposed by Resnik (1993). Selectional association measures the strength of the connection between a predicate (c) and a term (e) by:

$$A(c,e) = \frac{Pr(e|c)\log\frac{Pr(e|c)}{Pr(e)}}{S(c)}, \tag{13}$$

where

$$\begin{aligned} S(c) &= KL(Pr(e|c)||Pr(e)) \\ &= \sum_e Pr(e|c)\log\frac{Pr(e|c)}{Pr(e)} \end{aligned}$$

Given an NP-predicate pair, if its SA score is less than a threshold α (set to 10^{-4} by empirics), then the pair is recognized as a metaphor context.

Second competing method (called CP) is the *contextual preference* approach (Resnik, 1993) introduced in Section 4.2. To establish context preference distributions, we randomly select 100 million sentences from the web corpus, parse each sentence using Stanford parser (Group, 2013) to obtain all subject-predicate-object triples, and aggregate the triples to get 33,236,292 subject-predicate pairs and 38,890,877 predicate-object pairs. The occurrences of these pairs are used as context preference. Given a pair of NP-predicate pair, if its context preference score is less than a threshold β (set to 10^{-5} by empirics [7]), then the pair is considered as metaphoric.

The third competing method (called VH) is a variant of our own algorithm with Γ_m replaced by a metaphor database derived from the Slip Net proposed by Veale and Hao (2008), which we call Γ_{VH}. We built a Slip Net containing 21,451 concept nodes associated with 27,533 distinct talking points. We consider two concepts to be metaphoric if they are at most 5 hops apart on the Slip Net The choice of 5 hops is a trade-off between precision and recall for Slip Net. We thus created Γ_{VH} with 5,633,760 pairs of concepts.

We sampled 1,000 sentences from the BNC dataset (Clear, 1993) as follows. We prepare a list of 2,945 frequent verbs (and their different forms). For each verb, we obtain at most 5 sentences from BNC dataset which contain this verb as a predicate. At this point, we obtain a total of 22,601 sentences and randomly sample 1,000 sentences to form a test

[6] For example, "*this man* is an *animal*!".

[7] The authors didn't specify the choice of α and β, and we pick values which optimize the performance of their algorithms.

set. Each sentence in the set is then manually labeled as being "metaphor" or "non-metaphor". We label them according to this procedure:

1. for each verb, we collect the intended use, i.e., the categories of its arguments (subject or object) according to Marriam Webster's dictionary;
2. if the argument of the verb in the sentence belongs to the intended category, the sentence is labeled "non-metaphor";
3. if the argument and the intended meaning form a metonymy which uses a part or an attribute to represent the whole object, the pair is labeled as "non-metaphor";
4. else the sentence is labeled as "metaphor".

Table 3: Type 2/3 metaphor recognition

	Precision	Recall	F1
SA	23%	20%	21%
CP	50%	20%	26%
VH	11%	86%	20%
Our Approach	**65%**	**52%**	**58%**

The results for type 2 and 3 metaphor recognition are shown in Table 3. Our knowledge-based approach significantly outperforms the other peers by F-1 measure. Although VH achieves a good recall, its precision is poor. This is because i) Slip Net construction makes heavy use of sibling terms on the WordNet but sibling terms are not necessarily similar terms; ii) many pairs generated by slipping over the Slip Net are in theory related but are not commonly uttered due to the lack of practical context.

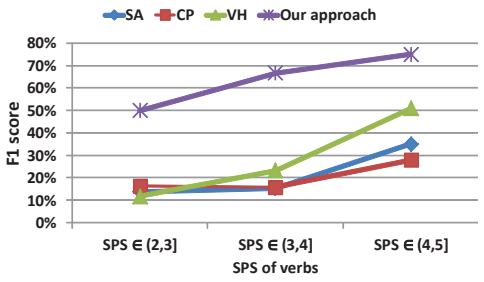


Figure 2: Metaphor recognition of type 2 and 3

Fig. 2 compares the four methods on verbs with different *selectional preference strength*, which indicates how strong a verb's arguments are restricted to a certain scope of nouns.[8] Again, our method shows a significant advantage across the board.

We explain why our approach works better using the examples in Table 4. In sentence $AAU200$, $shatters$ is a metaphoric usage because silence is not a thing that can be broken into pieces. SA and CP scores for shatters-silence pair are high because this word combination is quite common, and hence these methods incorrectly treat it as literal expression. The situation is similar with stalk-company pair in $ABG2327$. On the other hand, for $AN81309$, manipulate-life is considered rare combination and hence has low SA and CP scores and is deemed a metaphor while in reality it is a literal use. A similar case occurs for work-concur pair. In all these cases, our knowledge bases Γ_m and Γ_H are comprehensive and accurate enough to correctly identify metaphors vs. non-metaphors. On the contrary, the metaphor database Γ_{VH} covers way too many pairs that it treats every pair as a metaphor.

Besides our own dataset, we also experiment on TroFi Example Base[9], which consists of 50 verbs and 3,736 sentences containing these verbs. Each sentence is annotated as literal and nonliteral use of the verb. Our algorithm is used to classify the subjects and the objects of the verbs. We use Stanford dependency parser to obtain collapsed typed dependencies of these sentences, and for each sentence, run our algorithm to classify the subjects and objects related to the verb, if the verb acts as a predicate. Results show that our approach achieves 77.5% precision but just under 5% in recall. The low recall is because, i) non-literal uses in the TroFi dataset include not only metaphor but also metonymy, irony and other anomalies; ii) our approach currently focuses on subject-predicate and predicate-object dependencies in a sentence only, but the target verbs do not act as predicate in many of the example sentences; iii) the Stanford dependency parser is not robust enough so half of the sentences are not parsed correctly.

[8]Note that no verb has SPS larger than 5.

[9]TroFi Example Base is available at `http://www.cs.sfu.ca/~anoop/students/jbirke/`.

Table 4: Metaphor recognition for some example sentences from BNC dataset (HM: Human, M: Metaphor, L : Literal).

ID	Sentence	HM	SA	CP	VH	Ours
AAU 200	Road-block salvo *shatters* Bucharest's fragile **silence**.	M	L	L	M	M
ABG 2327	Obstruction and protectionism do not *stalk* only **big companies**.	M	L	L	M	M
AN8 1309	But when science proposes to *manipulate* the **life** of a human baby,	L	M	M	M	L
ACH 1075	Nevertheless, recent **work** on Mosley and the BUF has *concurred* about their basic unimportance.	L	M	M	M	L

5.4 Metaphor Explanation

In this experiment, we use the classic labeled metaphoric sentences from (Lakoff and Johnson, 1980). Lakoff provided 24 metaphoric mappings, and for each mapping there are about ten example sentences. In total, there are 214 metaphoric sentences. Among them, we focus on 83 sentences whose metaphor is expressed by subject-predicate or predicate-object relation, as this paper focuses on verb centric context preferences.

We evaluate the results of competing algorithms by the following labeling criteria. We consider an output (i.e. a pair of concept mapping) as a *match*, if the produced pair exactly matches the ground truth pair, of if the pair is subsumed by the ground truth pair. For example, the ground truth for the sentence *Let that idea simmer on the back burner* is *ideas* $\rightarrow$ *foods* according to Lakoff (Lakoff and Johnson, 1980). If our algorithm outputs *idea* $\rightarrow$ *stew*, then it is considered a *match* since *stew* belongs to the *food* category. An output pair is considered *correct* if it is not a *match* to the ground truth but is otherwise considered metaphoric by at least 2 of the 3 human judges.

Given a sentence, since our algorithm returns a list of possible explanations for the missing concept, ranked by the probability, we evaluate the results by three different metrics:

Match Top 1: result considered correct if there is a *match* with the top explanation;

Match Top 3: result considered correct if there is a *match* in the top 3 ranked explanations;

Correct Top 3: result considered correct if there is a *correct* in the top 3 explanations.

Table 5: Precision of metaphor explanation using differen metaphor databases

	Match Top 1	Match Top 3	Correct Top 3
Γ_{VH}	26%	49%	54%
Γ_m	**43%**	**67%**	**78%**

Comparison with Slip Net

We compare the result of our algorithm (from Section 4.3) against the variant which uses Γ_{VH} obtained in Section 5.3.

Table 5 summarizes the precisions of the two algorithms under three different metrics. Some of these sentences and the top explanations given by our algorithm are listed in Table 6. The concept to be explained is italicized while the explanation that is a match or correct is bolded or bold-italicized, respectively. The explanations are ordered from left to right by the score.

Comparison with paraphrasing

While we define metaphor explanation as a task to recover the missing noun-based concept in a source-target mapping, an alternative way to explain a metaphor (Shutova, 2010) is to find the paraphrase of the verb in the metaphor. Here we evaluate paraphrasing task on verbs in metaphoric sentence by Shutova et al(Shutova, 2010). For a metaphoric verb V in a sentence, Shutova et al. select a set of verbs that probabilistically best matches grammar relations of V, and then filter out those verbs that are not related to V according to the WordNet, and eventually re-rank remaining verbs based on selection association.

In some sense, Shutova's work uses a similar framework as ours: first restrict the target paraphrasing set using a knowledge, then select the most

Table 6: Metaphor sentences explained by the system

Metaphor mapping	Sentence	Explanation
Ideas are food	Let that *idea* simmer on the back burner.	**stew; carrot; onion**
	We don't need to spoon-feed our students *with knowledge*.	**egg roll; acorn**; word
Eyes are containers	His *eyes* displayed his compassion.	**window**; symbol; **tiny camera**
	His *eyes* were filled with anger.	**hollow ball; water balloon; balloon**
Emotional effect is physical contact	His *mother's death* hit him hard.	***enemy***; **monster**
	That *idea* bowled me over.	**punch; stew; onion**
Life is a container.	Her *life* is crammed with activities.	***tapestry***; **beach**; dance
	Get the most out of *life*.	***game***; ***journey***; **prison**

proper word based on the context. The difference is that the target of (Shutova, 2010) is the verb in sentence, while our approach focuses on the noun.

To implement algorithm by Shutova, we extract and count each grammar relation in 1 billion sentences. These counts are used to calculate context matching in (Shutova, 2010), and are also used to calculate selection association. We perform Shutova's paraphrasing on verbs in 83 sentences, of which only 25 finds a good paraphrases in Shutova's top 3 results. After removing 17 sentences which contain light verbs (e.g., take, give, put), the algorithm finds 21 good paraphrases in top 3 results. One reason for the low recall is that Wordnet is inadequate in providing candidate metaphor mapping. This is also the reason why our metaphor base is better than the metaphor base generated by talking points.

6 Conclusion

Knowledge is essential for a machine to identify and understand metaphors In this paper, we show how to make use of two probabilistic knowledge bases automatically acquired from billions of web pages for this purpose. This work currently recognizes and explains metaphoric mappings between nominal concepts with the help of selectional preference of just subject-predicate or predicate-object contexts. An immediate next step is to extend this framework to more general contexts and a further improvement will be to identify mappings between any source and target domains.

7 Acknowledgements

Kenny Q. Zhu was partially supported by Google Faculty Research Award, and NSFC Grants 61100050, 61033002 and 61373031.

References

Rodrigo Agerri. 2008. Metaphor in textual entailment. In *COLING (Posters)*, pages 3–6.

John Barnden, Sheila Glasbey, Mark Lee, and Alan Wallington. 2002. Reasoning in metaphor understanding: the att-meta approach and system. In *COLING '02*, pages 1–5.

Eric P. S. Baumer, James P. White, and Bill Tomlinson. 2010. Comparing semantic role labeling with typed dependency parsing in computational metaphor identification. In *CALC '10*, pages 14–22.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *In Proceedings of EACL-06*, pages 329–336.

Jeremy H. Clear. 1993. The digital word. chapter The British national corpus, pages 163–187.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16.3:235–250.

James Geary. 2011. *I is an Other: The Secret Life of Metaphor and How It Shapes the Way We See the World*. Harper.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *In Workshop On Scalable Natural Language Understanding*.

Stanford NLP Group. 2013. The Stanford parser. `http://nlp.stanford.edu/software/lex-parser.shtml`.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING '92*, pages 539–545.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *In Language Resources and Evaluation*.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 13–20, Rochester, New York, April. ACL.

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

J. H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc.

Zachary J. Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Comput. Linguist.*, 30:23–44, March.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38:39–41, November.

Srinivas Sankara Narayanan. 1997. Knowledge-based action representations for metaphor and aspect (karma). Technical report.

Philip Stuart Resnik. 1993. *Selection and information: a class-based approach to lexical relationships*. Ph.D. thesis.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *COLING '10*, pages 1002–1010.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *HLT '10*, pages 1029–1037.

Catherine Smith, Tim Rumbell, John Barnden, Bob Hendley, Mark Lee, and Alan Wallington. 2007. Don't worry about metaphor: affect extraction for conversational agents. In *ACL '07*, pages 37–40.

P.D. Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33(1):615–655.

Tony Veale and Yanfen Hao. 2008. A fluid knowledge representation for understanding and generating creative metaphors. In *COLING*, pages 945–952.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197 – 223.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *SIGMOD Conference*, pages 481–492.

Li Zhang. 2010. Metaphor interpretation and context-based affect detection. In *COLING (Posters)*, pages 1480–1488.