Meta4meaning: Automatic Metaphor Interpretation Using Corpus-Derived Word Associations

Ping Xiao[1]
ping.xiao@cs.helsinki.fi

Khalid Alnajjar[1]
alnajjar@cs.helsinki.fi

Mark Granroth-Wilding[2]
mark.granroth-wilding@cl.cam.ac.uk

Kat Agres[3]
kathleen.agres@qmul.ac.uk

Hannu Toivonen[1]
hannu.toivonen@cs.helsinki.fi

[1]Dept. of Computer Science and HIIT, University of Helsinki, Finland; [2]Computer Laboratory, University of Cambridge, UK
[3]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Abstract

We propose a novel metaphor interpretation method, *Meta4meaning*. It provides interpretations for nominal metaphors by generating a list of properties that the metaphor expresses. Meta4meaning uses word associations extracted from a corpus to retrieve an approximation to properties of concepts. Interpretations are then obtained as an aggregation or difference of the saliences of the properties to the tenor and the vehicle. We evaluate Meta4meaning using a set of human-annotated interpretations of 84 metaphors and compare with two existing methods for metaphor interpretation. Meta4meaning significantly outperforms the previous methods on this task.

Introduction

Metaphor has various linguistic manifestations, such as the metaphorical use of nouns, verbs, adjectives and adverbs, as well as at larger conceptual frames, for instance, an entire poem or story in a metaphor with something outside it. The present work focuses on interpreting nominal metaphors of the form 'NOUN$_1$ is [a] NOUN$_2$', where, following Richards (1936), NOUN$_1$ is called the *tenor* and NOUN$_2$ the *vehicle*.

The meaning of a metaphor is not fixed. It arises from the interaction between at least two conceptual spaces, the tenor's and the vehicle's (Black 1962), but often also the context's (Ortony et al. 1978). Metaphors are different with respect to the number and the saliences of the individual interpretations, such as a few salient meanings, a few or many non-salient meanings, and no meaning. The meaning distribution of a metaphor is sensitive to context, which increases or decreases the saliences of certain meanings.

Metaphor meanings have been most often talked about in terms of *properties* – which properties of the tenor have been highlighted or newly attributed to it (Glucksberg 2001; Moreno 2004). A key objective of a metaphor interpretation program is to identify those *highlighted* properties.

Metaphor interpretation relies on knowledge about the tenor and vehicle. In this work, we propose *Meta4meaning*, a novel method for metaphor interpretation. Meta4meaning derives *word associations* from a large text corpus in order to obtain a concept's properties and their saliences. To identify the properties highlighted by a metaphor, Meta4meaning measures either the aggregation or the difference of the saliences of a property to the tenor and the vehicle, capturing distinct ways in which metaphor may function. Furthermore, we test two hypotheses regarding the saliences of properties involved in metaphor understanding, the salience imbalance hypothesis (Ortony 1979) and the requirement of pre-existing saliences of interpretations to the tenor and the vehicle.

Meta4meaning is evaluated against the interpretations of 84 metaphors acquired from human subjects by Roncero and Almeida (2014). The performance is also compared with two existing methods developed by Terai and Nakagawa (2008) and Veale and Li (2012). We find Meta4meaning using word associations to be the most successful method.

In the remainder of this paper, we first give a formal definition of the metaphor interpretation problem, and review the related work. The Meta4meaning method is described next in two parts: first, how it acquires concept properties, and second, how it uses the properties of the tenor and vehicle to provide metaphor interpretations. Then, we report an evaluation of the methods and discuss the results.

Problem Formalization

Consider a nominal metaphor of the form 'NOUN$_1$ is [a] NOUN$_2$', such as 'alcohol is a crutch'. An interpretation of the metaphor is a property that the vehicle NOUN$_2$ (crutch) expresses about the tenor NOUN$_1$ (alcohol). In a study by Roncero and Almeida (2014), interpretations of 'alcohol is crutch' included properties such as 'helpful' and 'addictive'.

Given a nominal metaphor, the objective of metaphor interpretation is to produce a ranked list of possible interpretations, such that highly ranked interpretations are likely to be considered interpretations by humans.

Related Work

Kintsch (2000) applied Latent Semantic Analysis (LSA) as a knowledge source for the computational modeling of nominal metaphor interpretation. A vector approximation of the Construction-Integration (CI) model is used for finding the representations of metaphor meanings. The author only uses the *term vectors* of LSA. The meaning of a metaphor is represented by the centroid of a set of vectors, including the tenor, the vehicle, and a few terms related to both (k terms most related to the tenor are selected among m terms most related to the vehicle). The composed metaphor vector does not directly give the properties highlighted by a metaphor.

Terai and Nakagawa (2008) extended this work. They built a generative probabilistic model based on the dependency counts between nouns and adjectives and between nouns and verbs, treating the adjectives and verbs as the properties of the nouns. The statistical model captures the latent classes where the nouns and their properties are connected, and the latent classes are used as vector dimensions. To interpret a metaphor, a meaning vector is first constructed by applying the method of Kintsch (2000), which is subsequently used to assign saliences to the properties in the latent classes. As an additional step, the properties and their assigned saliences are used to construct a recurrent neural network, in order to model the dynamic interaction between properties. The properties with the highest activation, until the network converges, are taken as the metaphor meanings.

The system *Metaphor Magnet* developed by Veale and Li (2012) is based on the idea that metaphor interpretation works by stereotype expansion and property overlap. For each of the tenor and vehicle concepts, the concept is first expanded with a set of *stereotypes* that are commonly used to describe it. The stereotypes are obtained from Google n-grams using linguistic patterns, such as "NOUN$_1$ is [a] NOUN$_2$". Then, the union of the properties of the concept and its associated stereotypes are all attributed to the concept. The properties, in the forms of adjectives, VERB+*ing*s and VERB+*ed*s, are harvested from the Web using another set of linguistic patterns. In addition, manual filtering was involved in constructing both knowledge sources. The properties highlighted by a metaphor are at the intersection of the tenor's and the vehicle's properties.

Meta4meaning differs from the above literature in both knowledge acquisition and modeling metaphor interpretation. We will compare it with the method of Terai and Nakagawa (2008) and Metaphor Magnet in the evaluation.

Acquiring Knowledge for Metaphor Interpretation

In this section, we describe the Meta4meaning method for interpreting metaphors. The method has two major components. First, a text corpus is analyzed for associations between pairs of words. Then, for each metaphor to be interpreted, plausible properties (interpretations) are ranked.

Extracting Word Associations

Meta4meaning extracts word associations from corpora based on the statistical significance of their co-occurrence. We consider the associated words as an approximation of a concept's properties, and their association strengths as the properties' saliences to the concept.

There are different ways of extracting word relations depending on what exactly is being searched for (Rapp 2002). Concepts and their properties are more likely to have syntagmatic than paradigmatic relations. *Syntagmatic* relations are between co-occurring words, e.g., 'the shark has six fins' (shark is *related* to fins). *Paradigmatic* relations in turn exist between words that appear in similar context but usually do not co-occur, e.g., 'shark' and 'sawfish' (shark and sawfish are *similar*). Statistical association measures are suitable for extracting syntagmatic associations (Rapp 2002; Evert 2008). LSA has also been used to this end (Sahlgren 2006); nevertheless, the bag-of-words distributional models seem more appropriate for capturing paradigmatic associations (Rapp 2002; Peirsman, Heylen, and Geeraerts 2008).

In acquiring word associations, we start with a basic method, applying association measures to the co-occurrence counts of words. We use a 2 billion word web text corpus, *ukWaC*[1], and follow a standard process of acquiring co-occurrence counts. Lemmatization and punctuation removal are first applied to the corpus. The co-occurrence of words is counted within a symmetrical window of size 4, i.e. allowing at most 3 words between the two words, and further limited by sentence boundaries (Lapesa, Evert, and Schulte im Walde 2014). The most frequent 50,000 words are selected as vocabulary, excluding closed class words. We use the *log-likelihood* measure of Evert (2008), more specifically the one for surface co-occurrence, as our association measure, with all negative values set to zero. Finally, the score of an association is normalized by the *L1-norm* of the scores of all the associations of a concept (McGregor et al. 2015).

Moreover, we experiment with two methods of dimensionality reduction, Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF), in the hope of achieving better semantic representations. The rationale behind dimensionality reduction is to remove noise and to generalize individual word co-occurrences to associations between related concepts.

SVD and NMF both produce two matrices. One matrix has words as rows and the reduced dimensions as columns (henceforth, *term-dimension-matrix*). The other has the reduced dimensions as rows and the context terms as columns (*dimension-property-matrix*). A concept and its properties are connected via the reduced dimensions.

For both SVD and NMF, we employed the implementations provided by *Scikit-learn*[2], with default parameters. The SVD model has 900 dimensions, and was obtained with an oversampling factor of 2. When querying the SVD model, no dimension is skipped. The NMF model is built similarly, also with 900 dimensions.

Word associations of a concept, extracted from a large corpus in the above way, cover a long list of words with a big variance in association strength, including small values just above zero. This is in contrast with the properties produced by human subjects in psychological experiments, which is generally a much shorter list of salient properties (Roncero and Almeida 2014). We next describe how Meta4meaning further extracts the best features for metaphor processing.

[1] http://wacky.sslmit.unibo.it/doku.php?id=corpora.

[2] http://scikit-learn.org.

Focusing on Metaphor Activated Properties

Adjectives typically denote properties (Murphy 2010), and metaphor processing frequently involves *abstraction* (Glucksberg 2001). We take advantage of these two characteristics to separate properties that are most likely to be activated in metaphor interpretation from those less likely.

Adjective properties have a dominant presence in metaphor processing. The three existing metaphor interpretation methods introduced in the Related Work section all include adjectives in their knowledge sources. On the other hand, for noun and verb properties, abstraction is often required for metaphors to work (Glucksberg 2001; Utsumi and Sakamoto 2011). We use a combination of the part-of-speech (POS) information of single words and a large-scale rating of term abstractness to only retain adjective, abstract noun and abstract verb associations.

Turney et al. (2011) derived the abstractness ratings of 114,501 WordNet terms with supervised learning. The abstractness ratings are between 0 and 1, 1 being more abstract. We determine the most frequent POS of a word by consulting WordNet and SUBTLEX-UK (Subtitle-based word frequencies for British English)[3].

All adjectives, as well as nouns and verbs with abstractness ratings above 0.5, are retained as metaphor-activated properties. As an example, the resulting highly rated properties for *shark* include 'bask', 'loan', 'white', 'attack', 'see', 'grey', 'marine', 'large', etc.

Interpreting Metaphors

In Meta4meaning, the association strength between a concept and a property is regarded as an approximation of the property's salience to the concept. We use the saliences of a property to the tenor and the vehicle respectively to devise a metric, in order to rank the properties for their likelihood of being the interpretations of the metaphor.

In a metaphor, a set of properties are *transferred* or *emerge* in the interaction between the tenor and vehicle's conceptual spaces (Black 1962; Glucksberg 2001; Moreno 2004). In some metaphors, it seems that the salient properties of the vehicle are transferred to the tenor. In others, however, the metaphor meaning is not among the salient properties of either the vehicle or the tenor: these are called 'emergent properties'.

In this work, we only consider properties associated with both the tenor and the vehicle. While this may seem similar to the 'common features' (Becker 1997) elicited in psychological experiments, there is an important difference: with word associations derived from a corpus, Meta4meaning can also find overlap of low-salience properties.

Ranking of Properties

Meta4meaning ranks properties by their saliences to both the tenor and the vehicle. Taking the product of the saliences emphasizes properties that are strongly associated with both.

Definition 1. The *product of saliences* p_i of property i is

$$p_i = t_i \cdot v_i,$$

where t_i and v_i are the association strengths of the property to the tenor and the vehicle, respectively.

The *salience imbalance hypothesis* predicts that the properties highlighted by a metaphor are among the common properties of the tenor and the vehicle and are more salient to the vehicle than to the tenor (Ortony 1979). Based on this, one can further hypothesize that a larger difference $v_i - t_i$ correlates with higher metaphor aptness.

Definition 2. The *difference of saliences* d_i of property i is

$$d_i = v_i - t_i.$$

We compare the salience difference d_i experimentally to the product of saliences p_i. These two weights measure different, possibly complementary aspects of metaphor properties; we therefore also consider their combination. To avoid making assumptions about the distributions of the values, we produce a combined measure by considering the *ranks* of properties with respect to p_i and d_i and associate the property with the better of these.

Definition 3. The *combined metaphor rank* c_i of property i is

$$c_i = min(rank(i, p), rank(i, d))$$

where $rank(i, x) = |\{j \mid x_j \leq x_i\}|$.

Alternative Measures

An alternative to taking the product of saliences t_i and v_i is to take their sum. Addition promotes properties that are salient for at least one of the tenor and vehicle, and is consistent with those metaphors that highlight a salient property of either one. When the property is more salient to the vehicle than the tenor, the behavior pairs well with the salience imbalance hypothesis.

Definition 4. The *sum of saliences* p_i^+ of feature i is

$$p_i^+ = t_i + v_i.$$

The *combined metaphor sum rank* c_i^+ of property i is

$$c_i^+ = min(rank(i, p^+), rank(i, d)).$$

We use p_i^+ and c_i^+ to emphasize that these are the variants of measures p_i and c_i. Which one is more appropriate depends, among other things, on how the association strengths have been derived and processed.

We also compare the performances of all the above measures when SVD or NMF has been used to reduce the dimensionality of the word co-occurrence matrix. The above measures are then applied to the term-dimension-matrix, and the obtained score vector is multiplied by the dimension-property-matrix. The resulting vector contains the saliences of the properties to the metaphor.

Matrices factorized using SVD contain both positive and negative values, meaning that the above intuition regarding the pointwise product of vectors does not apply. For this reason, we do not apply p_i or the combined c_i to SVD matrices.

[3] http://crr.ugent.be/archives/1423.

Testing Hypotheses

We also empirically test two popular hypotheses about metaphor interpretation. First, the salience imbalance hypothesis (Ortony 1979), that metaphor interpretations are among the common properties of the tenor and the vehicle and are more salient to the vehicle than to the tenor, corresponds in our setting to $v_i > \ t_i \ > \ 0$.

Definition 5. The *salience imbalance hypothesis* is that

$$v_i > t_i > 0.$$

Second, emergent properties, not associated with either the tenor or the vehicle, have been observed dominating metaphor interpretations in psychological experiments. We investigate whether it is still the case for association strengths derived from large text corpora through statistical association measures, which cover both strong and weak associations.

Definition 6. The *emergent property hypothesis* in its weakest form is that

$$v_i \neq 0 \text{ and } t_i \neq 0.$$

Comparison to Other Methods

We compare the performance of Meta4meaning to two others from the literature: the method of Terai and Nakagawa (2008), and Metaphor Magnet by Veale and Li (2012).

The method of Terai and Nakagawa (2008) consists of two processes: categorization followed by dynamic interaction. Meta4meaning is comparable to the categorization process, which composes the metaphor vector by adding together not only the tenor and the vehicle vectors but also a few vectors corresponding to words related to both the tenor and the vehicle. To find the related words, the cosine similarity of vectors has to be computed across the entire vocabulary, and a few parameters (m, k) have to be tuned.

In implementing the categorization method of Terai and Nakagawa (2008), we use the current NMF model, which is similar to the generative probabilistic model employed in the original work, as well as SVD. In both cases, we use $m = 250$ and $k = 5$ following the original paper.

Metaphor Magnet (Veale and Li 2012) is available via an application programming interface[4]. Given a metaphor of the form '*tenor* is [a] *vehicle*', it returns a list of entries in the format of 'property:stereotype(score)'. *Score* indicates the salience of the *property* to the *stereotype*. We tallied the scores of each *unique* property (stereotypes may overlap on properties), and ranked the properties based on their accumulated scores.

Evaluation

We evaluate Meta4meaning using metaphor interpretations acquired from human subjects by Roncero and Almeida (2014). Our empirical goals are to test the various ways of deriving interpretation rankings (Definitions 1–3). We compare them to the alternative measures described above – Definition 4 and use of SVD or NMF for dimensionality reduction – and to other methods proposed in the literature.

Table 1: Example metaphor interpretations from Roncero and Almeida (2014).

Metaphor	Interpretation	Frequency
Alcohol is a crutch	Helpful	5
	Dependable	4
	Addictive	3
	Support	3
	Disability	2
	Aid	2
	Problem	2

Evaluation Dataset

Roncero and Almeida (2014) collected interpretations of 84 nominal metaphors. Subjects were given a metaphor such as 'alcohol is a crutch' and were then asked to provide up to three properties which "the vehicle word [crutch] was expressing about the topic [alcohol]". For each metaphor, the dataset of Roncero and Almeida (2014) provides those interpretations (properties) that were mentioned at least twice, together with their frequencies. Table 1 shows the relevant information for the metaphor 'alcohol is a crutch'.

We use this dataset in our evaluation as follows. First, we only use the salient interpretations. Following Roncero and Almeida (2014), we consider as salient those interpretations that have been mentioned by at least 25% (i.e. 4) of the twenty participants. There are eight metaphors which have no salient interpretations, leaving 76 metaphors.

Second, we carried out some simple linguistic processing by hand on the terms (tenors, vehicles and interpretations) in the dataset and the properties output by Meta4meaning. For the tenors and vehicles, we decapitalized all terms except proper nouns, changed plurals to singulars, and converted phrases to single words when possible (e.g., from "hard cover" to "hardcover"). The tenors and vehicles do not include complex phrases so they are amenable to the methods for nominal metaphors described in this paper. We then stemmed both the human-given interpretations and the properties produced by Meta4meaning before comparing them. The same has been done in the work of Kintsch and Bowles (2002) and Roncero and Almeida (2014).

In total, 76 metaphors are used in the evaluation, each of which has from one to four salient interpretations (together accounting to 145 interpretations). Therefore, a metaphor has slightly under two salient interpretations on average .

Evaluation Metric

We measure the performances of metaphor interpreters using *recall*. Given the human interpretations of a metaphor (from Roncero and Almeida (2014), as described above) and a list of properties given by the method (predicted interpretations), the recall for the metaphor is defined as the fraction of human interpretations that were also predicted by the computer. If a metaphor has n human interpretations and n' of them are among the predicted properties, the recall is n'/n. For the whole set of metaphors, the recall is computed as the average over all individual metaphors.

Since the automated methods described in this paper and elsewhere can produce long lists of possible interpretations,

[4] http://ngrams.ucd.ie/metaphor-magnet-acl/.

Table 2: Examples of metaphor interpretations by Meta4meaning (salient interpretations of the evaluation dataset in bold).

	Metaphor		
Ranking	cloud is cotton	life is a joke	alcohol is a crutch
	(per p_i)	(per d_i)	(per c_i)
1	**white**	**funny**	psychological
2	cover	story	dependence
3	thick	make	drug
4	black	anecdote	emotional
5	blue	good	addiction
6	**fluffy**	humour	week
7	thin	hilarious	**help**
8	**soft**	trivia	mental
9	layer	cruel	cope
10	heavy	fun	dependent

it is relevant to ask how good the most highly ranked properties are. We therefore report *recall at k* (or "@k" for short), defined as the recall when using only the top k properties of a computer-generated ranking.

For instance, recall @10 considers for each metaphor the ten properties ranked highest by the computer, and calculates how many of the salient interpretations are included in these ten. Table 2 shows the top ten interpretations for three different metaphors and for different measures introduced above, representing successful results of the method. For the metaphor 'cloud is cotton', Meta4meaning (with measure p_i) provides all three salient interpretations and thus has a recall of 100% @10. Recall @5 is 33% since only one out of three salient interpretations is among the top five properties. In the experiments, we report the average recalls @5, @10, @15, @25 and @50.

Results and Analysis

We first report on the performance of Meta4meaning, and then compare it to other methods. To compare the recall performances of different measures, we use the Wilcoxon signed-rank test, which tells whether the difference between the average recalls of two measures is significant or not. To test for a statistical difference between recalls, two measures are compared by pairing the recalls for every metaphor at each of the five ks.

Metaphor Interpretation Performance Recall of metaphor interpretations by Meta4meaning, using the product of saliences p_i, the difference of saliences d_i, and the combined rank c_i are given in the first three rows of Table 3.

Among the three measures, the combined rank achieved the best recalls, followed by the product of saliences, while the recalls obtained with the difference of saliences are consistently inferior ($p < .05$ in both tests). For the combined rank, the recall @10 of 0.303 indicates that about 30% of the salient interpretations are among the top ten properties listed by this variant. This can be considered a strong result given the difficulty of the metaphor interpretation task. While the difference of saliences is on its own inferior, the good performance of the combined rank suggests that the difference of saliences and the product of saliences are complementary and recognize different interpretations.

Table 3: Recall of metaphor interpretations by Meta4meaning (best performance in bold).

	Recall				
Meta4meaning variant	**@5**	**@10**	**@15**	**@25**	**@50**
Product of saliences p_i	0.215	0.274	0.304	0.325	**0.466**
Difference of saliences d_i	0.193	0.227	0.27	0.308	0.391
Combined rank c_i	**0.221**	**0.303**	**0.339**	**0.397**	0.454
Sum of saliences p_i^+	0.164	0.239	0.316	0.368	0.41
Combined sum rank c_i^+	0.184	0.254	0.299	0.384	0.462

The two last rows of Table 3 show the recalls of metaphor interpretations by the alternative measures using the sum of saliences p_i^+ and the respective combined sum rank c_i^+ rather than the product of saliences. The combined sum rank performs significantly worse than the combined rank ($p < .01$). Other differences between the measures are not statistically significant.

Effect of Dimensionality Reduction We now evaluate the effect of dimensionality reduction (SVD or NMF) on the metaphor interpretation performance. Tables 4 and 5 show the recall of metaphor interpretations when using SVD and NMF. The results are clearly inferior to the results obtained without dimensionality reduction in Table 3 above. For instance, to reach a recall of about 20%, one needs to consider around top 50 predicted properties for each metaphor, while top five were sufficient in Table 3.

An analysis of the relative performances of different measures within tables shows some interesting results. In the case of SVD, the sum of saliences performs significantly better than the combined sum rank ($p < .01$), whereas there are no significant differences between measures for NMF. Clearly, the way the association strengths of properties have been obtained has a strong effect on how to best combine them.

Table 6 gives an overview of the best performing variants with and without dimensionality reduction. The variants with dimensionality reduction perform much worse than the variant without dimensionality reduction ($p < .001$ in both tests).

According to a preliminary observation, dimensionality reduction seems to select certain aspects of a concept and generalize those over the vocabulary, i.e. properties not salient to the concept per se but similar to the salient ones, rise to the top. At the same time, the unselected dimensions are downplayed. Moreover, the accuracy of selecting the salient aspects of a concept subjects to the methods chosen for dimensionality reduction and the associated parameter setting. To achieve optimal results, it requires systematic experiments, which is out of the scope of this paper. As to be discussed below, it is possible for Meta4meaning to capture those interpretations that are at least salient to one of the tenor and the vehicle. In the unlucky cases where dimensionality reduction actually reduces the saliences of the interpretations (properties), the recalls decrease consequently, although the contrary may happen – the saliences of the interpretations are raised by dimensionality reduction.

Table 4: Recall of metaphor interpretations using SVD (best performance in bold).

	Recall				
Method	**@5**	**@10**	**@15**	**@25**	**@50**
SVD, Difference of saliences d_i	0.037	0.064	0.075	0.088	0.188
SVD, Sum of saliences p_i^+	**0.057**	**0.081**	**0.099**	**0.145**	**0.226**
SVD, Combined sum rank c_i^+	0.053	0.061	0.079	0.121	0.191

Table 5: Recall of metaphor interpretations using NMF (best performance in bold).

	Recall				
Method	**@5**	**@10**	**@15**	**@25**	**@50**
NMF, Product of saliences p_i	0.069	0.076	0.091	0.113	0.144
NMF, Difference of saliences d_i	0.054	0.061	0.061	0.096	0.114
NMF, Combined rank c_i	0.073	0.078	0.089	0.107	0.16
NMF, Sum of saliences p_i^+	**0.076**	**0.098**	0.098	**0.115**	**0.192**
NMF, Combined sum rank c_i^+	0.073	0.08	**0.102**	0.11	0.173

It seems that the former occurred more frequently than the latter in this evaluation.

Comparison to Other Methods The recall performance of salient metaphor interpretations by the categorization method of Terai and Nakagawa (2008) (T&N) and by Metaphor Magnet are given in Table 7, together with the combined rank results for Meta4meaning.

T&N performs better using SVD than using NMF (which is closer to the original version) ($p < .01$). T&N with the current SVD model is not as good as Metaphor Magnet ($p < .01$), and Meta4meaning outperforms all the other methods ($p < .001$).

Testing the Salience Imbalance Hypothesis and the Emergent Property Hypothesis Above, we formulated two hypotheses for metaphor interpretations: the salience imbalance hypothesis ($v_i > t_i > 0$) and the emergent property hypothesis $v_i \neq 0$ and $t_i \neq 0$.

We now study how often the hypotheses hold for the association strengths. Of the 76 metaphors, four have either the tenor or the vehicle missing from our 50k vocabulary: ‘cigarette is a timebomb’, ‘desk is a junkyard’, ‘tree trunk is a straw’, and ‘sermon is a sleeping pill’, which together have nine human interpretations in the dataset. In addition, there are two other human interpretations not in our vocabulary: ‘breakable’ for the metaphor ‘health is glass’, and ‘extinguished’ for ‘typewriter is a dinosaur’. We used the remaining 134 metaphor interpretations in testing the hypotheses, looking at the association strengths of the interpretations with respect to the corresponding tenor (t_i) and vehicle (v_i).

Among the 72 cases where $t_i > 0$ and $v_i > 0$, there are 20 cases (28%) where the salience imbalance hypothesis does not hold, i.e., where the property is actually more salient to the tenor than to the vehicle ($t_i > v_i > 0$). This leads to the conclusion that a metaphor interpretation does not have to be more salient to the vehicle than to the tenor, at least for properties acquired from corpora using this method.

Table 6: Recall of metaphor interpretations, effect of dimensionality reduction (best performance in bold).

	Recall				
Method	**@5**	**@10**	**@15**	**@25**	**@50**
Meta4meaning (c_i)	**0.221**	**0.303**	**0.339**	**0.397**	**0.454**
SVD, Sum of saliences p_i^+	0.057	0.081	0.099	0.145	0.226
NMF, Sum of saliences p_i^+	0.076	0.098	0.098	0.115	0.192

Table 7: Recall of metaphor interpretations (with best performance in bold), compared to the method of Terai and Nakagawa (2008) (T&N) and to Metaphor Magnet (Veale and Li 2012).

	Recall				
Method	**@5**	**@10**	**@15**	**@25**	**@50**
T&N, NMF	0.053	0.061	0.068	0.121	0.195
T&N, SVD	0.053	0.072	0.094	0.167	0.252
Metaphor Magnet	0.102	0.155	0.181	0.193	0.239
Meta4meaning (c_i)	**0.221**	**0.303**	**0.339**	**0.397**	**0.454**

The emergent property hypothesis was tested in a similar fashion. Among the 134 metaphor interpretations, there are 48 cases (36%) where the salience of the interpretation to the tenor is $t_i = 0$ and 24 cases (18%) where the salience to the vehicle is $v_i = 0$. In ten of these cases (7%), $t_i = 0 = v_i$. Altogether, 62 metaphor interpretations (46%) do not appear in the corpus-derived list of properties of either the tenor or the vehicle. This result may be evidence for emergent properties of metaphors. It also highlights issues with our current approach. We address them next.

Error Analysis

Let us now take a closer look at the cases where Meta4meaning is not successful in recalling metaphor interpretations.

First, some of the properties proposed by Meta4meaning are actually semantically the same as or very similar to interpretations given in the dataset. For instance, the metaphor ‘city is a jungle’ has interpretation ‘crowded’, while Meta4meaning suggests ‘dense’, at rank 3. Other examples of semantically similar interpretation–property pairs include ‘scary’/‘fear’, ‘challenging’/‘difficult’, ‘destructive’/‘destroy’. These properties can be considered to be correct interpretations, and one could argue that the issue is more in the evaluation methodology than in the metaphor interpretation method.

As mentioned previously, four metaphors and additionally two metaphor interpretations are not included in our 50k vocabulary, so Meta4meaning has no way to provide (correct) interpretations for these. Increasing the size of the vocabulary could help here, but it could also add noise and reduce recall. Computationally, a larger vocabulary is not a problem for calculating association scores but might impose a challenge for dimensionality reduction.

Recall that 46% of metaphor interpretations have zero salience to either the tenor or the vehicle. They are thus entirely missed by all the variants of Meta4meaning, since Meta4meaning only considers overlap properties ($t_i > 0$ and $v_i > 0$). Dropping the overlap requirement might potentially increase recall, but it would add a lot of noise as well.

As an example, the metaphor 'the woman is a cat' has only one interpretation – 'independent'. Independent, in our method, is not associated with 'cat' at all, but 'feral' and 'wild' are among the salient properties of cat (but neither one is associated with 'woman' too). Clearly, feral and wild both touch upon independent, but they are in the language of talking about cats, not women. A method allowing one to find analogies could be helpful to solve some cases like this, by finding the property ('independent') of women that is similar to the properties 'feral' and 'wild' of cats.

In 7% of metaphor interpretations, $v_i = 0 = t_i$, and Meta4meaning has no means of identifying them. Examples of such interpretations are 'annoying' for the metaphor 'obligation is a shackle' and 'life' for 'money is oxygen'. Part of this can be considered a failure at Meta4meaning's word association extraction step. One might expect matrix factorization to help with these cases; more work is needed here.

54% of the metaphor interpretations are among the overlap properties of the tenors and the vehicles. Meta4meaning achieved a recall of about 30% when only considering the top ten ranked properties. The variants of Meta4meaning promote the properties that are relatively salient, among the overlap properties, to both the tenor and the vehicle (i.e. the overlap properties may not be salient to the tenor and the vehicle after all), and partly the ones relatively salient to either the tenor or the vehicle. Metaphor interpretations of such characteristics have the chance to be captured. Nevertheless, Meta4meaning can not spot the interpretations that have relatively low saliences to both the tenor and the vehicle.

Conclusions and Future Work

We have described Meta4meaning, a method for interpreting metaphors. Meta4meaning uses corpus-derived word associations so it has a large vocabulary and can potentially be applied to languages other than English. We evaluated Meta4meaning empirically using salient human interpretations of metaphors and compared its performance to other leading methods. The results indicate that Meta4meaning has high recall performance, considering the difficulty of the task, and substantially outperforms other methods.

We proposed and compared several ways of combining the salience of a property to the tenor with its salience to the vehicle in order to rank the properties as possible metaphor interpretations. The combinations are based on three principles: salience aggregation (the product or sum of saliences), salience difference, and combining the results of the two. Salience aggregation captures more correct metaphor interpretations than the difference. Combining the two improves the results. However, the current combination method is simple, and a more sophisticated method may bring further improvement. Moreover, future research could be dedicated to better understanding when salience aggregation and difference work best.

In addition to direct co-occurrence-based word associations, we also experimented with dimensionality reduction (SVD and NMF). However, the results we obtained with them were inferior to those obtained directly with association strength. Further work is needed to investigate how to make better use of the co-occurrence matrix in the context of metaphor interpretation, possibly with dimensionality reduction.

Our analysis of the emergent property hypothesis shows that it holds for our corpus-derived word associations: almost half of the interpretations have no association at all with the tenor or the vehicle. It would be interesting to discover a mechanism of how non-salient properties emerge as interpretations of metaphors.

Given a metaphor, Meta4meaning provides a list of interpretations with varying weights. The interpretations cover multiple aspects of the tenor and vehicle, including various linguistic forms and closely related meanings. Such a multitude of interpretations can be a great benefit at least in two ways. First, it offers opportunity for context adaption. Metaphors are always used in a context, and this context could potentially be used to increase the weights of context-relevant properties so that different contexts result in different interpretations. Second, when Meta4meaning is used as part of a creative system, such as a computer novelist, its rich repository of semantically adjacent words can help find suitable metaphors.

Acknowledgments

This work has been supported by the European Commission under the FET grant 611733 (ConCreTe) and by the Academy of Finland under grant 276897 (CLiC).

References

Becker, A. H. 1997. Emergent and common features influence metaphor interpretation. *Metaphor and Symbol* 12(4):243–259.

Black, M. 1962. *Models and Metaphors: Studies in Language and Philosophy*. New York: Cornell University Press.

Evert, S. 2008. Corpora and collocations. In Lüdeling, A., and Kytö, M., eds., *Corpus Linguistics. An International Handbook*, volume 2. Berlin: Mouton de Gruyter. 1212–1248.

Glucksberg, S. 2001. *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press.

Kintsch, W., and Bowles, A. R. 2002. Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol* 17(4):249–262.

Kintsch, W. 2000. Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review* 7(2):257–266.

Lapesa, G.; Evert, S.; and Schulte im Walde, S. 2014. Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics*, SEM '14, 160–170.

McGregor, S.; Agres, K.; Purver, M.; and Wiggins, G. 2015. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence* 6(1):55–86.

Moreno, R. E. V. 2004. Metaphor interpretation and emergence. *UCL Working Papers in Linguistics* 16:297–322.

Murphy, M. L. 2010. *Lexical Meaning*. Cambridge University Press.

Ortony, A.; Schallert, D. L.; Reynolds, R. E.; and Antos, S. J. 1978. Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior* 17(4):465–477.

Ortony, A. 1979. The role of similarity in similes and metaphors. In Ortony, A., ed., *Metaphor and Thought*. Cambridge University Press. 186–201.

Peirsman, Y.; Heylen, K.; and Geeraerts, D. 2008. Size matters: Tight and loose context definitions in english word space models. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 34–41.

Rapp, R. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the Nineteenth International Conference on Computational Linguistics*, 1–7.

Richards, I. A. 1936. *The Philosophy of Rhetoric*. London: Oxford University Press.

Roncero, C., and Almeida, R. G. 2014. Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior Research Methods* 47(3):800–812.

Sahlgren, M. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. Dissertation, University of Stockholm, Stockholm, Sweden.

Terai, A., and Nakagawa, M. 2008. A corpus-based computational model of metaphor understanding incorporating dynamic interaction. In *Proceedings of The Eighteenth International Conference on Artificial Neural Networks*, ICANN '08, 443–452.

Turney, P. D.; Neuman, Y.; Assaf, D.; and Cohen, Y. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 680–690.

Utsumi, A., and Sakamoto, M. 2011. Indirect categorization as a process of predicative metaphor comprehension. *Metaphor and Symbol* 26(4):299–313.

Veale, T., and Li, G. 2012. Specifying viewpoint and information need with affective metaphors: A system demonstration of the metaphor magnet web app/service. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, 7–12.