Measuring metaphoricity

Jonathan Dunn
Department of Computer Science / Illinois Institute of Technology
jonathan.edwin.dunn@gmail.com

Abstract

This paper presents the first computationally-derived scalar measurement of metaphoricity. Each input sentence is given a value between 0 and 1 which represents how metaphoric that sentence is. This measure achieves a correlation of 0.450 (Pearson's R, p <0.01) with an experimental measure of metaphoricity involving human participants. While far from perfect, this scalar measure of metaphoricity allows different thresholds for metaphoricity so that metaphor identification can be fitted for specific tasks and datasets. When reduced to a binary classification evaluation using the VU Amsterdam Metaphor Corpus, the system achieves an F-Measure of 0.608, slightly lower than the comparable binary classification system's 0.638 and competitive with existing approaches.

1 Introduction

Metaphor is a cognitive phenomenon (Lakoff & Johnson, 1980, 1999) which has a significant impact on human reasoning abilities (Casasanto & Jasmin, 2012; Johansson Falk & Gibbs, 2012) and which, as a result, commonly appears in language in the form of metaphoric expressions (e.g., Deignan, 2005). The most comprehensive non-computational study of metaphoric expressions in large corpora (Steen, et al., 2010) found that up to 18.5% of words in the British National Corpus were used metaphorically. This means that metaphorically used words not only have very different interpretations than literally used words, but they are also common enough to pose a significant challenge for computational linguistics.

Starting with Wilks (1978), the problem of metaphor has been approached as an identification task: first identify or detect metaphoric expressions and then (1) prevent them from interfering with computational treatments of literal expressions and (2) use them to gain additional insight about a text (e.g., Carbonell, 1980; Neuman & Nave, 2009). The identification or detection task has been approached as a binary classification problem: for a given unit of language (e.g., word, phrase, sentence) decide whether it is metaphoric or non-metaphoric. Wilks (1978) used selectional restrictions for this purpose; Mason (2004) used hand-crafted knowledge resources to detect similar selectional mismatches; another approach is to detect selectional mismatches using statistically created resources (e.g., Shutova, et al. 2013; Shutova & Sun, 2013). A second general approach to the binary classification problem has been to use mismatches in properties like abstractness (Gandy, et al., 2013; Assaf, et al., 2013; Tsvetkov, et al., 2013; Turney, et al., 2011), semantic similarity (Li & Sporleder, 2010; Sporleder & Li, 2010), and domain membership (Dunn, 2013a, 2013b) to identify metaphoric units of language. A third approach has been to use forms of topic modelling to identify linguistic units which represent both a metaphoric topic and a literal topic (Strzalkowski, 2013; Bracewell, et al, 2013; Mohler, et al., 2013).

The single constant across all of these approaches is that the task is viewed as a binary classification problem of distinguishing metaphoric language from non-metaphoric language. This binary distinction assumes a clear boundary between the two; in other words, it assumes that metaphoricity is a discrete property. However, three strands of theoretical research show that metaphoricity is not a discrete property. First, psycholinguistic studies of metaphor processing show that there is no difference between the processing of metaphoric and non-metaphoric language (Coulson & Matlock, 2001; Gibbs, 2002; Evans, 2010). The most plausible interpretation

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 745–751,
Baltimore, Maryland, USA, June 23-25 2014.

of this psycholinguistic evidence is that most linguistic units fall somewhere between metaphoric and literal, so that metaphoricity is a scalar value which influences processing gradually (and is difficult to uncover because of related factors like salience; Giora, 2002). Second, linguistic studies of metaphor have found that the metaphoricity of a linguistic unit can be predicted given certain factors (Dunn, 2011, 2013c). Third, the high frequency of metaphorically used language implies that it is hard to set a boundary beyond which a word is used metaphorically. In other words, it seems clear that 18.5% of the BNC is not highly metaphoric, but rather is the sort of slightly metaphoric language that speakers are not consciously aware of because it is used so frequently.

This paper introduces a system for producing a scalar measurement of metaphoricity which places sentences anywhere between 0 (literal) and 1 (highly metaphoric). The goal is to produce a computationally derived measurement that models the gradient nature of metaphoricity, with the result that metaphors which are clearly and consciously seen as metaphors score closer to 1 and metaphors which are not realized by speakers to be metaphoric score further from 1. This scalar measurement approach has two advantages: (1) it adheres more closely to the current theoretical understanding of metaphor, thus being more cognitively accurate; (2) it allows applications to control the threshold of metaphoricity when identifying metaphor, thus allowing the treatment of metaphor to be optimized for a given task.

2 Measuring Gradient Metaphoricity

An experiment was conducted to set a standard for evaluating scalar measurements of metaphoricity. A corpus of 60 sentences of varying metaphoricity, drawn equally from four top-level domains (PHYSICAL, MENTAL, SOCIAL, and ABSTRACT), was created using the Corpus of Contemporary American English. Each domain was represented by five verbs and each verb by three sentences: one literal, one slightly metaphoric, and one very metaphoric (as judged by the author).

The selection of various domains, verbs, and hypothesized metaphoricity levels helps to control for other factors, like abstractness, which might be only indirectly related to metaphoricity. It also ensures that the experiment covers a wide-range of metaphors. It should be noted that the purpose of the experiment is not to (1) test a three-way distinction between metaphoricity levels (which is simply used to ensure a representative selection of metaphors) or (2) test the author's intuitions of metaphoricity. Rather, the purpose is to have a representative selection of metaphors rated for metaphoricity against which to test scalar measurements of metaphoricity.

Three survey tasks were used. The first tested speakers' ability to consistently separate metaphoric and non-metaphoric sentences. Participants were given a sentence and asked to identify it as "Literal" or "Metaphoric." The second task tested speakers' ability to consistently label a given sentence as "Not Metaphoric", "Slightly Metaphoric", and "Very Metaphoric." The additional label was added in order to provide participants with a middle ground between metaphoric and literal. The third task tested speakers' ability to consistently rank three sentences according to their metaphoricity. In order to ensure comparability, each set of three sentences contained a literal, a slightly metaphoric, and a very metaphoric use of a single verb (e.g., three uses of "butcher"). The purpose of this task was to allow participants to directly compare different uses of the same verb.

The surveys were conducted using the MechanicalTurk platform. Each participant took a particular survey only once and the sentences to be rated were drawn randomly from the corpus. Participants were given eight questions for the identification and labeling tasks and four questions for the ranking task. This was done in order to keep the survey short and prevent participants from losing interest. All participants were asked if they had attended a primary or elementary school conducted in English in order to ensure consistent language ability. Further, a test question was positioned part way through the survey to ensure that participants read the prompts correctly. Only answers valid according to these two tests are considered in the following results. Each task had 100 unique participants who gave valid answers, for a total of 300 participants. Participants did not see any domain information for the sentence prompts.

For the first task, the binary identification task, the metaphoricity of a sentence was computed by taking the percentage of participants who identified it as metaphoric. Thus, if all participants agreed that a sentence was metaphoric, then it receives a 1, while if half of the participants agreed,

then it receives a 0.5. The idea here is that high metaphoricity is consciously available to participants, so that the more agreement there is about metaphor the more the participants are aware of the sentence's metaphoricity and thus the higher its metaphoricity value should be. The results of this first experiment are summarized in Table 1 with the mean, standard deviation, and range of the metaphoricity measurements. The results are strong on the low end of the scale, with every domain having sentences with either 0 values or close to 0 values. The high end is more problematic, with the highest values in each domain being below 0.9. This is a result of not having perfect agreement across all participants. However, in spite of this, the measure makes a good distinction between utterances. For example, it assigns the metaphoricity value of 0.833 to the sentence in (1), but a metaphoricity value of only 0.153 to the sentence in (2). This reflects a distinction in metaphoricity, although the extreme top and bottom of the scale are problematic.

(1) "A lady on high heels clacked along, the type my mother says invests all of her brainpower in her looks."

(2) "The banks and the corporations in America today have lots of money that they can invest right now."

Domain	Mean	Std. Dev.	Range
Abstract	0.373	0.282	0.065–0.833
Mental	0.289	0.278	0.000–0.888
Physical	0.417	0.331	0.000–0.846
Social	0.389	0.351	0.000–0.812
All	0.367	0.316	0.000–0.888

Table 1: Metaphoricity by identification.

The second experiment asks participants to label metaphoricity, this time including a distinction between slightly metaphoric and highly metaphoric sentences. The purpose of this is not to test a three-way distinction in metaphoricity values, but rather to improve the scale by moving intermediate sentences out of the Literal or Metaphoric categories. The metaphoricity values for this experiment were calculated in the same way: the percentage of participants who rated a sentence as highly metaphoric. Thus, this measurement also is based on the idea that more participants will be consciously aware of highly metaphoric sentences, with a third category available to allow an extra distinction to be made. This measurement, summarized in Table 2, is more accurate at the lower end of the scale, with many sentences receiving a 0 because participants were able to choose a category other than metaphoric. At the same time, the values tend to be further from 1 at the upper end of the scale. The sentence in (2) above, for example, received a 0; however, the sentence in (1) above received only a 0.571, which, while high given the range of values, is still far from 1. Thus, while the measurement makes distinctions at the top of the scale, it does not approach 1.

Domain	Mean	Std. Dev.	Range
Abstract	0.170	0.165	0.000–0.571
Mental	0.096	0.119	0.000–0.455
Physical	0.220	0.248	0.000–0.778
Social	0.258	0.281	0.000–0.769
All	0.186	0.222	0.000–0.778

Table 2: Metaphoricity by labelling.

The third task gathered ordering information by presenting participants with three sentences, all of which contained the same main verb. The participants were asked to order the sentences from the least metaphoric to the most metaphoric. The purpose of this experiment was to give participants context in the form of other uses of a given verb against which to make their judgments. The metaphoricity value was computed by taking the percentage of participants who identified a sentence as the most metaphoric of the three given sentences. This measurement, shown in Table 3, has similar averages across domains, unlike the previous measurements. It tends to be better than the previous measures on the upper bound, likely because of the contextual comparison it allows. However, because sentences with the same main verb were forced into a three-way ordering, participants could not, for example, label two of the sentences as equally metaphoric. Thus, it is possible that some of this advantage on the upper bound is a result of the task itself.

Given these three experiments for measuring the metaphoricity of sentences, Table 4 shows the correlations between each measure using Pearson's R. Each correlation is significant at the 0.01 level (2-tailed). The highest correlation is between the first and second tasks, at 0.819. The lowest is between the first and third (which differ in the

Domain	Mean	Std. Dev.	Range
Abstract	0.333	0.211	0.056–0.773
Mental	0.331	0.175	0.071–0.632
Physical	0.331	0.235	0.050–0.941
Social	0.327	0.280	0.050–0.783
All	0.331	0.227	0.050–0.941

Table 3: Metaphoricity by ordering.

number of distinctions allowed) at 0.699. However, this is still a high correlation.

Task	Identify	Label	Order
Identify	–	0.819	0.699
Label	0.819	–	0.702
Order	0.699	0.702	–

Table 4: Correlation between measurements.

This section has put forward a robust series of scalar measurements of metaphoricity. Each experiment had 100 participants and operationalized the task of rating metaphoricity in different ways across a representative section of domains, verbs, and metaphoricity levels. The resulting highly correlated measures show that we have a good standard of metaphoricity against which to evaluate computational models which produce scalar measurements of metaphoricity. The next section introduces such a system.

3 Description of the System

We approach the problem by starting with an existing binary identification system and converting it to a scalar system. In principle any of the property-based systems listed above could be converted in this way. We have chosen to start with the domain interaction system (Dunn, 2013a, 2013b), which performed competitively in an evaluation with other systems (Dunn, 2013b). The original system uses the properties of domain-membership and event-status of concepts to identify metaphors at the sentence-level using a logistic regression classifier. The scalar version of the system will have to evaluate the features in a different way.

The first step is to increase the robustness of the system's representation of sentences by adding additional properties. We split the original system's domain membership feature into two: the domain of a word's referent and the domain of a word's sense. The idea is to capture cases like MINISTER, in which a physical object (a human) is defined by its social role (being a minister). The event-status property is unchanged.

Several additional properties are added; these properties were not used in the original system. First, animacy-status allows a distinction to be made between inanimate objects like rocks and stones and animate or human objects. Second, the fact-status property allows a distinction to be made between objects which exist as such independently of humans (e.g., rocks and stones) and those which exist to some degree dependent on human consciousness (e.g., laws and ideas). Third, the function-status property allows a distinction to be made between objects which encode a function (e.g., a screwdriver is specifically an object meant to turn screws) and those which do not encode a function (e.g., rocks are simply objects). A finer distinction within the function-status property distinguishes social functions (e.g., laws) from physical-use functions (e.g., screwdrivers).

Following the original system, these properties are taken from a knowledge-base and used to create feature vectors. The text is first processed using Apache OpenNLP for tokenization, named entity recognition, and part of speech tagging. Morpha (Minnen, et al., 2001) is used for lemmatization. At this point word sense disambiguation is performed using SenseRelate (Pedersen & Kolhatkar, 2009), mapping the lexical words to the corresponding WordNet senses. These WordNet senses are first mapped to SynSets and then to concepts in the SUMO ontology, using existing mappings (Niles & Pease, 2001, 2003).

Thus, each sentence is represented by the SUMO concepts which it contains and each concept is represented by its six concept properties. The features used are computed as follows: First, the relative frequency of each value of each concept property in the sentence is determined; Second, the number of instances of the most common value for each property is determined, as well as the number of instances of all other values (both relativized to the number of concepts present in the sentence). Third, the number of types of values for each concept property is determined, relative to the number of possible types. This gives a total of 41 features for each sentence.

These features were computed for each of the sentences used in the experiments and then

the correlation between the features and the metaphoricity measurements were computed using Pearson's R. Those features which had a significant positive relationship with the experimental results, shown in Table 5, were added together to create a rough computational measure of metaphoricity and then converted so that they fall between 0 and 1. The resulting computationally-derived measure correlates significantly with each of the experiments: 0.450, 0.416, and 0.337.

Properties	Values
Domain of the Referent	Mental
Domain of the Referent	Other / Concepts
Event-Status	State
Animacy-Status	Undetermined
Animacy-Status	Other / Concepts
Fact-Status	Physical
Function-Status	None
Domain of the Referent	Types / Possible
Event-Status	Types / Possible
Animacy-Status	Types / Possible
Function-Status (negative)	Types / Possible

Table 5: Predictive features.

4 Evaluation

A scalar measurement of metaphoricity allows the threshold for metaphor in metaphor identification tasks to be fitted for specific purposes and datasets. The scalar system was evaluated on the VU Amsterdam Metaphor Corpus (Steen, et al., 2010) which consists of 200,000 words from the British National Corpus divided into four genres (academic, news, fiction, and spoken; performance on the spoken genre was not evaluated for this task because it consists of many short fragmentary utterances) and manually annotated for metaphor by five raters. Previous evaluations using this corpus (Dunn, 2013b) concluded that prepositions annotated as metaphoric in the corpus should not be considered metaphoric for computational purposes. Thus, metaphorically used prepositions have been untagged as metaphoric. Further, we have also untagged the ambiguously metaphoric sentences. Sentences with an insufficiently robust conceptual representation were removed (e.g., fragments). The evaluation dataset thus consists of 6,893 sentences, distributed as shown in Table 6.

For the purposes of this evaluation, the threshold for metaphor was set independently for each genre and tied to the number of sentences containing metaphorically used words, as rated by the annotators of the corpus. Thus, for the number x of metaphors in the genre, the x sentences with the top metaphoricity values were identified as metaphoric. This illustrates the flexibility of such a scalar approach to metaphor identification. The baseline results are taken from a binary classification evaluation of the corpus using the full set of 41 features produced by the system and evaluated using the logistic regression algorithm with 100-fold cross-validation.

Subset	Literal	Metaphor	Total
Academic	759	1,550	2,309
Fiction	1,215	1,389	2,604
News	366	1,614	1,980
Total	2,340	4,553	6,893

Table 6: Size of evaluation dataset in sentences.

System	Subset	Prec.	Recall	F-Meas.
Scalar	Acad.	0.578	0.686	0.578
Binary	Acad.	0.649	0.682	0.623
Scalar	News	0.712	0.822	0.712
Binary	News	0.750	0.812	0.748
Scalar	Fict.	0.554	0.582	0.554
Binary	Fict.	0.632	0.633	0.630
Scalar	All	0.608	0.703	0.608
Binary	All	0.663	0.685	0.638

Table 7: Evaluation results.

The binary classification system, with access to the full range of features, out-performs the scalar measurement in most cases. It is important to note, however, that the binary classification system requires labelled training data and is restricted to a single threshold of metaphoricity, in this case the threshold embedded in the corpus by the raters. The scalar system, however, was trained only on the experimental data and was not influenced by the evaluation corpus (except, of course, that it had access to the number of metaphoric sentences in the dataset, which is a parameter and not part of the model itself). Further, it can be applied to any English text without the need for labelled training data. Thus, the scalar approach performs competitively on a binary task (0.608 vs. 0.638 F-Measure) but can also produce scalar identifications, which binary systems cannot produce.

References

Assaf, D., Neuman, Y., Cohen, Y., Argamon, S., Howard, N., Last, M., Koppel, M. 2013. Why "dark thoughts" aren't really dark: A novel algorithm for metaphor identification. *2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain*: 60–65. Institute of Electrical and Electronics Engineers.

Bracewell, D. B., Tomlinson, M. T., Mohler, M. 2013. Determining the Conceptual Space of Metaphoric Expressions. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing, Volume I*: 487–500. Berlin, Heidelberg: Springer-Verlag.

Carbonell, J. 1980. Metaphor - A Key to Extensible Semantic Analysis. *Proceedings of the 18th Meeting of the Association for Computational Linguistics*: 17–21. Association for Computational Linguistics.

Casasanto, D., Jasmin, K. 2012. The Hands of Time: Temporal gestures in English speakers. *Cognitive Linguistics*, 23(4): 643–674.

Coulson, S., Matlock, T. 2001. Metaphor and the space structuring model. *Metaphor & Symbol*, 16(3), 295-316.

Deignan, A. 2005. *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Dunn, J. 2011. Gradient Semantic Intuitions of Metaphoric Expressions. *Metaphor & Symbol*, 26(1), 53-67.

Dunn, J. 2013a. Evaluating the premises and results of four metaphor identification systems. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing, Volume I*: 471-486. Berlin, Heidelberg: Springer-Verlag.

Dunn, J. 2013b. What metaphor identification systems can tell us about metaphor-in-language. *Proceedings of the First Workshop on Metaphor in NLP*: 1-10. Association for Computational Linguistics.

Dunn, J. 2013c. How linguistic structure influences and helps to predict metaphoric meaning. *Cognitive Linguistics*, 24(1), 33-66.

Evans, V. 2010. Figurative language understanding in LCCM Theory. *Cognitive Linguistics*, 21(4), 601-662.

Gandy, L., Allan, N., Atallah, M., Frieder, O., Howard, N., Kanareykin, S., Argamon, S. 2013. Automatic Identification of Conceptual Metaphors With Limited Knowledge. *Proceedings of the 27th Conference on Artificial Intelligence*: 328–334. Association for the Advancement of Artificial Intelligence.

Gibbs Jr., R. W. 2002. A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, 34(4), 457-486.

Giora, R. 2002. Literal vs. figurative language: Different or equal? *Journal of Pragmatics*, 34(4), 487-506.

Johansson Falck, M., Gibbs, Jr., R. W. 2012. Embodied motivations for metaphorical meanings. *Cognitive Linguistics*, 23(2): 251–272.

Lakoff, G., Johnson, M. 1980. *Metaphors we live by*. Chicago: University Of Chicago Press.

Lakoff, G., Johnson, M. 1999. *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Chicago: University Of Chicago Press.

Li, L., Sporleder, C. 2010a. Linguistic Cues for Distinguishing Literal and Non-literal Usages. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*: 683-691. Association for Computational Linguistics.

Li, L., Sporleder, C. 2010b. Using Gaussian Mixture Models to Detect Figurative Language in Context. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*: 297–300. Association for Computational Linguistics.

Mason, Z. 2004. CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics*, 30(1), 23-44.

Minnen, G., Carroll, J., Pearce, D. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3), 207-223.

Mohler, M., Bracewell, D., Tomlinson, M., Hinote, D. 2013. Semantic Signatures for Example-Based Linguistic Metaphor Detection. *Proceedings of the First Workshop on Metaphor in NLP*: 27-35. Association for Computational Linguistics.

Neuman, Y., Nave, O. 2009. Metaphor-based meaning excavation. *Information Sciences*, 179, 2719-2728.

Niles, I., Pease, A. 2001. Towards a standard upper ontology. *Proceedings of the International Conference on Formal Ontology in Information Systems*: 2-9. Association for Computing Machinery.

Niles, I., Pease, A. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*: 412-416. World Congress in Computer Science, Computer Engineering, and Applied Computing.

Pedersen, T., Kolhatkar, V. 2009. WordNet::SenseRelate::AllWords - A broad coverage word sense tagger that maximimizes semantic relatedness. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North*

American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session: 17-20. Association for Computational Linguistics.

Shutova, E., Sun, L. 2013. Unsupervised Metaphor Identification using Hierarchical Graph Factorization Clustering. *Proceedings of Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*: 978-988. Association for Computational Linguistics.

Shutova, E., Teufel, S., Korhonen, A. 2013. Statistical Metaphor Processing. *Computational Linguistics*, 39(2), 301-353.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T. 2010. Metaphor in usage. *Cognitive Linguistics*, 21(4), 765-796.

Strzalkowski, T., Broadwell, G. A., Taylor, S., Feldman, L., Shaikh, S., Liu, T., Elliot, K. 2013. Robust Extraction of Metaphor from Novel Data. *Proceedings of the First Workshop on Metaphor in NLP*: 67-76. Association for Computational Linguistics.

Tsvetkov, Y., Mukomel, E., Gershman, A. 2013. Cross-Lingual Metaphor Detection Using Common Semantic Features. *Proceedings of the First Workshop on Metaphor in NLP*: 45-51. Association for Computational Linguistics.

Turney, P. D., Neuman, Y., Assaf, D., Cohen, Y. 2011. Literal and Metaphorical Sense Identification Through Concrete and Abstract Context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: 680-690. Association for Computational Linguistics.

Wilks, Y. 1978. Making preferences more active. *Artificial Intelligence*, 11(3), 197-223.