Modeling the interaction between sensory and affective meanings for detecting metaphor

Andrew Gargett
School of Computer Science
University of Birmingham
United Kingdom
A.D.Gargett@cs.bham.ac.uk

John Barnden
School of Computer Science
University of Birmingham
United Kingdom
J.A.Barnden@cs.bham.ac.uk

Abstract

Concreteness and imageability have long been held to play an important role in the meanings of figurative expressions. Recent work has implemented this idea in order to detect metaphors in natural language discourse. Yet, a relatively unexplored dimension of metaphor is the role of affective meanings. In this paper, we will show how combining concreteness, imageability and sentiment scores, as features at different linguistic levels, improves performance in such tasks as automatic detection of metaphor in discourse. By gradually refining these features through descriptive studies, we found the best performing classifier for our task to be random forests. Further refining of our classifiers for part-of-speech, led to very promising results, with $F1$ scores of .744 for nouns,.799 for verbs, .811 for prepositions. We suggest that our approach works by capturing to some degree the complex interactions between external sensory information (concreteness), information about internal experience (imageability), and relatively subjective meanings (sentiment), in the use of metaphorical expressions in natural language.

1 Introduction

Figurative language plays an important role in "grounding" our communication in the world around us. Being able to talk about "the journey of life", "getting into a relationship", whether there are "strings attached" to a contract, or even just "surfing the internet", are important and useful aspects of everyday meaning-making practices. Much recent work on modeling metaphor, especially using computational techniques, has concentrated on more inter-subjective aspects of such meanings, such as the way that figurative expressions are apparently used to inject meanings that are somehow more "concrete" into daily discourse (Turney et al., 2011; Tsvetkov et al., 2013).On such an account, describing love as a journey, or life as a test, is a way of casting a fairly abstract idea, such as love or life, in more concrete and everyday terms, such as a journey or a test. Related dimensions of figurative meanings, such as imageability, having to do with how readily the concept expressed by some linguistic item brings an image to mind, have also been investigated (Cacciari and Glucksberg, 1995; Gibbs, 2006; Urena and Faber, 2010).

Work across a range of disciplines has begun examining the complex interaction between metaphor and the intra-subjective emotional meanings expressed at all levels of language (Kövecses, 2003; Meier and Robinson, 2005; Strzalkowski et al., 2014), although modelling such interaction has proved to be somewhat challenging. For example, while a native speaker of some language can be expected to consistently and reliably rate isolated words for their levels of valence ("pleasantness"), arousal ("emotional intensity") and dominance ("control") (Warriner et al., 2013), the same cannot be expected for more complex expressions such as "the journey of life" or "strings attached". Whether there are indeed systematic and stable patterns for the intra-subjective meanings of such expressions is still an open question.

Linking these two components of figurative meaning, while it has been understood for some time that concreteness and imageability very strongly correlate (Paivio et al., 1968), recent work has suggested strong reasons for rethinking this. On the contrary, (Dellantonio et al., 2014) suggest that concreteness and imageability are in fact quite different psychological constructs, and the basis for this difference is that imageability involves both "external sensory information" as well as "internal bodily-related sensory experience," whereas concreteness ratings capture only external sensory information. From this apparent difference internal vs. external sensory information, they derive an index of the "weight" such internal sensory information has in relation to individual word meaning, which can be derived as the difference between the concreteness and imageability of a word. Labelling this weight as **w**, we could symbolise this idea as follows:

$$\mathbf{w} = |(\text{CONC - IMAG})|$$

This will allow us to more clearly separate concreteness from imageability in our modelling, and so better examine the interactions of each with sentiment in processing metaphor.

There has been much work on the interaction between metaphor and sentiment (Fainsilber and Ortony, 1987; Fussell and Moss, 1998; Littlemore and Low, 2006). Metaphor researchers have long recognised that metaphor and affective communication are central to each other: metaphor is a central way of conveying affect, and conversely conveying affect is a central function of metaphor. However, it is important to distinguish between (a) using metaphor to describe an emotion (e.g. "anger swept through me") vs. (b) emotion being conveyed through connotations of source terms (e.g. "terrorism is a form of cancer", where negative affect about cancer carries over to terrorism).(Barnden, 2015) However, there is still much more work to do in elucidating these complex connections.

Motivated by such considerations, we focus here on how concreteness, imageability, and affective meaning, interact in metaphorical expressions, and to this end we have examined a large corpus annotated for metaphor, the Vrije University Amsterdam Metaphor Corpus (VUAMC) (Steen et al., 2010), with respect to such fetures as imageability and concreteness, as well as valence, arousal and dominance. The background for these studies is our ongoing work on devising a computational tool for detecting, and to some degree, also understanding, metaphor.[1]

2 Method

2.1 Data

Our data comes from the Vrije University Amsterdam Metaphor Corpus (VUAMC), consisting of over 188,000 words selected from the British National Corpus-Baby (BNC-Baby), and annotated for metaphor using the Metaphor Identification Procedure (MIP) (Steen et al., 2010). The MIP involves annotators considering individuals words from the corpus, and answering the question (somewhat simplified here): does this word have a more "basic" meaning[2] than its current "contextual" meaning, with the latter also being understandable in comparison with the former? If the answer is "yes", the current item is used metaphorically, else it is used non-metaphorically.

The corpus itself has four registers, of between 44,000 and 50,000 words each: academic texts, news texts, fiction, and conversations, with over 23,600 words were annotated as metaphorical across the 4 registers.[3] Table (1) lists statistics for the VUAMC from (Steen et al., 2010), specifically presenting standardised residuals (SRs) for counts of metaphorical vs. non-metaphorical nouns, verbs and prepositions lexical units, and of[4] SRs usefully enable pinpointing interesting deviations of the observed frequency for items occurring in specific categories in our sample from the frequency we actually expect for them given their overall frequency in the entire sample.[5] For example, while there are

[1] For more information, see: `http://www.cs.bham.ac.uk/~gargetad/genmeta-index.html`

[2] Defined in terms of being *more concrete*, *related to bodily actions*, *more precise*, and *historically older*, see (Steen et al., 2010) for details.

[3] The VUAMC is available from: `http://ota.ahds.ac.uk/desc/2541.html`

[4] More strictly, they refer to so-called *metaphor-related words*, i.e. a word the use of which "may potentially be explained by some form of cross-domain mapping from a more basic meaning of that word."

[5] For a proper appreciation of the statistics, please see the relevant sections of (Steen et al., 2010).

far fewer nouns in all registers except Conversations, prepositions occur with far greater than expected frequency in all registers, and verbs are similar to prepositions, although not as extreme, in occurring with greater than expected frequency in all registers. The VUAMC is usefully balanced across 4 registers, making it highly useful for our ongoing work on automatic metaphor annotation.

2.2 Procedure

2.2.1 Pre-processing

We have enriched the VUAMC in several ways. First, we have parsed the corpus using the graph-based version of the Mate tools dependency parser (Bohnet, 2010), adding rich syntactic information.[6] Second, we have incorporated the MRC Psycholinguistic Database (Wilson, 1988), a dictionary of 150,837 words, with different subsets of these words having been rated by human subjects in psycholinguistic experiments.[7] Of special note, the database includes 4,295 words rated with degrees of concreteness, these ratings ranging from 158 (meaning *highly abstract*) to 670 (meaning *highly concrete*), and also 9,240 words rated for degrees of imageability, which is taken to indicate how easily a word can evoke mental imagery, these ratings also ranging between 100 and 700 (a higher score indicating greater imageability). The concreteness scores (and to some extent the imageability ones also) have been used extensively for work on metaphor, e.g. (Turney et al., 2011; Tsvetkov et al., 2013). Finally, we have incorporated the work by (Warriner et al., 2013) on the Affective Norms for English Words (ANEW), which provides 13,915 English content words, rated for: valence (measuring the "pleasantness" of the word), arousal ("emotional intensity" of the word) and dominance (the degree of "control" evoked by the thing that the word denotes). This latter dataset includes rich statistical information (such as means and variance) for these scores, which we make use of in our work.

Combining these resources, we extend the VUAMC with information about dependency relations, concreteness, imageability, valence, arousal and dominance. However, this combination is not without problems, for example, the VUAMC data set is much larger than the MRC data set, so that many VUAMC words have no MRC scores, and we need a smoothing procedure; a similar disparity in size exists between the VUAMC corpus and the ANEW scores. Now, a key finding in the literature is that POS strongly correlates with metaphor; Table (1) illustrates this quite well, and we have carried out various studies in this direction (see below). As a first approximation, we smooth such discrepancies between the VUAMC and MRC, by calculating an average MRC score for each POS across the entire corpus, as follows: first, from VUAMC words with MRC scores, we calculated an average MRC score (concreteness/imageability) by POS across all the the VUAMC data, second, those VUAMC words without MRC scores (i.e. missing from the MRC database) could then be assigned a score based on their POS. We did the same for the ANEW scores, but this time not in terms of POS, which is missing from ANEW: we maintained average ANEW scores, and gave these to VUAMC items not represented in the ANEW dataset. However, this kind of naive "global" average is not very discriminative of the key difference we are trying to model between metaphorical vs. non-metaphorical expressions, and we are currently re-implementing our smoothing strategy.[8]

2.2.2 Experimental design

We carried out a preliminary study, followed by three main studies, using the pre-processed data described in Section (2.2.1). Below we list the aims, hypotheses and procedures for these studies.

Preliminary study. This initial study aimed to select features for use in subsequent machine learning studies. In particular, we covered features considered important for the task in previous literature. We were seeking to discover the optimal combination of features for discriminating between literal and non-literal words.

[6] Note this includes POS tagging, POS tags rich enough to capture such distinctions as common nouns vs. personal nouns, participles vs. independent verbs vs. copula verbs, etc – see: `https://code.google.com/p/mate-tools/`.

[7] `http://ota.oucs.ox.ac.uk/headers/1054.xml`

[8] Thanks to the reviewers for this workshop for noting this issue; although, it we should point out that the planning phase for re-implementing this aspect of our work pre-dates submission of this paper.

POS	Academic		News		Fiction		Conversation	
	Lit.	Met.	Lit.	Met.	Lit.	Met.	Lit.	Met.
Nouns	82.4 (1.2)	17.6 (-2.5)	86.8 (4.0)	13.2 (-9.1)	89.5 (1.4)	10.5 (-3.8)	91.7 (-0.4)	8.3 (1.5)
Prepositions	57.5 (-21.4)	42.5 (45.0)	61.9 (-17.0)	38.1 (38.5)	66.6 (-14.9)	33.4 (40.6)	66.2 (-13.5)	33.8 (46.9)
Verbs	72.3 (-9.2)	27.7 (19.3)	72.4 (-10.9)	27.6 (24.6)	84.1 (-4.2)	15.9 (11.6)	90.9 (-1.7)	9.1 (5.7)

Table 1: Percentages of POS and metaphors per register, with standardised residuals in brackets (n=47934)

Study 1. This study aimed to find a suitable learning algorithm, for predicting literal vs. nonliteral expressions. In addition, we also examined the relative importance of particular independent variables, for predicting literal vs. nonliteral expressions, by sampling a range of standard machine learning algorithms,[9] and from this we arrived at a smaller set of more viable learning algorithms, specifically, random forests (**rf**), gradient boosting machines (**gbm**), k nearest neighbours (**knn**), and support vector machines (**svm**). In addition, we considered different combinations of the features collected in the preliminary studies. The resulting models (learning algorithms, plus combinations of features) were chosen because they showed promising performance, and adequately represented the range of models used for similar tasks in other studies elsewhere. This study coincided with the initial phase in developing our system for automatically annotating metaphor, and for this early development version of our system, we constructed a random sample covering 80% of the VUAMC.

For evaluation, we compared results for each target model against a baseline model, this latter being the best single variable model we found in earlier studies, which can predict whether a word is metaphorical or not, based simply on the concreteness score for that word. Results consisted of comparing confusion matrices for ground truth vs. the output of each model, trained on a training set of 60% of the data, then these models were tuned with a testing set of 20% of the data. Finally using a validation set of the remaining 20% of the data, accuracy, precision, recall and F1 scores were calculated for each model.

Study 2. For this next study, focusing on the Random Forests algorithm, we extended our preliminary studies in a quite natural way, and separated the classifiers according to POS, separately training random forest classifiers on our original data set split into nouns, verbs and prepositions. The training setup used here was largely the same as the one used for Study 1, except that the entire VUAMC data set was employed for this study.

Study 3. Finally, having identified various dimensions of metaphorical meaning, and having set out to validate the interaction between these dimensions in Studies 1 and 2, we next turned to possible explanations of the patterns we observed across, for example, different POS. Starting from the random forest classifiers we had trained on the VUAMC in study 2, it was apparent that the sheer number of trees used by such classifiers means they are far from being readily interpretable; nevertheless, we explored the use of recent tools for attempting to improve the interpretability of these kinds of ensemble classifiers.[10]

3 Results

3.1 Preliminary study

In earlier work (Gargett et al., 2014), we examined the role of concreteness and imageability in cap-

[9] Specifically, we considered linear discriminant analysis, k nearest neighbours, naive bayes, random forest, gradient boosting machines, logistic regression, support vector machines.

[10] In particular, we employed the "inTrees" R package for this (Deng, 2014) – see also: `http://cran.r-project.org/web/packages/inTrees/index.html`.

turing the variation between nonliteral vs. literal expressions, for heads vs. their dependents. Figures (1) and (2) suggest that making this kind of fine-grained distinction within our data set between heads and their dependents, enables capturing variation between literal and nonliteral items for some POS; for example, nonliteral head nouns appear to have higher MRC scores than their dependents, distinct from literal head nouns (verbs appear to make no such distinction). While literal and nonliteral head prepositions both seem indistinguishable from their dependents in terms of concreteness scores, nonliteral head prepositions seem to have imageability scores quite distinct from their dependents.

As can be seen from Figures (1) and (2), our initial study failed to capture variation between verbs. As a follow-up, we incorporated features from the ANEW data set; initial results are plotted in Figure (3), and the variation exibited across all POS in this plot suggest, e.g., a possible role for arousal in distinguishing literal from nonliteral verbs.

3.2 Study 1

Next, we focused on selecting a learning algorithm, for predicting literal vs. nonliteral expressions. The features used here are drawn from our earlier studies, directly incorporating the various scores from the MRC and ANEW databases. Results are displayed in Table (2), with the boxed cell in this table showing the strongest performing combination of learning model and features, which turned out to be all features from the MRC and ANEW scores, trained using random forests.

3.3 Study 2

In Table (3), we present the results for our study of different random forest models by POS. The metrics we used here are standard: harmonic mean of recall and precision, or **F1**, and the overall agreement rate between model and validation set, or **accuracy**. A clear effect when including the "weight" term w can be seen (recall this was the difference between concreteness and imageability). The clear winners in each vertical comparison (e.g. between F1 for **Verbs** vs. **Verbs**$_w$) is shown in this table. We will come back to a discussion of the significance of these results in Section (4) below.

3.4 Study 3

Our next study sets out to try to interpret in some way the results from Study 2, and Tables (4) to (6) present the rule conditions extracted from a sample of the first 100 trees for each random forest model for each POS. Important measures of the performance of the classifiers given here include **err**, the so-called out-of-bag (OOB) error estimate, and **freq**, the proportion of occurrences of instances of this rule. OOB error rate is a statistic calculated internally while running the algorithm, and has been shown to be unbiased: while constructing trees, a bootstrap sample is taken from the original data, with some proportion (e.g. about a third) left out, which can be used as a test set while a particular tree is being built, and in this way, a test classification can be obtained for each case in this same proportion (say, about a third) of trees. The error rate is then the proportion of times this test classification was wrong.

Ensemble methods such as random forests are notoriously opaque, largely due to the sheer volume of trees constructed during model building, thereby making clear interpretation of these models problematic (Breiman, 2002; Zhang and Wang, 2009); and yet, they have also emerged as one of the best performing learning models,[11] and work very well for classification tasks such as ours. Consequently, there is broad interest in better interpretation of such models, and development in this direction is ongoing.

Many of the trees built by such ensemble methods,[12] typically contain not only trees crucial for classification performance, but also many which could be removed without significant impact on performance. Proceeding along these lines, "pruning" the forest can result in a smaller, and thus more interpretable, set of trees, without significant impact on performance; from this smaller set, it is more feasible that a relatively transparent set of rules for constructing some significant proportion of trees could be extracted – this smaller set of rules could in principle be used to attempt an interpretation of the re-

[11] As witnessed in several recent Kaggle competitions, `https://www.kaggle.com/wiki/RandomForests`.

[12] Other relevant methods we are currently investigating include gradient boosting machines.

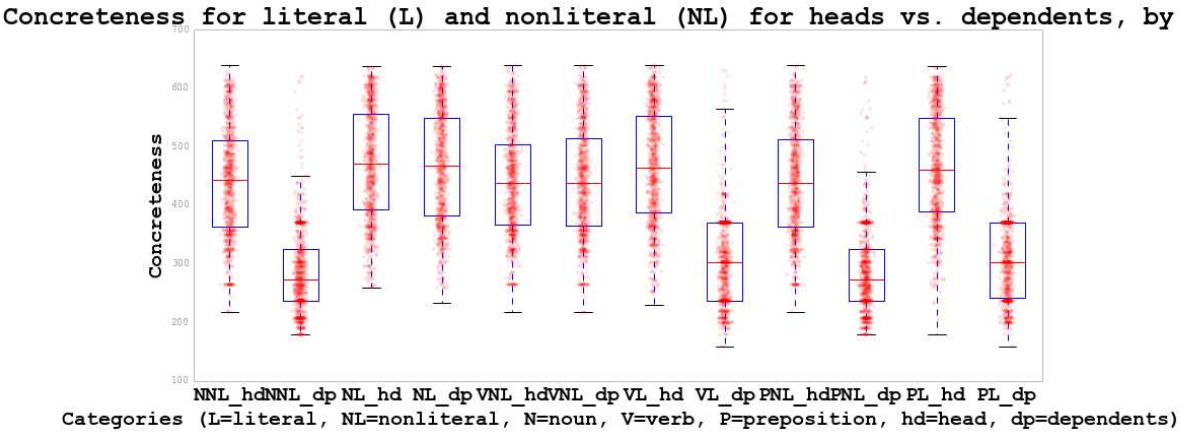

Figure 1: Plot of concreteness scores for literal vs. nonliteral/metaphorical heads vs. their dependents, in the VUAMC, grouped by parts of speech

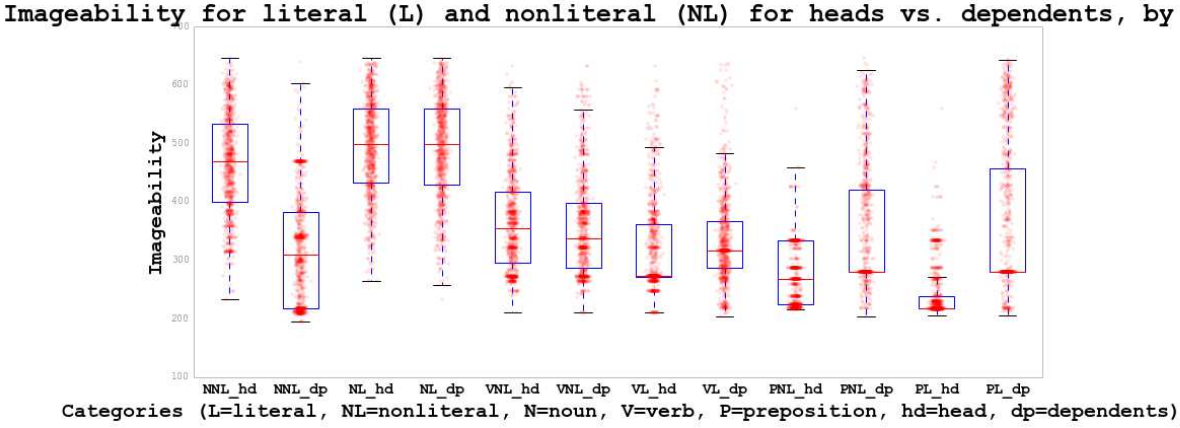

Figure 2: Plot of imageability scores for literal vs. nonliteral/metaphorical heads vs. their dependents, in the VUAMC, grouped by parts of speech

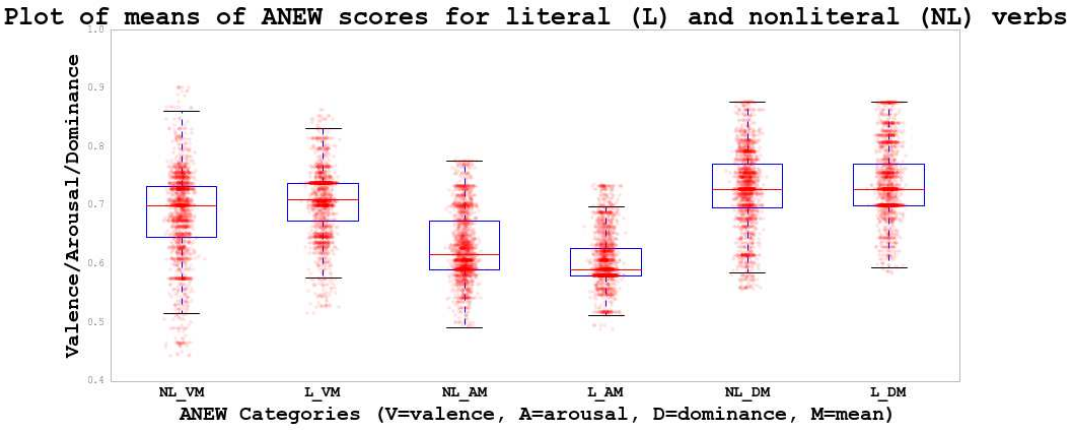

Figure 3: Plot of anew scores for literal vs. nonliteral/metaphorical verbs in the VUAMC

sulting model. The results presented here are illustrative only, be-

Models	**Full**	**Minimal**	**MRC**	**ANEW**	**Base**
gbm	0.7542	0.7364	0.7266	0.7051	0.6673
knn	0.6945	0.6793	0.6861	0.6823	0.6802
rf	0.7813	0.7362	0.7275	0.7144	0.6906
svm	0.6787	0.6689	0.6396	0.6690	0.6348

Table 2: Results (F1) of evaluating models with different combinations of features, for predicting non/literal items (n=3574)

POS	**Accuracy**	**F1**	**n**
Nouns	0.733	0.737	1470
Nouns$_{\mathbf{w}}$	0.743	0.744	1470
Verbs	0.7870	0.799	2216
Verbs$_{\mathbf{w}}$	0.7866	0.798	2216
Prepositions	0.785	0.801	2294
Prepositions$_{\mathbf{w}}$	0.790	0.811	2295

Table 3: Results (Accuracy, F1) for random forest models for different POS, for predicting non/literal items, for models with and without the "weight" term **w**

len	**freq**	**err**	**condition**	**prediction**
1	0.059	0.019	**pos** $\leq$ 1.037	L
1	0.213	0.367	**imag** $>$ 0.544	L
3	0.071	0.194	**VSDSum** $\leq$ -0.469 **&** **ASDSum** $\leq$ -0.891 **&** **DRatSum** $\leq$ -0.1925	NL
3	0.225	0.396	**AMeanSum** $\leq$ -0.597 **&** **DMeanSum** $>$ -0.272 **&** **w** $\leq$ 0.121	L
2	0.012	0.477	**conc** $>$ 1.030 **&** **DMeanSum** $\leq$ -0.480	NL

Table 4: Rules extracted from random forest models for Nouns (len=length of condition, freq=proportion of data instances meeting condition, err=proportion of incorrectly classified instances as over instances meeting condition, SD=standard deviation)

ing based on a sample of the first 100 trees from each classifier built for each POS. Looking across these tables, we see some evidence of commonality for some features across different POS; for example, extensive use of the "weight" term **w** is made across POS (see Section (1) about this).[13] On the other hand, there are also quite distinct combinations of concreteness, imageability, **w**, and a small but interesting subset of the ANEW categories, across POS. For example, while nouns make good use of concreteness, imageability and **w**, combinations of im-

[13]Note also the use of the feature **pos**, which utilises the finer POS distinctions made by the Mate tools (see Section (2.2.1) about this).

len	freq	err	condition	prediction
2	0.442	0.316	**imag** $>$ -1.945 **&** **depslist_DSDSum_mean** $>$ -3.530	NL
2	0.102	0.21	**imag** $\leq$ -1.578 **&** **w** $\leq$ -0.013	L
2	0.073	0.192	**DMeanSum** $>$ 1.146 **&** **depslist_DSDSum_mean** $\leq$ -3.530	L
1	0.116	0.414	**DMeanSum** $>$ 1.274	L
2	0.218	0.4	**ARatSum** $\leq$ -0.019 **&** **DRatSum** $>$ -0.222	NL
2	0.534	0.395	**imag** $>$ -1.827 **&** **depslist_imag_mean** $>$ -0.990	NL

Table 5: Rules extracted from random forest models for Verbs (len=length of condition, freq=proportion of data instances meeting condition, err=proportion of incorrectly classified instances for instances meeting condition, SD=standard deviation, depslist=list of dependents)

len	freq	err	condition	prediction
1	0.73	0.349	**imag** $>$ -1.124	NL
2	0.413	0.313	**w** $>$ -0.0881 **&** **w** $\leq$ 0.518	NL
2	0.039	0.217	**w** $\leq$ 0.110 **&** **w** $>$ 0.066	L
1	0.53	0.336	**imag** $>$ -1.023	NL

Table 6: Rules extracted from random forest models for Prepositions (len=length of condition, freq=proportion of data instances meeting condition, err=proportion of incorrectly classified instances for instances meeting condition)

ageability and **w** are prevalent for verbs and prepositions. Further, nouns and verbs, being content words, seem to make good use of the ANEW features, but prepositions make no use of such features, perhaps due to their status as function words. More careful study of a wider range of rules, as well as possible conditioning environments is required, and such suggestions remain tentative. Note also that one complication in all of this is that there are extensive errors for most of the extracted rules, and close study of possible sources of such errors is planned for future work.

4 Discussion

In this paper, we presented results from various studies we conducted to help us refine features, and determine suitable training algorithms for our automatic metaphor detection system. The training regime included training separate classifiers for distinct POS (nouns, verbs, prepositions), and also implements suggestions from psycholinguistics, (Dellantonio et al., 2014), to model the interaction between concreteness, imageability and sentiment as dimensions of figurative meaning, in particular, distinguishing concreteness from imageability as the feature **w** (i.e. the difference between concreteness and imageability scores for individual lexical items).

Incorporating **w** led to marked improvement in our classifier performance, and we reported very competitive performance for this system: achieving an **FI** of over .81 for prepositions, and just below .80 for verbs, with nouns achieving just under .75. Finally, we have attempted to go beyond detection, toward trying to interpret the models we are using, which has led to a tentative proposal regarding function vs. content words in our approach, in terms of the features being used for classification: whereas content words such as nouns and verbs use the full range of the MRC and ANEW scores, function words like prepositions tend to use a much sparer combinations of features, such as the derived score **w** together with imageability. We are currently trying to exploit these and other insights to further improve system performance.

Acknowledgments

Thanks go to the organisers of the workshop; as well as to the anonymous reviewers who provided very helpful feedback, although, of course, we alone remain responsible for the final version. We acknowledge financial support through a Marie Curie International Incoming Fellowship (project 330569) awarded to both authors (A.G. as fellow, J.B. as P.I.).

References

John Barnden. 2015. Open-ended elaborations in creative metaphor. In *Computational Creativity Research: Towards Creative Machines*, pages 217–242. Springer.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China*.

Leo Breiman. 2002. Looking inside the black box. Technical report, Department of Statistics, California University. Wald Lecture II.

Christina Cacciari and Sam Glucksberg. 1995. Imaging idiomatic expressions: literal or figurative meanings. *Idioms: Structural and psychological perspectives*, pages 43–56.

Sara Dellantonio, Claudio Mulatti, Luigi Pastore, and Remo Job. 2014. Measuring inconsistencies can lead you forward. the case of imageability and concreteness ratings. *Language Sciences*, 5:708.

Houtao Deng. 2014. Interpreting tree ensembles with intrees. Technical report, Intuit. arXiv:1408.5456.

Lynn Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol*, 2(4):239–250.

Susan R Fussell and Mallie M Moss. 1998. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141.

Andrew Gargett, Josef Ruppenhofer, and John Barnden. 2014. Dimensions of metaphorical meaning. In Michael Zock, Reinhard Rapp, and Chu-Ren Huang, editors, *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (COGALEX)*, pages 166–173.

Raymond W Gibbs. 2006. Metaphor interpretation as embodied simulation. *Mind & Language*, 21(3):434–458.

Zoltán Kövecses. 2003. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.

Jeannette Littlemore and Graham Low. 2006. *Figurative thinking and foreign language learning*. Palgrave Macmillan Houndmills/New York.

Brian P Meier and Michael D Robinson. 2005. The metaphorical representation of affect. *Metaphor and Symbol*, 20(4):239–257.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1, pt.2):1–25.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, and T. Krennmayr. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company.

Tomek Strzalkowski, Samira Shaikh, Kit Cho, George Aaron Broadwell, Laurie Feldman, Sarah Taylor, Boris Yamrom, Ting Liu, Ignacio Cases, Yuliya Peshkova, et al. 2014. Computing affect in metaphors. *ACL 2014*, page 42.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia, June. Association for Computational Linguistics.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.

Jose Manuel Urena and Pamela Faber. 2010. Reviewing imagery in resemblance and non-resemblance metaphors. *Cognitive Linguistics*, 21(1):123–149.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Heping Zhang and Minghui Wang. 2009. Search for the smallest random forest. *Statistics and its Interface*, 2(3):381.