Multilingual Affect Polarity and Valence Prediction in Metaphor-Rich Texts

Zornitsa Kozareva
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
kozareva@isi.edu

Abstract

Metaphor is an important way of conveying the affect of people, hence understanding how people use metaphors to convey affect is important for the communication between individuals and increases cohesion if the perceived affect of the concrete example is the same for the two individuals. Therefore, building computational models that can automatically identify the affect in metaphor-rich texts like "*The team captain is a rock.*", "*Time is money.*", "*My lawyer is a shark.*" is an important challenging problem, which has been of great interest to the research community.

To solve this task, we have collected and manually annotated the affect of metaphor-rich texts for four languages. We present novel algorithms that integrate triggers for cognitive, affective, perceptual and social processes with stylistic and lexical information. By running evaluations on datasets in English, Spanish, Russian and Farsi, we show that the developed affect polarity and valence prediction technology of metaphor-rich texts is portable and works equally well for different languages.

1 Introduction

Metaphor is a figure of speech in which a word or phrase that ordinarily designates one thing is used to designate another, thus making an implicit comparison (Lakoff and Johnson, 1980; Martin, 1988; Wilks, 2007). For instance, in

"*My lawyer is a* ***shark***"

the speaker may want to communicate that his/her lawyer is strong and aggressive, and that he will attack in court and persist until the goals are achieved. By using the metaphor, the speaker actually conveys positive affect because having an aggressive lawyer is good if one is being sued.

There has been a substantial body of work on metaphor identification and interpretation (Wilks, 2007; Shutova et al., 2010). However, in this paper we focus on an equally interesting, challenging and important problem, which concerns the automatic identification of affect carried by metaphors. Building such computational models is important to understand how people use metaphors to convey affect and how affect is expressed using metaphors. The existence of such models can be also used to improve the communication between individuals and to make sure that the speakers perceived the affect of the concrete metaphor example in the same way.

The questions we address in this paper are: "*How can we build computational models that can identify the polarity and valence associated with metaphor-rich texts?*" and "*Is it possible to build such automatic models for multiple languages?*". Our main contributions are:

- We have developed multilingual metaphor-rich datasets in English, Spanish, Russian and Farsi that contain annotations of the *Positive* and *Negative* polarity and the valence (from -3 to $+3$ scale) corresponding to the intensity of the affect conveyed in the metaphor.
- We have proposed and developed automated methods for solving the polarity and valence tasks for all four languages. We model the polarity task as a classification problem, while the valence task as a regression problem.
- We have studied the influence of different information sources like the metaphor itself, the context in which it resides, the source and

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 682–691,
Sofia, Bulgaria, August 4-9 2013.

target domains of the metaphor, in addition to contextual features and trigger word lists developed by psychologists (Tausczik and Pennebaker, 2010).

- We have conducted in depth experimental evaluation and showed that the developed methods significantly outperform baseline methods.

The rest of the paper is organized as follows. Section 2 describes related work, Section 3 briefly talks about metaphors. Sections 4 and 5 describe the polarity classification and valence prediction tasks for affect of metaphor-rich texts. Both sections have information on the collected data for English, Spanish, Russian and Farsi, the conducted experiments and obtained results. Finally, we conclude in Section 6.

2 Related Work

A substantial body of work has been done on determining the affect (sentiment analysis) of texts (Kim and Hovy, 2004; Strapparava and Mihalcea, 2007; Wiebe and Cardie, 2005; Yessenalina and Cardie, 2011; Breck et al., 2007). Various tasks have been solved among which polarity and valence identification are the most common. While polarity identification aims at finding the *Positive* and *Negative* affect, valence is more challenging as it has to map the affect on a $[-3, +3]$ scale depending on its intensity (Polanyi and Zaenen, 2004; Strapparava and Mihalcea, 2007).

Over the years researchers have developed various approaches to identify polarity of words (Esuli and Sebastiani, 2006), phrases (Turney, 2002; Wilson et al., 2005), sentences (Choi and Cardie, 2009) even documents (Pang and Lee, 2008). Multiple techniques have been employed, from various machine learning classifiers, to clustering and topic models. Various domains and textual sources have been analyzed such as Twitter, Blogs, Web documents, movie and product reviews (Turney, 2002; Kennedy and Inkpen, 2005; Niu et al., 2005; Pang and Lee, 2008), but yet what is missing is affect analyzer for metaphor-rich texts.

While the affect of metaphors is well studied from its linguistic and psychological aspects (Blanchette et al., 2001; Tomlinson and Love, 2006; Crawdord, 2009), to our knowledge the building of computational models for polarity and valence identification in metaphor-rich texts is still a novel task (Smith et al., 2007; Veale, 2012; Veale and Li, 2012; Reyes and Rosso, 2012; Reyes et al., 2013). Little (almost no) effort has been put into multilingual computational affect models of metaphor-rich texts. Our research specifically targets the resolution of these problems and shows that it is possible to build such computational models. The experimental result provide valuable contributions and fundings, which could be used by the research community to build upon.

3 Metaphors

Although there are different views on metaphor in linguistics and philosophy (Black, 1962; Lakoff and Johnson, 1980; Gentner, 1983; Wilks, 2007), the common among all approaches is the idea of an interconceptual mapping that underlies the production of metaphorical expressions. There are two concepts or conceptual domains: the target (also called topic in the linguistics literature) and the source (or vehicle), and the existence of a link between them gives rise to metaphors.

The texts "*Your claims are **indefensible**.*" and "*He **attacked** every weak point in my argument.*" do not directly talk about argument as a war, however the winning or losing of arguments, the attack or defense of positions are structured by the concept of war. There is no physical battle, but there is a verbal battle and the structure of an argument (attack, defense) reflects this (Lakoff and Johnson, 1980).

As we mentioned before, there has been a lot of work on the automatic identification of metaphors (Wilks, 2007; Shutova et al., 2010) and their mapping into conceptual space (Shutova, 2010a; Shutova, 2010b), however these are beyond the scope of this paper. Instead we focus on an equally interesting, challenging and important problem, which concerns the automatic identification of affect carried by metaphors. To conduct our study, we use human annotators to collect metaphor-rich texts (Shutova and Teufel, 2010) and tag each metaphor with its corresponding polarity (*Positive/Negative*) and valence $[-3, +3]$ scores. The next sections describe the affect polarity and valence tasks we have defined, the collected and annotated metaphor-rich data for each one of the English, Spanish, Russian and Farsi languages, the conducted experiments and obtained results.

4 Task A: Polarity Classification

4.1 Problem Formulation

Task Definition: Given metaphor-rich texts annotated with *Positive* and *Negative* polarity labels, the goal is to build an automated computational affect model, which can assign to previously unseen metaphors one of the two polarity classes.

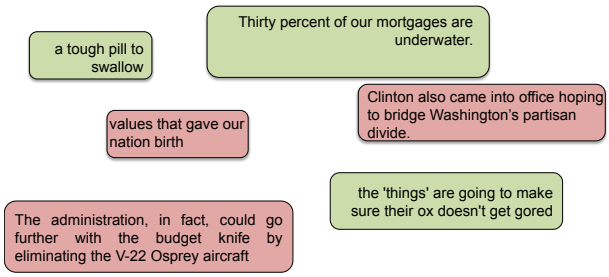

Figure 1: Polarity Classification

Figure 1 illustrates the polarity task in which the metaphors were classified into *Positive* or *Negative*. For instance, the metaphor "*tough pill to swallow*" has *Negative* polarity as it stands for something being hard to digest or comprehend, while the metaphor "*values that gave our nation birth*" has a *Positive* polarity as giving birth is like starting a new beginning.

4.2 Classification Algorithms

We model the metaphor polarity task as a classification problem in which, for a given collection of N training examples, where m_i is a metaphor and c_i is the polarity of m_i, the objective is to learn a classification function $f : m_i \rightarrow c_i$ in which 1 stands for positive polarity and 0 stands for negative polarity. We tested five different machine learning algorithms such as Nave Bayes, SVM with polynomial kernel, SVM with RBF kernel, AdaBoost and Stacking, out of which AdaBoost performed the best. In our experimental study, we use the freely available implementations in Weka (Witten and Frank, 2005).

Evaluation Measures: To evaluate the goodness of the polarity classification algorithms, we calculate the f-score and accuracy on 10-fold cross validation.

4.3 Data Annotation

To conduct our experimental study, we have used annotated data provided by the Language Computer Corporation (LCC)[1], which developed annotation toolkit specifically for the task of metaphor detection, interpretation and affect assignment. They hired annotators to collect and annotate data for the English, Spanish, Russian and Farsi languages. The domain for which the metaphors were collected was *Governance*. It encompasses electoral politics, the setting of economic policy, the creation, application and enforcement of rules and laws. The metaphors were collected from political speeches, political websites, online newspapers among others (Mohler et al., 2013).

The annotation toolkit allowed annotators to provide for each metaphor the following information: the metaphor, the context in which the metaphor was found, the meaning of the metaphor in the source and target domains from the perspective of a native speaker. For example, in the **Context**: *And to all nations, we will speak for the values that gave our nation birth.*; the annotators tagged the **Metaphor:** *values that gave our nation birth*; and listed as **Source:** *mother gave birth to baby*; and **Target:** *values of freedom and equality motivated the creation of America*. The same annotators also provided the affect associated with the metaphor. The agreements of the annotators as measured by LCC are: .83, .87, .80 and .61 for the English, Spanish, Russian and Farsi languages.

In our study, the maximum length of a metaphor is a sentence, but typically it has the span of a phrase. The maximum length of a context is three sentences before and after the metaphor, but typically it has the span of one sentence before and after. In our study, the source and target domains are provided by the human annotators who agree on these definitions, however the source and target can be also automatically generated by an interpretation system or a concept mapper. The generation of source and target information is beyond the scope of this paper, but studying their impact on affect is important. At the same time, we want to show that if the technology for source/target detection and interpretation is not yet available, then how far can one reach by using the metaphor itself and the context around it. Later depending on the availability of the information sources and toolkits one can decide whether to integrate such information or to ignore it. In the experimental sections, we show how the individual information sources and their combination affects the resolution of the metaphor polarity and valence prediction tasks.

Table 1 shows the positive and negative class

[1] http://www.languagecomputer.com/

distribution for each one of the four languages.

	Negative	Positive
ENGLISH	2086	1443
SPANISH	196	434
RUSSIAN	468	418
FARSI	384	252

Table 1: Polarity Class Distribution for Four Languages

The majority of the the annotated examples are for English. However, given the difficulty of finding bilingual speakers, we still managed to collect around 600 examples for Spanish and Farsi, and 886 examples for Russian.

4.4 N-gram Evaluation and Results

N-gram features are widely used in a variety of classification tasks, therefore we also use them in our polarity classification task. We studied the influence of unigrams, bigrams and a combination of the two, and saw that the best performing feature set consists of the combination of unigrams and bigrams. In this paper, we will refer from now on to n-grams as the combination of unigrams and bigrams.

Figure 2 shows a study of the influence of the different information sources and their combination with n-gram features for English.

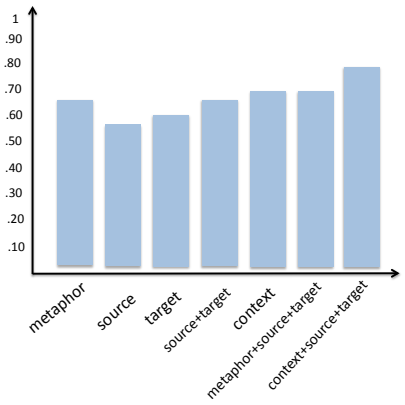

Figure 2: Influence of Information Sources for Metaphor Polarity Classification of English Texts

For each information source (metaphor, context, source, target and their combinations), we built a separate n-gram feature set and model, which was evaluated on 10-fold cross validation. The results from this study show that for English, the more information sources one combines, the higher the classification accuracy becomes.

Table 2 shows the influence of the information sources for Spanish, Russian and Farsi with the n-gram features. The best f-scores for each language are shown in bold. For Farsi and Russian high performances are obtained both with the context and with the combination of the context, source and target information. While for Spanish they reach similar performance.

	SPANISH	RUSSIAN	FARSI
Metaphor	71.6	71.0	62.4
Source	67.1	62.4	55.4
Target	68.9	67.2	62.4
Context	73.5	**77.1**	**67.4**
S+T	**76.6**	68.7	62.4
M+S+T	76.0	75.4	64.2
C+S+T	**76.5**	**76.5**	**68.4**

Table 2: N-gram features, F-scores on 10-fold validation for Spanish, Russian and Farsi

4.5 LIWC as a Proxy for Metaphor Polarity

LIWC Repository: In addition to the n-gram features, we also used the Linguistic Inquiry and Word Count (LIWC) repository (Tausczik and Pennebaker, 2010), which has 64 word categories corresponding to different classes like emotional states, psychological processes, personal concerns among other. Each category contains a list of words characterizing it. For instance, the LIWC category *discrepancy* contains words like *should*, *could* among others, while the LIWC category *inhibition* contains words like *block*, *stop*, *constrain*. Previously LIWC was successfully used to analyze the emotional state of bloggers and tweeters (Quercia et al., 2011) and to identify deception and sarcasm in texts (Ott et al., 2011; González-Ibáñez et al., 2011). When LIWC analyzes texts it generates statistics like number of words found in category C_i divided by the total number of words in the text. For our metaphor polarity task, we use LIWC's statistics of all 64 categories and feed this information as features for the machine learning classifiers. LIWC repository contains conceptual categories (dictionaries) both for the English and Spanish languages.

LIWC Evaluation and Results: In our experiments LIWC is applied to English and Spanish metaphor-rich texts since the LIWC category dictionaries are available for both languages. Table 3 shows the obtained accuracy and f-score results in English and Spanish for each one of the information sources.

	ENGLISH		SPANISH	
	Acc	Fscore	Acc	Fscore
Metaphor	**98.8**	**98.8**	87.9	87.2
Source	**98.6**	**98.6**	**97.3**	**97.3**
Target	**98.2**	**98.2**	**97.9**	**97.9**
Context	91.4	91.4	**93.3**	**93.2**
S+T	98.0	98.0	76.3	75.5
M+S+T	95.8	95.7	86.8	86.0
C+S+T	87.9	88.0	79.2	78.5

Table 3: LIWC features, Accuracy and F-scores on 10-fold validation for English and Spanish

The best performances are reached with individual information sources like metaphor, context, source or target instead of their combinations. The classifiers obtain similar performance for both languages.

LIWC Category Relevance to Metaphor Polarity: We also study the importance and relevance of the LIWC categories for the metaphor polarity task. We use information gain (IG) to measure the amount of information in bits about the polarity class prediction, if the only information available is the presence of a given LIWC category (feature) and the corresponding polarity class distribution. IG measures the expected reduction in entropy (uncertainty associated with a random feature) (Mitchell, 1997).

Figure 3 illustrates how certain categories occur more with the positive (in red color) vs negative (in green color) class. With the positive metaphors we observe the LIWC categories for present tense, social, affect and family, while for the negative metaphors we see LIWC categories for past tense, inhibition and anger.

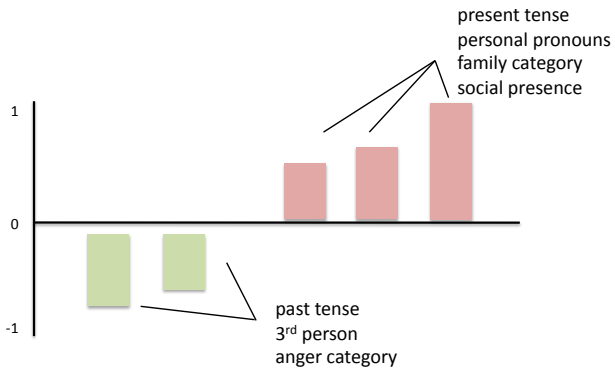

Figure 3: LIWC category relevance to Metaphor Polarity

In addition, we show in Figure 4 examples of the top LIWC categories according to IG ranking for each one of the information sources.

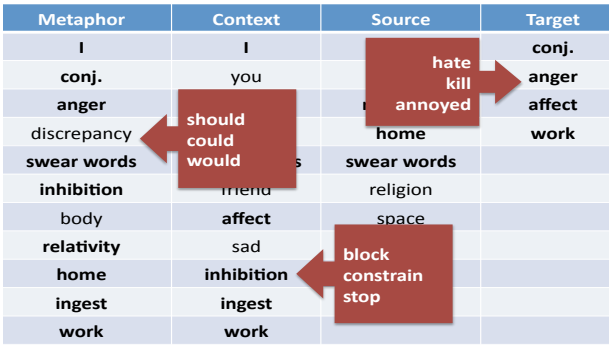

Figure 4: Example of LIWC Categories and Words

For metaphor texts, these categories are *I*, *conjuntion*, *anger*, *discrepancy*, *swear words* among others; for contexts the categories are pronouns like *I*, *you*, *past tense*, *friends*, *affect* and so on. Our study shows that some of the LIWC categories are important across all information sources, but overall different triggers activate depending on the information source and the length of the text used.

4.6 Comparative study

Figure 5 shows a comparison of the accuracy of our best performing approach for each language. For English and Spanish these are the LIWC models, while for Russian and Farsi these are the n-gram models. We compare the performance of the algorithms with a majority baseline, which assigns the majority class to each example. For instance, in English there are 3529 annotated examples, of which 2086 are positive and 1443 are negative. Since the positive class is the predominant one for this language and dataset, a majority classifier would have .59 accuracy in returning the positive class as an answer. Similarly, we compute the majority baseline for the rest of the languages.

	Accuracy	Majority Baseline	Difference
English	98.80	59.11	+39.69
Spanish	97.90	68.88	+29.02
Russian	77.00	52.82	+24.18
Farsi	72.20	60.30	+11.90

Figure 5: Best Accuracy Model and Comparison against a Majority Baseline for Metaphor Polarity Classification

As we can see from Figure 5 that all classifiers significantly outperform the majority base-

line. For Farsi the increment is $+11.90$, while for English the increment is $+39.69$. This means that the built classifiers perform much better than a random classifier.

4.7 Lessons Learned

To summarize, in this section we have defined the task of polarity classification and we have presented a machine learning solution. We have used different feature sets and information sources to solve the task. We have conducted exhaustive evaluations for four different languages namely English, Spanish, Russian and Farsi. The learned lessons from this study are: (1) for n-gram usage, the larger the context of the metaphor, the better the classification accuracy becomes; (2) if present source and target information can further boost the performance of the classifiers; (3) LIWC is a useful resource for polarity identification in metaphor-rich texts; (4) analyzing the usages of tense like past vs. present and pronouns are important triggers for positive and negative polarity of metaphors; (5) some categories like *family*, *social presence* indicate positive polarity, while others like *inhibition*, *anger* and *swear words* are indicative of negative affect; (6) the built models significantly outperform majority baselines.

5 Task B: Valence Prediction

5.1 Problem Formulation

Task Definition: Given metaphor-rich texts annotated with valence score (from -3 to $+3$), where -3 indicates strong negativity, $+3$ indicates strong positivity, 0 indicates neural, the goal is to build a model that can predict without human supervision the valence scores of new previously unseen metaphors.

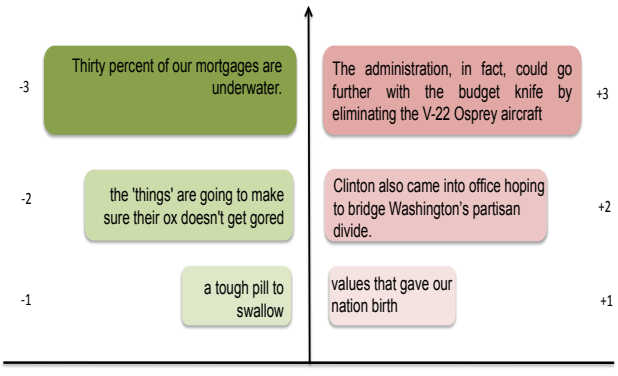

Figure 6: Valence Prediction

Figure 6 shows an example of the valence prediction task in which the metaphor-rich texts must be arranged by the intensity of the emotional state provoked by the texts. For instance, -3 corresponds to very strong negativity, -2 strong negativity, -1 weak negativity (similarly for the positive classes). In this task we also consider metaphors with neutral affect. They are annotated with the 0 label and the prediction model should be able to predict such intensity as well. For instance, the metaphor "*values that gave our nation birth*", is considered by American people that giving birth sets new beginning and has a positive score $+1$, but "*budget knife*" is more positive $+3$ since tax cut is more important. As any sentiment analysis task, affect assignment of metaphors is also a subjective task and the produced annotations express the values, believes and understanding of the annotators.

5.2 Regression Model

We model the valence task a regression problem, in which for a given metaphor m, we seek to predict the valence v of m. We do this via a parametrized function $f{:}\hat{v} = f(m; w)$, where $w \in R^d$ are the weights. The objective is to learn w from a collection of N training examples $\{< m_i, v_i >\}_{i=1}^{N}$, where m_i are the metaphor examples and $v_i \in R$ is the valence score of m_i.

Support vector regression (Drucker et al., 1996) is a well-known method for training a regression model by solving the following optimization problem:

$$\min_{w \in R^s} \frac{1}{2}||w||^2 + \frac{C}{N}\sum_{i=1}^{N} \underbrace{\max(0, |v_i - f(m_i; w)| - \epsilon)}_{\epsilon\text{-insensitive loss function}}$$

where C is a regularization constant and ϵ controls the training error. The training algorithm finds weights w that define a function f minimizing the empirical risk. Let h be a function from seeds into some vector-space representation $\subseteq R^d$, then the function f takes the form: $f(m; w) = h(m)^T w = \sum_{i=1}^{N} \alpha_i K(m, m_i)$, where f is re-parameterized in terms of a polynomial kernel function K with dual weights α_i. K measures the similarity between two metaphoric texts. Full details of the regression model and its implementation are beyond the scope of this paper; for more details see (Schölkopf and Smola, 2001; Smola et al., 2003). In our experimental study, we use the freely available implementation of SVM in Weka (Witten and Frank, 2005).

Evaluation Measures: To evaluate the quality of the valence prediction model, we compare the actual valence score of the metaphor given by human annotators denoted with y against those valence scores predicted by the regression model denoted with x. We estimate the goodness of the regression model calculating both the correlation coefficient $cc_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$ and the mean squared error $mse_{x,y} = \frac{\sum_{i=i}^{n}(x-\hat{x})}{n}$. The two evaluation measures should be interpreted in the following manner. Intuitively the higher the correlation score is, the better the correlation between the actual and the predicted valence scores will be. Similarly the smaller the mean squared error rate, the better the regression model fits the valence predictions to the actual score.

5.3 Data Annotation

To conduct our valence prediction study, we used the same human annotators from the polarity classification task for each one of the English, Spanish, Russian and Farsi languages. We asked the annotators to map each metaphor on a $[-3, +3]$ scale depending on the intensity of the affect associated with the metaphor.

Table 4 shows the distribution (number of examples) for each valence class and for each language.

	-3	-2	-1	0	+1	+2	+3
ENGLISH	1057	817	212	582	157	746	540
SPANISH	106	65	27	17	40	132	262
RUSSIAN	118	42	308	13	202	149	67
FARSI	147	117	120	49	91	63	98

Table 4: Valence Score Distribution for Each Language

5.4 Empirical Evaluation and Results

For each language and information source we built separate valence prediction regression models. We used the same features for the regression task as we have used in the classification task. Those include n-grams (unigrams, bigrams and combination of the two), LIWC scores. Table 5 shows the obtained correlation coefficient (CC) and mean squared error (MSE) results for each one of the four languages (English, Spanish, Russian and Farsi) using the dataset described in Table 4.

The Farsi and Russian regression models are based only on n-gram features, while the English and Spanish regression models have both n-gram and LIWC features. Overall, the CC for English and Spanish is higher when LIWC features are used. This means that the LIWC based valence regression model approximates the predicted values better to those of the human annotators. The better valence prediction happens when the metaphor itself is used by LIWC. The MSE for English and Spanish is the lowest, meaning that the prediction is the closest to those of the human annotators. In Russian and Farsi the lowest MSE is when the combined metaphor, source and target information sources are used. For English and Spanish the smallest MSE or so called prediction error is 1.52 and 1.30 respectively, while for Russian and Farsi is 1.62 and 2.13 respectively.

5.5 Lessons Learned

To summarize, in this section we have defined the task of valence prediction of metaphor-rich texts and we have described a regression model for its solution. We have studied different feature sets and information sources to solve the task. We have conducted exhaustive evaluations in all four languages namely English, Spanish, Russian and Farsi. The learned lessons from this study are: (1) valence prediction is a much harder task than polarity classification both for human annotation and for the machine learning algorithms; (2) the obtained results showed that despite its difficulty this is still a plausible problem; (3) similarly to the polarity classification task, valence prediction with LIWC is improved when shorter contexts (the metaphor/source/target information source) are considered.

6 Conclusion

People use metaphor-rich language to express affect and often affect is expressed through the usage of metaphors. Therefore, understanding that the metaphor "*I was **boiling inside** when I saw him.*" has *Negative* polarity as it conveys feeling of anger is very important for interpersonal or multicultural communications.

In this paper, we have introduced a novel corpus of metaphor-rich texts for the English, Spanish, Russian and Farsi languages, which was manually annotated with the polarity and valence scores of the affect conveyed by the metaphors. We have studied the impact of different information sources such as the metaphor in isolation, the context in which the metaphor was used, the source and target domain meanings of the metaphor and

	RUSSIAN N-gram		FARSI N-gram		ENGLISH N-gram		SPANISH N-gram		ENGLISH LIWC		SPANISH LIWC	
	CC	MSE	CC	MSE	CC	MSE	CC	MSE	CC	MSE	CC	MSE
Metaphor	.45	1.71	.25	2.25	.36	2.50	.37	2.54	.74	1.52	.87	1.20
Source	.22	1.89	.11	2.42	.40	2.27	.22	2.43	.81	1.30	.85	1.28
Target	.25	1.91	.15	2.47	.37	2.41	.32	2.36	.72	1.56	.85	1.29
Context	.43	1.83	.32	2.38	.37	2.59	.40	2.37	.40	2.16	.67	1.92
S+T	.29	1.83	.18	2.38	.40	2.40	.41	2.19	.70	1.60	.78	1.53
M+S+T	.45	1.62	.29	2.13	.43	2.34	.43	2.14	.67	1.67	.78	1.53
C+S+T	.42	1.85	.26	2.61	.43	2.52	.39	2.41	.44	2.08	.64	1.96

Table 5: Valence Prediction, Correlation Coefficient and Mean Squared Error for English, Spanish, Russian and Farsi

their combination in order to understand how such information helps and impacts the interpretation of the affect associated with the metaphor. We have conducted exhaustive evaluation with multiple machine learning classifiers and different features sets spanning from lexical information to psychological categories developed by (Tausczik and Pennebaker, 2010). Through experiments carried out on the developed datasets, we showed that the proposed polarity classification and valence regression models significantly improve baselines (from 11.90% to 39.69% depending on the language) and work well for all four languages. From the two tasks, the valence prediction problem was more challenging both for the human annotators and the automated system. The mean squared error in valence prediction in the range $[-3, +3]$, where -3 indicates strong negative and $+3$ indicates strong positive affect for English, Spanish and Russian was around 1.5, while for Farsi was around 2.

The current findings and learned lessons reflect the properties of the collected data and its annotations. In the future we are interested in studying the affect of metaphors for domains different than *Governance*. We want to conduct studies with the help of social sciences who would research whether the tagging of affect in metaphors depends on the political affiliation, age, gender or culture of the annotators. Not on a last place, we would like to improve the built valence prediction models and to collect more data for Spanish, Russian and Farsi.

Acknowledgments

The author would like to thank the reviewers for their helpful comments as well as the LCC annotators who have prepared the data and made this work possible. This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0025. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

Max Black. 1962. *Models and Metaphors*.

Isabelle Blanchette, Kevin Dunbar, John Hummel, and Richard Marsh. 2001. Analogy use in naturalistic settings: The influence of audience, emotion and goals. *Memory and Cognition*, pages 730–735.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 2683–2688. Morgan Kaufmann Publishers Inc.

Yejin Choi and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 590–598.

Elizabeth Crawdord. 2009. Conceptual metaphors of affect. *Emotion Review*, pages 129–139.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. Support vector regression machines. In *Advances in NIPS*, pages 155–161.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, pages 417–422.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Roberto González-Ibáñez, Smaranda Muresa n, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 581–586.

Alistair Kennedy and Diana Inkpen. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, pages 110–125.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04.

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

James H. Martin. 1988. Representing regularities in the metaphoric lexicon. In *Proceedings of the 12th conference on Computational linguistics - Volume 1*, COLING '88, pages 396–401.

Thomas M. Mitchell. 1997. *Machine Learning*. McGraw-Hill, Inc., 1 edition.

Michael Mohler, David Bracewell, David Hinote, and Marc Tomlinson. 2013. Semantic signatures for example-based linguistic metaphor detection. In *The Proceedings of the First Workshop on Metaphor in NLP, (NAACL)*, pages 46–54.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *In: Proceedings of the American Medical Informatics Association 2005 Annual Symposium*, pages 570–574.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Livia Polanyi and Annie Zaenen. 2004. Contextual lexical valence shifters. In Yan Qu, James Shanahan, and Janyce Wiebe, editors, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. AAAI Press. AAAI technical report SS-04-07.

Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2011. In the mood for being influential on twitter. In *the 3rd IEEE International Conference on Social Computing*.

Antonio Reyes and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decis. Support Syst.*, 53(4):754–760, November.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268, March.

Bernhard Schölkopf and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. In *International Conference on Language Resources and Evaluation*.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1002–1010.

Ekaterina Shutova. 2010a. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 1029–1037.

Ekaterina Shutova. 2010b. Models of metaphor in nlp. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 688–697.

Catherine Smith, Tim Rumbell, John Barnden, Bob Hendley, Mark Lee, and Alan Wallington. 2007. Don't worry about metaphor: affect extraction for conversational agents. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 37–40. Association for Computational Linguistics.

Alex J. Smola, Bernhard Schlkopf, and Bernhard Sch Olkopf. 2003. A tutorial on support vector regression. Technical report, Statistics and Computing.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74. Association for Computational Linguistics, June.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, March.

Marc T. Tomlinson and Bradley C. Love. 2006. From pigeons to humans: grounding relational learning in concrete examples. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, AAAI'06, pages 199–204. AAAI Press.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424.

Tony Veale and Guofu Li. 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, pages 7–12.

Tony Veale. 2012. A context-sensitive, multi-faceted model of lexico-conceptual affect. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 75–79.

Janyce Wiebe and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation (formerly Computers and the Humanities*.

Yorick Wilks. 2007. A preferential, pattern-seeking, semantics for natural language inference. In *Words and Intelligence I*, volume 35 of *Text, Speech and Language Technology*, pages 83–102. Springer Netherlands.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition.

Ainur Yessenalina and Claire Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 172–182.