Supervised Metaphor Detection using Conditional Random Fields

Sunny Rai*, Shampa Chakraverty*, and Devendra K. Tayal**
*Dept. of CoE, Netaji Subhas Institute of Technology, University of Delhi, India
**Dept. of CSE, Indira Gandhi Delhi Technical University for Women, India
{post2srai, apmahs.nsit}@gmail.com, dev_tayal2001@yahoo.com

Abstract

In this paper, we propose a novel approach for supervised classification of linguistic metaphors in an open domain text using Conditional Random Fields (CRF). We analyze CRF based classification model for metaphor detection using syntactic, conceptual, affective, and word embeddings based features which are extracted from MRC Psycholinguistic Database (MRCPD) and WordNet-Affect. We use word embeddings given by Huang *et al.* to capture information such as coherence and analogy between words. To tackle the bottleneck of limited coverage of psychological features in MRCPD, we employ synonymy relations from WordNet®. A comparison of our approach with previous approaches shows the efficacy of CRF classifier in detecting metaphors. The experiments conducted on VU Amsterdam metaphor corpus provides an accuracy of more than 92% and F-measure of approximately 78%. Results shows that inclusion of conceptual features improves the recall by 5% whereas affective features do not have any major impact on metaphor detection in open text.

1 Introduction

According to Merriam-Webster dictionary, *Metaphor*[1] is defined as "a figure of speech in which a word or a phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them". As a literary tool, metaphor is popular and widely appreciated for the novelty and vividness it provides to an otherwise bland text. Contrary to the popular belief that metaphor is a tool for poetry, we see vast usage of metaphors in common parlance and even in scientific discourses like economics, security and politics. To achieve human computer interaction in the true sense, Natural Language Processing (NLP) systems need to be enhanced to process and understand metaphors and indeed, all kinds of figurative speeches in a continuous and open-domain text.

Interpreting metaphor in a language is a complicated task because the metaphorical meaning of a sentence depends on what a speaker actually wants to convey irrespective of the literal meaning of the words used (Searle, 1985). Thus, the literal meaning of words needs to be modified according to the contextual presentation of the metaphor (Gibbs, 1984). A linguistic metaphor indicates a domain that seems incongruous with the surrounding context but presents the underlying belief by re-conceptualization (Lynne and Deignan, 2003). Interestingly, we observe that unlike the process of metaphor interpretation, a linguistic metaphor can be detected by analyzing word properties and surrounding words i.e. context, to identify the incongruity it causes in the sentence without even examining the reconceptualization involved.

One of the earliest criteria for metaphor detection in a text was proposed by Wilks (1978). He stated that a metaphorical interpretation is made when literal interpretation makes no sense or is out of context, a condition known as violation of selectional preference, e.g.

My car drinks gasoline. (Wilks, 1978) *(a)*

However, metaphors occur even without violation of selectional preference, e.g.

All men are animals. (b)

[1] Merriam-Webster Dictionary, Metaphor: http://www.merriam-webster.com/dictionary/metaphor

Proceedings of The Fourth Workshop on Metaphor in NLP, pages 18–27,
San Diego, CA, 17 June 2016. ©2016 Association for Computational Linguistics

We also observe that combinations such as <*car, drink*> and <*drink, gasoline*> have very low co-occurrence frequency in a corpus. Lakoff and Johnson (1980) introduced another perspective and termed it as *conceptual metaphor* which regards metaphor as a product of cognitive phenomenon, and a way to explain abstract concepts such as *argument* by mapping them to concrete ones such as *war*, e.g.

The committee shot down her ideas one by one. (c)

Here in (c), the concrete concept, shot is used to explain the outright rejection of an abstract concept namely, ideas. Thus, conceptual features like abstractness (Turney *et al.*, 2011) or concreteness (Klebanov *et al.*, 2015), familiarity or rarity (Schulder and Hovy, 2014), imageability (Strzalkowski *et al.*, 2013) and sensory features (Gargett and Barnden, 2015) can improve automatic metaphor detection. Recently, Hovy *et al.* (2013) utilized word embeddings by Collobert *et al.* (2011) for capturing coherence and contextual features for supervised metaphor detection.

In this paper, we strive to provide a solution to the fundamental problem of metaphor processing *i.e. metaphor detection in text*. We propose a novel approach for binary classification of continuous, open-domain text into Metaphor or otherwise, using Conditional Random Fields (CRF) (Lafferty *et al.*, 2001) and a hybrid feature set. CRF has widespread applications in computational linguistics. In contrast with (Gedigan *et al.*, 2006) and (Turney *et al.*, 2011) which focus on verb and adjective centric metaphors, we extend the applicability of CRF to detect all kinds of metaphor. The idea is to develop an effective and computationally inexpensive process to filter out metaphors in an open text.

Unlike the majority of earlier works which focused on utilizing a particular type of feature set, we propose a rich and hybrid feature set to cover different aspects of a metaphor. We include features that pertain to sentence structure, psychological/conceptual properties of a word, affective interpretation which provides information on whether a word conveys a mood, a mental state or an emotion and the context of a word's usage using word embeddings. We analyze their effect on metaphor detection individually and in combination. We use Medical Research Council Psycholinguistic Database (MRCPD) (Wilson, 1988) for extracting conceptual features and WordNet-Affect (Strapparava and Valitutti, 2004) for extracting affective features. Further, we employ WordNet® (Fellbaum, 1998) to expand the limited coverage of psychological features in MRCPD. We utilize word embeddings (Huang *et al.*, 2012) as contextual features which provides semantic information for a word such as coherence, context, relatedness and analogy.

The paper is organized as follows. Section 2 provides an overview of existing literature on detection of metaphor. Section 3 delineates our proposed approach for metaphor detection, the feature sets used and extension of MRCPD by using WordNet®. We demonstrate our experiments and results in Section 4 and conclude our work in Section 5.

2 Related Work

Some of the earliest works in metaphor detection used concept of *violation of selectional preference* given by Wilks (1978). These include the system MIDAS (Martin, 1990) and met* (Fass, 1991), both of which are hand-coded rule-based systems. The work in (Mason, 2004) introduced a corpus based approach *viz.* CorMet to identify conventional non-literal phrases from a domain specific text. Recently, Baumer *et al.* (2010) proposed an approach based on selectional preference given by Resnik (1993) to detect conceptual metaphors.

Several researchers used *clustering based techniques* to identify non-literalness in text. Birke and Sarkar (2006) proposed the Trope Finder (TroFi) system to recognize verbs with non-literal meaning using Word Sense Disambiguation (WSD) and clustering. However, TroFi doesn't identify the type of non-literalness *i.e.* whether it is a metaphor or some other non-literal category such as idiom, sarcasm etc. Shutova *et al.* (2010) pointed out that a target concept associated with the same source concept are more likely to co-occur in similar lexico-syntactic environments. They proposed a clustering based technique using grammatical relations and lexical features to detect metaphors. In (Birke and Sarkar, 2007), the authors introduced the concept of Active Learning using an existing similarity based WSD algorithm (Karov and Edelman, 1998) to annotate a corpus for non-literal language.

Some researchers used *knowledge resources* to strengthen the approaches for metaphor detection. The work in (Gedigian *et al.*, 2006) proposed a technique to detect metaphorical verb usage by utilizing existing knowledge sources such as WordNet® (Fellbaum, 1998), FrameNet (Baker et al., 1998) and PropBank (Kingsbury and Palmer, 2002). Krishnakumaran and Zhu (2007) proposed an approach on the basis of violations in co-occurring noun and verb phrases using WordNet® hierarchies. The works in (Mohler *et al.*, 2013) and (Wilks *et al.*, 2013) also utilized WordNet® ontology for recognizing metaphors. Dunn (2013) proposed a domain interaction system namely Measuring and Identifying Metaphor in Language (MIMIL) using SUMO ontology (Niles and Pease, 2011). The default sense of lexical items in a sentence were mapped to concepts from SUMO ontology on the basis of domain type and event status. Tsvetkov *et al.* (2014) and Bracewell et al. (2014) used the concept of hybrid feature set by using features from WordNet, MRCPD and vector representations. Klebanov *et al.* (2015) evaluated the effect of concreteness as a feature for metaphor detection using MRCPD. Gargett and Branden (2015) proposed using sensory features to enhance the process of metaphor detection using Affective Norms for English Word (ANEW) and MRCPD.

Ferrari (1996) proposed an approach based on the analysis of syntactic patterns and used textual cues such as *metaphorically, figuratively* and *like* to detect metaphors. The works in (Sardinha, 2002; Sardinha, 2006) suggested using corpus-based metrics such as collocations list, word frequency and semantic-distance for metaphor detection. Turney *et al.* (2011) observed that metaphorical usage is directly correlated with the degree of abstractness in a word's contextual usage. They proposed an approach to identify metaphorical verbs and adjectives using an algorithm to calculate abstractness given in (Turney and Littman, 2003). Hovy *et al.* (2013) proposed an approach based on features derived from syntactic patterns and word embeddings (Collobert *et al.*, 2011) to analyze the context of a word usage. They used Support Vector Machine (SVM) classifier with tree kernels to identify metaphorical words.

Strazalkowski *et al.* (2013) utilized topic chaining and imageability score to identify metaphorical usage in text. Neuman *et al.* (2013) developed a rule based approach which combines selectional preference and abstractness of words to identify metaphors. Schulder and Hovy (2014) proposed a statistical domain-independent approach based on a metric term relevance derived from Term Frequency-Inverse Document Frequency (TF-IDF). A term is categorized as metaphor if the domain relevance and common relevance for a given term is lower than a predetermined threshold.

Our work is different from earlier works in terms of improved and rich feature set that we employ. Unlike previous works which used a subset of conceptual features, we utilize a multifarious feature set. Along with that, we introduce different affective features which deal with abstract concepts like mental state, physical state, and behavioral characteristics. We systemically extend the coverage of MRCPD to include unavailable terms rather than taking average values as in (Gargett and Branden, 2015). To the best of our knowledge, CRF for metaphor classification has not yet been fully investigated, especially in the context of large, and overlapping feature set. Hovy *et al.* (2013) compared their proposed approach with CRF. However they did not emphasize on the feature set and relations which can be exploited while using a CRF model.

3 Proposed Methodology

Our methodology is based on an innovative application of CRF on an amalgamation of syntactic, conceptual, affective and word embeddings based features extracted by using corpus and knowledge resources. We now give a detailed description of the features that we have employed, our technique for expanding the coverage of MRCPD using WordNet® and an approach to build a CRF model for *Metaphor Detection*.

3.1 Feature Set

We have built a hybrid feature set drawing upon traditional features as well as on the basis of our inferences from the existing literature that are likely to enhance the process of metaphor detection. The feature set is categorized into four categories namely, Syntactic Feature set, F_S, Conceptual feature set, F_C, Affective feature set, F_A and Contextual feature set, F_X.

3.1.1 Syntactic Features

The set of syntactic features (SF) is defined as a feature vector F_S= {$F_{S_1}, F_{S_2}, \ldots, F_{S_5}$}, based on a word attributes, sentence structure and dependency between components of a sentence. We use Stanford CoreNLP (Manning et al., 2014) to extract following syntactic features:

1. *Lemma*: It includes lemmas i.e. the base form, for all tokens in the corpus.
2. *PoS*: It denotes grammatical word-category such as noun, verb etc. for each word in a sentence.
3. *Named_Entity_Type*: It marks named, numerical and temporal entities such as *person, location, money, number, date, time etc*.
4. *Dependency*: It provides information about the position of a word with respect to grammatical structure of a sentence including relations between words and modifiers.
5. *Stop_word*: It marks a word as *content* or *non-content* word, based on a list of stop-words.

3.1.2 Conceptual Features

The set of conceptual features (CF) is defined as a feature vector, F_C= {$F_{C_1}, F_{C_2}, \ldots, F_{C_5}$}, which consists of features based on the meaning, abstract attributes of a word and its measured impact on a large corpus. We use MRCPD (Wilson, 1988), a large psycholinguistics database to extract the conceptual features of a word enumerated below (Gilhooly and Logie, 1980; Paivio and Madigan, 1968; Toglia and Battig, 1978). These conceptual features have numerical values ranging from 100 to 700.

1. *Concreteness*: It is a measurement of the ability of a word to refer to concrete concepts.
2. *Familiarity*: It refers to a feeling of knowing the word or concept behind a word, depending upon how commonly the word is used.
3. *Imageability*: It refers to the expressiveness of a word to evoke a visual image of the concept behind the word.
4. *Frequency*: It refers to the frequency of occurrence of a word sense in the Brown corpus.
5. *Meaningfulness*: It is a measure of the association of a given word with other words. We utilize the Colerado based *Meaningfulness* rating, MEANC.

3.1.3 Affective Features

The set of affective features (AF) is defined as a feature vector, F_A= {$F_{A_1}, F_{A_2}, \ldots, F_{A_5}$}, based on the affective concepts correlated with an affective word. We use WordNet Affect (Strapparava and Valitutti, 2004) to extract following affective features:

1. *Cognitive_State*: It denotes mental state or feelings such as confusion. Labels *mood, sensation, emotional response* are merged with *cognitive state* label because of very few instances and indirect dependence.
2. *Physical_State*: It refers to physical or bodily state such as illness.
3. *Trait*: It denotes characteristics of personality such as aggressiveness.
4. *Attitude*: The label *behavior* and *attitude* are merged together as they are interrelated.
5. *Emotion*: It refers to emotional state or process such as joy, anger.

3.1.4 Contextual Features

Word Embeddings provide semantic information about a word and are capable of comparing words on basis of relatedness, analogy, coherence and context, the criteria we use for metaphor detection as well. We use word embeddings by Huang et al. (2012) as it captures local i.e. syntactic information as well global context based on word's usage in large corpus. CWE is trained using local context only. Word embeddings are used as contextual features (XF) and represented as a 50 dimensional vector, F_X= {$F_{X_1}, F_{X_2}, \ldots, F_{X_{50}}$}.

3.2 Extending MRCPD using WordNet

The MRCPD resource contains linguistic information for over 150,000 words. However, psycholinguistic features are available only for a limited number of words. A total of 8228 words are available for Concreteness rating, 9392 for Familiarity, 9240 for Imageability and 5450 words for Colerado norms based meaningfulness rating (Wilson, 1988). As a result, several words which are used frequently is not available in the database. Liu et al. (2014) proposed extending MRCPD for imageability rating using synonymy and hyponymy relations from WordNet. We strive to use a rich feature set comprising of other features such as concreteness, familiarity and meaningfulness, thus expanding MRCPD for all of these features.

WordNet®(Princeton University, 2010) is an online machine readable lexical database for English language developed by Christiane Fellbaum at Princeton University. In WordNet®, words are grouped on the basis of cognitive concepts known as *synset*, a set of synonymous words. Senses in a synset are arranged in a form of a list on the basis of decreasing frequency count in the Brown corpus (Francis and Kucera, 1979) For example, if we search the word, *man,* we observe that its first sense <*man#1, adult-male#1*> sense has occurred 749 times in the Brown corpus whereas its second sense <*homo#2*> occurred only 29 times. So, we can say, that highest ranked sense is the most frequently used sense for a word. We use this inference to prioritize senses in our technique to extend the coverage of words with psychological features in MRCPD. The hyponymy relation provide a more specific instance for a concept. E.g. for the word, beverage, we obtain {*milk, tea, wish-wash, hydromel, oenomel, soft drink* etc.} as hyponymy concepts. However, *hydromel*, *oenomel*, *wish-wash* are relatively less familiar, have lower imageability, and have a lower probability of co-occurrence with other concepts. Therefore, we find it more judicious to consider only synonymy relations from WordNet® while expanding MRCPD.

As an example, consider the word, *deviance* which is not in MRCPD. We extract its synset {*aberrance, aberration, deviation*} from WordNet in order of decreasing Brown frequency. Next, we check the availability of each of these synonyms in MRCPD until a match is found. In this case, the word, *aberration* is available in MRCPD. Therefore, the conceptual features for *aberration* is assigned to *deviance*. Similarly, in case of *courtroom*, the word *court* is selected as a substitute and its conceptual features are extracted.

3.3 Approach

The proposed approach is divided in three phases namely, Data Processing, Feature Extraction and Model Building.

1. Data Processing phase: This involves data cleaning, tokenization and conversion to CoNLL[2] data format. After data processing, we obtain a sequence of tokens derived from sentences of the corpus.

2. Feature Extraction phase: This is sub-divided into four parts: Syntactic Feature Extraction, Conceptual Feature Extraction, Affective features extraction and Contextual feature extraction. The extraction process is explained in subsections 3.1 and 3.2. We use a context window of 7 words i.e. the word itself, 3 previous words and 3 next words, whose features are used to predict the label/class of a token.

3. Model Building phase: CRF (Lafferty *et al.*, 2001) is a probabilistic sequence labelling algorithm which takes an input sequence *X,* and predict the class label, *Y.* Since we are performing binary classification, *Y* consists of only two labels namely {*M, NM*} where *M* stands for metaphor and *NM* for otherwise cases. CRF doesn't make the assumption that features are independent and it is capable of handling large number of inter-independent or overlapping features unlike joint probability based models such as HMM which becomes intractable in such cases.

In our approach, these dependencies are defined by analyzing their effect on precision and recall. Thus, creating a feature set more aligned towards precise detection of metaphor without compromising recall. Features are defined in a template file (in case of CRF++) along with the context window and inter-dependencies between different features such as concreteness/pos, imageability/concreteness, cognitive_state/concreteness and so on. $X_{1:N}$ is a feature vector which is generated on the basis of a given template file. The joint distribution for the label sequence *Y* given *X* is as in eq. (1):

$$p(y|x,\alpha) = \frac{1}{Z(x)} \exp(\sum_k \alpha_k \sum_i^n f(y_{i-1}, y_i, x, i)) \quad (1)$$

Where $f(y_{i-1}, y_i, x, i)$ is a feature/transition function, y_{i-1}, y_i are class labels, x is an input sequence and *i* is an input position, *Z(x)* is a Normalization factor and α_k is a parameter vector. The elements of the parameter vector, $\{\alpha_k\} \epsilon R^K$ are calculated from the training data to maximize the likelihood of *p(y|x)*.

[2] CoNLL Data Format: http://ilk.uvt.nl/conll/#dataformat

4 Experiments and Results

We used CRF++ (V0.58) (Kudo, 2005), an open source C++ implementation of CRF based classifier for our experiments and VU Amsterdam Metaphor Corpus as dataset.

4.1 Dataset

We used VU Amsterdam Metaphor corpus (Steen *et al.*, 2010) developed by MIP Pragglejaz group. The annotation for metaphor in a sentence is based on anomaly between basic meaning of the term and contextual meaning of the term in the given sentence (Pragglejaz Group, 2007). The corpus is a small set of BNC Baby manually annotated for metaphors in XML format. For our experimental purpose, we converted the XML dataset into CoNLL format. We have considered words marked with MRW-Met sub-class as *Metaphor* and all other categories as *Non-Metaphor*. In original files, total number of metaphors is 25496, however in few cases we have separated multi-word expressions and marked all of them as Metaphor or Not Metaphor (whichever is the initial case) which led to minor increase in total number of metaphors. We have performed 10-fold cross validation across every genre as well as on overall dataset say Dataset2. The percentage of metaphor in each dataset is given in Table1.

Dataset	**M**	**NM**	**Total Tokens**	**% of M**
News	8388	52207	60595	13.84
Academic	8416	63275	71691	11.74
Fiction	4883	45105	49988	9.77
Conversation	3854	54342	58196	6.62
Total	25541	214929	240570	10.62
Legends: M: Metaphor, NM: Not Metaphor				

Table1: Dataset

4.2 Experiments

We trained the CRF based classifier on categorized datasets in steps to analyze individual effect of every feature set. During feature extraction, we observed that affective features are too sparse i.e. exists in case of affective words only which is around 11% (refer Table2). Therefore, affective feature set is merged with conceptual feature set to analyze its effect on metaphor detection in an open domain dataset.

Dataset	**% of AW**	**% of AW-M**
News	11.34	33.47
Academic	10.93	24.11
Fiction	12.72	22.33
Conv	11.46	16.71
Total	11.53	24.74
Legends: AW: Affective words, M-Metaphor		

Table2: Coverage of Affective Words

However, we observed that 24.74% of total affective words in corpus are annotated as metaphors in dataset indicating a correlation between metaphors and affective words. Below (Table3-7) are the values obtained for accuracy, precision and recall.

Dataset	**A**	**P-M**	**P-N**	**R-M**	**R-N**
News	90.62	70.60	93.98	54.94	96.46
Academic	90.58	62.21	90.28	46.37	97.85
Fiction	92.46	67.47	90.60	44.15	98.47
Conv	94.95	65.82	96.40	49.52	97.90
Avg	92.15	66.52	92.82	48.74	97.67
Dataset2	91.80	63.44	92.86	54.55	97.19

Table 3: Metaphor Detection using CRF classifier and Syntactic Feature (SF) set

Dataset	**A**	**P-M**	**P-N**	**R-M**	**R-N**
News	91.09	71.09	94.60	59.76	96.44
Academic	90.77	62.06	91.28	51.30	97.64
Fiction	92.67	67.06	91.47	49.00	98.20
Conv	95.15	66.60	96.62	53.79	97.73
Avg	92.42	66.70	93.49	53.46	97.50
Dataset2	92.10	64.32	93.39	58.54	97.27

Table 4: Metaphor Detection using CRF classifier with Conceptual Feature (SF+CF) set

Dataset	**A**	**P-M**	**P-N**	**R-M**	**R-N**
News	91.13	71.09	94.73	60.25	96.40
Academic	90.72	61.82	91.23	51.31	97.58
Fiction	92.67	66.86	91.53	49.46	98.16
Conv	95.15	66.60	96.62	53.89	97.70
Avg	92.42	66.59	93.53	53.73	97.46
Dataset2	92.08	64.14	93.46	58.64	97.20

Table 5: Metaphor Detection using CRF classifier with Conceptual Feature and Affective Feature (SF+ CF + AF) set

Dataset	**A**	**P-M**	**P-N**	**R-M**	**R-N**
News	90.43	69.24	94.11	55.30	96.16
Academic	90.45	61.04	90.58	48.05	97.80
Fiction	92.45	65.58	90.38	47.93	98.06
Conv	94.96	64.52	96.50	53.32	97.61
Avg	92.07	65.10	92.89	51.15	97.41
Dataset2	91.73	62.20	93.10	56.49	96.88

Table 6: Metaphor Detection using CRF classifier with Contextual Feature (SF + XF) set

Dataset	A	P-M	P-N	R-M	R-N
News	90.90	70.49	94.60	58.86	96.48
Academic	90.57	61.06	91.05	50.87	97.57
Fiction	92.64	66.54	91.22	49.66	98.16
Conv	95.12	65.97	96.60	54.43	97.63
Avg	92.31	66.02	93.37	53.46	97.46
Dataset2	91.97	63.33	93.34	58.71	96.95

Legends: Conv: Conversation; A: Accuracy; P-M: Precision for Metaphor Class; P-N: Precision for Non Metaphor Class; R-M: Recall for Metaphor class and R-N: Recall for Non- Metaphor class; Avg: Average of results across every genre.

Table 7: Metaphor Detection using CRF classifier on Hybrid Feature (SF+CF+AF+XF) set

Analysis: Extension of MRCPD with WordNet led to higher coverage and fewer missing values in conceptual feature vector. From the results in Table 3-4, we observed that inclusion of conceptual features significantly improved recall (up to 5%) for metaphors however it didn't have any visible effect on precision. Due to sparse affective feature vector, we could not observe any major impact on either precision or recall on adding affective feature set (refer Table5). Contextual Features i.e. embeddings with Syntactic features was also effective in recognizing metaphors but did not supersede the feature set {SF+CF} (refer Table4 and Table6). The combination {SF, CF, AF} (refer Table5) and {SF, CF, AF, XF} (refer Table8) didn't show any significant improvement over feature set {SF+CF}.

We gathered that presence of open category words (non-content words) such as determiners, conjunction etc. limited the effect of inter-dependent features. Consider the following sentence (metaphors in bold):

It would be *also necessary 'to* ***smash*** *the* ***decrepit, effete*** constitution that allows a minority to **capture power**, and then **use** it ruthlessly **in** the **interests** of the privileged few'.

Here, *smash* is used metaphorically. Despite of considering a context window of 7 words i.e. *also necessary to* ***smash*** *the decrepit effete,* we were unable to capture object of interest, i.e. *constitution.*

4.3 Comparison

We compared our approach with approaches proposed by Dunn (2013), Klebanov *et al.* (2015) and Hovy *et al.* (2013). Dunn used logistic regression classifier implemented in WEKA (Hall, 2009) on features extracted from SUMO Ontology. Klebanov *et al.* (2015) used concreteness as a feature with baseline features (Klebanov *et al.*, 2014) and optimal weighting technique. The dataset used was VU Amsterdam Metaphor corpus. Hovy *et al.* (2013) used SVMlight Tree Kernel (TK) implementation by (Moschitti, 2006) with syntactic features and CWE. However, they conducted the experiments on their own dataset, say Dataset3.

Model	A	P-M	R-M	F-M	*F-M
Proposed model^	**92.31**	**66.02**	53.46	59.08	77.22
Proposed model^^	91.97	63.33	58.71	**60.93**	**78.02**
Klebanov (2015)**	NA	43.8	**66.9**	51.1	NA
(Dunn, 2013)***	57.26	53.75	39.40	45.47	56.1
Proposed model#	**93.96**	**76.19**	47.42	58.46	77.60
Hovy *et al.*(2013)	75	70	**80**	**75**	NA

Legends: A: Accuracy; P-M: Precision for Metaphor Class; R-M: Recall for Metaphor class ; F-M is F-measure for Metaphor class; *F-M is average F-measure for both classes; NA: Not Available

^average of results obtained in each genre for {SF+CF+AF+XF}

^^results across all genres, *i.e.* Dataset2 for {SF+CF+AF+XF}

**average values for optimal weighting with concreteness measure method recommended in paper

***Results for MRW-Met Sub-Class across all genres

#results on Hovy *et al.*(2013) data *i.e.* Dataset3 using {SF+CF+AF+XF}

Table 8: Comparison of proposed model with previous approaches

By using CRF as a classifier, there is a significant improvement in accuracy and precision in comparison to models used in previous approaches. From Table8, we observed that CRF based classifier with {SF+CF+AF+XF} feature set outperformed approaches proposed by Dunn (2013) and Klebanov *et al.* (2015) in terms of accuracy, precision and F-measure (refer Table8). Klebanov *et al.* (2015) approach was based on optimal weighting to obtain optimal F-score which lead to comparatively higher recall. It is worth noting that accuracy increased by 18.96% and precision by 6.19% with respect to system proposed by Hovy *et al.* (2013). However, Hovy *et al.* (2013) outperformed in terms of recall and F-measure for metaphor class. The problem of low recall is likely to be resolved by optimizing the model for optimal F-measure as in Klebanov *et al.* (2015).

5 Conclusion

We used a CRF based classifier to perform metaphor detection in text. Using a rich feature set and a wide context window, we demonstrated the advantage of using CRF classifier over other classifiers in terms of accuracy and precision. We analyzed the efficacy of using CRF classifier over various combinations of feature sets. We expanded the coverage of the conceptual features of MRCPD using WordNet® ontology, resulting in significant improvement in the recall of metaphor detection. In future, we would like to conduct an in-depth analysis of affective features, fine-tune the process of feature selection and develop a multi-stage model for metaphor processing.

Acknowledgments

We acknowledge the contribution of Yash Kukreti and Ayush Garg in coding and implementation of the paper. We would also like to thank anonymous reviewers for helping us clarify several points and their insightful comments.

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. "The Berkeley framenet project." *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics.

Baumer, Eric PS, James P. White, and Bill Tomlinson. 2010. "Comparing semantic role labeling with typed dependency parsing in computational metaphor identification." *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*. Association for Computational Linguistics.

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

Birke, Julia, and Anoop Sarkar. 2007. "Active learning for the identification of nonliteral language." *Proceedings of the Workshop on Computational Approaches to Figurative Language*. Association for Computational Linguistics.

Bracewell, David B., Marc T. Tomlinson, Michael Mohler, and Bryan Rink. 2014. "A tiered approach to the recognition of metaphor." In *Computational Linguistics and Intelligent Text Processing*, pp. 403-414. Springer Berlin Heidelberg.

Cameron, Lynne, and Alice Deignan. 2003. "Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse." *Metaphor and Symbol "*18.3 (2003): 149-160.

Christine Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural language processing (almost) from scratch." The Journal of Machine Learning Research 12 (2011): 2493-2537.

Dunn, Jonathan. 2013. "What metaphor identification systems can tell us about metaphor-in-language". In Proceedings of the *First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, Georgia.

Fass. D. 1991. met*: A method for discriminating metonymy and metaphor by computer. Computational Linguistics, 17(1):49–90

Ferrari, Stéphane. 1996. "Using textual clues to improve metaphor processing." Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics.

Francis, W. Nelson, and Henry Kucera. 1979. "Brown corpus manual." Brown University.

Gargett, Andrew, and John Barnden. "Modeling the interaction between sensory and affective meanings for detecting metaphor." *NAACL HLT 2015*(2015): 21.

Gedigan M., J. Bryant, S. Narayanan, and B. Ciric. 2006. Catching metaphors. In Proceedings of the 3rd Workshop on Scalable Natural Language Understanding, pages 41–48, New York

Gibbs, Raymond W. 1984. "Literal Meaning and Psychological Theory*." *Cognitive science "*8.3 (1984): 275-304.

Gilhooly, Kenneth J., and Robert H. Logie. 1980. "Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words." *Behavior Research Methods & Instrumentation* 12.4 (1980): 395-427.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.

Hovy, Dirk, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. "Identifying metaphorical word use with tree kernels."

In Proceedings of the First Workshop on Metaphor in NLP, pp. 52-57.

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.

Karov, Yael, and Shimon Edelman. 1998. "Similarity-based word sense disambiguation." Computational linguistics 24, no. 1 (1998): 41-59.

Kingsbury, Paul, and Martha Palmer. 2002. "From TreeBank to PropBank."*LREC*.

Klebanov, B.B., Leong, C.W., Heilman, M. and Flor, M., 2014, June. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP* (pp. 11-17).

Klebanov, Beata Beigman, Chee Wee Leong, and Michael Flor. "Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples." *NAACL HLT 2015* (2015): 11.

Krishnakumaran, S., and X. Zhu. 2007. Hunting elusive metaphors using lexical resources. In Proceedings of the Workshop on Computational Approaches to Figurative Language, pages 13–20, Rochester, NY.

Kudo, Taku. 2005. "CRF++: Yet another CRF toolkit." Software available at http://crfpp.sourceforge.net.

Kudo, Taku, and Yuji Matsumoto. 2005. "YamCha: Yet another multipurpose chunk annotator." *2005-09-05)[2009-02-25]. http://www. chasen. org/~ tAKu/software/yamcha*.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira. 2001. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data."

Lakoff, George, and Mark Johnson. 1980. "Metaphors we live by, University of Chicago Press." Chicago, IL.

Liu, Ting, Kit Cho, George Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah M. Taylor et al. 2014. "Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings." In *LREC*, pp. 2800-2805.

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Martin, James H. 1990. A computational model of metaphor interpretation. Academic Press Professional, Inc.

Mason, Zachary J. 2004. "CorMet: a computational, corpus-based conventional metaphor extraction system." Computational Linguistics 30, no. 1 (2004): 23-44.

Mohler, Michael, David Bracewell, David Hinote, and Marc Tomlinson. 2013. "Semantic signatures for example-based linguistic metaphor detection." In Proceedings of the First Workshop on Metaphor in NLP, pp. 27-35.

Moschitti, Alessandro, Daniele Pighin, and Roberto Basili. 2006. "Tree kernel engineering for proposition re-ranking." Proceedings of Mining and Learning with Graphs (MLG 2006).

Neuman, Yair, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013, "Metaphor identification in large texts corpora." e62343.

Niles, Ian, and Adam Pease. 2001. "Towards a standard upper ontology." Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001. ACM.

Paivio, A., Yuille, J. C. and Madigan, S. A. 1968. Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement* 76, (3, Part 2)

Pragglejaz Group. 2007. "MIP: A method for identifying metaphorically used words in discourse." *Metaphor and symbol* 22.1 (2007): 1-39.

Princeton University.2010. "About WordNet." WordNet. Princeton University. http://wordnet.princeton.edu

Rentoumi, Vassiliki, George A. Vouros, Vangelis Karkaletsis, and Amalia Moser. 2012. "Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective." ACM Transactions on Speech and Language Processing (TSLP) 9, no. 3 (2012): 6.

Resnik, P. 1993. Selection and Information: A Class-based Approach to Lexical Relationships. Ph.D. thesis, Philadelphia, PA, USA.

Searle, John R. 1985. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Sardinha, Tony Berber. 2002. "Metaphor in early applied linguistics writing: A corpus-based analysis of lexis in dissertations." In I Conference on Metaphor in Language and Thought.

Sardinha, Tony Berber. 2006. "Collocation lists as instruments for metaphor detection in corpora." DELTA: Documentação de Estudosem Lingüística Teóricae Aplicada 22, no. 2 (2006): 249-274.

Schulder, Marc, and Eduard Hovy. 2014. "Metaphor detection through term relevance." *ACL 2014* (2014): 18.

Shutova, Ekaterina, Lin Sun, and Anna Korhonen. 2010. "Metaphor identification using verb and noun clustering." *Proceedings of the 23rd International*

Conference on Computational Linguistics. Association for Computational Linguistics.

Steen, G.J., Dorst A.G., Herrmann, J.B., Kaal, A.A.,Krennmayr, T., Pasma, T. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU.*

Strapparava, Carlo, and Alessandro Valitutti. "WordNet Affect: an Affective Extension of WordNet." In *LREC*, vol. 4, pp. 1083-1086. 2004.

Strzalkowski, T., Broadwell, G.A., Taylor, S., Feldman, L., Yamrom, B., Shaikh, S., Liu, T., Cho, K., Boz, U., Cases, I. and Elliott, K., 2013. Robust extraction of metaphors from novel data. In Proceedings of the ACL-13 Workshop on Metaphor (p. 67).

Sutton, Charles, and Andrew McCallum. 2011. "An introduction to conditional random fields." Machine Learning 4, no. 4 (2011): 267-373.

Toglia, Michael P., and William F. Battig. 1978. *Handbook of semantic word norms*. Lawrence Erlbaum.

Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. "Metaphor detection with cross-lingual model transfer."

Turney, Peter D., 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, pp. 680-690.

Turney, Peter D., and Michael L. Littman. 2003. "Measuring praise and criticism: Inference of semantic orientation from association." *ACM Transactions on Information Systems (TOIS)* 21.4 (2003): 315-346.

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

Wilks, Yorick, Lucian Galescu, James Allen, and Adam Dalton. 2013. "Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction." Meta4NLP 2013 (2013): 36.

Wilson, M.D. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, 20(1), 6-11.