Semantically Motivated Hebrew Verb-Noun Multi-Word Expressions Identification

Chaya Liebeskind, Yaakov HaCohen-Kerner
Department of Computer Science, Jerusalem College of Technology, Lev Academic Center
21 Havaad Haleumi St., P.O.B. 16031
9116001 Jerusalem, Israel
liebchaya@gmail.com, kerner@jct.ac.il

Abstract

Identification of Multi-Word Expressions (MWEs) lies at the heart of many natural language processing applications. In this research, we deal with a particular type of Hebrew MWEs, Verb-Noun MWEs (VN-MWEs), which combine a verb and a noun with or without other words. Most prior work on MWEs classification focused on linguistic and statistical information. In this paper, we claim that it is essential to utilize semantic information. To this end, we propose a semantically motivated indicator for classifying VN-MWE and define features that are related to various semantic spaces and combine them as features in a supervised classification framework. We empirically demonstrate that our semantic feature set yields better performance than the common linguistic and statistical feature sets and that combining semantic features contributes to the VN-MWEs identification task.

1 introduction

Multi-word expressions (MWE) were defined by Sag et al. (2002) as "idiosyncratic interpretations that cross word boundaries (or spaces)" while Bouamor et al. (2012) defined a MWE as "a combination of words for which syntactic or semantic properties of the whole expression can not be obtained from its parts".

Jackendoff (1997) claimed that that the frequency of MWEs in a speaker's lexicon is of the same order of single words. Due to their relative high frequency and complexity, MWEs require high-quality treatment in many applications in natural language processing (NLP) such as data mining, machine translation (MT), information retrieval, natural language understanding, natural language generation, question answering (QA), text summarization, and word sense disambiguation (WSD).

The aim of this work is to explore Hebrew Verb-Noun MWEs (VN-MWEs). VN-MWEs are MWEs whose constitutents include a verb and a noun. The motivation of this research is to enable automatic identification of VN-MWEs for various NLP tasks such as MT, QA, and WSD, and to classify collocations that include verbs as VN-MWEs or non-VN-MWEs.

Most prior efforts to automatically classify MWEs focused on three approaches: (1) Statistical approaches, either frequency-based or co-occurrence-based (Dias et al., 1999; Deane, 2005; Pecina and Schlesinger, 2006). (2) Linguistic approaches that are based on NLP tools, such as taggers and parsers (Al-Haj, 2009; Bejcek et al., 2013; Green et al., 2013). (3) Hybrid approaches which combine statistical and linguistic approaches (Baldwin, 2005; Boulaknadel et al., 2008; Farahmand and Nivre, 2015).

In this paper, we claim that on top of standard linguistic and statistical metrics, MWE identification methods can greatly benefit from exploiting semantically motivated cues. For example, when a VN-MWE is highly idiomatic, the semantics of the verb and the noun are not likely to overlap.

The contribution of this paper is, in a first step, to combine semantic features in the framework of supervised MWE classification. We suggest a simple semantically-motivated indicator that helps to

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1242–1253, Osaka, Japan, December 11-17 2016.

detect VN-MWEs. Then, we define semantic features that implement our indicator in various semantic spaces and integrate them within our Machine Learning (ML) classification algorithm.

We show that combining semantic features improves the accuracy and F-score results of VN-MWE classification. Moreover, our analysis reveals that the semantic feature set yields better results than each one of the two other approaches, the statistical and the linguistic.

The rest of this paper is organized as follows: Section 2 introduces relevant background about MWEs in Hebrew and identification of MWEs using semantic features. Section 3 presents the linguistic, statistical, and semantic feature sets that were applied for the supervised VN-MWEs classification task. Section 4 introduces the experimental setting, the experimental results for nine ML methods, and their analysis. Finally, Section 5 summarizes the main findings and suggests future directions.

2 Background

2.1 MWEs in Hebrew

Al-Haj (2009) presented an architecture for lexical representation of MWEs written in Hebrew and a specification of the integration of MWEs into a morphological processor of Hebrew. He also introduced a system that extracts noun compounds from Hebrew raw text based on their idiosyncratic morphological and syntactic features. A support vector machine (SVM) classifier using these features identified noun-noun constructs with an accuracy of over 80%.

Al-Haj and Wintner (2010) created for each noun-noun construction, a vector of the 16 features: 12 linguistically-motivated features and 4 collocation measures. Their dataset includes 463 instances, of which 205 are noun compounds (positive examples) and 258 negative. They applied LIBSVM classifier (Chang and Lin, 2001) with a radial basis function kernel. The best combination of features yielded an accuracy of 80.77% and F-score of 78.85, representing a reduction of over one third in classification error rate compared with the baseline.

Tsvetkov and Wintner (2012) proposed a methodology for extracting MWEs in Hebrew-English corpora. MWEs of various types are extracted along with their translations, from small, word-aligned parallel corpora. They focused on misalignments, which typically indicate expressions in the source language that are translated to the target in a non-compositional way. They implemented a simple algorithm that proposes MWE candidates based on such misalignments, relying on 1:1 alignments as anchors that delimit the search space. Evaluation of the algorithm's quality demonstrates significant improvements over Naive alignment-based methods.

Tsvetkov and Wintner (2014) proposed a framework for identifying MWEs in texts using multiple sources of linguistic information. Their system enables identification of MWEs of various types and multiple syntactic constructions. Their methodology is unsupervised and language-independent; it requires relatively few language resources and is thus suitable for a large number of languages. They applied four ML methods. The system was tested on three languages: Hebrew, French, and English. Applying the Bayesian Network ML method on a combination of linguistically motivated features and feature interdependencies reflecting domain knowledge yielded the best results (Hebrew: accuracy of 76.82% and F-score of 0.77; French: accuracy of 79.04% F-score of 0.778; and English: accuracy of 83.52% and F-score of 0.835).

Sheinfux et al. (2015) introduced different types of verbal MWEs in Modern Hebrew. In addition, they proposed an analysis of these MWEs in the framework of HPSG, and they incorporated this analysis into HeGram, a deep linguistic processing grammar of Modern Hebrew. Their analysis covers various MWE types, including challenging phenomena such as (possessive) co-indexation and internal modification. The HeGram grammar produced two analyses for most MWEs, corresponding to their idiomatic and literal readings.

Liebeskind and HaCohen-Kerner (2016) presented a lexical resource containing 505 Verb-Noun MWEs (VN-MWEs) in Hebrew. These VN-MWEs (247 bigrams and 258 trigrams) were manually collected from five web resources and annotated. Following Al-Haj (2009), the authors classified the linguistic properties of these VN-MWEs along 3 dimensions: morphological, syntactic, and semantic. The major findings are: (1) the main characteristic properties of VN-MWEs are the semantic properties

of non-compositionality and lexical fixedness; (2) High degrees of idiomaticity (92%) and lexical fixedness (94%) were found for the VN-MWEs; (3) 82% of the VN-MWEs do not allow any changes in the constituent order; and (4) 87% have a non-compositional syntax.

2.2 Identification of MWEs using Semantic Features

Katz and Giesbrecht (2006) applied latent semantic analysis (LSA) vectors to distinguish compositional from non-compositional uses of German expressions. The LSA vectors of compositional and non-compositional meaning were constructed from a training set of example sentences. Afterwards, a simple nearest neighbor algorithm was applied on the LSA vectors of the tested MWEs. The LSA-based classifier obtained an average accuracy of 72%, which outperformed the simple maximum-likelihood baseline with accuracy of 58%.

Sporleder and Li (2009) proposed supervised and unsupervised methods to distinguish literal from non-literal usages of idiomatic expressions by measuring the semantic relatedness of an expression's component words to nearby words in the text. Their assumption was that if an expression is used literally, but not idiomatically, its component words will be related semantically to a few words in the surrounding discourse. If one or more of the expression's components were sufficiently related to enough surrounding words, the usage was classified as literal, otherwise as idiomatic. The supervised classifier method (90% F-score on literal uses) was better than the lexical chain classifier methods (60% F-score).

Biemann and Giesbrecht (2011) provided an overview of the shared task at the ACL-HLT 2011 DiSCo (Distributional Semantics and Compositionality) workshop. The authors described the motivation for the shared task, the acquisition of datasets, the evaluation methodology, and the results of participating systems. The evaluation shows that most systems outperformed simple baselines, yet have difficulties in reliably assigning a compositionality score that closely matches the gold standard. Generally, approaches based on word space models performed slightly better than approaches relying merely on statistical association measures.

Guevara (2011) proposed and evaluates a framework that models the semantic compositionality in computational linguistics based on the combination of distributional semantics and supervised ML. The applied method, Partial Least Squares (PLS) Regression, outperformed all the competing models in the reported experiments with Adjective-Noun (AN) pairs extracted from the BNC.

Salehi et al. (2015) introduced the first attempt to use word embeddings to predict the compositionality of MWEs. They considered both single- and multi-prototype word embeddings. Experimental results showed that, in combination with a back-off method based on string similarity, word embeddings are superior to, or competitive with state-of-the-art methods over 3 standard compositionality datasets ((1) English noun compounds ("ENCs"); (2) English verb particle constructions ("EVPCs"); and (3) German noun compounds ("GNCs")).

3 Supervised VN-MWEs Classification

In the previous section, we discussed linguistic properties of Hebrew VN-MWEs that may help in distinguishing coincidental word combinations from collocations. We next define them and describe how to incorporate these properties as features within a ML framework for classifying candidate VN-MWEs.

3.1 Feature Sets

We next detail how the semantic properties of VN-MWEs, as well as the linguistic and statistical properties found useful in prior work, are encoded as features. Then, in Section 4, we describe the supervised ML model and our feature analysis procedure. There are 206 features in our model, divided into 3 sets: linguistic, statistical and semantic. We defined the sets as Al-Haj and Wintner (2010) and Liebeskind and HaCohen-Kerner (2016) did. However, we note that semantic information is often defined as a sub-type of linguistic information and it might be more accurate to contrast morpho-syntactic information (i.e., parts-of-speech and syntactic parses) with semantic information.

3.1.1 Linguistic features

Most of our linguistic features are based on information extracted from a Part-Of-Speech (POS) tagger for the Hebrew language (Adler, 2007). Our linguistic features encode both morphological and syntactical properties of VN-MWEs. For each candidate VN-MWE, we compute counts that reflect the reasonableness of the candidate to represent at least one of its linguistic properties. We focus on the linguistic properties that Liebeskind and HaCohen-Kerner (2016) recognized as notable. Our linguistic properties include two families of properties: morphological and syntactical.

Partial Inflection Following Al-Haj and Wintner (2010), for each VN-MWE candidate, the following 8 features are defined: the number of occurrences of the candidate in which both constituents are in singular, the number of occurrences in which both constituents are in plural, the number of occurrences in which the verb is in singular and the noun is in plural, the number of occurrences in which the noun is in singular and the verb is in plural, the number of occurrences of the verb in plural, the number of occurrences of the verb in singular, the number of occurrences of the noun in plural and the number of occurrences of the noun in singular. Two additional features that we calculate are the number of verb suffixes, which indicate a conjugation of grammatical tense, possession or direct objects, as well as the number of noun suffixes, which indicate nouns number and gender (*ildi*[1] (my child), *ild**inw*** (our children), *ild**h*** (a girl)).

Syntactic Fixedness VN-MWEs are expected to appear in restricted syntactic forms. Fazly and Stevenson (2006) suggested that to quantify the syntactic fixedness of a VN-MWE candidate, we need to: (i) identify relevant syntactic patterns and (ii) translate the frequency distribution of the candidate in the identified patterns into a measure of syntactic fixedness. Following this approach, we define syntactic patterns and clues as features in our supervised framework.

We use the most frequent POS patterns found in Liebeskind and HaCohen-Kerner (2016)'s VN-MWEs lexical resource as relevant syntactic patterns and count the number of occurrences of the candidate in each of these patterns (7 features).

Since prepositions and definite articles frequently appear in these patterns, we counted the number of occurrences of the candidate in which it includes an article, a pronoun, a particle, a conjunction, an auxiliary or a negation (6 features). Then, considering the fact that some of these POS are often Hebrew prefixes, we also encoded prefixes' occurrences. The features that we calculate are the number of occurrences of verb and noun prefixes (2 features), the number of occurrences of prefixes which start with a certain frequent *formative letter* (7 features for each POS, verb and noun), the number of occurrences of a certain frequent prefix (36 features for each POS). Our Hebrew stopword list also include some particles. Therefore, we calculated an additional feature of the number of stopwords in the candidate VN-MWE (1 feature).

The syntactic property of compositionality is encoded by the difference between the number of occurrences of the candidate constituents with and without a *slot* (1 feature). The syntactic property of a number of syntactic structures that permit a change in the order of constituents is encoded by the difference between the number of occurrences of the candidate constituents in their original order and the number of occurrences of the candidate constituent in a reversed order (1 feature).

3.1.2 Statistical features

We define some statistical features based on frequency and co-occurrence affinity. Each of these features is separately calculated for two candidate representations: surface and lemmatized. First, we compute the raw frequency of the VN-MWE candidate and the raw frequency of its verb and noun constituents (6 features). Then, we utilize features that represent known association measures: Log-likelihood, Total mutual information, Pointwise mutual information and Poisson-Stirling measure. We calculate them for bigrams and trigrams separately (16 features). Finally, we define four statistical features based on two non-parametric methods, which does not make the independence assumption and allows scores to be

[1]To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are abgdhwzxTiklmns`pcqršt.

compared across n-grams of different length: Mutual Expectation (ME) (Dias et al., 1999) and Mutual Rank Ratio (MRR) (Deane, 2005) (4 features).

In addition, we calculate the number of words, number of characters and the average number of characters per word (3 features) for each candidate in its base form.

3.1.3 Semantic features

Liebeskind and HaCohen-Kerner (2016) observed that the most characteristic properties of VN-MWEs are the semantic properties of compositionality and lexical fixedness. To encode this property, we represent the meaning of the candidate's constituents by vectors in the same semantic space. Due to the idiomaticity of VN-MWEs, we expect the similarity of vectors of words in a non-VN-MWE to be greater than the similarity of vectors of words in a VN-MWE. For example, the VN-MWE *to eat one's hat* vs. the non-VN-MWE *to eat an apple*. We expect the vectors of *eat* and *apple*, which share a common context, to be closer than the vectors of *eat* and *hat* in a representative semantic space.

We construct semantic features from the following five different semantic spaces:

(1) Hyperspace Analogue to Language (HAL) (Lund and Burgess, 1996): The algorithm computes a word-by-word matrix, using a 10-word reading frame that moves incrementally through a corpus of text. The algorithm considers context only as the words that immediately surround a given word. Any time two words are simultaneously in the frame, the association between them is increased, that is, the corresponding cell in the matrix is incremented. The amount by which the association is incremented varies inversely with the distance between the two words in the frame; closer neighboring words are thought to reflect more of the focus word's semantics and so are weighted higher. The algorithm also records word-ordering information by treating the co-occurrence differently based on whether the neighboring words appeared before or after the focus word.

(2) Correlated Occurrence Analogue to Lexical Semantics (COALS) (Rohde et al., 2006): The algorithm constructs a word-by-word matrix where each element in the matrix represents how frequently word_i occurs with word_j in a certain window. The matrix is then normalized by correlation, and any negative values are set to zero and all other values are replaced by its square root. Then, optionally, the Singular Value Decomposition (SVD) is used to reduce the word co-occurrence matrix.

(3) Random Indexing (RI) (Sahlgren, 2005): The algorithm uses statistical approximations of the full word co-occurrence data to achieve dimensionality reduction. RI represents co-occurrence through index vectors. Each word is assigned a high-dimensional, random vector that is known as its index vector. These index vectors are very sparse, which ensures that the chance of any two arbitrary index vectors having an overlapping meaning is very low. Word semantics are calculated for each word by keeping a running sum of all of the index vectors for the words that co-occur.

(4) Reflective Random Indexing (RRI) (Cohen et al., 2010): The algorithm is a second-order iterative extension to the RI method. Reflective random indexing adds another cycle by restarting the construction of the term vectors using the basis of document vectors, and then creating the document vectors again using the term vectors. Such retraining has been found to improve the ability of RI to make indirect inferences, drawing meaningful associations between terms that do not occur together in any document.

(5) Word Embeddings (Mikolov et al., 2013): Word embedding is the collective name for neural-network based approaches in which words are embedded into a low dimensional space. In word embedding models, the contexts of each word are modeled by a d-dimensional vector of real numbers. The vector are meaningless on their own, but semantically similar words have similar vectors, and vector similarities are easy to compute.

Each of these five semantic spaces is generated for two word representations: surface and lemmatized. We use different measures to compute the similarity between two vectors. For the first four semantic spaces[2], we calculate Cosine similarity, Lin similarity, Euclidean distance, Pearson correlation, average common feature rank, Jaccard index, Tanimoto coefficient, and Spearman rank correlation. For the fifth semantic space[3], we calculate cosine similarity, euclidean distance, and manhattan distance.

[2]implemented by the S-Space Package `https://github.com/fozziethebeat/S-Space`

[3]implemented by the deeplearning4j word2vec package `http://deeplearning4j.org/word2vec`

Some of the measures did not yield a valid score for all the examples in our dataset. As a result, the total number of semantic features is 62.

An additional semantic feature, which measures lexical fixedness, counts the number of occurrences of the VN-MWE candidate in the Bible. MWEs from the Bible are citations that tend to be fixed, replacing any of their constituents by a semantically similar word generally results in an invalid or a literal expression.

We note that corpus-based statistics are used to calculate some of the linguistic features (e.g., the features which encode the Partial Inflection property). Additionally, some of the statistical features, such as Mutual Expectation (ME) and Mutual Rank Ratio (MRR), capture the semantic behavior of VN-MWEs.

4 Evaluation and Analysis

4.1 Experimental setting

Following Al-Haj and Wintner (2010), we used four of the MILA knowledge center[4]corpora: the Knesset corpus, which contains the Israeli parliament proceedings from 2004-2005; the Haaretz corpus that contains articles from the Haaretz newspaper from 1991; TheMarker corpus, which contains financial articles from the TheMarker newspaper from 2002; and the Arutz 7 corpus, which contains newswire articles from 2001-2006. From the morphologically disambiguated version of the corpora (Itai and Wintner, 2008; Yona and Wintner, 2008; Bar-haim et al., 2008), we extracted all word bigrams and trigrams that include a verb and a noun.

To evaluate our proposed supervised model, we constructed a labeled dataset. We selected all the word bigrams and trigrams that occur at least 25 times in the corpora. These candidates were annotated by two annotators, who were asked to classify them as a VN-MWE or a non-VN-MWE. We evaluated the inter-annotator agreement and observed a Kappa (Cohen, 1960) value of 0.59, which is considered as moderate (Landis and Koch, 1977). Thus, we considered a candidate as a VN-MWE or not only if both annotators agreed on its classification. This reduced the labeled data to 553 instances, of which 306 are VN-MWEs (256 bigrams and 50 trigrams) and 247 are non-VN-MWEs (157 bigrams and 90 trigrams).

4.2 Application of nine Machine Learning methods

We combined the features in a supervised classification framework using nine ML methods: Random Forest, Decision Tree, Bagging, Adaboost, Bayes Network, Supported Vector Machine (SVM), Logistic Regression and Multilayered Perceptron. The accuracy rate of each ML method was estimated by a 10-fold cross-validation test. We ran these ML methods by the WEKA platform (Witten and Frank, 2005; Hall et al., 2009) using the default parameters. Table 1 shows the performances of the different ML methods on the full feature set of 206 features, as described above. The best ML method was Random Forest. Therefore, we have performed further experiments using only this method. These experiments are presented in the next sub-section.

4.3 Further experimental results using the random forest method

In this research, we defined three types of feature sets (Section 3): linguistic, statistic and semantic. The classification results of the Random Forest algorithm (the best ML method in Table 1) on each of the sets are presented in the left side of Table 2. The semantic feature set yielded the best accuracy result (77.4%). The advantage of the semantic feature set over the linguistic and statistical feature sets is notable (3.5% and 5% respectively) and is statistically significant according to the McNamar test (McNemar, 1947) for the statistical feature set (p=0.017). The advantage is also almost statistically significant at level 0.05 for the linguistic feature set (p=0.056).

A hybrid approach, which combines the linguistic and statistical information, is commonly used in MWE extraction. Therefore, we investigated different combinations of feature sets. The results of our exploration are presented in the right side of Table 2. The best results were obtained using all the three sub-feature sets. However, the contribution of the linguistic feature set was negligible (80.47% vs.

[4]http://www.mila.cs.technion.ac.il/resources/

#	ML Method	Accuracy (%)	F-Measure
1	Random Forest	**80.47**	**0.795**
2	Decision Tree (J48)	71.25	0.708
3	Bagging	78.84	0.78
4	AdaBoost (M1)	74.32	0.736
5	Bayes Network	69.80	0.703
6	Logistic Regression	70.52	0.706
7	Multilayered Pereceptron	68.72	0.686
8	SVM (SMO)	76.13	0.76
9	SVM (LibSVM)	63.11	0.488

Table 1: Comparison of results obtained by nine ML methods

Feature Set	Accuracy (%)	F-Measure	Feature Sets	Accuracy (%)	F-Measure
Linguistic	73.96	0.721	Linguistic & Semantic	77.4	0.765
Statistical	72.51	0.712	Linguistic & Statistic	78.84	0.777
Semantic	**77.4**	**0.767**	Semantic & Statistic	**80.29**	**0.796**
			All	**80.47**	**0.796**

Table 2: Comparison of results for different combinations of feature sets

80.29%). As was found in previous studies (Justeson and Katz, 1995; Pecina, 2010), the approach of combining linguistic and statistical features works efficiently. Yet, combining linguistic and semantic features did not yield any improvement over using only the semantic feature set.

For each of the above feature set configurations, we tried to filter out non-relevant features using two well-known feature selection methods: Information gain (InfoGain, IG) (Hunt et al., 1966) and Correlation-based Feature Subset (CFS) (Hall, 1998). The use of these two feature selection methods did not improve the accuracy of any configuration. However, we used the information obtained by the IG selection method to better understand which features have more influence on the classification accuracy. Table 3 presents the features, which were selected by the IG method for the three feature sets (the number in parentheses is the feature rank). The linguistic properties of partial inflection and constituent order were found as important properties for distinguishing MWEs from non-MWEs. The two non-parametric statistical features, Mutual Expectation (ME) and Mutual Rank Ratio (MRR), outperformed other baseline association measures. The good performance of the algorithm using the semantic features is due to the combination of various semantic spaces and vector comparison measures.

Table 4 shows the features that were selected by the IG method for the different combinations of feature sets. For each feature sub-set of a combined configuration, Table 4 details how many and which of its selected features were also selected by the IG measure when each set was tested as a standalone feature set (see Table 3). While the same semantic features were selected for the standalone (Semantic) and combined configurations (Linguistic & Semantic and Semantic & Statistical), different linguistic and statistical features were selected by each of the configurations that include them.

We further analyze our suggested semantic feature set by comparing the performance of the different semantic spaces. Table 5 shows the classification results of the Random Forest algorithm on the various sub-sets of the semantic features. The semantic features that were constructed by the HAL semantic space outperformed the other semantic representations. The advantage of the HAL semantic space might be due to its sensitivity to word-ordering. This sensitivity enables the representation to model the important constituent order linguistic property of VN-MWE. The low performance of the word embedding space could be explained either by its low number of features or by the fact that these vector were constructed without any task-depended training.

Finally, we investigated the False Positive (FP) and False Negative (FN) classifications of our suggested semantic feature set. We found that some of the FPs were due to light verbs, such as ***lqbl*** *mid`* (to

Feature set	# of feat.	Feature list
Linguistic	10	SINGUALR_VERB_PLURAL_NOUN (1), COUNJUNCTION (2), CONSTITUENT_ORDER (5). PREFIX_VERB (*wmš* (4), starts with *m* (7), *kšb* (9)), PREFIX_NOUN (*mh* (3), *š* (8), *kšm* (10))
Statistical	5	TRIGRAMSPMI (1), MUTUALRANKRATIO (2), NOUNLFREQUENCY (3), MUTUALSCORE (4), TRIGRAMSLL (5)
Semantic	23	**COAL**: PEARSON (1), COSINE (5), AVERAGE_COMMON_FEATURE_RANK (ACFR) (13), **COAL_LEMMA**: PEARSON (2), COSINE (3), TANIMOTO (4), ACFR (10)
		HAL: EUCLIDEAN (8), ACFR (11), **HAL_LEMMA**: ACFR (6), LIN (17), TANIMOTO (19), PEARSON (22), COSINE (23)
		RI: EUCLIDEAN (7), **RI_LEMMA**: ACFR (12), TANIMOTO (15), LIN (16)
		RRI: EUCLIDEAN (9), **RRI_LEMMA**: SPEARMAN (18)
		Word Embeddings: EUCLIDEAN (20), COSINE (21), **Word Embeddings_LEMMA**: MANHATTAN (14)

Table 3: InfoGain feature selection of the linguistic, statistic and semantic feature sets

get information) and ***`wšh*** *`bwdh* (to make a work). The general meaning of the light verbs decreased the vectors comparison score and candidates with light verbs were wrongly classified as VN-MWEs. This might be because light verbs have little semantic content of their own and they are used in combination with various nouns. Thus, the semantic similarity between the light verb and a specific noun was relatively low. A possible solution to the light verb issue is to use a directional inclusion-based measure to compute the similarity between two vectors (Weeds and Weir, 2003; Clarke, 2009; Kotlerman et al., 2010).

Another interesting finding is that domain-specific VN-MWEs were often misclassified as FNs. VN-MWEs like *lhqim mmšlh* (to establish a government) and *lgbš `mdh* (to form an opinion) were wrongly classified as non-MWEs since they frequently co-occur in our political domain, so their semantic vectors are rather close.

5 Conclusions and Future Work

We presented a supervised classification model for identification of Hebrew VN-MWEs. Our semantic feature set yields better performance than the common linguistic and statistical feature sets and that combining semantic features contributes to the Hebrew VN-MWEs identification task.

Most previous related studies apply only one ML method. An exception was the study of Tsvetkov and Wintner (2014), which applied 4 ML methods. In this research, we applied 9 ML methods. Moreover, we have performed further experiments using only the Random Forest method, which has been found as the best ML method for our task. Our experiment over a manually labeled dataset showed that the semantic feature set outperforms the statistical and linguistic feature sets and that combining semantic features with the two other feature sets further improved the performance (especially with the statistical set).

In future work, we would like to investigate more sophisticated models for representing the semantic meaning of VN-MWEs. For example, we plan to extend the single-word vector representation to learn larger semantic composition representations (Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Socher et al., 2012). We also plan to investigate directional inclusion-based similarity measures for computing vector similarity.

In addition, we plan to adopt our model to under-resourced languages, many of them are found in the developing world where we lack the linguistic information.

Feature set	# of feat.	Sub-feature set	Overlapping features	Additional features in top10
Linguistic & Semantic	33	Linguistic	CONSTITUENT_ORDER (1/10)	None
		Semantic	all (23/23)	None
Linguistic & Statistical	15	Linguistic	CONSTITUENT_ORDER (1/10)	POSPATTERN (verb + preposition + noun), SINGULAR_VERB, PREFIX_NOUN (*wšb*, *kl*), PLURAL_VERB_SINGULAR_NOUN, PREFIX_VERB (b), PREFIX_NOUN_NUM
		Statistical	MUTUALRANKRATIO (1/5)	MUTUALSCORELEMMA, BIGRAMSTMI
Semantic & Statistical	28	Semantic	all (23/23)	None
		Statistical	None (0/5)	FREQUENCY

Table 4: IG selection results for the different combinations of feature sets

Semantic Space	# of feat.	Accuracy (%)	F-Measure	ROC Area
COAL	13	73.59	0.731	0.8
HAL	14	**75.04**	**0.738**	**0.82**
RI	14	72.87	0.717	0.782
RRI	15	72.87	0.716	0.781
Word Embedding	6	64.56	0.635	0.69

Table 5: Comparison of the results obtained by different semantic sub-spaces

Acknowledgements

We would like to express our deep gratitude to Avital Day, our research assistant, for her help in programming and carrying out the research experiments. We would also like to acknowledge the networking support by the COST Action IC1207: PARSEME: PARSing and Multi-word Expressions. This work was partially funded by an internal research grant from Jerusalem College of Technology, Lev Academic Center.

References

Meni Adler. 2007. *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*. Ph.D. thesis, Ben-Gurion University of the Negev.

Hassan Al-Haj and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 10–18, Beijing, China, August. Coling 2010 Organizing Committee.

Hassan Al-Haj. 2009. *Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition*. Ph.D. thesis, University of Haifa.

Timothy Baldwin. 2005. Deep lexical acquisition of verb-particle constructions. *Comput. Speech Lang.*, 19(4):398–414, October.

Roy Bar-haim, Khalil Sima'an, and Yoad Winter. 2008. Part-of-speech tagging of modern hebrew text. *Natural Language Engineering*, 14(2):223–251, April.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Eduard Bejcek, Pavel Stranák, and Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 106–115.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality*, pages 21–28. Association for Computational Linguistics.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *LREC*, pages 674–679.

Siham Boulaknadel, Be'atrice Daille, and Driss Aboutajdine. 2008. A multi-word term extraction program for arabic language. In *LREC*, pages 1485–1488. European Language Resources Association.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 112–119. Association for Computational Linguistics.

Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of biomedical informatics*, 43(2):240–256.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Paul Deane. 2005. A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 605–613. Association for Computational Linguistics.

Gaël Dias, Sylvie Guilloré, and JG Pereira Lopes. 1999. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *Traitement Automatique des Langues Naturelles, Institut dEtudes Scientifiques, Cargèse, France*, pages 333–339.

Meghdad Farahmand and Joakim Nivre. 2015. Modeling the statistical idiosyncrasy of multiword expressions. In *Proceedings of NAACL-HLT*, pages 34–38.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 337–344, Trento Italy.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227, March.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.

Emiliano Guevara. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 135–144, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

M. Hall. 1998. *Correlation-based feature subset selection for machine learning*. Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

E. B. Hunt, J. Marin, and P. J. Stone. 1966. *Experiments in Induction*. Academic Press, New York.

Alon Itai and Shuly Wintner. 2008. Language resources for hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Ray Jackendoff. 1997. *The architecture of the language faculty*. Number 28. MIT Press.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Chaya Liebeskind and Yaakov HaCohen-Kerner. 2016. A lexical resource of hebrew verb-noun multi-word expressions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC'16, pages 522–527, Portoroz, Slovenia, may. European Language Resources Association (ELRA).

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 651–658, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut. 2006. An improved model of semantic similarity based on lexical co-occurence. *Communications of the ACM*, 8:627–633.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, London, UK, UK. Springer-Verlag.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2015. Hebrew verbal multi-word expressions. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, pages 123–136.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics.

Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

Yulia Tsvetkov and Shuly Wintner. 2014. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, 40(2):449–468.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88. Association for Computational Linguistics.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Shlomo Yona and Shuly Wintner. 2008. A finite-state morphological grammar of hebrew. *Natural Language Engineering*, 14:173–190, 4.