Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources

Yulia Tsvetkov
Language Technologies Institute
Carnegie Mellon University
yulia.tsvetkov@gmail.com

Shuly Wintner
Department of Computer Science
University of Haifa
shuly@cs.haifa.ac.il

Abstract

We propose an architecture for expressing various linguistically-motivated features that help identify multi-word expressions in natural language texts. The architecture combines various linguistically-motivated classification features in a Bayesian Network. We introduce novel ways for computing many of these features, and manually define linguistically-motivated interrelationships among them, which the Bayesian network models. Our methodology is almost entirely unsupervised and completely language-independent; it relies on few language resources and is thus suitable for a large number of languages. Furthermore, unlike much recent work, our approach can identify expressions of various types and syntactic constructions. We demonstrate a significant improvement in identification accuracy, compared with less sophisticated baselines.

1 Introduction

Multi-word Expressions (MWEs) are lexical items that consist of multiple orthographic words (e.g., *ad hoc*, *by and large*, *New York*, *kick the bucket*). MWEs are numerous and constitute a significant portion of the lexicon of any natural language (Jackendoff, 1997; Erman and Warren, 2000; Sag et al., 2002). They are a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or preventing) the inflection of other constituents. In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Syntactically, some MWEs behave like words while other are phrases; some occur in one rigid pattern (and a fixed order), while others permit various syntactic transformations. Semantically, the compositionality of MWEs is gradual, ranging from fully compositional to idiomatic (Bannard et al., 2003).

Because of their prevalence and irregularity, MWEs must be stored in lexicons of natural language processing applications. Correct handling of MWEs has been proven beneficial for various applications, including information retrieval, building ontologies, text alignment, and machine translation.

We propose a novel architecture for identifying MWEs of various types and syntactic categories in monolingual corpora. Unlike much existing work, which focuses on a particular syntactic construction, our approach addresses MWEs of all types by focusing on the general idiosyncratic properties of MWEs rather than on specific properties of each sub-class thereof. While we only evaluate our methodology on bi-grams, it can in principle be extended to longer MWEs. The architecture uses Bayesian Networks (BN) to express multiple interdependent linguistically-motivated features.

First, we automatically generate a small (training) set of MWE and non-MWE bi-grams (positive and negative instances, respectively). We then define a set of linguistically-motivated features that embody observed characteristics of MWEs. We augment these by features that reflect collocation measures. Finally, we define dependencies among these features, expressed in the structure of a Bayesian Network model, which we then use for classification. This is a directed graph, whose nodes express the features used for classification, and whose edges de-

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 836–845,
Edinburgh, Scotland, UK, July 27–31, 2011.

fine causal relationships among these features. In this architecture, learning does not result in a black box, expressed solely as feature weights. Rather, the structure of the BN allows us to learn the impact of different MWE features on the classification. The result is a new unsupervised method for identifying MWEs of various types in text corpora. It combines statistics with a large array of linguistically-motivated features, organized in an architecture that reflects interdependencies among the features.

The contribution of this work is manifold. First, we show how to generate training material (almost) automatically, so the method is almost completely unsupervised. The methodology we advocate is thus language-independent, requiring relatively few language resources, and is therefore optimal for medium-density languages (Varga et al., 2005). Second, we propose several linguistically-motivated features that can be computed from data and that are demonstrably productive for improving the accuracy of MWE identification. These feature focus on the expression of linguistic idiosyncrasies of various types, a phenomenon typical of MWEs. We propose novel computational modeling of many of these features; in particular, we account for the morphological idiosyncrasy of MWEs using a histogram of the number of inflected forms, in a technique that draws from image processing. Third, we advocate the use of Bayesian Networks as a mechanism for expressing manually-crafted dependencies among features; the use of BN significantly improves the classification accuracy. Finally, we demonstrate the utility of our methodology by applying it to Hebrew.[1] Our evaluation shows that the use of linguistically-motivated features results in reduction of 23% of the errors compared with a collocation baseline; organizing the knowledge in a BN reduces the error rate by additional 8.7%.

After discussing related work in the next section, we describe in Section 3 the methodology we propose, including a detailed discussion of the features and their implementation. Section 4 provides a thorough evaluation of the results. We conclude with suggestions for future research.

2 Related Work

Early approaches to MWEs identification concentrated on their collocational behavior (Church and Hanks, 1990). Pecina (2008) compares 55 different association measures in ranking German Adj-N and PP-Verb collocation candidates. He shows that combining different collocation measures using standard statistical classification methods improves over using a single collocation measure. Other results (Chang et al., 2002; Villavicencio et al., 2007) suggest that some collocation measures (especially PMI and Log-likelihood) are superior to others for identifying MWEs.

Soon, however, it became clear that mere co-occurrence measurements are not enough to identify MWEs, and their linguistic properties should be exploited as well (Piao et al., 2005). Hybrid methods that combine word statistics with linguistic information exploit morphological, syntactic and semantic idiosyncrasies to extract idiomatic MWEs.

Ramisch et al. (2008) evaluate a number of association measures on the task of identifying English Verb-Particle Constructions and German Adjective-Noun pairs. They show that adding linguistic information (mostly POS and POS-sequence patterns) to the association measure yields a significant improvement in performance over using pure frequency.

Several works address the *lexical fixedness* or *syntactic fixedness* of (certain types of) MWEs in order to extract them from texts. An expression is considered lexically fixed if replacing any of its constituents by a semantically (and syntactically) similar word generally results in an invalid or literal expression. Syntactically fixed expressions prohibit (or restrict) syntactic variation. For example, Van de Cruys and Villada Moirón (2007) use lexical fixedness to extract Dutch Verb-Noun idiomatic combinations (VNICs). Bannard (2007) uses syntactic fixedness to identify English VNICs. Another work uses both the syntactic and the lexical fixedness of VNICs in order to distinguish them from non-idiomatic ones, and eventually to extract them from corpora (Fazly and Stevenson, 2006).

While these approaches are in line with ours, they require lexical semantic resources (e.g., a database that determines semantic similarity among words) and syntactic resources (parsers) that are unavail-

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

able for Hebrew (and many other languages). Our approach only requires morphological processing and a bilingual dictionary, which are more readily-available for several languages. Note also that these approaches target a specific syntactic construction, whereas ours is adequate for various types of MWEs.

Several properties of Hebrew MWEs are described by Al-Haj (2010); Al-Haj and Wintner (2010) use them in order to construct an SVM-based classifier that can distinguish between MWE and non-MWE *noun-noun constructions* in Hebrew. The features of the SVM reflect several morphological and morpho-syntactic properties of such constructions. The resulting classifier performs much better than a naïve baseline, reducing over one third of the errors. We rely on some of these insights, as we implement more of the linguistic properties of MWEs. Again, our methodology is not limited to a particular construction: indeed, we demonstrate that our general methodology, trained on automatically-generated, general training data, performs almost as well as the noun-noun-specific approach of Al-Haj and Wintner (2010) on the very same dataset.

Recently, Tsvetkov and Wintner (2010b) introduced a general methodology for extracting MWEs from bilingual corpora, and applied it to Hebrew. The results were a highly accurate set of Hebrew MWEs, of various types, along with their English translations. A major limitation of this work is that it can only be used to identify MWEs in the bilingual corpus, and is thus limited in its scope. We use this methodology to extract both positive and negative instances for our training set in the current work; but we extrapolate the results much further by extending the method to *mono*lingual corpora, which are typically much larger than bilingual ones.

Bayesian Networks have only scarcely been used for classification in natural language applications. For example, BN were used for POS tagging of unknown words (Peshkin et al., 2003); dependency parsing (Savova and Peshkin, 2005); and document classification (Lam et al., 1997; Calado et al., 2003; Denoyer and Gallinari, 2004). Very recently, Ramisch et al. (2010) have used BN for Portuguese MWE identification. The features used for classification were of two kinds: (1) various collocation measures; (2) bi-grams aligned together by an automatic word aligner applied to a parallel (Portuguese-English) corpus. A BN was used to combine the predictions of the various features on the test set, but the structure of the network is not described. The combined classifier resulted in a much higher accuracy than any of the two methods alone. However, the BN does not play any special role in this work, and its structure does not reflect any insights or intuitions on the structure of the problem domain or on interdependencies among features.

We, too, acknowledge the importance of combining different types of knowledge in the hard task of MWE identification. In particular, we also believe that collocation measures are highly important for this task, but cannot completely solve the problem: linguistically-motivated features are mandatory in order to improve the accuracy of the classifier. In this work we focus on various properties of different types of MWEs, and define general features that may accurately apply to some, but not necessarily all of them. An architecture of Bayesian Networks is optimal for this task: it enables us to define weighted dependencies among features, such that certain features are more significant for identifying some class of MWEs, whereas others are more prominent in identifying other classes. As we show below, this architecture results in significant improvements over a more naïve combination of features.

3 Methodology

3.1 Motivation

The task we address is identification of MWEs, of various types and syntactic constructions, in monolingual corpora.[2] Several properties of MWEs make this task challenging: MWEs exhibit idiosyncrasies on a variety of levels, orthographic, morphological, syntactic and of course semantic (Al-Haj, 2010). They are also extremely diverse: for example, on the semantic dimension alone, MWEs cover an entire spectrum, ranging from frozen, fixed idioms to free combinations of words (Bannard et al., 2003).

Such a complex task calls for a combination of multiple approaches, and much research indeed suggests "hybrid" approaches to MWE identification

[2] For simplicity, we focus on bi-grams of tokens (MWEs of length 2) in this work; the methodology, however, is easily extensible to longer n-grams.

(Duan et al., 2009; Weller and Fritzinger, 2010; Ramisch et al., 2010; Hazelbeck and Saito, 2010). We believe that Bayesian Networks provide an optimal architecture for expressing various pieces of knowledge aimed at MWE identification, for the following reasons (Heckerman, 1995):

- In contrast to many other classification methods, BN can learn (and express) causal relationships between features. This facilitates better understanding of the problem domain.
- BN can encode not only statistical data, but also prior domain knowledge and human intuitions, in the form of interdependencies among features. We do indeed use this possibility here.

3.2 Linguistically-motivated Features

Based on the observations of Al-Haj (2010), we define several linguistically-motivated features that are aimed at capturing some of the unique properties of MWEs. While many idiosyncratic properties of MWEs have been previously studied, we introduce novel ways to express those properties as computable features informing a classifier. Note that many of the features we describe below are completely language-independent; others are applicable to a wide range of languages, while few are specific to morphologically-rich languages, and can be exhibited in different ways in different languages. The methodology we advocate, however, is completely universal.

A common theme for all these features is *idiosyncracy*: they are all aimed at locating some linguistic property on which MWEs may differ from non-MWEs. Below we detail these properties, along with the features that we define to reflect them. In all cases, the feature is applied to a *candidate MWE*, defined here as a bi-gram of tokens (all possible bi-grams are potential candidates). To compute the features, we use a 46M-token monolingual Hebrew corpus (Itai and Wintner, 2008), which we pre-process as in Tsvetkov and Wintner (2010b). All statistics are computed from this large corpus. Likewise, we compute these features on a small training corpus, which we generate automatically (see Section 3.4).

Orthographic variation Sometimes, MWEs are written with dashes instead of inter-token spaces. We define a binary feature, DASH, whose value is 1 iff the dash character appears in some surface form of the candidate MWE. For example, *xd-cddi* (*one sided*) "unilateral".

Hapax legomena MWEs sometimes include constituents that have no usage outside the particular expression, and are hence not included in lexicons. We define a feature, HAPAX, whose value is a binary vector with 1 in the i-th place iff the i-th word of the candidate is not in the lexicon, and does not occur in other bi-grams at the same location. For example, *hwqws pwqws* "hocus-pocus". In order to filter out potential errors, candidates must occur at least 5 times in the corpus in order for this feature to fire.

Frozen form MWE constituents sometimes occur in one fixed, frozen form. We define a feature, FROZEN, whose value is a binary vector with 1 in the i-th place iff the i-th word of the candidate never inflects in the context of this expression. Example: *bit xwlim* (*house-of sick-people*) "hospital"; the noun *xwlim* must be in the plural in this MWE.

Partial morphological inflection In some cases, MWE constituents undergo a (strict but non-empty) subset of the full inflections that they would undergo in isolation. We capture this property with a technique that has been proven useful in the area of image processing (Jain, 1989, Section 7.3). We compute a histogram of the distribution in the corpus of all the possible surface forms of each constituent of an MWE candidate. Such histograms can compactly represent distributional information on morphological behavior, in the same way that histograms of the distribution of gray levels in a picture are used to represent the picture itself.

Our assumption is that the inflection histograms of non-MWEs are more uniform than the histograms of MWEs, in which some inflections may be more frequent and others may be altogether missing. Of course, restrictions on the histogram may stem from the part of speech of the expression; such constraints are captured by dependencies in the BN structure.

Since each MWE is idiosyncratic in its own way, we do not expect the histograms of MWEs to have some specific pattern, except non-uniformity. We therefore sort the columns of each histogram, thereby losing information pertaining to the specific inflections, and retaining only information about the idiosyncrasy of the histogram. Offline, we compute the average histogram for positive and negative examples: The average histogram of MWEs is shorter and less uniform than the average histogram of non-MWEs. We define as feature, HIST, the L_1 (Manhattan) distance between the histogram of the candidate and the closest average histogram.

For example, the MWE *bit mepv* (*house-of law*) "court" occurs in the following inflected forms: *bit hmepv* "the court" (75%); *bit mepv* "a court" (15%); *bti hmepv* "the courts" (8%); and *bti mepv* "courts" (2%). The histogram for this candidate is thus $(75, 15, 8, 2)$. In contrast, the non-MWE *txwm mepv* (*domain-of law*) "domain of the law", which is syntactically identical, occurs in nine different inflected forms, and its sorted histogram is $(59, 14, 7, 7, 5, 2, 2, 2, 2)$.

Context We hypothesize that MWEs tend to constrain their (semantic) context more strongly than non-MWEs. We expect words that occur immediately after MWEs to vary less freely than words that immediately follow other expressions. One motivation for this hypothesis is the observation that MWEs tend to be less polysemous than free combinations of words, thereby limiting the possible semantic context in which they can occur.

We define a feature, CONTEXT, as follows. We first compute a histogram of the frequencies of words following each candidate MWE. We trim the tail of the histogram by removing words whose frequency is lower than 0.1% (the expectation is that non-MWEs would have a much longer tail). Offline, we compute the same histograms for positive and negative examples and average them as above. The value of CONTEXT is 1 iff the histogram of the candidate is closer (in terms of L_1 distance) to the positive average.

For example, the histogram of *bit mepv* "court" includes 15 values, dominated by *bit mepv yliwn* "supreme court" (20%) and *bit mepv mxwzi* "district court" (13%), followed by contexts whose frequency ranges between 5% and 0.6%. In contrast, the non-MWE *txwm mepv* "domain-of law" has a much shorter histogram, namely $(12, 11, 6)$: over 70% of the words following this expression occur less than 0.1% and are hence in the trimmed tail.

Syntactic diversity MWEs can belong to various part of speech categories. We define as feature, POS, the category of the candidate, with values obtained by selecting frequent tuples of POS tags. For example, Noun-Noun, PropN-PropN, Noun-Adj, etc.

Translational equivalents Since MWEs are often idiomatic, they tend to be translated in a non-literal way, sometimes to a single word. We use a dictionary to generate word-by-word translations of candidate MWEs to English, and check the number of occurrences of the English literal translation in a large English corpus.[3] Due to differences in word order between the two languages, we create two variants for each translation, corresponding to both possible orders. We expect non-MWEs to have some literal translational equivalent (possibly with frequency that correlates with their frequency in Hebrew), whereas for MWEs we expect no (or few) literal translations. We define a binary feature, TRANS, whose value is 1 iff some literal translation of the candidate occurs more than 5 times in the corpus.

For example, the MWE *htxtn ym* (*marry with*) "marry" is literally translated as *with marry, marry with, together marry* and *marry together*, none of which occurs in the corpus.

Collocation As a baseline, statistical association measure, we use a heuristic variant of pointwise mutual information (PMI), promoting also collocations whose constituents are frequent (Tsvetkov and Wintner, 2010b). We define a binary feature, PMI, with values (*low* and *high*) reflecting the threshold that maximizes the accuracy of MWE classification in Tsvetkov and Wintner (2010b).

3.3 Feature Interdependencies Expressed as a Bayesian Network

A Bayesian Network (Jensen and Nielsen, 2007) is organized as a graph whose nodes are random variables and whose edges represent interdependencies among those variables. We use a particular type of BN, known as *causal* networks, in which directed edges lead to a variable from each of its direct *causes*. This facilitates the expression of domain knowledge (and intuitions, beliefs, etc.) as structural properties of the network. We use the BN as

[3] We use a 120M-token newspaper corpus.

a classification device: training amounts to computing the joint probability distribution of the training set, whereas classification maximizes the posterior probability of the particular node (variable) being queried.

For MWE identification we define a BN whose nodes correspond to the features described in Section 3.2. In addition, we define a node MWE for the complete classification task. Over these nodes we impose the structure depicted graphically in Figure 1. This structure, which we motivate below, is *manually* defined: it reflects our understanding of the problem domain and is a result of thorough experimentations. That said, it can of course be modified in various ways, and in particular, new nodes can be easily added to reflect additional features.

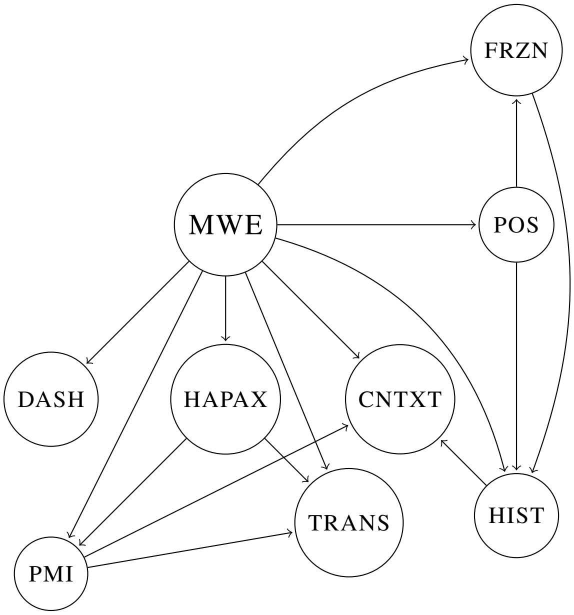

Figure 1: Bayesian Network for MWE identification

All nodes depend on MWE, as all are affected by whether or not the candidate is a MWE. The POS of an expression influences its morphological inflection, hence the edges from POS to HIST and to FROZEN. For example, Hebrew noun-noun constructions allow their constituents to undergo the full inflectional paradigm, but when such a construction is a MWE, inflection is severely constrained (Al-Haj and Wintner, 2010); similarly, when one of the constituents of a MWE is a conjunction, the entire expression is very likely to be frozen.

Hapaxes clearly affect all statistical metrics, hence the edge from HAPAX to PMI, and also the existence of literal translation, since if a word is not in the lexicon, it does not have a translation, hence the edge from HAPAX to TRANS. Also, we assume that there is a correlation between the frequency (and PMI) of a candidate and whether or not a literal translation of the expression exists, hence the edge from PMI to TRANS. The edges from PMI and HIST to CONTEXT are justified by the correlation between the frequency and variability of an expression and the variability of the context in which it occurs.

Once the structure of the network is established, the conditional probabilities of each dependency have to be determined. We compute the conditional probability tables from our training data (see below) using Weka (Hall et al., 2009), and obtain values for $P(X \mid X_1, \ldots, X_k)$ for each variable X and all variables X_i, $1 \leq i \leq k$, such that the graph includes an edge from X_i to X (parents of X). We then perform inference on the network in order to compute $P(X_{mwe} \mid X_1, \ldots, X_k)$, where X_{mwe} corresponds to the node MWE, and $X_1, \ldots, X_k$ are the variables corresponding to all *other* nodes in the network. Using Bayes Rule,

$$P(X_{mwe} \mid X_1, \ldots, X_k) \propto \\ P(X_1, \ldots, X_k \mid X_{mwe}) \times P(X_{mwe})$$

We define the prior, $P(X_{mwe})$, to be 0.41: this is the percentage of MWEs in WordNet 1.7 (Fellbaum, 1998). The conditional probabilities $P(X_1, \ldots, X_k \mid X_{mwe})$ are determined by Weka from the conditional probability tables:

$$P(X_1, \ldots, X_k \mid X_{mwe}) = \Pi_{i=1}^{k} P(X_i \mid \mathbf{pa}_i)$$

where k is the number of nodes in the BN (other than X_{mwe}) and $\mathbf{pa}_i$ is the set of parents of X_i.

3.4 Automatic Generation of Training Data

For training we need samples of positive and negative instances of MWEs, each associated with a vector of the values of all features discussed in Section 3.2. We generate this training material automatically. We use a small Hebrew-English bilingual corpus (Tsvetkov and Wintner, 2010a). We word-align the corpus with Giza++ (Och and Ney, 2003), and then apply the (completely unsupervised) algorithm of Tsvetkov and Wintner (2010b), which extracts MWE candidates from the aligned corpus and re-ranks them using statistics computed from a large monolingual corpus. The core idea behind this method is that MWEs tend to be translated in non-literal ways; in a parallel corpus, words that are 1:1 aligned typically indicate literal translations and are hence unlikely constituents of MWEs.

The result is a set of 134,001 Hebrew bi-gram types (from the bilingual corpus), classified as either 1:1 aligned (implying they are likely *not* MWEs) or unaligned (in which case they may or may not be MWEs). In addition, for each bi-gram we have a PMI score; naturally, higher PMI scores are indicative of MWEs. We thus divide the set into four classes: aligned bi-grams with high PMI score, aligned bi-grams with low PMI score, misaligned with high PMI and misaligned with low PMI. Aligned bi-grams, independently of their PMI score, are more likely non-MWEs; high-PMI misaligned bi-grams are very likely MWEs; and the status of low-PMI misaligned bi-grams is unclear, and must be further investigated. This is summarized in Table 1.

	Misaligned	Aligned
High PMI	MWE	non-MWE
Low PMI	unclear	non-MWE

Table 1: Classification of bi-grams

We set the threshold that separates low PMI from high PMI as in Tsvetkov and Wintner (2010b). The results of this classification is depicted in Table 2.

	Misaligned	Aligned	Total
High PMI	2,203	493	2,696
Low PMI	61,314	69,991	131,305
Total	63,517	70,484	134,001

Table 2: Statistics of the sample space from which the training set is generated

We assume that all bi-grams in the 'Aligned' column are non-MWEs. Additionally, we assume that the 2,203 misaligned bi-grams with high PMI scores are likely MWEs. As for the set of over 61,000 misaligned low-PMI bi-grams, certainly many of them are non-MWEs, but some may be MWEs, and we are interested in including them as positive examples of MWEs with low PMI scores. We therefore manually annotate a sample of 50 MWEs from this particular set (we had to manually go over a few thousands of bi-grams to select this sample). This is the only supervision provided in this work.

The remaining question is how to determine the sizes of samples from each of the other three classes. We use two guidelines: first, we would like the ratio of MWEs to non-MWEs in the training set to be $41 : 59$, reflecting the ratio in WordNet (the prior MWE probability). Second, we would like classification by PMI score only to yield a reasonable baseline; the baseline is defined as the ratio of the sum of high-PMI MWEs plus low-PMI non-MWEs to the size of the training set. We choose 67%, the PMI baseline reported by Al-Haj and Wintner (2010). As a result of these two considerations, we end up with training sets whose sizes are depicted in Table 3. We randomly select from the sample space this many instances for each class. Since much of the procedure of preparing training data is automatic, the results may be somewhat noisy. As Bayesian Network are known to be robust to noisy data, we expect the BN to compensate for this problem.

	MWE	non-MWE	Total
High PMI	300	232	532
Low PMI	50	272	322
Total	350	504	854

Table 3: Sizes of each training set

4 Results and Evaluation

We use the training set described above for training and evaluation: we perform 10-fold cross validation experiments, reporting Precision, Recall, Accuracy and F-measure in three setups: one (SVM) in which we train an SVM classifier[4] with the features described in Section 3.2; one (BN-auto) in which we train a BN but let Weka determine its structure (using the K2 algorithm); and one (BN) in which we train a Bayesian Network whose structure reflects manually-crafted linguistically-motivated knowledge, as depicted in Figure 1. The

[4] We use Weka SMO with the PolyKernel setup; experimentation with several other kernels yielded worse results.

results, along with the PMI baseline figures, are listed in Table 4.

	Accuracy	Prec.	Recall	F-score
PMI	66.98%	0.73	0.67	0.67
BN-auto	71.19%	0.71	0.71	0.71
SVM	74.59%	0.75	0.75	0.75
BN	**76.82%**	**0.77**	**0.77**	**0.77**

Table 4: 10-fold cross validation evaluation results

The linguistically-motivated features defined in Section 3.2 are clearly helpful in the classification task: the accuracy of the SVM, informed by these features, is close to 75%, reducing the error rate of the PMI baseline by 23%. The contribution of the Bayesian Network is also highly significant, reducing almost 7% more errors (8.7% of the errors made by the SVM classifier), or a total of almost 30% error-rate reduction with respect to the baseline. Interestingly, a BN whose structure does not reflect prior knowledge, but is rather learned automatically, performs poorly. It is the combination of linguistically-motivated features with feature interdependencies reflecting domain knowledge that contribute to the best performance.

As a further demonstration of the utility of our approach, we evaluate the algorithm on an additional test set that was used for evaluation in the past (Tsvetkov and Wintner, 2010b; Al-Haj and Wintner, 2010). This is a small annotated corpus, **NN**, of Hebrew noun-noun constructions. The corpus consists of 413 high-frequency bi-grams of the same syntactic construction; of those, 178 are tagged as MWEs (in this case, noun compounds) and 235 as non-MWEs. This corpus consolidates the annotation of three annotators: only instances on which all three agreed were included. Since it includes both positive and negative instances, this corpus facilitates a robust evaluation of precision and recall.

We train a Bayesian Network on the training set described in Section 3.4 and use it to classify the set **NN**. We compare the results of this classifier with a PMI baseline (using the same threshold as above), and also with the classification results reported by Al-Haj and Wintner (2010) (AW); the latter reflects 10-fold cross-validation evaluation using the entire set, so it should be considered an upper bound for any classifier that uses a general training corpus.

The results are depicted in Table 5. They clearly demonstrate that the linguistically-motivated features we define provide a significant improvement in classification accuracy over the baseline PMI measure. Note that our F-score, 0.77, is very close to the best result of 0.79 obtained by Al-Haj and Wintner (2010) as the average of 10-fold cross validation runs, using *only* high-frequency noun-noun constructions for training. We interpret this result as a further proof of the robustness of our architecture.

	Accuracy	Precision	Recall	F-score
PMI	71.43%	0.71	0.71	0.71
BN	**77.00%**	**0.77**	**0.77**	**0.77**
AW	80.77%	0.77	0.81	0.79

Table 5: Evaluation results: noun-noun constructions

Finally, we have used the trained BN to classify the entire set of bi-grams present in the (Hebrew side of the) parallel corpus described in Tsvetkov and Wintner (2010a). Of the 134,000 candidates, only 4,000 are classified as MWEs. We sort this list of potential MWEs by the probability assigned by the BN to the positive value of the variable X_{mwe}. The resulting sorted list is dominated by high-PMI bi-grams, especially proper names, all of which are indeed MWEs. The first non-MWE (false positive) occurs in the 50th place on the list; it is *crpt niqwla* “France Nicolas”, which is obviously a sub-sequence of the larger MWE, *neia crpt niqwla srqwzi* “French president Nicolas Sarkozy”. Similar sub-sequences are also present, but only five are in the top-100. Such false positives can be reduced when longer MWEs are extracted, as it can be assumed that a sub-sequence of a longer MWE does not have to be identified. Other false positives in the top-100 include some highly frequent expressions, but over 85 of the top-100 are clearly MWEs.

While more careful evaluation is required in order to estimate the rate of true positives in this list, we trust that the vast majority of the positive results are indeed MWEs.

5 Conclusions and future work

We presented a novel architecture for identifying MWEs in text corpora. The main insights we emphasize are sophisticated computational encoding of linguistic knowledge that focuses on the idiosyncratic behavior of such expressions. This is reflected in two ways in our work: by defining computable features that reflect different facets of irregularities; and by framing the features as part of a larger Bayesian Network that accounts for interdependencies among them. We also introduce a method for automatically generating a training set for this task, which renders the classification almost entirely unsupervised. The result is a nearly-unsupervised, language-independent classification method that can identify MWEs of various lengths, types and constructions. Evaluation on Hebrew shows significant improvement in the accuracy of the classifier compared with the state of the art.

The modular architecture of BN facilitates easy exploration with more features. We are currently investigating the contribution of various other sources of information to the classification task. For example, Hebrew lacks large-scale lexical semantic resources. However, it is possible to literally translate a MWE candidate to English and rely on the English WordNet for generating synonyms of the literal translation. Such "literal synonyms" can then be back-translated to Hebrew. The assumption is that if a back-translated expression has a high PMI, the original candidate is very likely not a MWE. While such a feature may contribute little on its own, incorporating it in a well-structured BN may improve performance.

While our methodology is applicable to MWEs of any length, we have so far only evaluated it on bigrams. In the future, we intend to extend the evaluation to longer n-grams. We also plan to apply the methodology to languages other than Hebrew.

Acknowledgments

This research was supported by THE ISRAEL SCIENCE FOUNDATION (grants No. 137/06, 1269/07). We are grateful to Gennadi Lembersky for his continuous help.

References

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. 2010. Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition. Master's thesis, University of Haifa, February.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

Pável Calado, Marco Cristo, Edleno Silva De Moura, Nivio Ziviani, Berthier A. Ribeiro-Neto, and Marcos André Gonçalves. 2003. Combining link-based and content-based methods for web document classification. In *Proceedings of CIKM-03, 12th ACM International Conference on Information and Knowledge Management*, pages 394–401, New Orleans, US. ACM Press, New York, US.

Baobao Chang, Pernilla Danielsson, and Wolfgang Teubert. 2002. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing*, pages 1–5, Morristown, NJ, USA. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Ludovic Denoyer and Patrick Gallinari. 2004. Bayesian network model for semi-structured document classification. *Information Processing and Management*, 40(5):807–827.

Jianyong Duan, Mei Zhang, Lijing Tong, and Feng Guo. 2009. A hybrid approach to improve bilingual multiword expression extraction. In Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 541–547. Springer, Berlin and Heidelberg.

Britt Erman and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text*, 20(1):29–62.

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.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

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

Gregory Hazelbeck and Hiroaki Saito. 2010. A hybrid approach for functional expression identification in a japanese reading assistant. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 81–84, Beijing, China, August. Coling 2010 Organizing Committee.

David Heckerman. 1995. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, March.

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

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, USA.

Anil K. Jain. 1989. *Fundamentals of digital image processing*. Prentice-Hall, Inc., NJ, USA.

Finn V. Jensen and Thomas D. Nielsen. 2007. *Bayesian Networks and Decision Graphs*. Springer, 2nd edition.

Wai Lam, Kon F. Low, and Chao Y. Ho. 1997. Using a bayesian network induction approach for text categorization. In Martha E. Pollack, editor, *Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence*, pages 745–750, Nagoya, JP. Morgan Kaufmann Publishers, San Francisco, US.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*.

Leonid Peshkin, Avi Pfeffer, and Virginia Savova. 2003. Bayesian nets in syntactic categorization of novel words. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2*, NAACL '03, pages 79–81, Morristown, NJ, USA. Association for Computational Linguistics.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech and Language*, 19(4):378–397.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Alline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*.

Carlos Ramisch, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria Finatto. 2010. A hybrid approach for multiword expression identification. In Thiago Pardo, António Branco, Aldebaro Klautau, Renata Vieira, and Vera de Lima, editors, *Computational Processing of the Portuguese Language*, volume 6001 of *Lecture Notes in Computer Science*, pages 65–74. Springer.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

Virginia Savova and Leonid Peshkin. 2005. Dependency parsing with dynamic bayesian network. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, pages 1112–1117. AAAI Press.

Yulia Tsvetkov and Shuly Wintner. 2010a. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3389–3392. European Language Resources Association (ELRA), May.

Yulia Tsvetkov and Shuly Wintner. 2010b. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, August.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, pages 590–596.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043.

Marion Weller and Fabienne Fritzinger. 2010. A hybrid approach for the identification of multiword expressions. In *Proceedings of the SLTC 2010 Workshop on Compounds and Multiword Expressions*, October.