Combining Linguistic Features for the Detection of Croatian Multiword Expressions

Maja Buljan and **Jan Šnajder**
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{maja.buljan,jan.snajder}@fer.hr

Abstract

As multiword expressions (MWEs) exhibit a range of idiosyncrasies, their automatic detection warrants the use of many different features. Tsvetkov and Wintner (2014) proposed a Bayesian network model that combines linguistically motivated features and also models their interactions. In this paper, we extend their model with new features and apply it to Croatian, a morphologically complex and a relatively free word order language, achieving a satisfactory performance of 0.823 F1-score. Furthermore, by comparing against (semi)naïve Bayes models, we demonstrate that manually modeling feature interactions is indeed important. We make our annotated dataset of Croatian MWEs freely available.

1 Introduction

Multiword expressions (MWEs) have attracted a great deal of attention in the natural language processing community. While MWEs span a wide range of types, common to all is the idiosyncrasy at the lexical, syntactic, semantic, pragmatic, or statistical level (Baldwin and Kim, 2010). A variety of models has been proposed for the automatic identification of MWE in corpora, including statistical (Church and Hanks, 1990; Lin, 1999; Pecina, 2010) and linguistic-based approaches (Cook et al., 2007; Baldwin, 2005; Green et al., 2011); see (Ramisch, 2015) for a recent overview. Sag et al. (2002) argued for a combination of the two approaches.

Recently, Tsvetkov and Wintner (2014) proposed an approach for the detection of MWE candidates that combines a number of statistical and linguistic features. The most interesting aspect of their work is that they explicitly model the linguistically motivated interactions between the features using a Bayesian network (BN). The advantages of BNs lie in their interpretability and the possibility to encode linguistic knowledge in the form of the network structure. Furthermore, unlike most previous work, Tsvetkov and Wintner address MWE of various types and flexible syntactic constructions. They show that the manually-designed BN outperforms a number of strong baselines, including an SVM model, on English, French, and Hebrew datasets. Another advantage of their model is that it is in principle language-independent, aside from a few language-specific features.

In this paper, we address the task of MWE detection (type-level MWE classification) for Croatian, a South Slavic language with a rich morphology and a relatively free word order. The starting point of our work is the model of Tsvetkov and Wintner (2014), which we extend with a number of features, including language-specific ones that account for the relatively free word order. Our main research question is whether modeling the interactions between features is important, and whether these can be learned automatically. Tsvetkov and Wintner (2014) showed that a manually-designed BN substantially outperforms the one whose structure is learned automatically, hypothesizing that the cause for this might be the increased model complexity. We conduct a similar experiment using a structure-learning algorithm, but also model the interactions using a simpler, semi-naive Bayes classifier, for which the number of parameters is restricted. Finally, we compare these models against a structure-free counterpart, a naïve Bayes classifier.

For the experiments, we compile a new manually annotated dataset of Croatian MWEs. Unlike Tsvetkov and Wintner (2014), who only consider bigrams, we consider MWEs of up to five words in length. We make the dataset freely available, along with all feature sets needed to replicate the experiments.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 194–199,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

2 Model

We adopt the BN model of Tsvetkov and Wintner (2014), but extend it with language-specific as well as semantically motivated features. Most newly added features were inspired by the analysis of Croatian MWEs of Blagus Bartolec (2008), and a sample-based analysis of a MWE from a dictionary of Croatian MWEs (Kovačević, 2012) and their occurrences in the hrWaC corpus (Ljubešić and Erjavec, 2011). The MWE candidates were POS-tagged using the tagger from (Pinnis et al., 2012).

2.1 Features

Original features. The model of Tsvetkov and Wintner (2014) uses nine statistically and linguistically motivated features, computed for each MWE candidate and designed to discriminate between MWEs and ordinary word sequences. We adopted eight of these features:[1] (1) *capitalization* (indicating which MWE constituents are capitalized), (2) *hyphenation* (which constituents are hyphenated), (3) *fossil word* (whether constituents also occur outside of the MWE), (4) *frozen form* (whether the MWE is morphologically frozen), (5) *partial morphological inflection* (whether MWE admits only limited inflection), (6) *syntactic pattern* (the MWE's part-of-speech pattern), (7) *semantic context*, and (8) *association measure*.

The values of statistical features were computed from hrWaC, a 1.2B-token Croatian web corpus compiled by Ljubešić and Erjavec (2011). All numeric features were discretized into five reference levels based on their average values in the corpus.

Interesting MWE examples from the corpus that showcase the above-mentioned statistical properties are *curriculum vitae*, which is made of fossil words, *hodati po jajima* (*to walk on eggshells*), which is a frozen form, and *zlatno doba* (*golden age*), which almost exclusively appears in the nominative and locative singular (partial inflection).

Modified features. In the original model, the semantic context feature computes the lexical variety of the words following a MWE candidate vary, the idea being that MWEs have a more restricted context. In our sample-based analysis of Croatian MWEs, we concluded that in many cases this restriction is not limited to the right context. Thus, we introduced two additional features: one for the left context and another considering a 5-word window around the MWE. Likewise, we used the Dice coefficient association measure, rather than PMI as used in the original model, as the former turned out to be more discriminative.

New features. We introduced six new features, four of which were inspired by our analysis of Croatian MWEs. The *simile* feature is motivated by the observation that many Croatian similes are MWEs, e.g., *plakati kao ljuta godina*, (*to cry like a bitter year* – to cry heavily). We consider a MWE to be a simile if it contains a preposition *kao* (*like*) or *poput* (*as*). We furthermore observe many Croatian MWEs contain loanwords. The *foreign word* feature indicates, for each MWE constituent, whether it has been tagged as a foreign word by the POS tagger.

We also introduced two features to account for the relatively free word order of Croatian: *constituent adjacency* and *constituent permutation*. The former is turned on if there are more contiguous than discontinuous MWE candidate occurrences, while the latter is turned on if the corpus contains five or more word permutations of the MWE candidate. While most MWEs in Croatian nominally do not allow intervening words between its components, in fact most types of MWEs will allow the insertion of copula and pronoun enclitics; e.g., *zadnji [je] čas* (*[is] last moment*). When searching for discontinuous MWE candidates of length n, we only consider n-grams for which the number of tokens between the first and final constituent is less than or equal to $2n$. On the other hand, permutation of MWE constituents is much less frequent, even for a relatively free word order language such as Croatian. Thus, there may be a benefit to capturing which types of MWE – presumably mostly characterized by their POS patterns – allow for permutations; e.g., *jednim udarcem ubiti dvije muhe / dvije muhe ubiti jednim udarcem, etc. (to kill two flies with one stone).*

Finally, inspired by a growing body of research on semantic non-compositionality of MWEs (Baldwin et al., 2003; Kim and Baldwin, 2006; Biemann and Giesbrecht, 2011; Krčmář et al., 2013), we introduced a simple *semantic opacity* feature. We opted for a simple approach proposed by (Mitchell and Lapata, 2008), and computed this feature by deriving distributional vectors from hrWaC for the MWE and the additive composition of its con-

[1] We omitted a feature that indicates the existence of a translation equivalent. Namely, Tsvetkov and Wintner (2014) use parallel bilingual corpora for acquiring the initial MWE candidates.

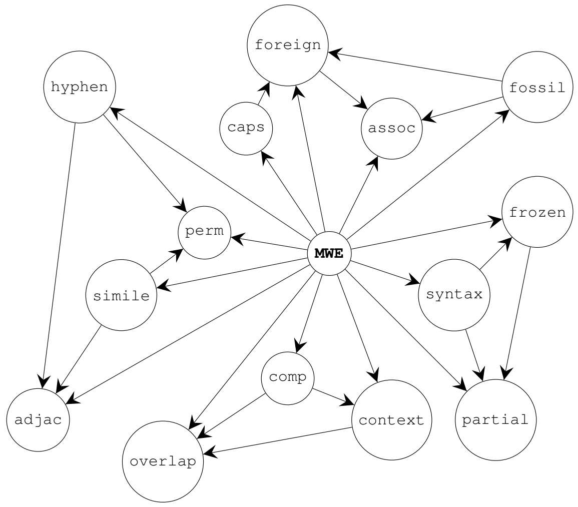

Figure 1: Bayesian network for MWE classification

stituents, and then computing the cosine between the two vectors. For opaque MWEs, we expect the cosine to be lower than for semantically transparent MWEs. Similarly as with other numeric features, we discretized the cosine scores into five levels.

2.2 Feature Interactions

The structure of a BN defines feature interactions by means of conditional independence assumptions between the variables. When constructed manually, the structure of the network essentially models our knowledge about the causal links between the features.

We extended the structure of the original BN model by introducing additional links for the newly added features. We primarily based our design choices on linguistic intuition, but also on experimental validation. To this end, we compiled a small validation set of 33 MWEs and 33 non-MWEs, for which we computed the features over 50K sentences from hrWaC. We used this dataset to verify whether adding an interaction link improves the accuracy of the model.

The resulting BN is shown in Fig. 1. All nodes depend on the `MWE` node, which is the label to be predicted.[2] We introduced feature interaction between the `caps` and `foreign` node, given that a high number of loanwords pertain to proper names. Additionally, we defined interactions between `comp`, `context`, and `overlap`, as the semantic opacity influences the general context of an expression, and the ratio of overlapping context words depends upon both features. Finally, since similes and hyphenated expressions signal a strict word order, we defined interactions between `perm`, `adjac`, `hyphen`, and `simile`.

3 Dataset

MWE definition. As there is no publicly available annotated datasets of Croatian MWEs, we decided to create one. We first established a working definition of Croatian MWEs, starting out from the taxonomy proposed by Blagus Bartolec (2008), and adopted it to the universal classification of Sag et al. (2002). We identified five major groups of MWEs: (1) *idioms*, semantically opaque expressions; (2) *fixed expressions*, common phrases whose meaning can clearly be gleaned from its constituents, but whose constituents are rarely replaced with synonyms in practice; (3) *technical terms*, expressions pertaining to the technical language of a particular profession; (4) *foreign terms*, any expression adopted from another language, as well as imaginary and nonsensical phrases; and (5) *proper names*, names of persons, institutions, geographical terms, etc., composed of two or more words.

Annotation. As a source of data for our dataset, we use hrMWELex, a lexicon of Croatian MWEs candidate n-grams compiled by Ljubešić et al. (2015). The lexicon was obtained by matching parse trees from hrWaC against a set of predefined syntactic patterns (POS patterns) for Croatian, yielding a high-recall, low-precision MWE lexicon. The resulting lexicon contains 12M n-grams with matching POS patterns.

We next sorted the n-grams by corpus frequency, and made a balanced 2-, 3-, and 4-gram selection from the most frequent candidates, selecting 4000 MWE candidates. We then asked four native speakers of Croatian to label the dataset. Each annotated all 4000 instances, presented in random order to minimize the effect of a context bias. We also included 124 gold positive MWEs, extracted from (Anić, 2003), to serve as a control set.

To measure the inter-annotator agreement, we calculated the Cohen's coefficient (Cohen, 1960) between all pairs of annotations (Table 1). The agreement ranges between 0.413 and 0.578, which, according to Landis and Koch (1977), is considered a moderate agreement.

Gold dataset. For the final dataset, we adjudicated the annotations by considering a MWE can-

[2] When using the BN model for MWE detection, we simply run a maximum a posteriori query on the `MWE` variable with all feature variables set to the observed values.

$\kappa(x, y)$	A	B	C	D
A	–	0.499	0.505	0.578
B	0.499	–	0.420	0.466
C	0.505	0.420	–	0.413
D	0.578	0.466	0.413	–

Table 1: Inter-annotator agreement on the MWE classification

	n-gram length				
	2	3	4	5	Total
Positive	338	76	44	3	461
Negative	233	150	78	–	461
Total	571	226	122	3	**922**

Table 2: Dataset breakdown by n-gram length

didate to be a true MWE if at least three annotators have labeled it as positive. Out of 4124 MWE candidates, 111 MWEs were labeled as positive by all four annotators, while 163 were labeled as positive by three annotators. To this set we add 187 positive MWEs extracted from a standard Croatian dictionary (Anić, 2003) and a dictionary of multiword expressions (Kovačević, 2012), yielding a total of 461 positive MWEs.[3] Finally, we add an equal number of n-grams annotated as negative MWE instances by at least three annotators, yielding a perfectly-balanced dataset of 922 n-grams. Table 2 shows a breakdown of positive and negative examples by n-gram length. For each n-gram from this dataset, we computed the feature values on a random sample of the hrWaC corpus comprising 200K sentences (∼5M tokens). We make the dataset and the precomputed features publicly available.[4]

4 Evaluation

We compare the BN model from Section 2 against two commonly used statistical baselines: Dice and PMI association measures. Furthermore, we compare the BN model to three variants of Bayes classifiers, differing in their ability to model feature interactions: a Naive Bayes classifier (NB), a tree-augmented Naive Bayes classifier (TAN) (Friedman et al., 1997), and a Bayesian network classifier trained using the K2 structure learning algorithm (BN-K2) (Cooper and Herskovits, 1992). The NB and TAN allow for no feature interaction or limited feature interaction, respectively. More precisely, a TAN cannot model circular feature dependencies, such as those among the `syntax`, `frozen`, and `partial` features in Fig. 1. The NB is even simpler, as it does not model any feature interactions at all, i.e., it assumes all feature pairs are conditionally independent within the MWE and non-MWE classes. In contrast, the BN and BN-K2 models can model (undirected) circular dependencies. The difference between them is that for the BN model the feature interactions were designed manually, based on linguistic insights, whereas in case of BN-K2 the interactions are learned from the train set.

	Acc	P	R	F_1
Dice	0.735	0.788	0.788	0.788
PMI	0.717	0.777	0.769	0.773
NB	0.783	0.795	0.761	0.778
TAN	0.804	0.808	**0.796**	0.805
BN-K2	0.809	0.850	0.751	0.797
BN	**0.832**	**0.867**	0.783	**0.823**

Table 3: Performance of Bayes classifiers and the baselines (scores averaged over ten folds)

Table 3 shows the MWE classification accuracy, precision, recall, and F1-scores of the two baselines and the four Bayes classifiers. All models were trained and tested using 10-fold cross-validation on the gold dataset. The threshold of the two baseline models was optimized on the train sets. We observe that all four Bayes classifiers outperform the baselines in terms of accuracy and F1-score, except for the NB model which performs worse than Dice in terms of F1-score. On the other hand, the BN model outperforms all considered models in terms of both accuracy and F1-score by a considerable margin. This demonstrates that manual modeling of feature interactions is indeed important for MWE detection, and that BN does a reasonably good job in modeling these interactions. The more simple NB and TAN models even out in terms of F1-score, but differ in precision and recall scores, while the BN-K2 model performs comparably to TAN.

5 Conclusion

We described the experiments on using a combination of linguistically motivated features for MWE detection in Croatian. We adopted the Bayesian network model of Tsvetkov and Wintner (2014) and extended it with new features and manually-designed feature interactions, inspired by an analysis of Croatian MWEs. To train and evaluate the model, we built a manually annotated

[3] We took care not to select any MWEs from the samples we used for designing the features or feature interactions.

[4] `http://takelab.fer.hr/cromwe`

dataset of Croatian MWEs. On this dataset, our model substantially outperforms statistical baselines, reaching a satisfactory performance of 0.823 F1-score on our dataset. The model also outperforms the (semi)naïve Bayes models, which limit the feature interactions, as well as a Bayesian network model with automatically learned feature interactions. Thus, the main finding of our work is that the model benefits from the linguistically motivated, manually-designed feature interactions, which proves that MWE features interact in rather intricate ways.

Acknowledgments

This work has been supported by the Croatian Science Foundation under the project UIP-2014-09-7312. We thank the annotators for their work and the anonymous reviewers for their insightful comments. We thank Nikola Ljubešić for his help with the hrMWELex lexicon.

References

V. Anić. 2003. *Veliki rječnik hrvatskoga jezika*. Novi Liber.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

Timothy Baldwin. 2005. Deep lexical acquisition of verb–particle constructions. *Computer Speech & Language*, 19(4):398–414.

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

Goranka Blagus Bartolec. 2008. Kolokacijske sveze u hrvatskom jeziku (s posebnim osvrtom na leksikografiju).

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48. Association for Computational Linguistics.

Gregory F Cooper and Edward Herskovits. 1992. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163.

Spence Green, Marie-Catherine De Marneffe, John Bauer, and Christopher D Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–735. Association for Computational Linguistics.

Su Nam Kim and Timothy Baldwin. 2006. Automatic identification of English verb particle constructions using linguistic features. In *Proc. of the Third ACL-SIGSEM Workshop on Prepositions*, pages 65–72. Association for Computational Linguistics.

Barbara Kovačević. 2012. *Hrvatski frazemi od glave do pete*. Institut za hrvatski jezik i jezikoslovlje.

Lubomír Krčmář, Karel Ježek, and Pavel Pecina. 2013. Determining compositionality of word expresssions using various word space models and methods. In *Proc. of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. Association for Computational Linguistics.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.

Nikola Ljubešić, Kaja Dobrovoljc, and Darja Fišer. 2015. MWELex – MWE lexica of Croatian, Slovene and Serbian extracted from parsed corpora. *Informatica*, 39(3):293.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

Marcis Pinnis, Nikola Ljubešić, Dan Stefanescu, Inguna Skadina, Marko Tadic, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, pages 20–21.

Carlos Ramisch. 2015. State of the art in mwe processing. In *Multiword Expressions Acquisition*, pages 53–102. Springer.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

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