Impact of MWE Resources on Multiword Recognition

Martin Riedl and **Chris Biemann**
Language Technology
Computer Science Department
Technische Universität Darmstadt
{riedl,biem}@cs.tu-darmstadt.de

Abstract

In this paper, we demonstrate the impact of Multiword Expression (MWE) resources in the task of MWE recognition in text. We present results based on the Wiki50 corpus for MWE resources, generated using unsupervised methods from raw text and resources that are extracted using manual text markup and lexical resources. We show that resources acquired from manual annotation yield the best MWE tagging performance. However, a more fine-grained analysis that differentiates MWEs according to their part of speech (POS) reveals that automatically acquired MWE lists outperform the resources generated from human knowledge for three out of four classes.

1 Introduction

Identifying MWEs in text is related to the task of Named Entity Recognition (NER). However, the task of MWE recognition mostly considers the detection of word sequences that form MWEs and are not Named Entities (NEs). For both tasks mostly sequence tagging algorithms, e.g. Hidden Markov Model (HMM) or Conditional Random Fields (CRF), are trained and then applied to previously unseen text. In order to tackle the recognition of MWEs, most approaches (e.g. (Schneider et al., 2014; Constant and Sigogne, 2011)) use resources containing MWEs. These are mostly extracted from lexical resources (e.g. WordNet) or from markup in text (e.g. Wikipedia, Wiktionary). While these approaches work well, they require respective resources and markup. This might not be the case for special domains or under-resourced languages.

On the contrary, methods have been developed that rank word sequences according to their multiwordness automatically using information from corpora, mostly relying on frequencies. Many of these methods (e.g. C/NC-Value (Frantzi et al., 1998), GM-MF (Nakagawa and Mori, 2002)) require previous filters, which are based on Part-of-Speech (POS) sequences. Such sequences, (e.g. Frantzi et al. (1998)) need to be defined and mostly do not cover all POS types of MWE.

In this work we do not want to restrict to specific MWE types and thus will use DRUID (Riedl and Biemann, 2015) and the Student's t-test as multiword ranking methods, which do not require any previous filtering. This paper focuses on the following research question: how do such lists generated from raw text compete against manually generated resources? Furthermore, we want to examine whether a combination of resources yields better performance.

2 Related Work

There is a considerable amount of research that copes with the recognition of word sequences, be it NE or MWE. The field of NER can be considered as subtask from the recognition of MWE. However, in NER additionally, single-worded names need to be recognized.

The experiments proposed in our paper are related to the ones performed by Nagy T. et al. (2011). Their paper focuses on the introduction of the Wiki50 dataset and demonstrates how the performance of the system can be improved by combining classifiers for NE and MWE. Here, we focus on the impact of different MWE resources.

An extensive evaluation of different measures for ranking word sequences regarding their multiwordness has been done before. Korkontzelos (2010) performs a comparative evaluation of MWE measures that all rely on POS filtering.

Riedl and Biemann (2015), in contrast, introduced a measure, relying on distributional similarities, that does not require a pre-filtering of candidate words by their POS tag. It is shown to compare favorably to an adaption of the t-test, which only relies on filtering of frequent words.

3 Datasets

For the evaluation we use the Wikipedia-based Wiki50 (Nagy T. et al., 2011) dataset. This dataset comprises of annotations for both NEs and MWEs as shown in Table 1.

MWE/NE	type	count
MWE	noun compound	2931
MWE	verb-particle construction	447
MWE	light-verb construction	368
MWE	adjective compound	78
MWE	other	21
MWE	idiom	19
NE	person	4099
NE	misc.	1827
NE	location	1562
NE	organization	1499

Table 1: Frequency of MWE types in the Wiki50 dataset.

The dataset primarily consists of annotations for NEs, especially for the person label. The annotated MWEs are dominated by noun compounds followed by verb-particle constructions, light-verb constructions and adjective compounds. Idioms and other MWEs occur only rarely.

4 Method

For detecting MWEs and NEs we use the CRF sequence-labeling algorithm (Lafferty et al., 2001). As basic features, we use a mixture of features used in previous work (Schneider et al., 2014; Constant and Sigogne, 2011). The variable i indicates the current token postion:

- token$_j$ with $j \in \{i-2, i-1, i, i+1, i+2\}$
- token$_j$ & token$_{j+1}$ with $j \in \{i-2, i-1, i, i+1, i+2\}$
- word shape of token$_i$, as used by Constant and Sigogne (2011)
- has token$_i$ digits
- has token$_i$ alphanumeric characters
- suffix of token$_i$ with length $l \in \{1, 2, 3, 4\}$
- prefix of token$_i$ with length $l \in \{1, 2, 3, 4\}$
- POS of token$_j$ with $j \in \{i-2, i-1, i, i+1, i+2\}$
- POS(token$_j$) & POS(token$_{j+1}$) with $j \in \{i-2, i-1, i, i+1, i+2\}$
- POS(token$_j$) & token$_j$ with $j \in \{i-2, i-1, i, i+1, i+2\}$
- lemma of token$_i$
- lemma of token$_j$ and lemma of token$_{j+1}$ with $j \in \{i-1, i\}$

For showing the impact of a MWE resource mr, we featurize the resource as follows:

- number of times token$_i$ occurs in mr
- token bigram: token$_j$ token$_{j+1}$ contained in mr with $j \in \{i-1, i\}$
- token trigram: token$_j$ token$_{j+1}$ token$_{j+2}$ occurence in mr with $j \in \{i-2, i-1, i\}$
- token 4-gram: token$_j$ token$_{j+1}$ token$_{j+2}$ token$_{j+3}$ occur in mr with $j \in \{i-3, i-2, i-1, i\}$

5 Multiword Expression Resources

For generating features from MWE resources, we distinguish between resources that are extracted from manually generated/annotated content[1] and resources that can be automatically computed based on raw text. First, we describe the resources extracted from manually annotated corpora or resources.

- **EnWikt**: This resource consists of 82,175 MWEs extracted from Wiktionary.
- **WordNet**: The WordNet resource is a list of 64,188 MWEs that are extracted from WordNet (Miller, 1995).
- **WikiMe**: WikiMe (Hartmann et al., 2012) is a resource extracted from Wikipedia that consists of 356,467 MWEs from length two to four that have been extracted using markup information.

[1]For this, we rely on the MWE resources that are provided here: `http://www.cs.cmu.edu/~ark/LexSem/mwelex-1.0.zip`.

- **SemCor**: This dataset consists of 16,512 MWE and was generated from the Semantic Concordance corpus (Miller et al., 1993).

Additionally, we select the best-performing measures for ranking word sequences according to their multiwordness as described in (Riedl and Biemann, 2015) that do not require any POS filtering:

- **DRUID**: We use the DRUID implementation[2], which is based on a distributional thesaurus (DT) and does not rely on any linguistic processing (e.g. POS tagging).
- **t-test**: The Student's t-test is a statistical test that can be used to compute the significance of the co-occurrence of tokens. For this it relies on the frequency of the single terms as well as the word sequence. As this measure favors to rank word sequences highest that begin and end with stopwords, we remove word sequences that begin and end with stopwords. As stopwords, we select the 100 most frequent words from the Wikipedia corpus.

6 Experimental Setting

We perform the evaluation, using a 10-fold cross validation and use the `crfsuite`[3] implementation of CRF as classifier. For retrieving POS tags, we apply the OpenNLP POS tagger[4]. The lemmatization is performed using the `WordNetLemmatizer`, contained in `nltk` (Loper and Bird, 2002).[5]

For the computation of automatically generated MWEs lists, we use the raw text from an English Wikipedia dump, without considering any markup and annotations. For applying them as resources, we only consider word sequences in the resource that are also contained in the Wiki50 dataset, both training and test data. Based on these candidates, we select the n highest ranked MWE candidates. The previous filtering does not influence the performance of the algorithm but enables an easier filtering parameter.

[2] `http://jobimtext.org/jobimtext/components/DRUID/`

[3] `http://www.chokkan.org/software/crfsuite`

[4] We use the version 1.6 available from: `https://opennlp.apache.org`.

[5] An implementation of the complete system is available at `http://maggie.lt.informatik.tu-darmstadt.de/files/mwe/MWE_TAGGER.tar.gz`.

7 Results

First, we show the overall performance for the Wiki50 dataset for recognizing labeled MWE and NE spans. We show the performance for training classifiers to predict solely NEs and MWEs and also the combination without the usage of any MWE resource. As can be observed (see Table 2), the detection of NE reaches higher scores than learning to predict MWE.

	precision	**recall**	**F1**
MWE +NE	80.83	75.29	77.96
MWE	77.51	57.89	66.28
NE	83.76	82.58	83.16

Table 2: Performance for predicting labels for MWE and NE without using MWE resources.

Comparing the performance between classifying solely NEs and MWEs, we observe low recall for predicting MWE. Next, we will conduct experiments for learning to predict MWE with the use of MWE resources.

In Table 3 we present results for the overall labeled performance for MWEs in the Wiki50 dataset. Using MWE resources, we observe consistent improvements over the baseline approach, which does not rely on any MWE resource (*None*). For manually constructed MWE resources, improvements of up to 3 points F1-measure on MWE labeling are observed, the most useful resource being WikiMe. The combination of manual resources does not yield improvements.

	precision	**recall**	**F1**
None	77.51	57.89	66.28
SemCor	78.28	59.78	67.79
WordNet	78.48	60.04	68.04
EnWikt	79.16	60.56	68.62
WikiMe	**79.35**	**61.54**	**69.32**
All resources	78.90	61.44	69.08
t-test 1,000	78.14	59.65	67.65
t-test 10,000	78.60	60.53	68.39
DRUID 1,000	78.42	60.30	68.18
DRUID 10,000	78.56	60.58	68.41
DRUID & t-test 10,000	78.56	60.30	68.23
All	79.06	60.79	68.73

Table 3: Overall performance on the labels for different MWE resources applied solely to the MWEs annotated in the Wiki50 dataset.

Using the top 1000 ranked word sequences that are contained in the Wiki50 corpus, we already obtain improvements for both unsupervised rank-

MWE	Noun Comp.			Verb-part. constr.			light-verb constr.			adj. comp.		
Resource	P	R	F1	P	R	F1	P	R	F1	P	R	F1
None	76.64	63.46	69.43	86.64	59.51	70.56	73.13	26.63	39.04	72.22	16.67	27.08
Semcor	77.25	65.23	70.74	86.83	61.97	72.32	76.34	27.17	40.08	78.26	23.08	35.64
WordNet	77.44	65.47	70.96	88.05	62.64	73.20	75.37	27.45	40.24	73.91	21.79	33.66
EnWikt	78.18	65.88	71.51	86.46	62.86	72.80	79.26	29.08	42.54	78.26	23.08	35.64
WikiMe	**78.41**	**67.28**	**72.42**	87.42	62.19	72.68	77.14	29.35	42.52	80.95	21.79	34.34
All resources	77.94	67.25	72.20	87.16	63.76	73.64	76.19	26.09	38.87	79.17	24.36	37.25
t-test 1,000	77.07	65.03	70.54	87.11	61.97	72.42	76.12	27.72	40.64	77.27	21.79	34.00
t-test 10,000	77.36	65.51	70.94	**88.20**	63.53	73.86	77.55	**30.98**	**44.27**	**81.82**	23.08	36.00
DRUID 1,000	77.30	65.64	71.00	87.97	62.19	72.87	77.37	28.80	41.98	74.07	**25.64**	**38.10**
DRUID 10,000	77.42	65.64	71.05	86.31	**64.88**	**74.07**	**79.70**	28.80	42.32	78.26	23.08	35.64
DRUID & t-test 10,000	77.60	65.37	70.96	86.50	63.09	72.96	76.55	30.16	43.27	78.26	23.08	35.64

Table 4: Detailed performance in terms of precision (P), recall (R) and F1-measure (F1) for the different MWE types. The experiments have been performed only on the MWE annotations.

ing measures. Whereas we observe improvements by around 1 points F1 for the t-test, we gain improvements of almost 2 points for DRUID. When extracting the top 10,000 MWEs, additional improvements can be obtained, which are close to the performances using the markup-based MWE resources. Here, using DRUID with the top 10,000 highest ranked MWEs achieves the third best improvements in comparison to all resources. Using more than the top 10,000 ranked word sequences does not result in any further performance improvement. Surprisingly, using MWE resources as features for MWE recognition improves the performance only marginally.

We assume that each resource focuses on different kinds of MWEs. Thus, we also show results for the four most frequent MWE types in Table 4. Inspecting the results using MWE lists, that are generated using human knowledge, we obtain the best performance for noun compounds using WikiMe. Verb-particle constructions seem to be better covered by the WordNet-based resource. For light-verb constructions the highest F1 measures are observed using EnWikt and WikiMe and for adjective compounds EnWikt achieves the highest improvements. We omit presenting results for the MWE classes other and idiom as only few annotations are available in the Wiki50 dataset.

Inspecting results for the t-test and DRUID, we obtain slightly higher F1 measures for noun-compounds using DRUID. Whereas for verb-particle constructions the t-test achieves the overall highest precision, recall and F1 measure of DRUID are higher. However, t-test achieves better results for light-verb constructions and using DRUID yields the highest F1 measure for adjective compounds.

Overall, only for noun compounds the best results are obtained using MWE lists that are generated from lexical resources or text annotations. For all remaining labels, the best performance is obtained using MWE lists that can be generated in an unsupervised fashion. However, as noun compounds constitutes the largest class, using unsupervised lists does not result to the best overall performance.

In addition, we performed the classification task of MWEs without labels, as shown in Table 5. In contrast to the overall labeled results (see Table 3) the performance drops. Whereas one might expect higher results for the unlabeled dataset, the labels help the classifier in order to use features according to the label. This is in accordance with the previous findings shown in Table 4.

	P	**R**	**F1**
None	74.47	58.20	65.34
SemCor	75.01	59.50	66.36
WordNet	75.32	59.47	66.46
EnWikt	**76.04**	60.35	**67.29**
WikiMe	75.78	**60.48**	67.27
All resources	76.07	61.44	67.97
t-test 1,000	74.89	58.59	65.75
t-test 10,000	75.81	60.20	67.11
DRUID 1,000	75.42	59.78	66.70
DRUID 10,000	75.17	**60.48**	67.03
DRUID & t-test 10,000	75.81	60.35	67.20
All	76.39	60.79	67.70

Table 5: Unlabeled results for MWEs recognition.

Furthermore, in this evaluation highest improvements are achieved with the EnWikt. Using MWE lists that are generated in an unsupervised fashion results in comparable scores to the EnWikt. Again, these resources have the third-

highest performance of all lists and outperform SemCor and WordNet.

8 Conclusion

In this paper, we have investigated whether unsupervisedly acquired MWE resources are comparable with knowledge-based or manual-annotation-based MWE resources for the task of MWE tagging in context. The highest overall performance, both for the labeled and unlabeled tagging task, is achieved using lists extracted from Wikipedia (WikiMe) and Wiktionary (EnWikt). However, for three out of four MWE types, resources that are extracted using unsupervised methods achieve the highest scores. In summary, using MWE lists for MWE recognition with sequence tagging is a feature that adds a few points in F-measure. In the case that high quality MWE resources exist, these should be used. If not, it is possible to replace them with unsupervised extraction methods such as the t-test or DRUID.

References

Matthieu Constant and Anthony Sigogne. 2011. MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World held in conjunction with ACL-2011*, pages 49–56, Portland, OR, USA.

Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. 1998. The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL 1998, pages 585–604, Heraklion, Greece.

Silvana Hartmann, György Szarvas, and Iryna Gurevych. 2012. Mining multiword terms from wikipedia. In *Semi-Automatic Ontology Development: Processes and Resources*, pages 226–258. IGI Global, Hershey, PA, USA.

Ioannis Korkontzelos. 2010. *Unsupervised Learning of Multiword Expressions*. Ph.D. thesis, University of York, UK.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML 2001, pages 282–289, Williams College, Williamstown, MA, USA.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, PA, USA.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 303–308, Princeton, New Jersey.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

István Nagy T., Gábor Berend, and Veronika Vincze. 2011. Noun Compound and Named Entity Recognition and their Usability in Keyphrase Extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 162–169, Hissar, Bulgaria.

Hiroshi Nakagawa and Tatsunori Mori. 2002. A Simple but Powerful Automatic Term Extraction Method. In *International Workshop on Computational Terminology held in conjunction with COLING-02*, COMPUTERM 2002, pages 1–7, Taipei, Taiwan.

Martin Riedl and Chris Biemann. 2015. A Single Word is not Enough: Ranking Multiword Expressions Using Distributional Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, pages 2430–2440, Lisboa, Portugal.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah Smith. 2014. Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.