Using a Rich Feature Set for the Identification of German MWEs

Fabienne Cap & Marion Weller
IMS
Universität Stuttgart
Pfaffenwaldring 5B
70174 Stuttgart
cap@ims.uni-stuttgart.de
wellermn@ims.uni-stuttgart.de

Ulrich Heid
IwiSt
Universität Hildesheim
Lübeckerstrasse 3
31141 Hildesheim
heid@uni-hildesheim.de

Abstract

Due to the formal variability and the irregular behaviour of MWEs on different levels of linguistic description, they are a potential source of errors for many NLP applications, e.g. Machine Translation. While most of the known approaches to MWE identification focus on one dimension of irregular behaviour, we present an approach that combines morpho-syntactic features (extracted from dependency parsed text) with semantic opacity features (approximated using word alignments). We trained supervised classifiers with different feature sub-sets and show that the combination of morpho-syntactic and semantic opacity features yields best overall results.

1 Introduction

The task of automatically identifying multiword expressions (MWEs) has gained considerable interest in NLP research in the past years (Sag et al., 2002).

Due to the formal variability and the irregular behaviour of MWEs on different levels of linguistic description, they are a potential source of errors for many NLP applications: consider for example Machine Translation, where MWEs with (partially) opaque semantics can hardly ever be transferred word by word to the target language:

(1)

zur Sprache bringen =	lit. "to bring to speech"
	idiom. "to address sth."

We present a method to identify German MWEs of the type preposition+noun+verb based on a rich feature set comprising morpho-syntactic features extracted from monolingual data and cross-lingual features obtained from word-aligned parallel data (DE-EN, DE-SE). The used features aim at modelling characteristic properties of MWEs, namely *fixedness* in terms of their disposition for variation with respect to e.g. number or type of determiner (morpho-syntactic features) and *irregular translational behaviour*, e.g. a broad variation in translational equivalents (cross-lingual features). Our experiments show that combining these different types of features leads to an improved classification accuracy.

Our approach consists of three main steps:

1. extraction of syntactically related multiword constructions
2. collect, sum and average feature values of all their occurrences
3. train a classifier on a hand-crafted dataset to distinguish unseen MWEs from regular combinations.

The remainder of this paper is structured as follows: In Section 2, we briefly describe our objectives and the specificities of MWE extraction for German. We give an overview of the morpho-syntactic and cross-lingual features and explain how we extract them in Section 3. Then, in Section 4, we describe the data and show how feature values are integrated to train the classifier. The experiments are presented in Section 5 and the results are discussed in Section 6. We report on related work in Section 7 and finally, we conclude in Section 8.

Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, September 3, 2013.
Monti, J., Mitkov, R., Corpas Pastor, G., Seretan, V., eds.

2 Background

2.1 Objectives and State of the Art

Our approach to the identification of German MWEs (of the type preposition+noun+verb) makes use of their morpho-syntactic properties and their semantic transparency vs. opacity. We classify MWEs on the level of lexical types into idiomatic ones vs. trivial (non-idiomatic) word combinations.

We start from the widely shared assumption that idiomaticity is often correlated with morpho-syntactic fixedness, e.g. Bannard (2007), Fazly and Stevenson (2006), Weller and Heid (2010) and that idiomaticity implies an element of non-compositionality, e.g. Baldwin et al. (2003). The former type of features is observable in morpho-syntactically analysed data; for the latter, which is not directly observable in monolingual corpora, we follow Villada Moirón and Tiedemann (2006) and Fritzinger (2010) and induce transparency vs. opacity from bilingual corpora.

We operate on MWE types, not on tokens; we consider individual occurrences and then sum up and average the feature values observed for one MWE type. Obviously, not all co-occurrences of lexemes that can be part of an MWE neccessarily can be interpreted as being idiomatic (cf. work on token-based analysis, e.g. Cook et al. (2008), Fritzinger et al. (2010)). However, in lack of respective hand-crafted data, a classification on token level, as in e.g. Diab and Bhutada (2009) is beyond the scope of the present paper.

2.2 Specificities of MWE Extraction for German

German has a relatively rich inflectional morphology, both in the nominal and verbal domain. Strong morpho-syntactic preferences in a word combination may thus indicate idiomatisation (i.e. MWE status). German also has a relatively free constituent order and, despite its morphological richness, substantial syncretism in nominal morphology (Evert et al., 2004); as a consequence, POS-pattern based approaches to MWE extraction tend to have low recall. As suggested by e.g. Seretan (2011), we thus use dependency parsing (Schiehlen, 2003) and extract MWE candidates from the parse output.

In this paper, we concentrate on the extraction of verb+PP collocations. Examples are *zur Sprache bringen* (lit.: "to bring to language", idiom.: to raise), cf. Example (1) above. We expect our results to be transferable to other MWE patterns, e.g. verb+direct object, verb+subject, adjective+noun, etc. Some of the candidates identified as verb+PP collocations may be part of larger patterns, such as *den Wind **aus** den **Segeln nehmen*** ("**to take** the wind **out of** so.'s **sails**").

3 Preprocessing: Feature Collection

In this section, we describe how all occurrences of the MWE candidates and their features are extracted. Later, in Section 4.2, we describe how the feature values of all occurrences of lexically identical MWE candidates are averaged and integrated into the classifier.

3.1 Candidate Extraction

As German allows for a flexible constituent order, the components of an MWE need not always occur adjacently[1]. Consider the following example sentence, where the verbal component of *im Raum stehen* (lit. "stand in the room", idiom. "to be dealt with") occurs 4 words to the left of the preposition and the noun:

(2)

Also Thus	**steht** stands	das the	Gerücht rumor	weiter still	**im** in the	**Raum** room
Thus	the rumor	is still	to be	dealt	with	

Thus the rumor is still to be dealt with

A deep syntactic analysis is thus required in order to reliably extract candidate triples, regardless of the actual constituent order or the distance of their component words. We use FSPAR (Schiehlen, 2003), a finite-state based dependency parser providing good lexical coverage and a full morpho-syntactic analysis (including POS, lemma, gender, number, case, compound splitting). Based on this annotation, the morpho-syntactic fixedness features (cf. Section 3.2) are extracted. While the dependency-parsed representation allows for the extraction of different syntactic patterns, we focus on preposition-noun-verb triples in the present paper.

3.2 Morpho-Syntactic Features

MWEs often exhibit a certain degree of fixedness with respect to morphological or syntacti-

[1] However, the words of semantically opaque MWEs mostly **do** occur adjacently and we use this adjacency as an additional fixedness feature, as described in Section 3.2 below.

	name	description	type
A	refl	verb having a reflexive pronoun	M
	n-adj	noun taking an adjectival modifier	S'
	det-fus	noun with fused prep+determiner	M
	neg	verb negated	S'
	vorf	expression occurring in the *vorfeld*	S
B	num	number of the noun	M
	det	determiner of the noun	M
C	adja	adjacency of component words	S

Table 1: Overview of morpho-syntactic features. M = morphological, S = syntactical, S' = syntactical in a broader sense.

cal variability (Sag et al., 2002). For example, the verb+PP constructions *hinter+Ohr+schreiben* (lit.: "behind+ear+write") has its idiomatic reading only if the number of the noun is plural (*Ohr**en***), the noun has a definite determiner (*die*) and the verb is reflexive (*sich*):
sich etw. hinter die Ohren schreiben
(idiom.: "to make sure to remember").

For German, such morpho-syntactic features can help to identify MWEs. A complete list of the features we use is given in Table 1.

Our feature set comprises morphological features (M), syntactic features (S), and features which are syntactically motivated in a broader sense (S'). We distinguish 3 different groups of morpho-syntactic features, depending on the possible values of the features: the first group (A) contains features for which we count their presence vs. absence regardless of the actual value. These features are represented as the ratio of the majority value to the total number of occurrences. For example, the *neg* feature indicates how often an expression occurs negated, but does not contain information about the type of negation (e.g. negation particle(s), verbal negation, negation of the noun).

In contrast, the values of the features of the second group (B) are summed up, i.e. we count how often the noun of a candidate expression occurred in *singular* vs. *plural* number or, in cases where the nouns take a determiner, how often it is a *definite* vs. *indefinite* or *quantifying* determiner.

Finally, the feature of the third group (C) indicates the adjacency of the expression's components: their sentence positions are summed and then divided by the position of the noun[2], with adjacent expressions (without any intervening words) scoring exactly 3. An adjacency score equal or close to 3 is regarded as indicator for an MWE.

While most features can be straightforwardly applied for many languages, the *fus*-feature (= preposition and determiner are melted into a fused form, e.g. *zur* = *zu+der* ="to the") is to be found in only a few languages. In comparison to Romance languages, where the fusion of certain preposition+article combinations is mandatory (de+le=du), the fusion of German articles and prepositions is optional in many cases and can thus be used as indicator for idiomaticity (strong preference for being fused or not fused). This is illustrated in Example (3), which is an invalid variation of the sentence given in (2): for the MWE *in+Raum+stehen*, the fusion of preposition and article is required. In contrast, for the regular combinations in Example (4), variation is possible.

(3)

*Also	steht	das	Gerücht	weiter	**in dem**	Raum
Thus	stands	the	rumor	still	in the	room

(4)

Im	Zimmer	steht	eine	Topfpflanze
In dem	Zimmer	steht	eine	Topfpflanze
In the	room	stands	a	potted plant

The *vorfeld*-feature applies to a syntactic characteristic of German only: there are different sentence structures of German (verb-initial vs. verb second vs. verb final sentences) and – contrary to PPs of fully compositional constructions – PPs of idiomatic MWEs only rarely occur in sentence initial (= *vorfeld*) position, even though gramatically possible, see Example (5).

(5)

?Im	Raum	steht	das	Gerücht	also	weiter
In the	room	stands	the	rumour	thus	still

3.3 Cross-Lingual Features

For our cross-lingual features, we adapt two metrics of (Villada Moirón and Tiedemann, 2006), both approximating the semantic opacity of MWEs using word alignment data: ***translational entropy*** (*te*) and the ***proportion of default alignments*** (*pda*). Both are measures of how "regular" and similar to non-idiomatic cases the translations of the respective lexical combinations are.

Translational entropy indicates the degree of variety of the candidate's translational equivalences. Regular combinations with a transparent or compositional semantics mostly have one or only very few different translations. In contrast,

[2]Example calculation of adjacency score: preposition at sentence position 5 + noun at 6 + verb at 7 = 18 / 6 (position of the noun) = 3.

semantically opaque MWEs show more different translations, i.e. much variation in equivalents: the lack of a respective (likely idiomatic) counterpart in the target language leads to translational variation which is recognisable in a broader variation of word alignments (and thus higher *te* scores). We use the following formula[3] to derive *translational entropy* scores from word alignments:

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) \log P(t|s)$$

where "T_s" is the set of all translation links from the source word "s" into different target words "t".

The *proportion of default alignments* indicates how often the words of a candidate expression have been translated literally. First, the four most frequent translational equivalences for each word of the corpus are collected (= default alignments). Then, the proportion of these default alignments among all alignments of a candiate expression is calculated (Villada Moirón and Tiedemann, 2006):

$$pda(S) = \frac{\sum_{s \in S} \sum_{d \in D_s} align\_freq(s,d)}{\sum_{s \in S} \sum_{t \in T_s} align\_freq(s,t)}$$

where "T_s" is the set of all translation links of "s" (the source word), "D_s" contains the word's default alignments and "$align\_freq(x, y)$" is the frequency of translation links from word x to word y in the context of the triple "S". Semantically opaque MWEs lead to low *pda* scores.

We calculate *te* and *pda* scores based on automatically generated word alignments using GIZA++ (Och and Ney, 2003) of the German section of Europarl to English, French and Swedish, following Fritzinger (2010) who showed that averaged scores based on alignments from several language pairs are more reliable than single language pair scores[4].

4 Experimental Setup

We use Conditional Random Fields (CRFs, Lafferty et al. (2001)) for the classification of verb+PP triples into MWEs vs. regular combinations[5] In this section we go into more detail about our data set, and we explain how features are extracted and transformed into a CRF-suitable format and how the classification accuracies of the CRFs are evaluated.

[3]Taken from Melamed (1997).

[4]This is in line with Lefever et al. (2013), who use data from several language pairs for WSD.

[5]Note however that we do not exploit the full potential of CRFs: for type-based MWE extraction, we do not take any sequential features into account.

group	interval	#all	#train	#dev	#test
high	>39	2,272	1,818	227	227
mid	18–39	2,367	1,893	237	237
low	4–17	2,124	1,700	212	212
		6,763	3,774	676	676
thereof MWEs:		862	697	75	90

Table 2: Distribution of data: 3 different frequency intervals, randomly extract 80% of each interval for training, 10% for development and 10% for testing.

4.1 Data

We start from a set of 10,276 preposition-noun-verb triple types, extracted from Europarl version 3 (Koehn, 2005). These are manually annotated as MWEs[6] (937) vs. regular combinations (9,339). From this data set, we use only triples that occur at least 4 times in order to get reliable fixedness scores. We consider this threshold necessary as low occurrence frequencies can lead to an inaccurate representation of morpho-syntactic preferences. Assume, for example, a regular combination with f=2 which randomly occurs with the same values of number, article, etc., even though variation is possible and to be expected in a larger set of occurrences. The cross-lingual features are based on word alignment which is a purely statistical method and thus to a certain extent inaccurate. As non-recurring alignment is used as indicator for idiomaticity, infrequent candidate triples do not provide a sufficient basis for reliable alignment statistics. For comparison, see Evert (2005) who proposes a threshold of f>=5. We set the threshold to f=4, which reduces the data to 6,763 triples, whereof 862 are MWEs and 5,901 are regular combinations.

The set of triples is divided into three different frequency intervals (high, mid and low-frequent) and of each interval, we randomly extract 80% for training, and 10% for developement and testing respectively, without allowing for overlap between these three sets, cf. Table 2 for details.

As can be seen, there are much more regular combinations than MWEs in each of the sets. In order to not work on a data set with an artificial distribution of MWEs vs. regular combinations, we decided to not balance the sets with regard to the number of MWEs they contain.

[6]Without considering different levels of opacity or fixedness.

triple	all	sg	pl	bkt.
an Ball bleiben	12	12 (100%)	0 (0%)	10
aus Auge verlieren	431	91 (21%)	340 (79%)	7
auf Gedanke bringen	7	4 (57%)	3 (43%)	5

Table 3: Example of how feature values (here: number feature) are grouped into suitable buckets (bkt.) for CRF training.

4.2 Features

Feature values are considered as strings in CRFs. In order to be able to abstract over the training data and predict idiomaticity on the (unseen) development and test sets, the features need to be represented in a suitable format.

For the morpho-syntactic fixedness features (except *adjacency*) given in Table 1 above, and for each triple type: we (1) add up the values of all occurrences of one triple, (2) take the percentage of the most frequent value and (3) pack that into buckets incrementing in 10% steps, rounding down to the next smallest bucket. For clarification, we give some calculation examples in Table 3.

The values of the adjacency feature are spread around 3.0; for each lexical triple, we sum and average the values of all occurrences, round them to one decimal and calculate the absolute value of their distance to 3.0 (using increments of 0.1), with low bucket scores indicating high fixedness[7].

Translational entropy values (of all three language pairs DE-EN, DE-FR, DE-SE) are summed and averaged for each distinct triple. Depending on the triple these range between 0.045 and 4.406, with higher scores indicating opaque semantics. They are packed into buckets of 0.5 increment.

Finally, the proportion of default alignment values range between 0 and 1. They are summed, averaged and packed into buckets of 0.1 increment.

In addition to these features, we also use the lexical form of the verb+PP triples themselves, because even those can be indicators for MWEs. This holds particularly for nouns, as can be seen from Table 4, where we give the most frequent nouns occuring in MWEs vs. regular combinations (both lists are derived from our training data, cf. Section 4.1).

[7]For example: averaged value is 2.78; rounded it is 2.8, distance to 3.0 is 0.2 (name of the bucket).

MWE	regular
Weg (way)	Jahr (year)
Hand (hand)	Bereich (range)
Auge (eye)	Rahmen (framework)
Tisch (table)	Bericht (report)
Leben (life)	Land (country)
Seite (side/page)	Herr (mister)

Table 4: Lists of most frequent nouns in verb+PP constructions, derived from the training data.

4.3 Evaluation

The accuracy of the different CRF classifiers is evaluated using *precision*, *recall* and *f-score*. These are calculated with regard to the number of valid MWEs from the 10% subsets of our manually annotated data found by the respective CRF.

$$Precision = \frac{\#correct-found}{\#all-found}$$

$$Recall = \frac{\#correct-found}{\#to-be-found}$$

$$F-Score = 2 * \frac{Precision*Recall}{Precision+Recall}$$

Note that a majority of 88% of the triples from the training data are regular combinations, while only roughly 12% are MWEs. To get an impression of the overall classification accuracy, we thus also calculate the percentage of all correct classifications (regardless if MWE or not).

5 Experiments

We use the developement set to experiment with different feature combinations and then, in a second series of experiments apply these combinations on the previously unseen test set.

We compare the results of our experiments to the following baselines[8]:

- **guess** use the frequency distribution of MWEs vs. regular triples derived from the training data (12% vs. 88%) to classify the data;
- **all triv** as there are many more regular triples than MWEs, classify everything as being regular;
- **pnv** use the constituent lemmas of the triples to train a CRF classifier.

[8]Note that only baseline nr.3 relies on a CRF for the classification; we used PERL scripts to realise baselines (1)+(2).

(a) Results on **development** set (75 MWEs to be found).

exp	TP	FP	FN	prec.	rec.	f-score
guess	8	69	67	10.39	10.67	10.53
pnv	21	15	54	58.33	28.00	37.84*
m-s.	26	19	49	57.78	34.67	43.33*
c-l.	27	11	48	71.05	36.00	47.79*
all	40	13	35	75.47	53.33	62.50***
best	42	12	33	77.78	56.00	65.12***

(b) Results on **test** set (90 MWEs to be found).

exp	TP	FP	FN	prec.	rec.	f-score
guess	11	73	79	13.10	12.22	12.64
pnv	25	8	65	75.76	27.78	40.65*
m-s.	41	14	49	74.54	45.56	56.55*
c-l.	47	11	43	81.03	52.22	63.51**
all	48	12	42	80.00	53.33	64.00**
best	46	10	44	82.14	51.11	63.01**

Table 5: F-scores with respect to the number of MWEs to be found. Statistical significance is calculated using chi-square, with * = significant at 0.001 level wrt. **guess**, ** = significant at 0.1 level wrt. **pnv**, *** = significant at 0.05 level wrt. **pnv**

We trained CRF classifiers for the following feature combinations (all include the preposition, noun and verb of the triples):

- **m-s** use all morpho-syntactic features: *refl, n-adj, det-fus, neg, vorf, num, det, adja*, cf. Table 1 above;
- **c-l** use the cross-lingual features *translational entropy* and *proportion of default alignments*;
- **all** use all morpho-syntactic features and all cross-lingual features;
- **best** use the "best" combination of features: by (1) adding each of the features independently to the current feature set, then (2) calculating the percentage of correct classifications on the development data and (3) permanently adding the best performing feature to the feature collection (4) repeat from (1) until performance drops. This lead to the following feature combination: *pnv+te+adja+fus+refl*;

We are aware that our "best" combination does not neccessarily represent the global maximum of all possible feature combinations. However, we believe it is a reasonable local maximum, given that the calculation of all combinations is too costly (in terms of both time and computing resources) to be realised.

(a) Results on **development** set
(676 triples, whereof 75 mwes / 601 literals).

exp.	# correct			% correct
	mwe	lit	all	
guess	8	532	540	79.89%
all triv	0	601	601	88.91%
pnv	21	586	607	89.79%
m-s	26	582	608	89.94%
c-l	27	590	617	91.27%
all	40	588	628	92.89%
best	42	589	631	93.34%

(b) Results on **test** set
(676 triples, whereof 90 mwes / 586 literals).

exp.	# correct			% correct
	mwe	lit	all	
guess	11	513	524	77.51%
all triv	0	586	586	86.69%
pnv	25	578	603	89.20%
m-s	41	572	613	90.68%
c-l	47	575	622	92.01%
all	48	574	622	92.01%
best	46	576	622	92.01%

Table 6: Percentage of correct classifications on the development and test sets .

6 Results

The accuracies of the different classification models in terms of f-scores (wrt. MWEs to be found) are given in Table 5, where "TP" (= true positives) designates valid MWEs identified by the classifier, "FP" (= false positives) are regular triples that were erroneously identified as MWEs and "FN" (= false negatives) are MWEs that should have been identified but were not found by the classifier.

The results in Table 5 show that any combination of features outperforms both baselines ("guess", "pnv"): while we can see a moderate improvement for the use of morpho-syntactic ("m-s") and cross-lingual ("c-l") features when used independently, the combinations of morpho-syntactic and cross-lingual features (cf. "all" and "best") lead to even higher f-scores with a statistically significant improvement with respect to the two baselines.

Similarly, the percentage of correct classification decisions in Table 6 shows that combining morpho-syntactic and cross-lingual features results in higher prediction accuracy.

The results in Table 6 and Table 5 confirms that features of different and independent dimensions (morpho-syntax vs. semantics) benefit from each other, and as a consequence, that their combination leads to an improved classification into MWEs and

approach	lang.	pattern?		classification			identification			
		yes	no	rank.	sup.	unsup.	frq.	m-s.	sem.	wa.
(Smadja, 1993)	EN	X		X			X			
(Bannard, 2007)	EN	X		X				X		
(Baldwin et al., 2003)	EN	X		X					X	
(Villada Moirón and Tiedemann, 2006)	NL	X		X						X
(Weller and Fritzinger, 2010)	DE	X		X				X		X
(Ramisch et al., 2010)	PT		X			X	X			X
(Fothergill and Baldwin, 2011)	JA		X		X			X	X	
(Tsvetkov and Wintner, 2011)	HE		X			X	X	X	X	X
present paper	DE	X			X			X		X

Table 7: Non-exhaustive overview of different approaches dealing with the identification of MWE types. The approaches are classified according to the following categories: *pattern?* (= is it restricted to MWEs of a certain syntactic pattern), *classification* (= ranking according to association measures or supervised, unsupervised classification), *identification* (= aspect of the MWE that is used for their identification: *frq.* = collocational behaviour, *m-s.* = morpho-syntactic features, *sem* = semantic features, *wa* = word alignment, i.e. translational behaviour.

regular triples.

Overall, the classification performance is similar on the development and test set, indicating that the built classifiers are robust and not over-fitting. However, the combination of all features seems to be more stable than the "best" combination obtained by searching for a local maximum on the development set.

7 Related Work

The task of automatically identifying MWEs has gained much attention in the NLP research community in the past. The approaches that emerged are just as multi-dimensional as the phenomenon of MWEs itself, each tackling one (or more) specific characteristics of MWEs. Consider Table 7 for a partial overview.

There are three of these characteristics that have been repeatedly implemented by different researchers to identify MWEs: i) word association measures, ii) morpho-syntactic fixedness, iii) semantic opacity.

Approaches based on word association measures exploit estimated vs. observed co-occurrence frequencies of an MWE's content words and the expression as a whole to identify valid MWEs (e.g. Church and Hanks (1990)). Such approaches proved to work well, but their performance can easily be enhanced by additionally checking for syntactic consistency of the MWEs. This can be realised either by restricting the candidate list to a certain syntactic MWE pattern beforehand (Evert and Krenn, 2001) or by filtering out syntactically inconsistent MWEs after having identified highly associated word pairs (Smadja, 1993).

Many types of MWEs exhibit a certain degree of morpho-syntactic fixedness: they do not allow for morphological or syntactic variation when used idiomatically. While some approaches investigate different types of syntactic variation (e.g. Bannard (2007) for English or Weller and Heid (2010) for German), others combine syntactic fixedness with limited lexical variability to identify MWEs (Fazly and Stevenson, 2006).

Finally, there are approaches tackling the opaque semantics of MWEs: they are based on the assumption that the semantics of the expression as a whole cannot be derived from the semantics of its constituent words. While Baldwin et al. (2003) use Latent Semantic Analysis for this task, Villada Moirón and Tiedemann (2006) present an approach that approximates the MWE's semantics by deriving translational equivalences from parallel text. While Villada Moirón and Tiedemann (2006) use word alignment only for ranking MWE candidates identified separately by means of syntactic patterns in parsed data, other approaches, e.g. Zarrieß and Kuhn (2009), de Caseli et al. (2010) use word alignment as basis for MWE extraction itself.

More recently, some approaches came up with combinations of features addressing different characteristics of MWEs. (Ramisch et al., 2010) combine word association measures with alignment-based approaches and use Bayesian Networks to predict MWEs, while (Weller and Fritzinger, 2010) combine morpho-syntactic fixedness with translational equivalences. In contrast, (Fothergill and Baldwin, 2011) combine morpho-syntactic fixedness with lexical hypernyms and (Tsvetkov

and Wintner, 2011) present a very feature-rich approach (using Bayesian Networks) that combines collocational behaviour with morpho-syntactic fixedness and translational equivalences[9].

Table 7 shows where our approach ranges in relation to the work just discussed. It is most comparable to the one of Weller and Fritzinger (2010). However, they use less morpho-syntactic features and their evaluation is restricted to different rankings (of 200 candiate triples) in order to find an optimal feature combination. While such rankings are useful to identify MWE candidates for lexicographical applications, the CRF models trained in the present paper allow for a more robust MWE identification that is easier to integrate into higher order applications.

To our knowledge, our approach allows to extract the most detailed morpho-syntactic data on MWEs for German, taking into account the rather intricate specificities of German morphology and syntax.

8 Conclusion and Future Work

We presented an approach for the identification of MWEs using morpho-syntactic fixedness (derived from deep syntactic analysis) and cross-lingual features (derived from automatic word alignment). We showed that combinations of these two feature sets, which both address different aspects of MWEs, clearly outperform the baselines, as well as the independent use of any of the feature sets in a supervised classification task.

Our approach could be applied prior to post-editing of SMT output, providing a comparatively accurate highlighting of MWEs (which are known to be potential sources of SMT errors).

In the future, we plan to investigate an even more fine-grained combination of features, e.g. in more linguistically motivated combinations. To give an example, the German verb+PP *in Gang kommen* means "to be set in motion" when the noun appears in singular form without a determiner, while the same lemmas used in plural form with a definite article *in die Gänge kommen*, bears the meaning "to get organised". Occurrences for which always the same two feature values are modified should have an additional impact.

As soon as manually annotated data on token-level become available, our approach can easily be trained on them. Moreover, we can then extend it to use sequential features, where appropriate.

So far, we focused on verb+PP constructions, but we plan to extend our approach to MWEs of different patterns in the future. Moreover, we intend to apply the CRF models we trained on the Europarl corpus to verb+PP constructions extracted from other domains.

[9]Note however that – in lack of available parallel corpora for Hebrew – they approximate the translational equivalences by combining dictionary entries with corpus lookups.

References

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003)*, pages 89–96.

Bannard, Colin. 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.

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

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the Workshop: Towards a shared task for multiword expressions (LREC 2008)*, pages 19–22.

de Caseli, Helena Medeiros, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, 44(1-2):59–77.

Diab, Mona T. and Pravin Bhutada. 2009. Verb noun construction mwe token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (ACL-JICNLP 2009)*, pages 17–22.

Evert, Stefan and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 188–195.

Evert, Stefan, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th international conference on language resources and evaluation (LREC 2004)*, pages 907–910.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. University of Stuttgart, PhD dissertation.

Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11st Conference of the European Chapter of the ACL (EACL 2006)*, pages 337–344.

Fothergill, Richard and Timothy Baldwin. 2011. Fleshing it out: a supervised approach to mwe-token and mwe-type classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 911–919.

Fritzinger, Fabienne, Marion Weller, and Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)*, pages 2908 –2914.

Fritzinger, Fabienne. 2010. Using parallel text for the extraction of german multiword expressions. *Lexis: E-Journal in English Lexicology*.

Koehn, Phillip. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit 2005)*, pages 79–86.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*.

Lefever, Els, Veronique Hoste, and Martine De Cock. 2013. Five languages are better than one: An attempt to bypass the data acquisition bottleneck for wsd. In *Proceedings of the 14th international conference on intelligent text processing and computational linguistics (CICLing 2013)*, pages 343–354.

Melamed, I. Dan. 1997. Measuring semantic entropy. In *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What and How*, pages 41–46.

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

Ramisch, Carlos, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. 2010. A hybrid approach for multiword expression identification. In *Computational Processing of the Portuguese Language*, pages 65–74. Springer.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 1–15.

Schiehlen, Michael. 2003. A cascaded finite state parser for german. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL 2003)*.

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(3):143–177.

Tsvetkov, Yulia and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 836–845.

Villada Moirón, Begoña and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Workshop on Multiword-Expressions in a multilingual context (EACL 2006)*, pages 33–40.

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

Weller, Marion and Ulrich Heid. 2010. Extraction of German multiword expressions from parsed corpora using context features. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)*.

Zarrieß, Sina and Jonas Kuhn. 2009. Exploiting translational correspondences for pattern-independent mwe identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30.