A Hybrid Approach for the Identification of Multiword Expressions

Marion Weller and Fabienne Fritzinger

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
Azenbergstr. 12 D 70174 Stuttgart
{wellermn|fritzife}@ims.uni-stuttgart.de

1. Introduction

According to Bannard (2007), *"A multiword expression is usually taken to be any word combination that has some feature (syntactic, semantic or purely statistic) that cannot be predicted on the basis of its component words and/or the combinatorial process of language."*.

We present a procedure for the identification of German multiword expressions (MWES) by making use of the *morpho-syntactic restrictions* of MWEs and of their semantic opacity which we approximate by their *translational behaviour*. As morpho-syntactic and translational features are independent, we combine their benefits assuming that they complement each other (hybrid approach). Based on the different features, we compute scores indicating a degree of idiomaticity. The procedure can be divided into two parts: candidate extraction from preprocessed text and feature evaluation, cf. Figure 1.

We concentrate on preposition-noun-verb (PNV) triples (e.g. *'unter (den) Teppich kehren'* – lit. 'to sweep under the rug', idiom. 'to hide sth.'), as this is a high-frequent pattern covering many idiomatic MWEs.

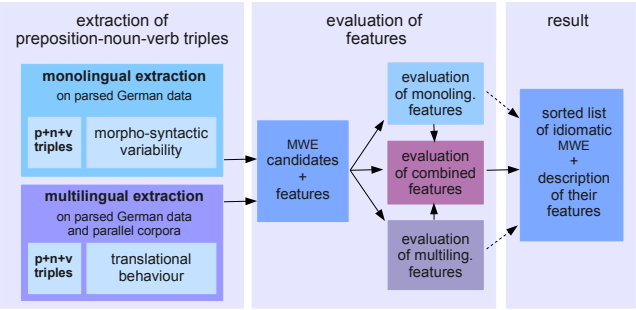

Figure 1: Architectural sketch of the procedure.

2. Methodology

MWE candidates (i.e. PNV-triples) are extracted from dependency parsed text (Schiehlen, 2003). For each candidate expression, we also extract an array of monolingual morpho-syntactic features to compute the degree of morpho-syntactic fixedness (Weller and Heid, 2010). Multilingual translational features are derived from parallel text.

2.1 Monolingual features

We use the morphological features *number*, *determiner* and *negation*; idiomatic MWEs tend to have a strong preference for one specific realization of these features. Additionally, we use two syntactically motivated criteria: (i) the *adjacency* of the components of the triples and (ii) *vorfeld* which is specific for German: since idiomatic MWEs are not likely to be used in the very beginning of a sentence, these occurrences are counted as an indicator for trivial triples.

2.2 Multilingual features

We exploit the fact that idiomatic MWEs frequently have non-compositional semantics. It is assumed that opaque word combinations are translated as a whole, whereas compositional uses would show regular, individual translations of the words involved.

Two different measures indicate semantic opacity based on word equivalences (cf. Villada Moirón and Tiedemann (2006)): *translational entropy* (*te*) shows the degree of diversity of translations, where higher diversity is expected for valid MWE candidates and the proportion of *default alignments* (*pda*) reflects how many of the MWE's translations contain trivial translations of the MWE's component words.

2.3 Computing idiomaticity scores

Adjacency, *te* and *pda* are averaged and normalized. For the remaining features *vorfeld*, *determiner*, *number* and *negation*, the percentage of their most common value is calculated. Scores are computed by summing up the values of the features and normalizing the result. As we expect idiomatic MWEs to be morpho-syntactically restricted and to exhibit idiosyncratic translational behaviour, they should rate higher than trivial triples and thus be on top of a list of candidates sorted by their scores. The quality of a sorted list is measured by the *uninterpolated average precision* (UAP) (Manning and Schütze (1999)).

3. Data

For the extraction of MWE candidates and their monolingual context features we used a maximum amount of available data (269 million words of German newspaper text). The test set of our experiments consists of the 1,013 most frequent PNV triples ($f \geq 60$). Translational features were derived from the Europarl corpus (Koehn, 2005) as this requires parallel data.

The combination of the two corpora led to a slight domain mismatch: candidates extracted from newspaper text which did not occur (in a sufficient number) in Europarl were assigned *te* and *pda* values of 0. As Europarl does not contain many of the MWE candidates that are common in newspaper text (e.g. *(to) threaten with knife*), this penalizes mostly trivial triples.

4. Evaluation

We experimented with different feature settings which are given in Table 1. While the individual features do not always lead to a better result than the baseline (sorting according to frequency, UAP = 6.51), the combination of different features substantially improves the UAP values. This applies to the combination of morphological and syntactic features (Table 1 (b)) and particularly to the combination of translational and morpho-syntactic features (Table 1 (d)).

Our assumption that monolingual and multilingual features complement each other is illustrated in Table 2. For example, *zu Schweigen bringen* is highly ranked by the monolingual features, but poorly when sorted by multilingual features; vice versa for *mit Leben füllen*. The two lower entries show that trivial triples highly ranked in the baseline are penalized by any of our scores.

(a)

feat.	num	det	neg	adjacency	vorfeld
UAP	0.607	0.650	0.643	0.780	0.685

(b)

feat.	M_1: det , num	M_2: det , num , neg	S: adja , vorf.	M_2 , S
UAP	0.658	0.753	0.799	0.847

(c)

feat.	te	pda	te , pda	B: 2·te , pda
UAP	0.777	0.590	0.813	0.824

(d)

feat.	all features	(M_2) , (S) , (te) , (pda)	(M_2) , (S) , (B)	(M_2 , S) , (B)
UAP	0.878	0.868	0.889	0.875

Table 1: UAP-values for separately computed or grouped monolingual features (a), (b) and translational features (c). In (d), monolingual and translational features are combined. The UAP value of the baseline is 0.651. Brackets indicate groupings.

PNV-triple	mono	multi	both	freq.
zu Schweigen bringen lit. *to silence put*: to silence so.	74	438	258	677
mit Leben erfüllen lit. *with life fill*: to animate sth.	473	66	166	921
in Erklärung heissen to be mentioned in statement	938	476	862	35
in Krankenhaus bringen to bring so. to hospital	705	929	785	41

Table 2: Ranks of candidates in the sorted lists. Idiomatic MWEs are **bold-faced**.

5. Conclusion and Future Work

We showed that scores based on characteristic features of MWEs are better suited for the identification of idioms than frequency. Our method benefits particularly from the combination of monolingual and multilingual features. As we only experimented with handcrafted feature combinations, we intend to further optimize feature settings by applying machine learning techniques in the future. Additionally, our method might be extended to distinguish between literal and non-literal usages of MWEs (Fritzinger et al. (2010)).

References

Bannard, C., 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In: Proceedings of ACL 2007 (Workshop).

Fritzinger, F., Weller, M., Heid, U., 2010. A survey of idiomatic preposition-noun-verb triples on token level. In: Proceedings of LREC 2010.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit 2005.

Manning, C. D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts.

Schiehlen, M., 2003. A cascaded finite state parser for german. In: Proceedings of EACL 2003.

Villada Moirón, B., Tiedemann, J., 2006. Identifying idiomatic expressions using automatic word-alignment. In: Proceedings of EACL 2006 (Workshop).

Weller, M., Heid, U., 2010. Extraction of German multiword expressions from parsed corpora using context features. In: Proceedings of LREC 2010.