

Extraction of multiword expressions from parsed corpora using context features

Marion Weller Ulrich Heid

[weller|heid]@ims.uni-stuttgart.de

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
70174 Stuttgart
Germany

Overview

- ▶ Extraction of multiword expressions
- ▶ Context features
 - Morphologically motivated
 - Syntactically motivated
 - Lexical choice
- ▶ Evaluation and experiments
 - Use of context features for idiom identification
 - Expanding basic patterns: Find preferences for adjectives
- ▶ Analysis of extraction errors
- ▶ Conclusion and Future Work

Extraction of multiword expressions

	word form	pos tag	lemma	morph.-synt. features	gover- nor	gramm. function	<i>engl.</i>
0	Spaniens	NE	Spanien	Gen:Sg	1	GL	Spain's
1	Regierungs chef	NN	Regierungs: chef	Nom:M:Sg	3	SUBJ	head of government
2	Felipe Gonzalez	NN	Felipe Gonzalez	Nom:M:Sg	1	APP	Felipe Gonzalez
3	gab	VVFIN	geben	3:Sg:Past	-1	TOP	gave
4	ebenfalls	ADV	ebenfalls		3 6	ADJ	also
5	grünes	ADJA	grün		6	ADJ	green
6	Licht	NN	Licht	Akk:N:Sg	3	OBJ _{acc}	light
7	.	.	.		-1	TOP	.

“The head of Spain’s government, Felipe Gonzalez, also gave his approval.”

Extraction of multiword expressions

	word form	pos tag	lemma	morph.-synt. features	gover- nor	gramm. function	<i>engl.</i>
0	Spaniens	NE	Spanien	Gen:Sg	1	GL	Spain's
1	Regierungs chef	NN	Regierungs: chef	Nom:M:Sg	3	SUBJ	head of government
2	Felipe Gonzalez	NN	Felipe Gonzalez	Nom:M:Sg	1	APP	Felipe Gonzalez
3	gab	VVFIN	geben	3:Sg:Past	-1	TOP	gave
4	ebenfalls	ADV	ebenfalls		3 6	ADJ	also
5	grünes	ADJA	grün		6	ADJ	green
6	Licht	NN	Licht	Akk:N:Sg	3	OBJ _{acc}	light
7	.	.	.		-1	TOP	.

“The head of Spain’s government, Felipe Gonzalez, also gave his approval.”

geben

Extraction of multiword expressions

	word form	pos tag	lemma	morph.-synt. features	gover- nor	gramm. function	<i>engl.</i>
0	Spaniens	NE	Spanien	Gen:Sg	1	GL	Spain's
1	Regierungs chef	NN	Regierungs: chef	Nom:M:Sg	3	SUBJ	head of government
2	Felipe Gonzalez	NN	Felipe Gonzalez	Nom:M:Sg	1	APP	Felipe Gonzalez
3	gab	VVFIN	geben	3:Sg:Past	-1	TOP	gave
4	ebenfalls	ADV	ebenfalls		3 6	ADJ	also
5	grünes	ADJA	grün		6	ADJ	green
6	Licht	NN	Licht	Akk:N:Sg	3	OBJ_{acc}	light
7	.	.	.		-1	TOP	.

“The head of Spain’s government, Felipe Gonzalez, also gave his approval.”

Licht geben

Extraction of multiword expressions

	word form	pos tag	lemma	morph.-synt. features	gover- nor	gramm. function	<i>engl.</i>
0	Spaniens	NE	Spanien	Gen:Sg	1	GL	Spain's
1	Regierungs chef	NN	Regierungs: chef	Nom:M:Sg	3	SUBJ	head of government
2	Felipe Gonzalez	NN	Felipe Gonzalez	Nom:M:Sg	1	APP	Felipe Gonzalez
3	gab	VVFIN	geben	3:Sg:Past	-1	TOP	gave
4	ebenfalls	ADV	ebenfalls		3 6	ADJ	also
5	grünes	ADJA	grün		6	ADJ	green
6	Licht	NN	Licht	Akk:N:Sg	3	OBJ _{acc}	light
7	.	.	.		-1	TOP	.

“The head of Spain’s government, Felipe Gonzalez, also gave his approval.”

grün Licht geben

Context features: assessing idiomaticity

Morphologically motivated features

Number and **determiner** are often fixed in idiomatic expressions, but can vary in trivial combinations:

	MWE	f	NUM		DET		<i>engl.</i>
			Sg	Pl	def	null	
-	in Jahr aussehen	271	121	150	129	85	in year look
+	auf Barrikade gehen	167	2	165	165	2	on barricade go: <i>to go on the warpath</i>

Values for determination: *definite, indefinite, demonstrative, possessive, null* and *quantifying*.

Context features: assessing idiomaticity

Morphologically motivated features

Number and **determiner** are often fixed in idiomatic expressions, but can vary in trivial combinations:

	MWE	f	NUM		DET		<i>engl.</i>
			Sg	Pl	def	null	
-	in Jahr aussehen	271	121	150	129	85	in year look
+	auf Barrikade gehen	167	2	165	165	2	on barricade go: <i>to go on the warpath</i>

Values for determination: *definite, indefinite, demonstrative, possessive, null* and *quantifying*.

Negation: relevant for the linguistic description of a subgroup of MWEs which occur only in negative contexts: *negative polarity items*

MWE	f	negated	<i>engl.</i>
aus dem Kopf gehen	47	47	<i>to get out of the head</i>

Context features: assessing idiomaticity

Syntactically motivated features: **Adjacency**

Parts of non-trivial MWEs are likely to be immediate or near neighbours. For *preposition-noun-verb* (PNV) triples, we compute a simple position-based adjacency measure:

$$\frac{\text{pos}(P) + \text{pos}(N) + \text{pos}(V)}{\text{pos}(N)} = 3$$

if noun, verb and preposition are immediately adjacent with the noun in the middle position.

Context features: assessing idiomaticity

Syntactically motivated features: **Adjacency**

Parts of non-trivial MWEs are likely to be immediate or near neighbours. For *preposition-noun-verb* (PNV) triples, we compute a simple position-based adjacency measure:

$$\frac{\text{pos}(P) + \text{pos}(N) + \text{pos}(V)}{\text{pos}(N)} = 3$$

if noun, verb and preposition are immediately adjacent with the noun in the middle position.

Sie glauben, dass dadurch die Wirtschaft wieder **in Fahrt kommt**.
They believe that thereby the economy again **in run comes**.
They believe that thereby the economy gets going again.

Context features: assessing idiomaticity

Syntactically motivated features: **Adjacency**

Parts of non-trivial MWEs are likely to be immediate or near neighbours. For *preposition-noun-verb* (PNV) triples, we compute a simple position-based adjacency measure:

$$\frac{\text{pos}(P) + \text{pos}(N) + \text{pos}(V)}{\text{pos}(N)} = 3$$

if noun, verb and preposition are immediately adjacent with the noun in the middle position.

Sie glauben, dass dadurch die Wirtschaft wieder **in Fahrt kommt**.
 They believe that thereby the economy again **in run comes**.
 They believe that thereby the economy gets going again.

Auf kleinen **Zetteln**, die an Bäume geklebt worden waren, **stand**: "Wilson kommt".
On small **notes**, that to trees glued been had, **stood**: "Wilson comes".
 On small notes that had been glued to trees, it read: "Wilson comes".

Context features: assessing idiomaticity

Syntactically motivated features: **Vorfeld**

German idiomatic PNV-triples rarely occur at the very beginning of a sentence (*vorfeld-position*), except in contrastive contexts. In this case, all parts of the triple must be in the *vorfeld*, i.e. the verb can't be moved out.

Context features: assessing idiomaticity

Syntactically motivated features: **Vorfeld**

German idiomatic PNV-triples rarely occur at the very beginning of a sentence (*vorfeld-position*), except in contrastive contexts.

In this case, all parts of the triple must be in the *vorfeld*, i.e. the verb can't be moved out.

In **Stellung gebracht** worden seien Raketen mit einer Reichweite von 200 km
In **position brought** been had missiles with a range of 200 km
Missiles with a range of 200 km had been positioned.

* In **Stellung** seien Raketen mit einer Reichweite von 200 km **gebracht** worden.

Context features: assessing idiomaticity

Syntactically motivated features: **Vorfeld**

German idiomatic PNV-triples rarely occur at the very beginning of a sentence (*vorfeld-position*), except in contrastive contexts.

In this case, all parts of the triple must be in the *vorfeld*, i.e. the verb can't be moved out.

In **Stellung gebracht** worden seien Raketen mit einer Reichweite von 200 km
In **position brought** been had missiles with a range of 200 km
Missiles with a range of 200 km had been positioned.

* In **Stellung** seien Raketen mit einer Reichweite von 200 km **gebracht** worden.

In die **Klinik** hatten die Eltern sie gegen ihren Willen **gebracht**.
In the **hospital** had the parents her against her will **brought**.
The parents took her into a hospital against her will.

Context features: length of MWES

Lexical choice: **Adjectives** and **objects**

MWES can have a strong preference for specific lexical elements or even require further components to form a valid idiomatic multiword expression.

Adjective

grünes Licht geben

für bare Münze nehmen

auf { taube
offene } Ohren stoßen

am ∅ Ball bleiben

Object

Kind mit Bad ausschütten

Wind aus Segel nehmen

ADJ Wert legen

OBJ in den Sand setzen

Evaluation: Analysis of morpho-syntactic features

Context features used to identify idiomatic MWES

1013 PNV-triples ($f \geq 210$) extracted from newspaper text,
manually annotated with respect to their idiomaticity.

For each triple, compute a **fixedness-score**:

- Based on the MWES averaged or most prominent features
- Represents the morpho-syntactic fixedness of an MWE

→ Sort list according to the resulting scores

uninterpolated average precision (UAP):

measure for the quality of a sorted list [Manning and Schütze, 1999]

UAP=1 when the list is perfectly sorted

Evaluation: Analysis of morpho-syntactic features

results: idiomaticity

feature	number	det	neg	adjacency	vorfeld
UAP	0.607	0.605	0.643	0.694	0.566

UAP-values for the morpho-syntactic features computed separately

grouped	M ₁ det+num	M ₂ det+num+neg	S adja+vorfeld	M ₂ + S
UAP	0.635	0.681	0.664	0.830

UAP-values: sorted according to scores based on combined features

Experiment: Expanding basic patterns

- ▶ Identify PNV-triples with a clear preference for
 - a specific adjective
 - no adjective at all

First step: find triples with a preference for no adjectives.

size of test set	1013 [all]	610 [ADJ \leq 0.1]	133 [ADJ=0]
idioms	513	390	99
UAP	0.833	0.892	0.937

Sorting based on morpho-syntactic criteria and percentage of adjectives.

Experiment: Expanding basic patterns

- ▶ Identify PNV-triples with a clear preference for
 - a specific adjective
 - no adjective at all

First step: find triples with a preference for no adjectives.

size of test set	1013 [all]	610 [ADJ \leq 0.1]	133 [ADJ=0]
idioms	513	390	99
UAP	0.833	0.892	0.937

Sorting based on morpho-syntactic criteria and percentage of adjectives.

Take account of *creative use of language*:

Threshold (ADJ \leq 0.1) allows for occasional adjectives with supposedly adjective free triples.

Dort geht es bei Schunkelmusik ... zur fröhlichen Sache.

With beer tent music ... there is a great ambiance.

Experiment: Expanding basic patterns

- ▶ Analysis of PNV-triples with a preference for adjectives

Second step: Divide remaining candidates into sets of

- (i) idioms with obligatory (specific) adjectives
- (ii) idioms where adjectives are common and not restricted
- (iii) trivial word sequences

	PNV-triple	adjective	ADJ
+	auf Bank schieben	lang	1
-	mit Wirkung bestellen	sofortig	1
-	zu Fixing verbilligen	frankfurter	1
+	auf Fuß setzen	frei	0.997
+	in Gang sein	voll	0.992

Candidate triples with their most frequent adjectives.

Error Analysis

Correctness of the extracted candidates

Ambiguity handling

NPs with case ambiguities: not used for extraction

PP-attachment: all options in the parse output are used for extraction

Error Analysis

Correctness of the extracted candidates

Ambiguity handling

NPs with case ambiguities: not used for extraction

PP-attachment: all options in the parse output are used for extraction

False positives: Word sequences that appear to be idiomatic but consist of a verb and an adjunct prepositional phrase:

e.g. *in Betrieb sein (to operate)*.

waren in 192 Betrieben knapp 20.000 Mitarbeiter in Lohn und Brot.
were in 192 companies almost 20.000 employees in pay and bread.
in 192 companies, almost 20.000 members of staff were employed.

Evaluation of 6690 sentences: 94 false positives,
mostly in combination with specific verbs or prepositional phrases.

Conclusion and future work

We extracted MWEs with their **context features** and analyzed the usefulness of the features for **idiom identification**; our experiments showed that combining morphologically and syntactically motivated features results in better idiom identification.

Conclusion and future work

We extracted MWEs with their **context features** and analyzed the usefulness of the features for **idiom identification**; our experiments showed that combining morphologically and syntactically motivated features results in better idiom identification.

Future work on the separation of longer and shorter versions of multiword expressions: taking into account **mutual associations** between the individual parts of a candidate MWE. [Zinsmeister and Heid, 2004].

Conclusion and future work

We extracted MWEs with their **context features** and analyzed the usefulness of the features for **idiom identification**; our experiments showed that combining morphologically and syntactically motivated features results in better idiom identification.

Future work on the separation of longer and shorter versions of multiword expressions: taking into account **mutual associations** between the individual parts of a candidate MWE. [Zinsmeister and Heid, 2004].

Additionally to the **monolingual features** presented here, **translational behaviour**, i.e. semantic transparency vs. opaqueness is also a suitable indicator for idiomaticity. [Villada Moirón, 2006], [Fritzing 2009]

References

- Michael Schiehlen, 2003: *A cascaded finite-state parser for German*. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL '03)*, Budapest
- Christopher D. Manning and Hinrich Schütze, 1999: *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts
- Begoña Villada Moirón and Jörg Tiedemann, 2006: *Identifying idiomatic expressions using automatic word-alignment*. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context*, Trento Italy
- Fabienne Fritzing, 2009: *Using parallel text for the extraction of German multiword expressions*. In *Lexis–E-journal in English Lexicology*. Issue 4: Corpus Linguistics and the Lexicon
- Heike Zinsmeister and Ulrich Heid, 2004: *Collocations of complex nouns: Evidence for lexicalization*. In *Proceedings of KONVENS-2004, Heidelberg*, Springer