Extraction of multiword expressions from parsed corpora using context features

Marion Weller   Ulrich Heid
[wellermn|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
"The head of Spain’s government, Felipe Gonzalez, also gave his approval."

<table>
<thead>
<tr>
<th>word form</th>
<th>pos tag</th>
<th>lemma</th>
<th>morph.-synt. features</th>
<th>governor</th>
<th>gramm. function</th>
<th>engl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spaniens</td>
<td>NE</td>
<td>Spanien</td>
<td>Gen:Sg</td>
<td>1</td>
<td>GL</td>
<td>Spain’s</td>
</tr>
<tr>
<td>Regierungs</td>
<td>NN</td>
<td>Regierungs:</td>
<td>Nom:M:Sg</td>
<td>3</td>
<td>SUBJ</td>
<td>head of government</td>
</tr>
<tr>
<td>chef</td>
<td></td>
<td>chef</td>
<td></td>
<td></td>
<td></td>
<td>Felipe</td>
</tr>
<tr>
<td>Felipe</td>
<td>NN</td>
<td>Felipe</td>
<td>Nom:M:Sg</td>
<td>1</td>
<td>APP</td>
<td>Gonzalez</td>
</tr>
<tr>
<td>Gonzalez</td>
<td></td>
<td>Gonzalez</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gab</td>
<td>VVFIN</td>
<td>geben</td>
<td>3:Sg:Past</td>
<td>-1</td>
<td>TOP</td>
<td>gave</td>
</tr>
<tr>
<td>ebenfalls</td>
<td>ADV</td>
<td>ebenfalls</td>
<td></td>
<td>3</td>
<td>6</td>
<td>ADJ</td>
</tr>
<tr>
<td>grünes</td>
<td>ADJA</td>
<td>grün</td>
<td></td>
<td>6</td>
<td>ADJ</td>
<td>green</td>
</tr>
<tr>
<td>Licht</td>
<td>NN</td>
<td>Licht</td>
<td>Akk:N:Sg</td>
<td>3</td>
<td>OBJ&lt;sub&gt;acc&lt;/sub&gt;</td>
<td>light</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td>-1</td>
<td>TOP</td>
<td>.</td>
</tr>
</tbody>
</table>
"The head of Spain’s government, Felipe Gonzalez, also gave his approval."

gaben
"The head of Spain’s government, Felipe Gonzalez, also gave his approval."

Licht geben
"The head of Spain’s government, Felipe Gonzalez, also gave his approval."

grünes Licht geben
Context features: assessing idiomaticity
Morphologically motivated features

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

<table>
<thead>
<tr>
<th>MWE</th>
<th>f</th>
<th>NUM</th>
<th>DET</th>
<th>engl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>in Jahr aussehen</td>
<td>271</td>
<td>121</td>
<td>129</td>
<td>in year look</td>
</tr>
<tr>
<td>auf Barrikade gehen</td>
<td>167</td>
<td>2</td>
<td>165</td>
<td>on barricade go: to go on the warpath</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>MWE</th>
<th>f</th>
<th>NUM</th>
<th>DET</th>
<th>engl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>in Jahr aussehen</td>
<td>271</td>
<td>121</td>
<td>150</td>
<td>in year look</td>
</tr>
<tr>
<td>auf Barrikade gehen</td>
<td>167</td>
<td>2</td>
<td>165</td>
<td>on barricade go:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>to go on the warpath</td>
</tr>
</tbody>
</table>

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*

<table>
<thead>
<tr>
<th>MWE</th>
<th>f</th>
<th>negated</th>
<th>engl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aus dem Kopf gehen</td>
<td>47</td>
<td>47</td>
<td>to get out of the head</td>
</tr>
</tbody>
</table>
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:

\[
\frac{pos(P) + pos(N) + pos(V)}{pos(N)} = 3
\]

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:

\[ \frac{pos(P) + pos(N) + pos(V)}{pos(N)} = 3 \]

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:

\[
\frac{pos(P) + pos(N) + pos(V)}{pos(N)} = 3
\]

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“.
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.
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.

\[
\text{In Stellung gebracht worden seien Raketen mit einer Reichweite von 200 km.}
\]

Missiles with a range of 200 km had been positioned.

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

\[
\text{In die Klinik hatten die Eltern sie gegen ihren Willen gebracht.}
\]

The parents took her into a hospital against her will.

* *In die Klinik hatten die Eltern sie gegen ihren Willen gebracht.*
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* \}
- 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 ≥ 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

<table>
<thead>
<tr>
<th>feature</th>
<th>number</th>
<th>det</th>
<th>neg</th>
<th>adjacency</th>
<th>vorfeld</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAP</td>
<td>0.607</td>
<td>0.605</td>
<td>0.643</td>
<td>0.694</td>
<td>0.566</td>
</tr>
</tbody>
</table>

UAP-values for the morpho-syntactic features computed separately

<table>
<thead>
<tr>
<th>grouped</th>
<th>$M_1$ det+num</th>
<th>$M_2$ det+num+neg</th>
<th>$S$ adja+vorfeld</th>
<th>$M_2 + S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAP</td>
<td>0.635</td>
<td>0.681</td>
<td>0.664</td>
<td>0.830</td>
</tr>
</tbody>
</table>

UAP-values: sorted according to scores based on combined features
Experiment: Expanding basic patterns

- Identify PNV-triples with a clear preference for
  (i) a specific adjective
  (ii) no adjective at all

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

<table>
<thead>
<tr>
<th>size of test set</th>
<th>1013 [all]</th>
<th>610 [ADJ≤0.1]</th>
<th>133 [ADJ=0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>idioms</td>
<td>513</td>
<td>390</td>
<td>99</td>
</tr>
<tr>
<td>UAP</td>
<td>0.833</td>
<td>0.892</td>
<td>0.937</td>
</tr>
</tbody>
</table>

Sorting based on morpho-syntactic criteria and percentage of adjectives.
Experiment: Expanding basic patterns

- Identify PNV-triples with a clear preference for
  (i) a specific adjective
  (ii) no adjective at all

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

<table>
<thead>
<tr>
<th>size of test set</th>
<th>1013 [all]</th>
<th>610 [ADJ≤0.1]</th>
<th>133 [ADJ=0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>idioms</td>
<td>513</td>
<td>390</td>
<td>99</td>
</tr>
<tr>
<td>UAP</td>
<td>0.833</td>
<td>0.892</td>
<td>0.937</td>
</tr>
</tbody>
</table>

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

Take account of *creative use of language*:
Threshold (ADJ ≤ 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

<table>
<thead>
<tr>
<th>PNV-triple</th>
<th>adjective</th>
<th>ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>auf Bank schieben</td>
<td>lang</td>
<td>1</td>
</tr>
<tr>
<td>mit Wirkung bestellen</td>
<td>sofortig</td>
<td>1</td>
</tr>
<tr>
<td>zu Fixing verbilligen</td>
<td>frankfurter</td>
<td>1</td>
</tr>
<tr>
<td>auf Fuß setzen</td>
<td>frei</td>
<td>0.997</td>
</tr>
<tr>
<td>in Gang sein</td>
<td>voll</td>
<td>0.992</td>
</tr>
</tbody>
</table>

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

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 employes 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.
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 employes 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.

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], [Fritzinger 2009]
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 identifier for idiomaticity. [Villada Moirón, 2006], [Fritzinger 2009]
References


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 Fritzinger, 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