Phrase processing for detecting collocations with KoKS*

*Korpusbasierte Kollokationssuche (corpus based search for collocations)

University of Osnabrück (Germany): KoKS-Project
contents

- detection of phrases
  - bla
- identifications of collocations
- evaluation (results)
system overview

Parallel corpora & dictionaries

detecting phrase correspondences

identification of collocations

KoKS database
System overview

Parallel corpora & dictionaries

Detecting phrase correspondences

Identification of collocations

KoKS database
used bilingual corpora

- DE-News
  - radio news broadcast
  - translated by volunteers

- EU-publications
  - press releases
  - political documents
  - contracts

- the four Harry Potter books
system overview

Parallel corpora & dictionaries

Detecting phrase correspondences

Identification of collocations

KoKS database

University of Osnabrück (Germany): KoKS-Project
system overview

- POS-tagging
- Alignment
- Parallel corpora & dictionaries
- Detecting phrase correspondences
- Identification of collocations
- KoKS database
alignment of sentences

- distance measure
  - bilingual dictionaries
  - character trigram to identify cognats
  - sentence length
alignment of sentences

It stared back.

Die Katze starrte zurück.

open class words

bilingual dictionaries
character trigram to identify cognats
sentence length
system overview

POS-tagging

alignment

detecting phrase correspondences

identification of collocations

Parallel corpora & dictionaries

KoKS database

University of Osnabrück (Germany): KoKS-Project
detecting phrase correspondences

- POS tags sequences
  - extracted from chunk-parsed monolingual corpora
  - distinguished by syntactic category

- example:
The school’s party was called off.

Die [Fete] zum Ferienbeginn fiel {ins} Wasser.
detecting phrase correspondences

- POS tags sequences
  - extracted from chunk-parsed monolingual corpora
  - distinguished by syntactic category
- pair matching phrases
- example:
<table>
<thead>
<tr>
<th>DT</th>
<th>NN</th>
<th>VBZ</th>
<th>IN</th>
<th>NN</th>
<th>VBD</th>
<th>VBN</th>
<th>RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>school</td>
<td>'s</td>
<td>out</td>
<td>party</td>
<td>was</td>
<td>called</td>
<td>{off}</td>
</tr>
</tbody>
</table>

Pair:

<table>
<thead>
<tr>
<th>ART</th>
<th>NN</th>
<th>APPART</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>Fete</td>
<td>zum</td>
<td>Ferienbeginn</td>
</tr>
</tbody>
</table>

Pair:

<table>
<thead>
<tr>
<th>VVFIN</th>
<th>APPART</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>fiel</td>
<td>{ins}</td>
<td>Wasser.</td>
</tr>
</tbody>
</table>

PP

VP

University of Osnabrück (Germany): KKS-Project
detecting phrase correspondences

- multiple NPs
- identify non-literal-phrases
- no word alignment is used
- all combinations are considered
- a predefined number of references is required
system overview

- POS-tagging
- Parallel corpora & dictionaries
- Alignment
- Detecting phrase correspondences
- Identification of collocations
- KoKS database
collocativity measure

- Breidt`s definition of collocations
  - compositional semantics
- translation as semantics
- distance measure used in sentence alignment
results

- detecting phrase correspondences
- collocatitivity measure
so fare, we processed
  – all sentences with at most 19 words
  – apprx. 70,000 sentence pairs

next table shows examples
  – ordered by frequency ($f$)
<table>
<thead>
<tr>
<th>rank</th>
<th>$f$</th>
<th>German</th>
<th>English</th>
<th>correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>30</td>
<td>Professor</td>
<td>Dumbledore</td>
<td>bad</td>
</tr>
<tr>
<td>23</td>
<td>30</td>
<td>die Tür (the door)</td>
<td>Harry</td>
<td>bad</td>
</tr>
<tr>
<td>24</td>
<td>29</td>
<td>Professor</td>
<td>Professor Lupin</td>
<td>near</td>
</tr>
<tr>
<td>25</td>
<td>29</td>
<td>Schloss</td>
<td>the castle</td>
<td>good</td>
</tr>
<tr>
<td>33</td>
<td>25</td>
<td>zu Harry</td>
<td>to Harry</td>
<td>good</td>
</tr>
<tr>
<td>34</td>
<td>24</td>
<td>will</td>
<td>do n't want</td>
<td>near</td>
</tr>
<tr>
<td>35</td>
<td>24</td>
<td>schien</td>
<td>seemed to be</td>
<td>good</td>
</tr>
<tr>
<td>36</td>
<td>24</td>
<td>ist</td>
<td>do n't know</td>
<td>bad</td>
</tr>
<tr>
<td>37</td>
<td>24</td>
<td>sagte (said)</td>
<td>'ve got</td>
<td>bad</td>
</tr>
<tr>
<td>38</td>
<td>23</td>
<td>Dementoren</td>
<td>the dementors</td>
<td>good</td>
</tr>
<tr>
<td>39</td>
<td>22</td>
<td>Kammer</td>
<td>the Chamber</td>
<td>good</td>
</tr>
<tr>
<td>rank</td>
<td>f</td>
<td>German</td>
<td>English</td>
<td>correspondence</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>-------------------</td>
<td>-------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>22</td>
<td>30</td>
<td>Professor</td>
<td>Dumbledore</td>
<td>bad</td>
</tr>
<tr>
<td>23</td>
<td>30</td>
<td>die Tür (the door)</td>
<td>Harry</td>
<td>bad</td>
</tr>
<tr>
<td>24</td>
<td>29</td>
<td>Professor</td>
<td>Professor Lupin</td>
<td>near</td>
</tr>
<tr>
<td>25</td>
<td>29</td>
<td>Schloss</td>
<td>the castle</td>
<td>good</td>
</tr>
<tr>
<td>33</td>
<td>25</td>
<td>zu Harry</td>
<td>to Harry</td>
<td>good</td>
</tr>
<tr>
<td>34</td>
<td>24</td>
<td>will</td>
<td>do n't want</td>
<td>near</td>
</tr>
<tr>
<td>35</td>
<td>24</td>
<td>schien</td>
<td>seemed to be</td>
<td>good</td>
</tr>
<tr>
<td>36</td>
<td>24</td>
<td>ist</td>
<td>do n't know</td>
<td>bad</td>
</tr>
<tr>
<td>37</td>
<td>24</td>
<td>sagte (said)</td>
<td>'ve got</td>
<td>bad</td>
</tr>
<tr>
<td>38</td>
<td>23</td>
<td>Dementorenen</td>
<td>the dementors</td>
<td>good</td>
</tr>
<tr>
<td>39</td>
<td>22</td>
<td>Kammer</td>
<td>the Chamber</td>
<td>good</td>
</tr>
</tbody>
</table>
### Results (Phrase Detection) 2/3

<table>
<thead>
<tr>
<th>Rank</th>
<th>f</th>
<th>German</th>
<th>English</th>
<th>Correspondence</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>30</td>
<td>Professor</td>
<td>Dumbledore</td>
<td>bad</td>
</tr>
<tr>
<td>23</td>
<td>30</td>
<td>die Tür (the door)</td>
<td>Harry</td>
<td>bad</td>
</tr>
<tr>
<td>24</td>
<td>29</td>
<td>Professor</td>
<td>Professor Lupin</td>
<td>near</td>
</tr>
<tr>
<td>25</td>
<td>29</td>
<td>Schloss</td>
<td>the castle</td>
<td>good</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>25</td>
<td>zu Harry</td>
<td>to Harry</td>
<td>good</td>
</tr>
<tr>
<td>34</td>
<td>24</td>
<td>will</td>
<td>do n't want</td>
<td>near</td>
</tr>
<tr>
<td>35</td>
<td>24</td>
<td>schien</td>
<td>seemed to be</td>
<td>good</td>
</tr>
<tr>
<td>36</td>
<td>24</td>
<td>ist</td>
<td>do n't know</td>
<td>bad</td>
</tr>
<tr>
<td>37</td>
<td>24</td>
<td>sagte (said)</td>
<td>'ve got</td>
<td>bad</td>
</tr>
<tr>
<td>38</td>
<td>23</td>
<td>Dementoren</td>
<td>the dementors</td>
<td>good</td>
</tr>
<tr>
<td>39</td>
<td>22</td>
<td>Kammer</td>
<td>the Chamber</td>
<td>good</td>
</tr>
</tbody>
</table>
• candidate set with \( f > 6 \)
  – does not contain any collocations according to Breidt (human annotators)
  – a lot of compositional compounds
  – only a few non-compositional translations

• useless to apply collocativity measure
results (collocativity measure)

- manually aligned phrase pairs
  - 250 phrase pairs
  - 83 with non-compositional translation
  - 45 with non-compositional semantics
    (Breidt’s definition of collocation)
  - agreement of two annotators
  - 31 unresolved disagreements
**results (collocativity measure)**

<table>
<thead>
<tr>
<th>variant</th>
<th>ignores words with high $f$</th>
<th>uses length of phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>no</td>
<td>only if very different</td>
</tr>
<tr>
<td>01</td>
<td>no</td>
<td>always</td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>only if very different</td>
</tr>
<tr>
<td>11</td>
<td>yes</td>
<td>always</td>
</tr>
</tbody>
</table>
results (collocativity measure)

precision (compositional translation)

250 candidates
results (collocativity measure)

recall (compositional translation)

250 candidates

measure 00
measure 01
measure 10
measure 11
results (collocativity measure)

precision (compositional semantics)

University of Osnabrück (Germany): oKoS-Project
results (collocativity measure)

recall (compositional semantics)

measure 00
measure 01
measure 10
measure 11

250 candidates
• improve phrase correspondences
  – use proper chunking to find phrases
  – use word alignment
  – weight phrase pairs according to their correspondence probability
  – replace simple counts with advanced statistics (associations measure)
  – exploit substring relations among phrases
outlook

- improve collocativity measure
  - decompose composita
  - find translation equivalences accross word classes
  - better combine the different parts
discussion / questions / contact

- Norman Kummer, norman@VauDePe.de
- Joachim Wagner, jowagner@uos.de

University of Osnabrück
Institute of Cognitive Science
49078 Osnabrück
Germany

Link:
http://www.cl-ki.uos.de/~koks/
alignment of sentences (extra 1)

Die Katze starrte zurück.

It stared back.
alignment of sentences (extra 2)
system overview

- POS-tagging
- Parallel corpora & dictionaries
- alignment
- detecting phrase correspondences
- identification of collocations

University of Osnabrück (Germany): KoKS-Project
Application

- CALL-context
- Provides help to L2 learner in text understanding
- Web based interface
current KoKS demo application (screen-shot)
other possible applications

- intelligent lexicon lookup (iKoS)
- translation memory in CAT (computer assisted translation)
- full text search based on the lemmas
Hinsichtlich der Begründung, warum wir manuell arbeiten, ist Recall eigentlich ausschlaggebend. (Keine Kollokation gefunden, Obwohl vermutlich welche vorhanden.) -> ans Ende, falls Fragen