Term and Collocation Extraction by means of complex Linguistic Web Services

Ulrich Heid, Fabienne Fritzinger, Erhard Hinrichs, Marie Hinrichs, Thomas Zastrow

Institut für maschinelle Sprachverarbeitung, Universität Stuttgart
and Seminar für Sprachwissenschaft, Universität Tübingen
Germany

Linguistic Resources and Evaluation Conference, 2010: Valletta, Malta
Overview

• Objectives and scenarios addressed
• Data used for experimentation
• Procedures to extract single word term candidates
• Procedures to extract collocation candidates
• Combining the tools for both extraction tasks
• The extraction as a web service: Architecture – technical issues addressed – open questions
• Conclusion – Future Work
Overview

- Objectives and scenarios addressed
Overview

- Objectives and scenarios addressed
- Data used for experimentation
Overview

- Objectives and scenarios addressed
- Data used for experimentation
- Procedures to extract single word term candidates
Overview

- Objectives and scenarios addressed
- Data used for experimentation
- Procedures to extract single word term candidates
- Procedures to extract collocation candidates
Overview

- Objectives and scenarios addressed
- Data used for experimentation
- Procedures to extract single word term candidates
- Procedures to extract collocation candidates
- Combining the tools for both extraction tasks
Overview

- Objectives and scenarios addressed
- Data used for experimentation
- Procedures to extract single word term candidates
- Procedures to extract collocation candidates
- Combining the tools for both extraction tasks
- The extraction as a web service:
  Architecture – technical issues addressed – open questions
Overview

- Objectives and scenarios addressed
- Data used for experimentation
- Procedures to extract single word term candidates
- Procedures to extract collocation candidates
- Combining the tools for both extraction tasks
- The extraction as a web service:
  - Architecture – technical issues addressed – open questions
- Conclusion – Future Work
Objectives

- Provision of computational linguistic tools for
- Term candidate extraction
- Collocation candidate extraction
- Extraction of regionalism candidates
- Tools based on standard corpus processing techniques: Tagging – parsing – pattern-based extraction – lexicostatistics
- Tools wrapped and provided as chains of web services:
  - to assess possibilities of creating complex linguistic web services
  - to test the processing of non-trivial amounts of data via web services
Objectives

- Provision of computational linguistic tools for
  - Term candidate extraction
  - Collocation candidate extraction
  - Extraction of regionalism candidates
Objectives

- Provision of computational linguistic tools for
  - Term candidate extraction
  - Collocation candidate extraction
  - Extraction of regionalism candidates

- Tools based on standard corpus processing techniques:
  Tagging – parsing – pattern-based extraction – lexicostatistics
Objectives

• Provision of computational linguistic tools for
  • Term candidate extraction
  • Collocation candidate extraction
  • Extraction of regionalism candidates

• Tools based on standard corpus processing techniques:
  Tagging – parsing – pattern-based extraction – lexicostatistics

• Tools wrapped and provided as chains of web services:
  • to assess possibilities of creating complex linguistic web services
  • to test the processing of non-trivial amounts of data via web services
Scenarios addressed

- **Type I:** single word term candidate extraction
  - to find specialized terms of a specific domain of knowledge
  - to find lexical material specific of a given region: Germany – Austria – Switzerland – South Tyrol

- **Type II:** extraction of multiword expressions (MWEs)
  - to find collocations (cf. Weller & Heid, this session)
  - to find multiword terms and phraseology of specialized domains
  - to find collocations typical of a “region” (D – A – CH – ST)
Scenarios addressed

- **Type I**: single word term candidate extraction
  - to find specialized terms of a specific domain of knowledge
  - to find lexical material specific of a given region:
    German of: Germany – Austria – Switzerland – South Tyrol
Scenarios addressed

- **Type I**: single word term candidate extraction
  - to find specialized terms of a specific domain of knowledge
  - to find lexical material specific of a given region: German of: Germany – Austria – Switzerland – South Tyrol

- **Type II**: extraction of multiword expressions (MWEs)
  - to find collocations (cf. Weller & Heid, this session)
  - to find multiword terms and phraseology of specialized domains
  - to find collocations typical of a “region” (D – A – CH – ST)
Data used in the experiments

Work on German texts
Data used in the experiments

Work on German texts

- General Language: newspaper texts
  - *Die Zeit* (1999 - 2005) 50 M
  - total newspapers ca. 254 M

- Specialized language (taken from the OPUS Website):
  - *European Medecine Agency (EMEA)*: pharmaceuticals tests 10 M

- National or regional variants of German:
  - Austria (excerpts from the DeReKo corpus of IdS Mannheim) 180 M
  - Switzerland (dito: DeReKo) 180 M
  - South Tyrol (Eurac/Athesia publishers) ca. 60 M
Data used in the experiments
Work on German texts

- General Language: newspaper texts
  - *Die Zeit* (1999 - 2005) 50 M
  - total newspapers ca. 254 M

- Specialized language (taken from the OPUS Website):
  - European Medecine Agency (EMEA): pharmaceuticals tests 10 M
Data used in the experiments

Work on German texts

- General Language: newspaper texts
  - Frankfurter Rundschau (1992/1993) 40 M
  - Die Zeit (1999 - 2005) 50 M
  - Stuttgarter Zeitung (1992/1993) 36 M
  - total newspapers ca. 254 M

- Specialized language (taken from the OPUS Website):
  - European Medecine Agency (EMEA): pharmaceuticals tests 10 M

- National or regional variants of German:
  - Austria (excerpts from the DeReKo corpus of IdS Mannheim) 180 M
  - Switzerland (dito: DeReKo) 180 M
  - South Tyrol (Eurac/Athesia publishers) ca. 60 M
Procedures for single word term candidate extraction
Based of relative frequency relationships

“Weirdness scores”

Ahmad et al. 1992
Procedures for single word term candidate extraction
Based of relative frequency relationships

“Weirdness scores”

- Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere

Ahmad et al. 1992
Procedures for single word term candidate extraction
Based on relative frequency relationships

“Weirdness scores”

- Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere
- Calculation: for each noun, verb, adjective from the specialized text:
  - RS: Relative frequency in the specialized text:
    number of occurrences / corpus size (by POS) of the specialized text
  - RG: Relative frequency of the same item in general language text:
    newspapers taken to be without bias for a given domain
  - Relationship RS/RG

- Output:
  1. items occurring only in the specialized text
  2. items more frequent in the specialized text than elsewhere
Procedures for single word term candidate extraction
Based of relative frequency relationships

“Weirdness scores”

- Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere
- Calculation: for each noun, verb, adjective from the specialized text:
  - RS: Relative frequency in the specialized text:
    number of occurrences / corpus size (by POS) of the specialized text
  - RG: Relative frequency of the same item in general language text:
    newspapers taken to be without bias for a given domain
  - Relationship RS/RG

Output:
1. items occurring only in the specialized text
2. items more frequent in the specialized text than elsewhere

Ahmad et al. 1992
Procedures for single word term candidate extraction
Based of relative frequency relationships

“Weirdness scores”

- **Intuition:**
  Terms from a domain are more frequent in domain-specific texts than elsewhere

- **Calculation:** for each noun, verb, adjective from the specialized text:
  - **RS:** Relative frequency in the specialized text:
    number of occurrences / corpus size (by POS) of the specialized text
  - **RG:** Relative frequency of the same item in general language text:
    newspapers taken to be without bias for a given domain

Ahmad et al. 1992
Procedures for single word term candidate extraction
Based of relative frequency relationships

“Weirdness scores”

- Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere

- Calculation: for each noun, verb, adjective from the specialized text:
  - RS: Relative frequency in the specialized text: number of occurrences / corpus size (by POS) of the specialized text
  - RG: Relative frequency of the same item in general language text: newspapers taken to be without bias for a given domain
  - Relationship RS/RG

Ahmad et al. 1992
Procedures for single word term candidate extraction
Based on relative frequency relationships

“Weirdness scores”

- Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere

- Calculation: for each noun, verb, adjective from the specialized text:
  - RS: Relative frequency in the specialized text:
    number of occurrences / corpus size (by POS) of the specialized text
  - RG: Relative frequency of the same item in general language text:
    newspapers taken to be without bias for a given domain
  - Relationship RS/RG

- Output:

Ahmad et al. 1992
Procedures for single word term candidate extraction
Based of relative frequency relationships

“Weirdness scores”

- Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere

- Calculation: for each noun, verb, adjective from the specialized text:
  - RS: Relative frequency in the specialized text: number of occurrences / corpus size (by POS) of the specialized text
  - RG: Relative frequency of the same item in general language text: newspapers taken to be without bias for a given domain
  - Relationship RS/RG

- Output:
  1 items occurring only in the specialized text

Ahmad et al. 1992
Procedures for single word term candidate extraction

Based of relative frequency relationships

“Weirdness scores”

Ahmad et al. 1992

• Intuition:
  Terms from a domain are more frequent in domain-specific texts than elsewhere

• Calculation: for each noun, verb, adjective from the specialized text:
  • RS: Relative frequency in the specialized text:
    number of occurrences / corpus size (by POS) of the specialized text
  • RG: Relative frequency of the same item in general language text:
    newspapers taken to be without bias for a given domain
  • Relationship RS/RG

• Output:
  1. items occurring only in the specialized text
  2. items more frequent in the specialized text than elsewhere
### Procedures for single word term candidate extraction

**Scenario type I: typical results – term candidates from EMEA**

<table>
<thead>
<tr>
<th>term candidates</th>
<th>f (abs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durchstechflasche</td>
<td>5638</td>
</tr>
<tr>
<td>Injektionsstelle</td>
<td>3489</td>
</tr>
<tr>
<td>Pharmakokinetik</td>
<td>3426</td>
</tr>
<tr>
<td>Hämoglobinwert</td>
<td>3395</td>
</tr>
<tr>
<td>Fertigspritze</td>
<td>3271</td>
</tr>
<tr>
<td>Ribavirin</td>
<td>3234</td>
</tr>
<tr>
<td>Gebrauchsinformation</td>
<td>2801</td>
</tr>
<tr>
<td>Dosisanpassung</td>
<td>2580</td>
</tr>
<tr>
<td>Epoetin</td>
<td>2302</td>
</tr>
<tr>
<td>Hydrochlorothiazid</td>
<td>2128</td>
</tr>
</tbody>
</table>

### Only EMEA (not FR)

<table>
<thead>
<tr>
<th>term candidates</th>
<th>weirdness</th>
<th>f (abs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filmtabletten</td>
<td>25522</td>
<td>6389</td>
</tr>
<tr>
<td>Injektionslösung</td>
<td>19854</td>
<td>4970</td>
</tr>
<tr>
<td>Packungsbeilage</td>
<td>14710</td>
<td>7365</td>
</tr>
<tr>
<td>Niereninsuffizienz</td>
<td>14233</td>
<td>3563</td>
</tr>
<tr>
<td>Verkehrstüchtigkeit</td>
<td>13558</td>
<td>3394</td>
</tr>
<tr>
<td>Leberfunktion</td>
<td>8385</td>
<td>2099</td>
</tr>
<tr>
<td>Hypoglykämie</td>
<td>8353</td>
<td>2091</td>
</tr>
<tr>
<td>Toxizität</td>
<td>7957</td>
<td>1992</td>
</tr>
<tr>
<td>Einnehmen</td>
<td>7035</td>
<td>7045</td>
</tr>
<tr>
<td>Hypotonie</td>
<td>6823</td>
<td>1708</td>
</tr>
</tbody>
</table>

### EMEA and FR
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative
Procedures for collocation candidate extraction
Why not use a flat approach – dependency parsing as an alternative

- English: pattern-based extraction + sorting by AMs  Kilgarriff et al. 2004
  - configurational: subject < verb < object
  - little morphological form variation
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative

- **English**: pattern-based extraction + sorting by AMs  
  - configurational: subject < verb < object  
  - little morphological form variation  
  Kilgarriff et al. 2004

- **German**:  
  Problems in transferring the Sketch Engine approach  
  Ivanova et al. 2008
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative

- English: pattern-based extraction + sorting by AMs  
  - configurational: subject < verb < object  
  - little morphological form variation

- German:
  Problems in transferring the Sketch Engine approach  
  - three models of word order ⇒ need three sets of patterns

Kilgarriff et al. 2004
Ivanova et al. 2008
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative

- **English**: pattern-based extraction + sorting by AMs  
  - configurational: subject < verb < object  
  - little morphological form variation  
  Kilgarriff et al. 2004

- **German**:  
  Problems in transferring the Sketch Engine approach  
  - three models of word order ⇒ need three sets of patterns  
  - constituent order in the topological Mittelfeld: rather free  
  ⇒ need to permute the patterns  
  Ivanova et al. 2008
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative

- English: pattern-based extraction + sorting by AMs  
  - configurational: subject < verb < object
  - little morphological form variation  
  Kilgarriff et al. 2004

- German:  
  Problems in transferring the Sketch Engine approach
  - three models of word order ⇒ need three sets of patterns
  - constituent order in the topological Mittelfeld: rather free
    ⇒ need to permute the patterns
  - case syncretism of German:
    only 22% of all German NPs in Negra are unambiguous
  ⇒ low precision of flat analysis  
  Ivanova et al. 2008

Evert 2004
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative

- English: pattern-based extraction + sorting by AMs
  - configurational: subject < verb < object
  - little morphological form variation

- German:
  Problems in transferring the Sketch Engine approach
  - three models of word order ⇒ need three sets of patterns
  - constituent order in the topological Mittelfeld: rather free
    ⇒ need to permute the patterns
  - case syncretism of German:
    only 22% of all German NPs in Negra are unambiguous
    ⇒ low precision of flat analysis

- Alternative: Dependency parsing
Procedures for collocation candidate extraction

Why not use a flat approach – dependency parsing as an alternative

- **English**: pattern-based extraction + sorting by AMs
  - configurational: subject < verb < object
  - little morphological form variation

- **German**:
  Problems in transferring the Sketch Engine approach
  - three models of word order ⇒ need three sets of patterns
  - constituent order in the topological Mittelfeld: rather free ⇒ need to permute the patterns
  - case syncretism of German:
    - only 22% of all German NPs in Negra are unambiguous ⇒ low precision of flat analysis

- **Alternative**: Dependency parsing

```
Corpus I  Parsing  Corpus I (parsed)  Collocation Extraction  Collocations  Calculation of associative strength  Significant Collocations
```
Procedures for collocation candidate extraction

Sample dependency analysis

Use of FSPar

Schiehlen 2003

Die zweite Studie lieferte ähnliche Ergebnisse.
Procedures for collocation candidate extraction
Scenario type II: typical results – verb+object pairs from Swiss newspapers

| Abklärung  | treffen  | 96 |
| Abklärung  | vornehmen| 91 |
| Anlaß      | besuchen| 73 |
| Anlaß      | durchführen| 199 |
| Anlaß      | organisieren| 367 |
| Beschwerde | gutheißen| 88 |
| Bilanz     | deponieren| 82 |
| Busse      | aussprechen| 72 |
| Defizit    | budgetieren| 94 |
| Einsitz    | nehmen| 295 |
| Einsprache | erheben| 262 |
| Entscheid  | fällen| 79 |
| Gegensteuer| geben| 143 |
| Gesuch     | bewilligen| 90 |
Combining the two scenarios
Extraction of specialized collocations

Steps:
1. Find relevant single word terms (e.g. from EMEA or regional texts)
2. Extract collocation candidates only for these items
3. Output: candidates:
   - EMEA: domain-specific collocations
   - collocations of regionalisms (e.g. from CH)
The extraction as a web service

Framework

D-SPIN web service tool chain: *WebLicht*  
- Experiments with chaining of different corpus processing tools  
- Joint effort: Universities of Tübingen, Leipzig, BBAW Berlin and others 

Hinrichs et al. 2010
The extraction as a web service

Architecture principles

- Tool and resource wrappers: tools unchanged with respect to stand-alone version
- Slim format for data exchange between chained components: D-SPIN Text Corpus Format, TCF
- WebLicht used as:
  - Chaining tool and interface
  - Workflow infrastructure

Heid et al. 2010
The extraction as a web service

Technical problems to be addressed wrt the extraction scenarios

• Scenario I: comparison of two corpora
  • Uploading both corpora (e.g. in one 'file')
  • Or: keeping comparison data (e.g. from one journal) as an internal resource

• Scenario II: parsing of large amounts of data
  • Time-consuming (10 M words on a LINUX PC: ca. 30 min)
  • Web service should alert user when processing is done
The extraction as a web service

Technical problems to be addressed wrt the extraction scenarios

- Scenario I: comparison of two corpora
  - Uploading both corpora (e.g. in one 'file')
  - Or: keeping comparison data (e.g. from one journal) as an internal resource
The extraction as a web service

Technical problems to be addressed wrt the extraction scenarios

- Scenario I: comparison of two corpora
  - Uploading both corpora (e.g. in one 'file')
  - Or: keeping comparison data (e.g. from one journal) as an internal resource

- Scenario II: parsing of large amounts of data
  - Time-consuming (10 M words on a LINUX PC: ca. 30 min)
  - Web service should alert user when processing is done
The extraction as a web service
Open problems: parameterizing a complex web service

Users may wish to select options

- Tool-related options:
  - parser – association measures – collocation types ... to be used
  - Parameters to be given to the individual component tools

- Output-related options:
  - sorting of collocation candidates – format of the output
  - Possibly need for extra post-processing components
Conclusion – Future Work

• Computational linguistic tools for term and collocation extraction, based on standard corpus processing components

• Experiments of web service use:
  • works fine (version at IMS Stuttgart)
  • needs to be registered for WebLicht
    Hinrichs et al. 2010

• Future Work
  • Further development of extraction components
    Weller/Heid 2010
  • Integration of components into specific tool chains, e.g. for provision of raw material to lexicographers
  • Web service parameterization and pertaining user interfaces
Conclusion – Future Work

- Computational linguistic tools for term and collocation extraction, based on standard corpus processing components
Conclusion – Future Work

- Computational linguistic tools for term and collocation extraction, based on standard corpus processing components
- Experiments of web service use:
  - works fine (version at IMS Stuttgart)
  - needs to be registered for *WebLicht* (Hinrichs et al. 2010)
  - open questions wrt parameterization

Future Work

- Further development of extraction components (Weller/Heid 2010)
- Integration of components into specific tool chains, e.g. for provision of raw material to lexicographers
- Web service parameterization and pertaining user interfaces
Conclusion – Future Work

• Computational linguistic tools for term and collocation extraction, based on standard corpus processing components
• Experiments of web service use:
  • works fine (version at IMS Stuttgart)
  • needs to be registered for WebLicht
  • open questions wrt parameterization
• Future Work
  • Further development of extraction components
  • Integration of components into specific tool chains, e.g. for provision of raw material to lexicographers
  • Web service parameterization and pertaining user interfaces

Hinrichs et al. 2010

Weller/Heid 2010