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

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

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

Scenarios addressed

- Type I: single word term candidate extraction
  - to find specialilzed 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 specialilzed 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
  - *Frankfurter Rundschau* (1992/1993) 40 M
  - *Frankfurter Allgemeine Zeitung* (1995 - 1998) 78 M
  - *Die Zeit* (1999 - 2005) 50 M
  - *Stuttgarter Zeitung* (1992/1993) 36 M
  - *Handelsblatt* (1995 - 1998) 50 M
  - total newspapers ca. 254 M

Data used in the experiments

Work on German texts

- General Language: newspaper texts
  - *Frankfurter Rundschau* (1992/1993) 40 M
  - *Frankfurter Allgemeine Zeitung* (1995 - 1998) 78 M
  - *Die Zeit* (1999 - 2005) 50 M
  - *Stuttgarter Zeitung* (1992/1993) 36 M
  - *Handelsblatt* (1995 - 1998) 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
  - *Frankfurter Allgemeine Zeitung* (1995 - 1998) 78 M
  - *Die Zeit* (1999 - 2005) 50 M
  - *Stuttgarter Zeitung* (1992/1993) 36 M
  - *Handelsblatt* (1995 - 1998) 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

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" Ahmad et al. 1992

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

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:

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

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

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

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:

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

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

term candidates	f (abs.)
Durchstechflasche	5638
Injektionsstelle	3489
Pharmakokinetik	3426
Hämoglobinwert	3395
Fertigspritze	3271
Ribavirin	3234
Gebrauchsinformation	2801
Dosisanpassung	2580
Epoetin	2302
Hydrochlorothiazid	2128

Only EMEA (not FR)

term candidates	weirdness	f (abs.)
Filmtablette	25522	6389
Injektionslösung	19854	4970
Packungsbeilage	14710	7365
Niereninsuffizienz	14233	3563
Verkehrstüchtigkeit	13558	3394
Leberfunktion	8385	2099
Hypoglykämie	8353	2091
Toxizität	7957	1992
Einnehmen	7035	7045
Hypotonie	6823	1708

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 Kilgarriff et al. 2004
  - configurational: subject $<$ verb $<$ object
  - little morphological form variation
- German: Ivanova et al. 2008
  Problems in transferring the Sketch Engine approach

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
- German: Ivanova et al. 2008
  Problems in transferring the Sketch Engine approach
  - three models of word order $\Rightarrow$ need three sets of patterns

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
- German: Ivanova et al. 2008
  Problems in transferring the Sketch Engine approach
  - three models of word order $\Rightarrow$ need three sets of patterns
  - constituent order in the topological Mittelfeld: rather free
    $\Rightarrow$ need to permute the patterns

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
- German: Ivanova et al. 2008
  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 Evert 2004
    ⇒ low precision of flat analysis

Procedures for collocation candidate extraction

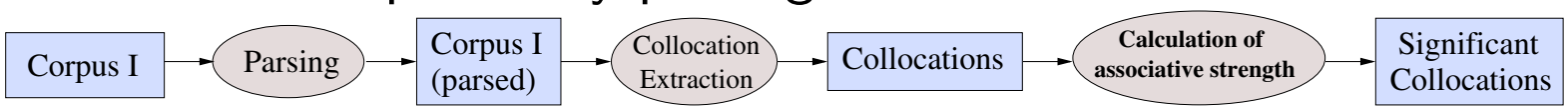
Why not use a flat approach – dependency parsing as an alternative

- English: pattern-based extraction + sorting by AMs Kilgarriff et al. 2004
  - configurational: subject < verb < object
  - little morphological form variation
- German: Ivanova et al. 2008
  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 Evert 2004
    ⇒ 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 Kilgarriff et al. 2004
  - configurational: subject < verb < object
  - little morphological form variation
- German: Ivanova et al. 2008
  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 Evert 2004
    ⇒ low precision of flat analysis
- Alternative: Dependency parsing

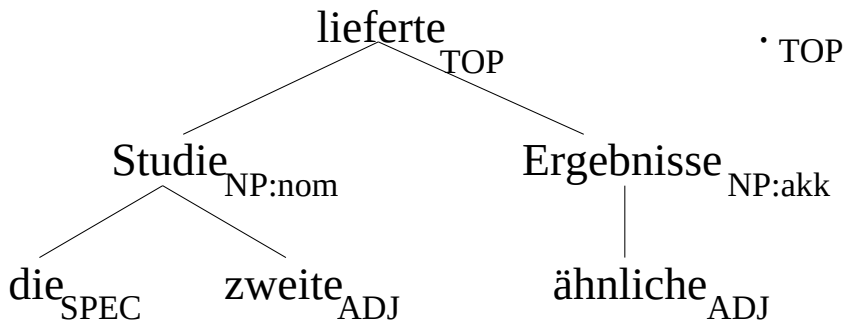


Procedures for collocation candidate extraction

Sample dependency analysis

Use of FSPar

Schiehlen 2003



0	Die	ART	d	|	2	SPEC
1	zweite	ADJA	2.	|	2	ADJ
2	Studie	NN	Studie	Nom:F:Sg	3	NP:1
3	lieferte	VVFIN	liefern	3:Sg:Past:Ind*	−1	TOP
4	ähnliche	ADJA	ähnlich	|	5	ADJ
5	Ergebnisse	NN	Ergebnis	Akk:N:Pl	3	NP:8
6	.	$.	.	|	−1	TOP

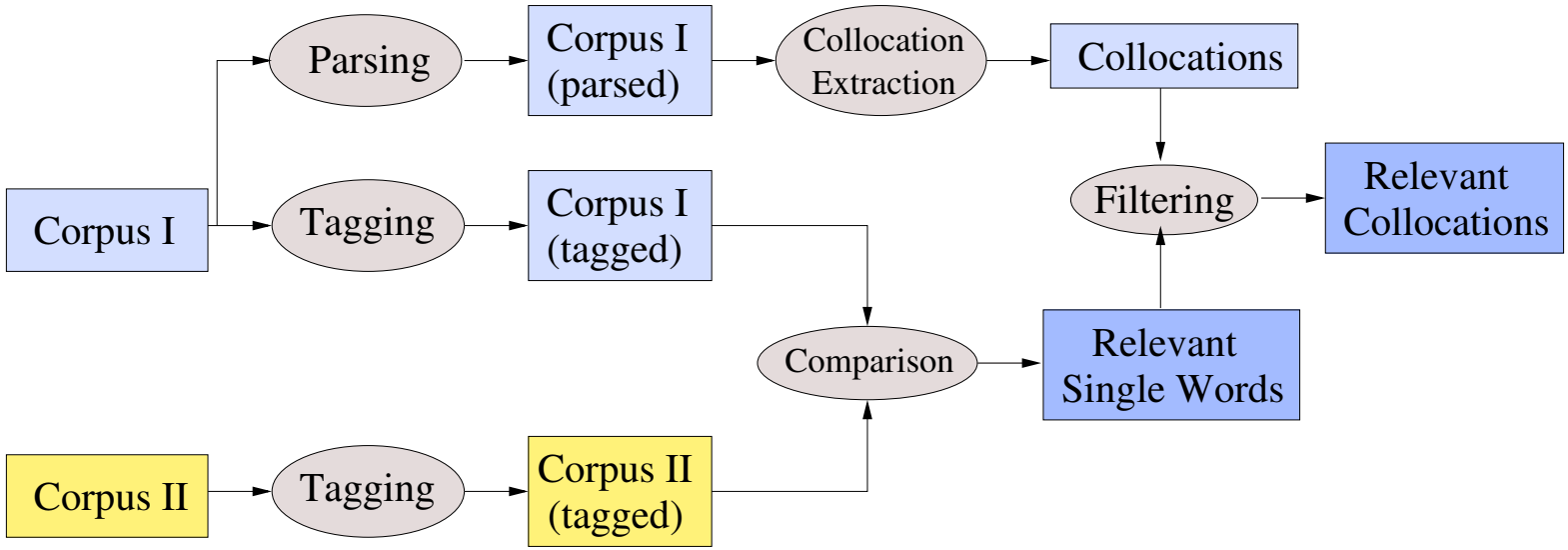
Procedures for collocation candidate extraction

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:

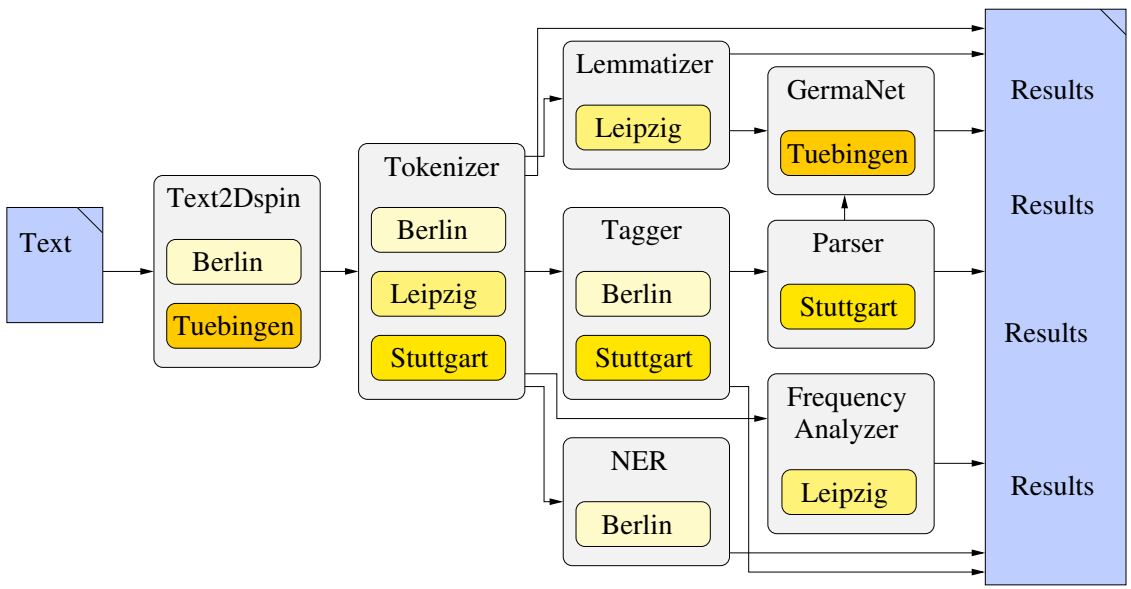
1. Find relevant single word terms (e.g. from EMEA or regional texts)
2. Extract collocation candidates only for these items
3. Output: candidates:
   - EMEA: domain-specific collocations
   - collocations of regionalisms (e.g. from CH)

The extraction as a web service

Framework

D-SPIN web service tool chain: *WebLicht* Hinrichs et al. 2010

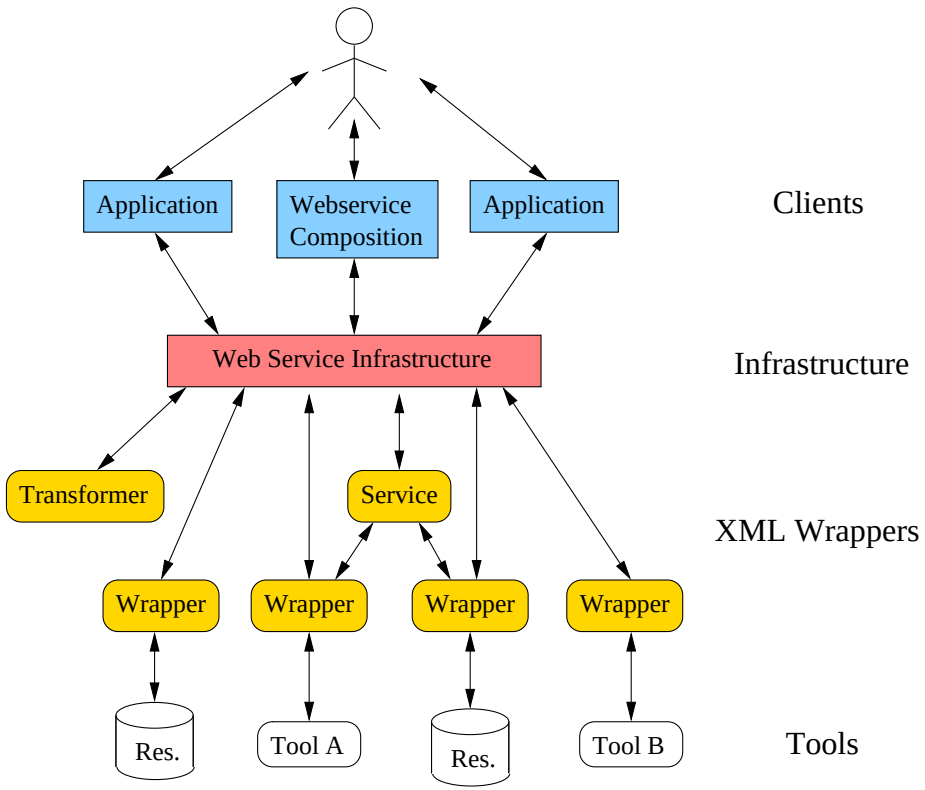
- Experiments with chaining of different corpus processing tools
- Joint effort: Universities of Tübingen, Leipzig, BBAW Berlin and others



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 Heid et al. 2010
- *WebLicht* used as:
  - Chaining tool and interface
  - Workflow infrastructure



The extraction as a web service

Technical problems to be addressed wrt the extraction scenarios

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

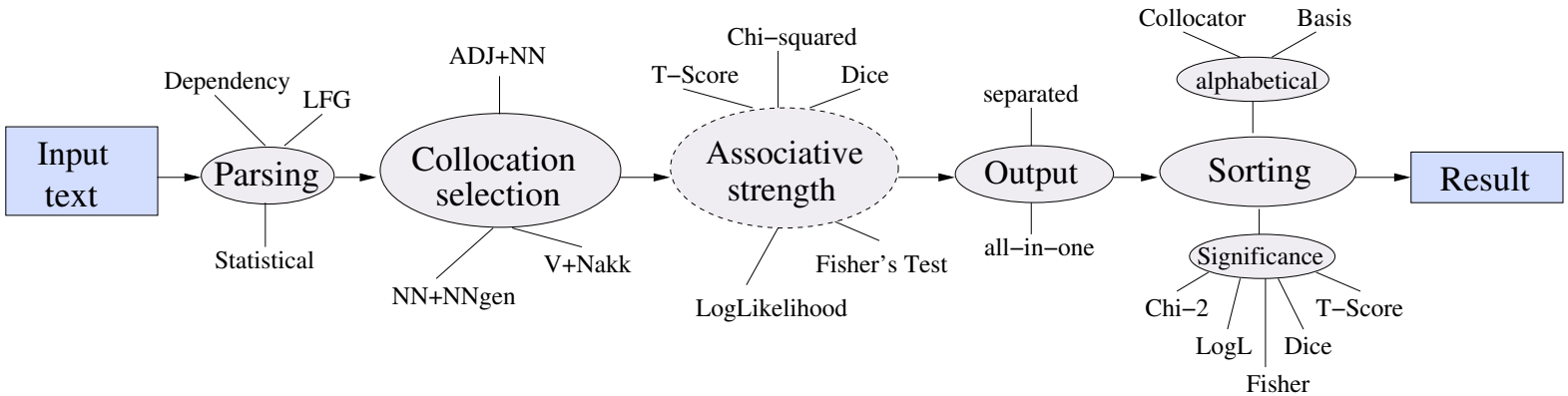
The extraction as a web service

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
  $\Longrightarrow$ Parameters to be given to the individual component tools
- Output-related options:
  sorting of collocation candidates – format of the output
  $\Longrightarrow$ Possibly need for extra post-processing components

Conclusion – Future Work

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

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