Introduction to Computational Linguistics

Resources

Jan-Philipp Söhn

jp.soehn@uni-tuebingen.de

February 13th, 2008

Today's Topic

- Prerequisite for NLP: large amounts of language data
  - Empirically founded research and theory construction
  - Development of processing techniques for text and speech
- Various sources (WWW, wordnets, etc.)
- Representation should adhere to international standards (e.g. XML)
- Collecting, processing/formatting, managing and providing language resources is a separate research area in computational linguistics

Today's Topic

- Prerequisite for NLP: large amounts of language data
  - Empirically founded research and theory construction
  - Development of processing techniques for text and speech
- Various sources (WWW, wordnets, etc.)
- Representation should adhere to international standards (e.g. XML)
- Collecting, processing/formatting, managing and providing language resources is a separate research area in computational linguistics

Today's Topic

- Prerequisite for NLP: large amounts of language data
  - Empirically founded research and theory construction
  - Development of processing techniques for text and speech
- Various sources (WWW, wordnets, etc.)
- Representation should adhere to international standards (e.g. XML)
- Collecting, processing/formatting, managing and providing language resources is a separate research area in computational linguistics

Today's Topic

- Prerequisite for NLP: large amounts of language data
  - Empirically founded research and theory construction
  - Development of processing techniques for text and speech
- Various sources (WWW, wordnets, etc.)
- Representation should adhere to international standards (e.g. XML)
- Collecting, processing/formatting, managing and providing language resources is a separate research area in computational linguistics

Today's Topic

- Prerequisite for NLP: large amounts of language data
  - Empirically founded research and theory construction
  - Development of processing techniques for text and speech
- Various sources (WWW, wordnets, etc.)
- Representation should adhere to international standards (e.g. XML)
- Collecting, processing/formatting, managing and providing language resources is a separate research area in computational linguistics

Today's Topic

- Prerequisite for NLP: large amounts of language data
  - Empirically founded research and theory construction
  - Development of processing techniques for text and speech
- Various sources (WWW, wordnets, etc.)
- Representation should adhere to international standards (e.g. XML)
- Collecting, processing/formatting, managing and providing language resources is a separate research area in computational linguistics

World Wide Web

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar "languages"
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar “languages”
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar "languages"
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar “languages”
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar "languages"
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar “languages”
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
  - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
  - Extensible Markup Language XML: arbitrary tags
  - Drawback: parallel development of similar "languages"
  - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

World Wide Web – Some Basics

- Client-Server-Technology: webserver – browsers
- Download via HTTP, FTP, etc.
- Links
- Markup languages
    - Hypertext Markup Language HTML (<body>,<p>,<i>,<h1>): predefined tags for structure and formatting
    - Extensible Markup Language XML: arbitrary tags
    - Drawback: parallel development of similar "languages"
    - → Development of XML-based annotation standards (e.g. the Resource Description Framework)

XML: An Example

```
<?xml version="1.0"?>
<s>
    <np case="nom">
        <det>Der</det>
        <n>Mann</n>
    </np>
    <vp>
        <v person="3" tense="pres">sieht</v>
        <np case="acc">
            <det>die</det>
            <n>Katze</n>
        </np>
    </vp>
</s>
```

Annotation in XML

- Text Encoding Initiative (TEI): development of document-type definitions for language annotation and different text types such as poems, plays, lexica, etc.
- MULTEXT-Project (Ide/Veronis, 1994): morphologic, syntactic, structural (chapter, volume) and bibliographic annotation
- Annotated material can easily be converted into HTML for web-publishing

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW

- Nagao/Hasida (1998): Global Document Annotation
- Ide/Veronis (1994): SGML-based markup languages, 4-level annotation:
  transfer of their ideas towards an automated process
  - Integrate linguistic information into HTML documents
  - Use robots (spiders) to automatically compile a corpus (e.g. for minority languages)
  - Use heuristics for adding additional information (keyword extraction using <em> etc.)
- Berners-Lee (1998): Semantic Web (*"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"*)
- Inclusion of Metadata via e.g. RDF

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Making use of the WWW (2)

- What is RDF?
  - RDF stands for Resource Description Framework
  - RDF is a framework for describing resources on the web
  - RDF is written in XML
    → Information can easily be exchanged between different types of computers, operating systems, and application languages.
- RDF - Examples of Use
  - Describing properties for shopping items, such as price and availability
  - Describing time schedules for web events
  - Describing information about web pages, such as content, author, created and modified date
  - Describing content and rating for web pictures
- RDF and the Semantic Web
  - Web information has exact meaning
  - Web information can be understood and processed by computers
  - Computers can integrate information from the web

Text Corpora

Types and their Content

- Text corpora (written or transcribed spoken text)
- Speech corpora (audio signal with phonetic/linguistic annotation)
- Multimodal corpora (Speech or text plus prosody, gesture, or mouse movements)
- Treebanks → see own section

Text Corpora

- Consist of $>$ 100 million tokens
- Corpora vary in
  - Kind of text (newspaper, scientific, etc.)
  - Domain (sports, linguistics,...)
  - Age
  - Quality
- Can be homogeneous or heterogeneous

Building a Corpus

No automatic procedure: texts consist of content and formatting which yields itself certain information

1. Standardize structure and meta-information
2. Tokenize and detect sentence boundaries
3. Encode (in XML) and store (text or binary format)
4. Tag
5. Chunk or parse

Querying a Corpus

- See extra slides for screenshots of some corpus query tools

Treebanks

Treebanks – Basics

- Special kind of text corpus
- Encode detailed linguistic information
- Basic unit is a sentence, not a token

Purpose and Properties of Treebanks

- Trainings- and test material for syntactic parsers
- Extraction of linguistic information (e.g. valency of verbs)
- Retrieval of specific phenomena (e.g. collocations within a syntactic domain)

- Annotation of linguistic information (depth of syntactic structure, informative labelling of wordforms or categories, anaphora resolution)
- Data structure (tree vs. graph)
- Theory independence
- Correctness and coherence
- Documentation

Examples of Treebanks

- Penn Treebank (1989–2000): Detailed syntactic analysis
- NeGra Project (1999): First German treebank (20.600 sents)
- Tiger Project (1999–): NeGra + 30.000 sents, unordered trees, secondary edges
  `http://www.ims.uni-stuttgart.de/projekte/TIGER`

Wordnets

Wordnets as a Resource

- See session about Wordnets...

Lexica for Multi-modal Systems

Lexica for Multi-modal Systems

- Fundament of every NLP system
- Continuum of complexity:
  acoustic machine control vs. translators' dictionary
- Wordbook vs. hypermedia lexicon

Structure of a Lexicon

- Macrostructure: list, tree, semasiological vs. onomasiological
- Mesostructure: present common properties only once (e.g. verb)
- Microstructure: structure of a single lexical entry
- Metadata: Definitions of the three structures, bibliographic information, etc.

Building a Lexicon

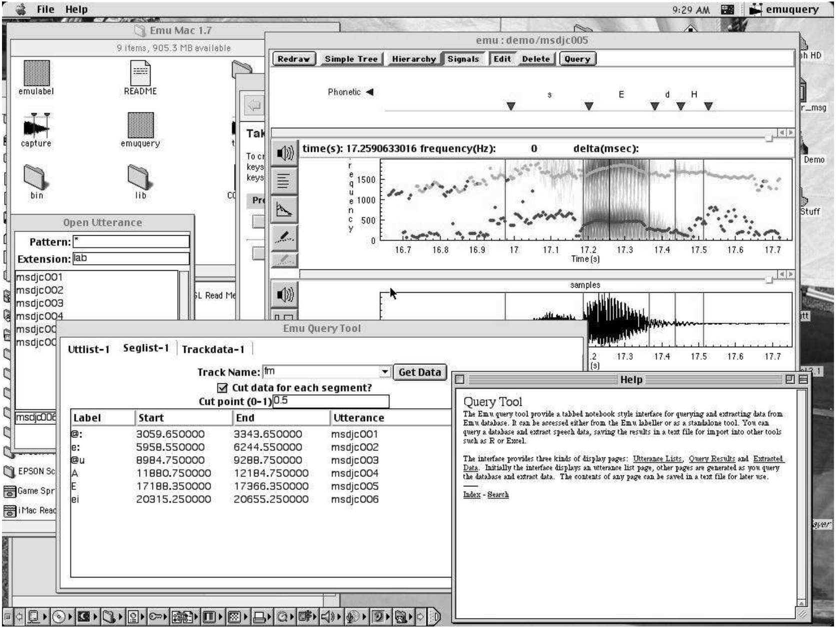
1. Define resource (corpus, transcriptions, etc.)
2. Standardize format
3. Identify words (tokenize)
4. Extract word list (lemmatize, analyze grammar, idioms)
5. Analyze co-occurrence patterns statistically
6. Transform format (print, concordance, hypertext, etc.)

Speech Databases

Speech Databases

- Structured collection of spoken language in digital form
- Contains: speech signals as audio-, sensor-, or video files, annotation and license agreements
- Example: EMU (`http://emu.sourceforge.net/`)

  > *EMU is a collection of software tools for creation, manipulation and analysis of speech databases. At the core of EMU is a database search engine which allows the researcher to find various speech segments based on the sequential and hierarchical structure of the utterances in which they occur. EMU includes an interactive labeller which can display spectrograms and other speech waveforms, and which allows the creation of hierarchical, as well as sequential, labels for a speech utterance.*



time(s): 17.2590633016 frequency(Hz): 0 delta(msec):

Pattern: *

Extension: lab

msdjc001
msdjc002
msdjc003
msdjc004
msdjc00
msdjc00
msdjc006

Track Name: fm

☑ Cut data for each segment?

Cut point (0-1) 0.5

Label	Start	End	Utterance
@:	3059.650000	3343.650000	msdjc001
e:	5958.550000	6244.550000	msdjc002
@u	8984.750000	9288.750000	msdjc003
A	11880.750000	12184.750000	msdjc004
E	17188.350000	17366.350000	msdjc005
ei	20315.250000	20655.250000	msdjc006

Query Tool

The Emu query tool provide a tabbed notebook style interface for querying and extracting data from Emu database. It can be accessed either from the Emu labeller or as a standalone tool. You can query a database and extract speech data, saving the results in a text file for import into other tools such as R or Excel.

The interface provides three kinds of display pages: Utterance Lists, Query Results and Extracted Data. Initially the interface displays an utterance list page, other pages are generated as you query the database and extract data. The contents of any page can be saved in a text file for later use.

Index - Search

The end...

Thank you!