Dienstag: Ressourcen

3 Ressourcen
- Einführung
- Korpora

Ressourcen

Ressourcen

Ressourcen

Korpora

- Annotierte und nicht-annotierte Textmengen
- BNC, ANC, Europarl, FrameNet, Salsa, Negra, Wikipedia, ...
- Kein einheitliches Format!

Tools

- Programme für die linguistische (und andere) Verarbeitung
- POS-Tagger, Parser, Sentence Splitter, Machine Learning Toolkits, ...
- Jedes Programm verhält sich anders!

Dokumentation I

https://wiki.cl.uni-heidelberg.de/foswiki/bin/view/Main/Resources/WebHome

Webseite

- Login mit ICL-Account
- Wikiseite enthält Index und detaillierte Beschreibung der verfügbaren Ressourcen

Dokumentation II

https://wiki.cl.uni-heidelberg.de/foswiki/bin/view/Main/Resources/WebHome

Index

- Oben: Kategorien
- Unten: Alphabetische Liste der Ressourcen nach Kategorien mit
  - Links zur Wikiseite (enthält Link zur Homepage und Dokumentation)
  - Kurzbeschreibung
  - Verzeichnispfad

Dokumentation III

Weiterführende Dokumentation zu einzelnen Tools

- README
- man-pages
- doc-Verzeichnis
- ...

Vertraulichkeitsvereinbarung I

- Non Disclosure Agreement (NDA)
- Für manche Ressourcen ist eine Unterschrift nötig, bevor damit gearbeitet werden kann.
- Damit übernehmen Sie die Verantwortung dafür, dass die Ressourcen nicht über Sie an unberechtigte Dritte gelangen.
- Die betroffenen Ressourcen sind mit einem kleinen Schloss-Symbol gekennzeichnet.
- `https://wiki.cl.uni-heidelberg.de/foswiki/bin/view/Main/Resources/NDA`

Vertraulichkeitsvereinbarung II

Howto

1. NDA bei Gruppe Technik abgeben (`http://www.cl.uni-heidelberg.de/gruppetechnik/`)
2. In die Gruppe **`resuser`** aufgenommen werden

Ressourcen

Kategorien I

Statistics/ML Tools für Statistik und Machine Learning

Processors Software für *einen* Verarbeitungsschritt: Parser, Tagger, ...

Preprocessors Software zur Vorverarbeitung (z.B. HTML-Extraktion)

Platforms Toolkits/Frameworks, die verschiedene Verarbeitungsschritte beinhalten (z.B. NLTK)

Ontologies Ontologien und WordNet

Kategorien II

Lingware	Linguistische Software: Grammatiken, Morphologien, ...
Corpora	Alle Korpora
APIs	*Application Programming Interfaces* – Interfaces, um aus Programmen auf Ressourcen zuzugreifen
Annotation	Annotationswerkzeuge

Unterkategorien für Korpora

- Manche Kategorien haben Unterkategorien

Corpora and Data

- Monolingual Corpora
- Multilingual Corpora
- Speech
- Viewers

Ressourcen

Setup

- Einstellungen für manche Tools
- Oft: Umgebungsvariablen (`$PATH`, `$LIBRARY_PATH`, ...)
- Datei `setup` in jedem Ressourcen-Verzeichnis
- Aktivierung mittels `source`

source

- Eingebaut in die Shell
- Erwartet als Argument eine Datei, in der Shell-Kommandos stehen
- Kommandos werden der Reihe nach ausgeführt *ohne eine Subshell zu starten*
- Beispiel:

```
:~$ source /resources/path/to/resource/setup
```

Troubleshooting

Was tun wenn es nicht klappt?

- *Don't Panic*!
- Häufigster Fehler: **`setup`**-Skript nicht ausgeführt
- Viele mögliche Ursachen
- Dokumentation lesen und nachvollziehen
- Kontakt: `resources@cl.uni-heidelberg.de` (Englisch)

Fehlerbeschreibungen

Was sollte eine sinnvolle Fehlerbeschreibung enthalten?

1. *Alles, was man braucht, um das Problem zu reproduzieren*
   1. Vor allem: die Fehlermeldung!
   2. Der Code, der den Fehler erzeugt
   3. Auf welchem Rechner passiert das ganze?
      Betriebssystem, Username, Verzeichnis, ...
2. Was haben Sie bereits versucht? Was haben Sie als letztes geändert?
3. Benutzen Sie einen *sinnvollen Betreff*
4. `http://www.cl.uni-heidelberg.de/computerpool/technikinfo/`

Ressourcen

Ressourcen

Korpora

- Entweder mit Annotationen versehen oder als reiner Text.
- Meistens in Abschnitte unterteilt (z.B. einzelne Dokumente, Sitzungen, Gespräche, Quellen).
- Verschiedene Annotationsstile sind gängig, je nachdem, was annotiert wird.
- Es hilft im Kopf zu behalten, wie ein Korpus erstellt wurde!

Unannotierte Korpora

- Daten kaum vorverarbeitet.
- Nützliche Information in der Herkunft eines Dokuments:
  - Datum
  - Ort
  - Autor
  - ...

Annotierte Korpora I

- Zwei Annotationsweisen:

Inline Annotation

- Annotation direkt ins Dokument eingefügt:
  `the <noun>dog</noun> barks.`

Stand-Off Annotation

- Annotation getrennt vom Dokument:
- Zeichen- oder Wortpositionen verweisen auf die Stelle im Text
  `<noun start="4" end="7" />`

Inline Annotation I

- Verändert das Dokument

Beispiel

- `the dog barks`
- + POS Tags: `<d>the</d> <n>dog</n> <v>barks</v>`
- + Chunks:
  `<np><d>the</d> <n>dog</n></np>`
  `<vp><v>barks</v></vp>`

Inline Annotation II

Nachteile

- Je mehr Annotation, umso schwieriger, die richtige Stelle zu finden.
- Nicht möglich, z.B. zwei verschiedene Parsebäume darzustellen
- Überschneidungen/Überlappungen sind schwer zu modellieren (`<a>just <a>an</a> example</a>`)

Stand-Off Annotation

- Originaldokument bleibt unverändert.
- Beliebige Annotation möglich.
- Annotatierte Daten schwer lesbar (für menschliche Leser).

Unbedingt beachten:

- Bezieht sich die Annotation auf das letzte Zeichen innerhalb einer Spanne oder auf das erste Zeichen außerhalb der Spanne?
- Bezieht sich die Nummerierung auf Bytes (unicode!), Zeichen oder Wörter?

Stand-Off Annotation

- Originaldokument bleibt unverändert.
- Beliebige Annotation möglich.
- Annotatierte Daten schwer lesbar (für menschliche Leser).

Unbedingt beachten:

- Bezieht sich die Annotation auf das letzte Zeichen innerhalb einer Spanne oder auf das erste Zeichen außerhalb der Spanne?
- Bezieht sich die Nummerierung auf Bytes (unicode!), Zeichen oder Wörter?

Gemischte Annotation

- Es wird kompliziert, wenn beide Annotationsstile vermischt werden.
- Beeinflusst Inline Annotation die Zeichenpositionen?
  - ja: Wenn Inline Annotation hinzugefügt wird müssen Zeichenpositionen neu berechnet werden!
  - nein: Man kann das annotierte Dokument parsen, um die ursprünglichen Positionen zu extrahieren!

- Tipp: Bei gemischter Annotation alle Inline Annotationen in Stand-Off Annotation konvertieren.

Gemischte Annotation

- Es wird kompliziert, wenn beide Annotationsstile vermischt werden.
- Beeinflusst Inline Annotation die Zeichenpositionen?
  - ja: Wenn Inline Annotation hinzugefügt wird müssen Zeichenpositionen neu berechnet werden!
  - nein: Man kann das annotierte Dokument parsen, um die ursprünglichen Positionen zu extrahieren!

- Tipp: Bei gemischter Annotation alle Inline Annotationen in Stand-Off Annotation konvertieren.

Ressourcen

Erstellung von Korpora

- (Die meisten) Texte werden nicht zum Zwecke der Korpuserstellung geschrieben/gesprochen.
- Ausnahme: Korpora für gesprochene Sprache häufig von bezahlten Sprechern in kontrollierten Szenarien erstellt (=€€€!)
- Textkorpora oft aus verschiedenen Quellen zusammengestellt.

Quellen

- Bücher, Artikel, sonstige Veröffentlichungen
- Webseiten
- Tonaufnahmen
- Zeitungs- und Nachrichtentexte

Konvertierung von Formaten

- Dateien in einem Ordner genügen nicht.
- Das Korpus muß in einem nutzbaren Format sein.
- Originalformat muss häufig konvertiert werden.

Konvertierung	Tool
PDF $\rightarrow$ TXT	**`pdftotext`** aber: Bindestriche, Text in Spalten
HTML $\rightarrow$ TXT	**`python`** , **`perl`** , **`bash`** , ...

Tabelle: Konvertierungswerkzeuge

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Aufbereitung

- Mehrere Schritte, je nach Bedarf
- 1: Sentence splitting
- 2: Tokenisierung
- 3: POS-Annotation, Lemmatisierung
- 4: Annotation von Eigennamen
- 5: Annotation der Satzstruktur
- n: ...
- = Pipeline

Pipeline-Werkzeuge

- Bestimmte Frameworks können helfen:
  - Datenstrukturen
  - Interfaces
  - Parallelisierung

Beispiel

- UIMA: `http://incubator.apache.org/uima/`
- OpenNLP: `http://opennlp.sourceforge.net/`
- GATE: `http://gate.ac.uk/`
- Heart of Gold: `http://heartofgold.dfki.de/`

Pipeline-Werkzeuge

- Bestimmte Frameworks können helfen:
  - Datenstrukturen
  - Interfaces
  - Parallelisierung

Beispiel

- UIMA: `http://incubator.apache.org/uima/`
- OpenNLP: `http://opennlp.sourceforge.net/`
- GATE: `http://gate.ac.uk/`
- Heart of Gold: `http://heartofgold.dfki.de/`

Ressourcen

Unannotierte Korpora

- Projekt Gutenberg: Sammlung von Texten mit abgelaufenem Urheberrecht.
- WAC: gecrawlte Webseiten von verschiedenen Domains (.de, .uk, .it, ...)
  Teilweise aufbereitete Korpora vorhanden (PukWAC)
- Web1t: 5-gramme aus dem Netz, erstellt von Google (1 TB!)
- Wortschatz: 3 Millionen deutsche Sätze

Unannotierte Korpora

- Projekt Gutenberg: Sammlung von Texten mit abgelaufenem Urheberrecht.
- WAC: gecrawlte Webseiten von verschiedenen Domains (.de, .uk, .it, ...)
  Teilweise aufbereitete Korpora vorhanden (PukWAC)
- Web1t: 5-gramme aus dem Netz, erstellt von Google (1 TB!)
- Wortschatz: 3 Millionen deutsche Sätze

Unannotierte Korpora

- Projekt Gutenberg: Sammlung von Texten mit abgelaufenem Urheberrecht.
- WAC: gecrawlte Webseiten von verschiedenen Domains (.de, .uk, .it, ...)
  Teilweise aufbereitete Korpora vorhanden (PukWAC)
- Web1t: 5-gramme aus dem Netz, erstellt von Google (1 TB!)
- Wortschatz: 3 Millionen deutsche Sätze

Unannotierte Korpora

- Projekt Gutenberg: Sammlung von Texten mit abgelaufenem Urheberrecht.
- WAC: gecrawlte Webseiten von verschiedenen Domains (.de, .uk, .it, ...)
  Teilweise aufbereitete Korpora vorhanden (PukWAC)
- Web1t: 5-gramme aus dem Netz, erstellt von Google (1 TB!)
- Wortschatz: 3 Millionen deutsche Sätze

Annotatierte Korpora

- BNC: Standardkorpus, „*balanced*", manuell annotiert mit POS und Lemma.
- Penn TreeBank: Nachrichtentexte mit manueller syntaktischer Annotation
- Negra: Deutsche Zeitungstexte, manuell annotiert mit POS (keine Lemmata) und syntaktischer Struktur.
- FrameNet: Englische Zeitungstexte, annotiert mit semantischen Rollen nach dem FrameNet Paradigma.
- TimeBank: Nachrichtentext mit Events, Temporalausdrücken und temporalen Relationen zwischen Events.

Annotatierte Korpora

- BNC: Standardkorpus, „*balanced*", manuell annotiert mit POS und Lemma.
- Penn TreeBank: Nachrichtentexte mit manueller syntaktischer Annotation
- Negra: Deutsche Zeitungstexte, manuell annotiert mit POS (keine Lemmata) und syntaktischer Struktur.
- FrameNet: Englische Zeitungstexte, annotiert mit semantischen Rollen nach dem FrameNet Paradigma.
- TimeBank: Nachrichtentext mit Events, Temporalausdrücken und temporalen Relationen zwischen Events.

Annotatierte Korpora

- BNC: Standardkorpus, „*balanced*", manuell annotiert mit POS und Lemma.
- Penn TreeBank: Nachrichtentexte mit manueller syntaktischer Annotation
- Negra: Deutsche Zeitungstexte, manuell annotiert mit POS (keine Lemmata) und syntaktischer Struktur.
- FrameNet: Englische Zeitungstexte, annotiert mit semantischen Rollen nach dem FrameNet Paradigma.
- TimeBank: Nachrichtentext mit Events, Temporalausdrücken und temporalen Relationen zwischen Events.

Annotatierte Korpora

- BNC: Standardkorpus, „*balanced*", manuell annotiert mit POS und Lemma.
- Penn TreeBank: Nachrichtentexte mit manueller syntaktischer Annotation
- Negra: Deutsche Zeitungstexte, manuell annotiert mit POS (keine Lemmata) und syntaktischer Struktur.
- FrameNet: Englische Zeitungstexte, annotiert mit semantischen Rollen nach dem FrameNet Paradigma.
- TimeBank: Nachrichtentext mit Events, Temporalausdrücken und temporalen Relationen zwischen Events.

Annotatierte Korpora

- BNC: Standardkorpus, „*balanced*", manuell annotiert mit POS und Lemma.
- Penn TreeBank: Nachrichtentexte mit manueller syntaktischer Annotation
- Negra: Deutsche Zeitungstexte, manuell annotiert mit POS (keine Lemmata) und syntaktischer Struktur.
- FrameNet: Englische Zeitungstexte, annotiert mit semantischen Rollen nach dem FrameNet Paradigma.
- TimeBank: Nachrichtentext mit Events, Temporalausdrücken und temporalen Relationen zwischen Events.

Multilinguale Korpora

- Einige Korpora beinhalten Text in mehreren Sprachen.
- Multilinguale Quellen:
  - Wikipedia
  - Internationale Organisationen (EU, UN)
  - Übersetzte Texte (z.B. Nachrichtenagenturen, Projekt Gutenberg, mehrsprachige Webseiten)
- Parallel: Derselbe Text in verschiedenen Sprachen.
  - Europarl, OpenSubtitles, etc.
- Vergleichbar: Texte zum selben Thema in verschiedenen Sprachen
  - Reuters, WikiXML, ISI

Multilinguale Korpora

- Einige Korpora beinhalten Text in mehreren Sprachen.
- Multilinguale Quellen:
  - Wikipedia
  - Internationale Organisationen (EU, UN)
  - Übersetzte Texte (z.B. Nachrichtenagenturen, Projekt Gutenberg, mehrsprachige Webseiten)
- Parallel: Derselbe Text in verschiedenen Sprachen.
  - Europarl, OpenSubtitles, etc.
- Vergleichbar: Texte zum selben Thema in verschiedenen Sprachen
  - Reuters, WikiXML, ISI

Alignierungen

- Eine Alignierung (*alignment*) verbindet korrespondierende Passagen in verschiedenen Dokumenten.
- Parallele Korpora können auf verschiedenen Ebenen aligniert werden
- Dokument-, Absatz-, Satz-, Phrasen- und Wortalignierung
- Alignierung normalerweise implizit dargestellt durch IDs von Textpassagen.

Übung 4