MarMot und Morfessor

Christoph Teichmann

Saarland University

Neues Wort Ohne Kontext

Morbiditätsrisikoausgleichszahlungsgesetze

- Was ist Part-Of-Speech?
- Was ist Themenbereich?
- Was ist Bedeutung?
- ...

Bedeutung für die CL

Unbekannte Worte Kernproblem in CL – mehr Morphologie $\rightarrow$ mehr unbekannte Worte

Sentiment: "Der Film war w_{932}"

Bedeutung für die CL

Unbekannte Worte Kernproblem in CL – mehr Morphologie $\rightarrow$ mehr unbekannte Worte

Sentiment: "Der Film war Attahöhlenschlecht"

- Regelbasierte Systeme können Regeln basierend auf Morphologie haben
- Machine Learning wird einfacher wenn es wenige, oft wiederholte Muster gibt

Morphologie Aufgaben

- Segmentierung:
  Attahöhlenschlecht / Atta-höhle-n-schlecht (heute)

Morphologie Aufgaben

- Segmentierung:
  Attahöhlenschlecht / Atta-höhle-n-schlecht (heute)
- Tagging: trug / VT|3pers|sing|prät (heute)

Morphologie Aufgaben

- Segmentierung:
  Attahöhlenschlecht / Atta-höhle-n-schlecht (heute)
- Tagging: trug / VT|3pers|sing|prät (heute)
- Lemmatisierung: trug / tragen (in NLTK)

Morphologie Aufgaben

- Segmentierung:
  Attahöhlenschlecht / Atta-höhle-n-schlecht (heute)
- Tagging: trug / VT|3pers|sing|prät (heute)
- Lemmatisierung: trug / tragen (in NLTK)
- Flektion: tragen+VT|3pers|sing|prät / trug

Morphologie Aufgaben

- Segmentierung:
  Attahöhlenschlecht / Atta-höhle-n-schlecht (heute)
- Tagging: trug / VT|3pers|sing|prät (heute)
- Lemmatisierung: trug / tragen (in NLTK)
- Flektion: tragen+VT|3pers|sing|prät / trug
- ...

MarMot

- Morphologischer Tagger von LMU CL

MarMot

- Morphologischer Tagger von LMU CL
- Kommandozeilenbasiert

MarMot

- Morphologischer Tagger von LMU CL
- Kommandozeilenbasiert
- Wie Part-Of-Speech Tagger aber optimiert für viele Tags/strukturierte Tags

MarMot

- Morphologischer Tagger von LMU CL
- Kommandozeilenbasiert
- Wie Part-Of-Speech Tagger aber optimiert für viele Tags/strukturierte Tags
- Machine Learning basiert

MarMot

- Morphologischer Tagger von LMU CL
- Kommandozeilenbasiert
- Wie Part-Of-Speech Tagger aber optimiert für viele Tags/strukturierte Tags
- Machine Learning basiert
- Vortrainierte Modelle / einfach neue Modelle zu trainieren

MarMot

Ausprobieren:

MarMot download `http://cistern.cis.lmu.de/marmot/bin/CURRENT/` and model for German `http://cistern.cis.lmu.de/marmot/models/CURRENT/`

MarMot

Ausprobieren:

Mit existierendem Modell taggen

java -cp marmot-2015-10-22.jar marmot.morph.cmd.Annotator -model-file <wo-model> -test-file form-index=<index-wort>,<Daten> -pred-file <Resultat>

MarMot

Ausprobieren:

Neues Modell trainieren

java -Xmx2G -cp marmot-2015-10-22.jar marmot.morph.cmd.Trainer -train-file form-index=<index-wort>,tag-index=<index-pos-tag>,morph-index=<index-morph-tag>,<Daten> -tag-morph true -model-file <speicherort>

Morfessor

- Tool für morphologische Segmentierung

Morfessor

- Tool für morphologische Segmentierung
- Unüberwacht $\rightarrow$ trainiert aus Daten ohne Annotation

Morfessor

- Tool für morphologische Segmentierung
- Unüberwacht $\rightarrow$ trainiert aus Daten ohne Annotation
- Nicht notwendigerweise linguistisch korrekt

Morfessor

- Tool für morphologische Segmentierung
- Unüberwacht $\rightarrow$ trainiert aus Daten ohne Annotation
- Nicht notwendigerweise linguistisch korrekt
- Versucht weiderkehrende Elemente zu finden

Morfessor

- Tool für morphologische Segmentierung
- Unüberwacht $\rightarrow$ trainiert aus Daten ohne Annotation
- Nicht notwendigerweise linguistisch korrekt
- Versucht weiderkehrende Elemente zu finden
- Minimum Description Length basiert

Morfessor

- Tool für morphologische Segmentierung
- Unüberwacht $\rightarrow$ trainiert aus Daten ohne Annotation
- Nicht notwendigerweise linguistisch korrekt
- Versucht weiderkehrende Elemente zu finden
- Minimum Description Length basiert
- Nützlich für downstream Machine Learning

Morfessor

Ausprobieren:

Clonen mit Git von:

`https://github.com/aalto-speech/morfessor`

Morfessor

Ausprobieren:

Neues Modell trainieren

morfessor-train - -traindata-list -s <speicherort> <data>

Morfessor

Ausprobieren:

Mit existierendem Modell segmentieren

morfessor-segment -l <model> <data>

Übungen

- Annotieren sie einen kurzen Text mit MarMot
- Schreiben sie ein kleines Trainingsfile für MarMot und trainieren sie ein neues Modell
- Schreiben sie ein kleines Trainingsfile für Mofesssor (viele Morphemwiederholungen) und trainieren sie ein Modell
- Segmentieren sie ihr File mit Morfessor
- Versuchen sie ihr Trainingsfile mit hilfreichen Beispielen zu erweitern