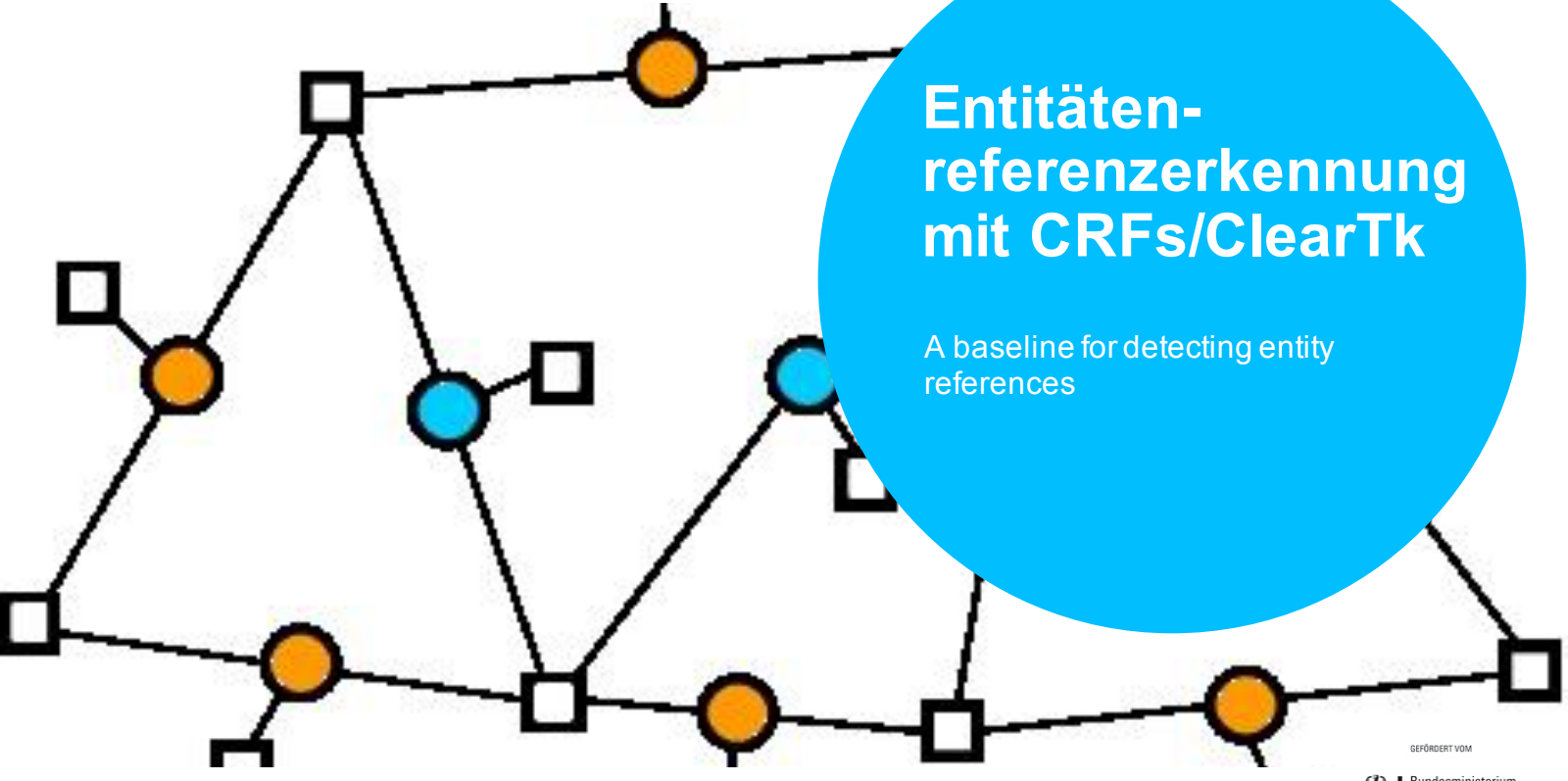
Universität Stuttgart
Center for Reflected Text Analytics

CRETA
CENTER FOR REFLECTED TEXT ANALYTICS



Entitäten-referenzerkennung mit CRFs/ClearTk

A baseline for detecting entity references

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Outline

- Conditional Random Fields
- Features
- Implementation & Realisation
- Ergebnisse

Conditional Random Fields

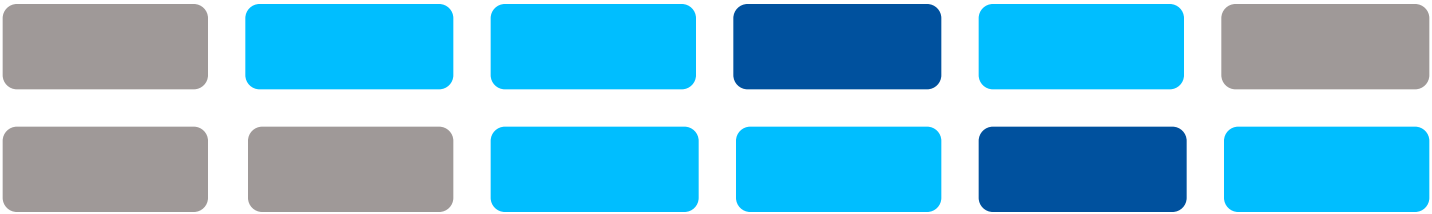
Ein graphisches Model zum Sequence Labeling

- Sequence Labeling
  - Sequenz: Wörter im Text
  - Annotation von Wörtern mit Kategorien im Kontext
- Conditional Random Fields
  - Entscheidungen für vorherige Wörter können revidiert werden
  - Forschungsstand für Named Entity Recognition (NER)
    - z.B. in Stanford NER
- Überwachtes Maschinelles Lernverfahren

Features

Kontexte

- Extraktion von Features für jedes Wort
- Vier Wörter steuern Features bei:
  - Das aktuelle Wort
  - Beide vorherige Wörter
  - Ein folgendes Wort



- BIO-Klassifikation: Jedes Wort ist entweder **B**egin einer Entity, **I**n einer Entity oder **O**ußerhalb einer Entity
  - Damit können Mehr-Wort-Entitäten erkannt werden

Features

Merkmale für das Maschinelle Lernverfahren

Beispieltext: „Geschichte des armen Werther“

Feature	Beschreibung	Beispiel
Oberfläche	Das Wort selbst	„armen“
Wortart	Die automatisch zugewiesene Wortart (mit einem neuhochdeutschen PoS-Tagger)	„ADJA“
Zeichenmuster	Muster an Unicode-Properties	„ll“

Features

Zeichenmuster

- Unicode definiert Eigenschaft für Zeichen
- Alle in einem Wort vorkommenden Eigenschaften werden zu einer neuen Zeichenfolge zusammengefasst

Eigenschaft	Beschreibung	Beispiele
Letter, lowercase	Kleinbuchstaben	a ц ώ 𝕞
Number, digit	Ziffern	0-9
Punctuation, start	Öffnende Zeichen	([{ <
Punctuation, other	Andere Zeichen	. ! ? : ;

Vollständige Liste: http://www.fileformat.info/info/unicode/category/index.htm

- Aus „1860München“ wird also „NdLuLl“

Implementierung

Frameworks to the rescue!

- Apache UIMA http://uima.apache.org
  - Framework für NLP-Pipelines
    - (s.a. Hellrich et al., 2017; Poster am Donnerstag)
- ClearTk https://cleartk.github.io/cleartk
  - UIMA + Machine Learning
  - Fokus auf Feature-Extraktion für Training, Test und Anwendung
    - Feature-Extraktoren für häufig verwendete Features
  - Integriert mit Weka, mallet, liblinear, libsvm, opennlp-maxent
- Mallet http://mallet.cs.umass.edu
  - Java machine learning Bibliothek für NLP
  - Implementierung für CRF, Topic Modeling, maximum entropy, ...

Ergebnisse

Zusammenfassung

- Präzision
  - vergleichsweise robust
- Recall
  - Lange spans sind schwierig, aufgrund des begrenzten Kontexts
  - Manchmal verhindert Kontext, Eigennamen zu erkennen
    - Die eigentlich exakt so in den Trainingsdaten sind

- Weiterentwicklungen
  - Namensliste
  - Existenz des Wortes mit anderer Schreibung (groß/klein)
  - „Active Learning“

Universität Stuttgart

Vielen Dank!