KobRA @ Classroom:

Ein interdisziplinäres Projektseminar zum Einsatz von Data Mining bei der korpusgestützten Analyse internetbasierter Kommunikation

Michael Beißwenger · Christian Pölitz · Lena Meyer

technische universität dortmund

Sommersemester 2014 **Korpusgestützte Analyse internetbasierter Kommunikation mit Hilfe von Data-Mining**			**Michael Beißwenger** (Germanistik) **Christian Pölitz** (Informatik)
Hauptseminar	2 SWS	Di 10-12 Uhr, R. U331	Beginn: 8.4.2014
LABG 2009: BLS1 / BaMaLa 2005: F6 / B.a./M.a. Angewandte Sprachwissenschaft: M2, M3, M7			

Kurzbeschreibung:

Ziel des Seminars ist es, anhand ausgewählter sprachwissenschaftlicher Fragestellungen den Einsatz innovativer Informatikmethoden („Data-Mining", maschinelles Lernen) für die empirische korpusgestützte Analyse internetbasierter Kommunikation zu erproben.

Um Besonderheiten der Sprachverwendung in der internetbasierten Kommunikation auf der Basis großer Datensammlungen (Korpora) quantitativ und qualitativ untersuchen zu können, bedarf es automatischer Verfahren, die die Daten vorsortieren, klassifizieren und nach potenziell interessanten Belegen durchforsten. Um solche Verfahren entwickeln zu können, bedarf es sprachwissenschaftlichen Know-Hows, das in Form sogenannter „Annotationen" in kleine Datensets eingebracht wird.

Im Seminar werden wir in kleinen Analyseprojekten solche Annotations- und Entwicklungsprozesse durchspielen. Dabei arbeiten Studierende der Germanistik mit Studierenden der Informatik zusammen, wobei die Germanistik-Studierenden Daten auf der Grundlage sprachwissenschaftlicher Konzepte analysieren und annotieren und die Informatik-Studierenden Ansätze dafür entwickeln, die linguistische Analyse durch automatische Verfahren zu unterstützen. Dabei wird deutlich, in welcher Weise die empirische, datengestützte Forschung in den Geisteswissenschaften von einer Zusammenarbeit mit der Informatik profitieren kann. Zugleich bietet das Seminar eine hervorragende Möglichkeit, Erfahrungen mit interdisziplinärer Zusammenarbeit zu einem aktuellen, hoch aktiven Forschungsbereich zu sammeln.

Studierende der Lehrämter Deutsch und der angewandten Studiengänge sind gleichermaßen angesprochen.

Teilnahmemodalitäten/Voraussetzungen:

Das Seminar ist auf 40 Plätze begrenzt. 20 Plätze werden an Studierende der Germanistik, 20 an Studierende der Informatik vergeben.

Seminarprojekt: Automatische Eliminierung von Pseudotreffern und Finden von „Nadeln im Heuhaufen" für große Trefferlisten zu ausgewählten sprachlichen Phänomenen internetbasierter Kommunikation:

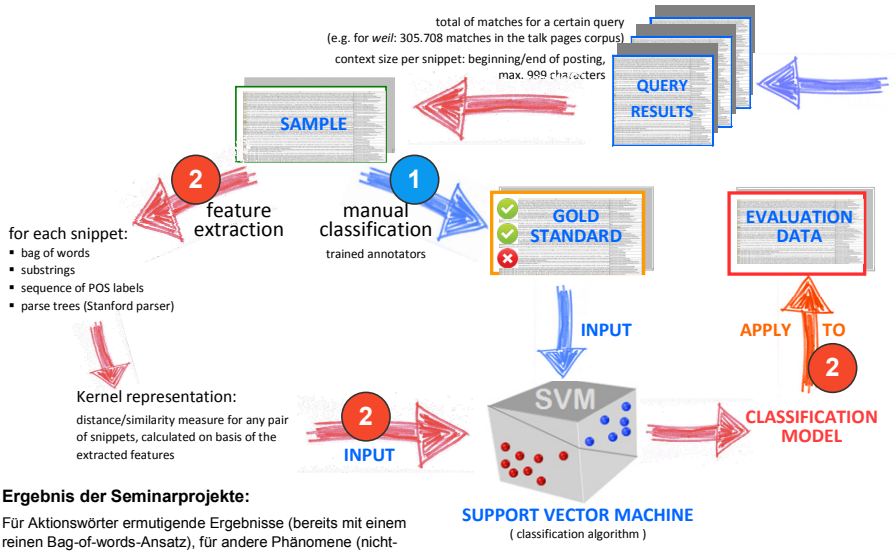
- Aktionswörter: *freu*, *lach*, *schmunzel*, *ganzfiesgrins*, …
- nicht-kanonische Verwendungen von *weil* (V2 anstelle von V1): *ja toll aber so richtig steht es nicht drin weil damals sollten wir nämlich eine arbeit in informatik machen über das dualsystem*
- Sprechsprachliche Verschmelzungen: *vorm* (< *vor dem/einem*), *inne* (< *in die/eine*), …

1 Workflow aus Sicht der Germanistik-Studierenden:

- Die Studierenden bearbeiten in Teams Listen mit Korpustreffern, die als Ergebnis einer Volltextrecherche im Wikipedia-Diskussionsseiten-Korpus generiert wurden (1 Snippet = 1 Posting).
- Sie klassifizieren jedes einzelne Vorkommen des Suchausdrucks unter linguistischem Gesichtspunkt und notieren ihr Analyseergebnis zu den einzelnen Vorkommen – zum Beispiel:
  - „Der Treffer in diesem Snippet ist ein / kein Aktionswort."
  - „Der Treffer für *weil* in diesem Snippet leitet einen / keinen Satz mit Verbletzt-Stellung ein."
- Die Daten werden in einem Tabellenformat bereitgestellt. Für die Annotation der Analyseergebnisse zu den Datenbeispielen sind **keinerlei Programmier- oder Informatikkenntnisse erforderlich**: Das Ergebnis der intellektuellen Klassifikation wird direkt in die Tabelle eingetragen und später für die Entwicklung der Lernverfahren von den Informatik-Studierenden in das benötigte Input-Format konvertiert.
- Nach erfolgter Analyse werden die annotierten Datenbeispiele an die Informatik-Studierenden übergeben. Dann beginnt der Informatik-Teil der Arbeit, in dessen Rahmen Lernverfahren entwickelt werden, die aus den "händisch" klassifizierten Beispielen Regeln ableiten, mit denen weitere Datenbeispiele gleichen Typs automatisch klassifiziert werden können.

2 Workflow aus Sicht der Informatik-Studierenden:

- Die Germanistik-Studierenden liefern Tabellen, in denen eine Auswahl an Belegen für die einzelnen Korpusabfragen manuell klassifiziert wurden (= Gold-Standard).
- Transformation der Tabellen in ein internes Format und Extraktion wichtiger Merkmale und Muster.
- Lernen einer Entscheidungsfunktion, die auf Basis der Merkmale und Muster vorhersagt, ob die relevante Merkmalhaftigkeit (z.B. "enhält ein Aktionswort" oder "enthält weil mit Verbzweitstellung") bei einem Snippet vorhanden ist oder nicht.
- Die von den Linguistik-Studierenden gelieferten Daten dienen als Grundlage für das Lernen und Testen der Entscheidungsfunktion.
- Am Ende des Prozesses steht eine Datei, die unbekannten Snippets das Merkmal +1 oder -1 zuordnet („+1" = Merkmal vorhanden, „-1" = Merkmal nicht vorhanden).



Ergebnis der Seminarprojekte:

Für Aktionswörter ermutigende Ergebnisse (bereits mit einem reinen Bag-of-words-Ansatz), für andere Phänomene (nicht-kanonische Verwendungen von ‚weil') wenig befriedigende Ergebnisse

⇨ Die sprachtechnologischen Ressourcen (z.B. PoS-Tagger), die bei der Feature-Extraktion genutzt werden, sind nicht auf Phänomene nicht-standardisierter Schriftlichkeit in Genres internetbasierter Kommunikation angepasst. Eine Verbesserung der Lernverfahren setzt eine Anpassung der genutzten Sprachverarbeitungswerkzeuge voraus.

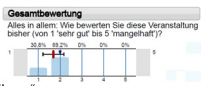
⇨ GSCL-AK „Social Media / Internetbasierte Kommunikation": http://gscl.org/ak-ibk.html

⇨ GSCL-Shared-Task zur automatischen linguistischen Annotation internetbasieter Kommunikation: https://sites.google.com/site/empirist2015/

Seminarevaluation (Germanistik-Studierende):

Was fanden Sie besonders gut an der Veranstaltung?

- „Interessante Diskussionen"
- „Interdisziplinarität !"
- „Interdisziplinäre Ausrichtung"
- „Die Kooperation mit den Informatikern"
- „Eine Seminarform, die es öfter geben sollte. Erweitert den Blickwinkel."



Was fanden Sie eher schlecht?

- „hatte gehofft, noch tiefer in die Informatikmaterie einzutauchen"

Das Korpus

Wikipedia-Korpus 2013 in DeReKo
(Kupietz & Lüngen 2014)
http://www.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html

	Articles	Talk pages
I5 file size	~16G	~4.8G
# pages	~1,6M	~550K
# postings	--	~5,5M
# tokens	~678M	~264M

Das Korpus ist repäsentiert in **I5** (Sperberg-McQueen & Lüngen 2012), der Customization des **P5**-Encoding-Formats der **Text Encoding Initiative** (TEI, http://tei-c.org) für DeReKo. I5 integriert Modelle für die Strukturbeschreibung von Genres internetbasierter Kommunikation aus dem adaptierten TEI-Schema von Beißwenger et al. (2012).

```
<div n="2" type="thread">
  <head type="cross">
    <s>Totensonntag in der DDR</s>
  </head>
  <posting indentLevel="0" who="WU00000000">
    <p>
      <s>Hallo, weiß jemand ob es auch einen
      Totensonntag in der DDR Gab?? Danke
      </s>
    </p>
  </posting>
  <posting indentLevel="1" synch="t00121163"
  who="WU00006525">
    <p>
      <s broken="yes">Warum sollte es den dort
      nicht gegeben haben?</s>
      <s>Auch in der DDR hörte das Kirchenjahr
      mit dem Ewigkeitssonntag/Totensonntag auf
      und das neue fing mit dem 1. Advent wieder
      an.</s>
      <s>--<autoSignature/> 23:23, 5. Dez. 2006
      (CET) </s>
    </p>
  </posting>
  [...]
```

Die Grundlage für das Seminar bildet das Diskussionsseiten-Teilkorpus. Die Konvertierung nach I5 ist beschrieben in Margaretha & Lüngen (2014).

http://www.kobra.tu-dortmund.de

michael.beisswenger@tu-dortmund.de christian.poelitz@tu-dortmund.de lena2.meyer@tu-dortmund.de