berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Extraktion guter Belege aus Textkorpora durch Kombination eines regelbasierten Verfahrens mit maschinellem Lernen

Lothar Lemnitzer, Alexander Geyken

Berlin-Brandenburgische Akademie der Wissenschaften

Neue Wege in der Nutzung von Korpora, Berlin, 30.10. 2015

Gliederung

- Motivation: Krise der Lexikographie
- Automatisierung der Extraktion gute Belege
  - Europäischer Kontext
  - Kontext DWDS
- Beschreibung „Gute-Belege-Extraktor“ im DWDS
- Erweiterung mit ML-Techniken
- Evaluierung
- Fazit

Kontext: Krise der Lexikographie

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- Verlage reduzieren die Zahl ihrer Mitarbeiter (dramatisch) oder schließen ganz
  - The digital revolution is changing the way readers consume news and search form information. People are moving away from printed reference books and going online, where, generally, they expect to get their information for free“ (press declaration Chambers Harrap, 2009)
  - OED3, Macmillan: only online publications
- Wissenschaftliche Lexikographie: zeitliche, planerische und finanzielle Probleme

Kontext: Krise der Lexikographie

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Prozess der Wörterbucherstellung muss neu gedacht und konzipiert werden, insbesondere

- Korpusaufbau
- Automatische Extraktion lexikogr. Informationen
- Repräsentation lexikogr. Daten
- Redaktionssysteme
- Integration kollaborativer Arbeitsschritte
- Planung und Projektmanagement

Automatische Extraktion guter Belege aus Korpora

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- Startpunkt: GdEX (Kilgarriff 2008)
- GdEx: Good Dictionary EXample
  - „gut“: klar, verständlich ohne weiteren Kontext
- Beispiel Schnellpresse (Wort des Tages)
  - **„Schnellpressen** der HHrn.“ [Dingler 1826/21, S. 473-480]
  - „Die **Schnellpresse** und die Rotationspresse wurden erfunden; das Telefon.“ [opalkatze.wordpress.com, 30.9.2012]
  + „Technologische Durchbrüche wie die Erfindung der Stereotypie und die Einführung der **Schnellpresse** ermöglichen die massenhafte Herstellung von Büchern und Zeitschriften.“ [Spiegel, 13.2.2006, Nr. 7]

Automatische Extraktion guter Belege aus Korpora

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- GdEx keine Forschungsaufgabe sondern eine praktische / projektspezifische Aufgabe (Jakubicek 2015)
- GdEx soll die Extraktion von Wörterbuchbelegen eines Projekts beschleunigen

- Quelle: ENEL-Cost action – 12.2.15 (Wien)
  http://www.elexicography.eu/working-groups/working-group-3/wg3-workshops/automatic-extraction-of-good-dictionary-examples/

Kontext DWDS

Ziel: Schaffung eines großen lexikalischen Informationssystems (aus vielen Quellen), um die deutsche Sprache gegenwartsbezogen mit hist. Tiefe darzustellen

Phase 1 (2007-2012): Vorbereitungsphase

- Aufbau der Arbeits- und Rechercheplattform

Phase 2+3 (2013-2024):

- Neueinträge und Revision bestehender Einträge

Website: www.dwds.de (http://zwei.dwds.de)

Kontext DWDS

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Quellen

- Wörterbücher: WDG, Grimm, Etym-Wb, GWDS-99 (Duden): ~450.000 entries
- Korpora:
  - Referenzkorpora: DWDS-Kern, DTA: 200 M Token
  - Zeitungskorpora: 4,5 Mrd. Token
  - Blogs: cc-Anteil: 100 M (cc); Gesamt: 2 Mrd.

    Davon öffentlich zugänglich: DWDS-Kern, DTA sowie Blogs-cc, ZEIT, Tagesspiegel, Berliner Zeitung (ca. 1,5 Mrd. Token)

Kontext DWDS

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

GdEx – Ziel im DWDS

1. „Gute“ Verwendungsbeispiele + Überarbeitung durch Lexikographen für alle Neueinträge (20.000) und zu überarbeitenden Bestandseinträge (geplant 30.000)
2. „Gute“ Verwendungsbeispiele ohne lexikographische Nachbearbeitung für Einträge, die aus zeitlichen Gründe nicht überarbeitet werden können oder keine voll lexikographische Beschreibung erhalten (‚Basiseinträge‘)

E-Mail; eMail (war und ist ungültig) – Substantiv
Femininum, -, -s
süddeutsch, österreichisch, schweizerisch Neutrum, -s, -s
Worttrennung: E-Mail (computergeneriert)

Bedeutungen | Thesaurus | Typische Verbindungen

Bedeutungen

DWDS (Vollartikel), 2015

1. **(briefähnliche) elektronische Nachricht**

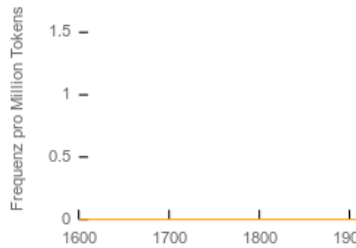
KOLLOKATIONEN:

- *mit Adjektivattribut:* eingehende, eintreffende, verschickte, weitergeleitete, gefälschte, unerwünschte, signierte **E-Mails**
- *als Akkusativobjekt:* eine **E-Mail** schicken, verschicken, senden, empfangen, abrufen, verschlüsseln, abfangen, mitlesen
- *in Präpositionalgruppe/-objekt:* etw. per **E-Mail** mitteilen, per **E-Mail** jmdn. benachrichtigen, kommunizieren, jmdn. mit **E-Mails** bombardieren
- *in Koordination:* **E-Mails** und Briefe
- *als Aktivsubjekt:* eine **E-Mail** kursiert, trudelt ein
- *in Präpositionalgruppe/-objekt:* eine **E-Mail** mit Betreff, Betreffzeile, Anhang, an einen Adressaten, im Postfach
- *als Genitivattribut:* der Anhang, Dateianhang, die Betreffzeile, der Absender, Versender, der Empfänger einer **E-Mail**
- *in vergleichender Wort-/Nominalgruppe:* etw. als **E-Mail** versenden, verschicken

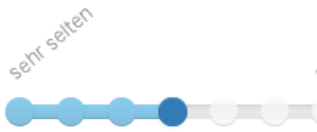
BEISPIELE:

- [...] »QWERTYIOP« lautete der Überlieferung nach der Inhalt der ersten **E-Mail**, die 1971 den Rechner des Amerikaners Ray Tomlinson verließ. [Die Zeit, 02.11.2009 (online)]
- In der päpstlichen Mailbox des Vatikans laufen so viele **E-Mails** Jugendlicher auf wie nie zuvor. [Bild, 18.08.2005]
- Die rund [...] 170.000 **E-Mails** waren im vergangenen Jahr bei einem Hackerangriff unbekannter Täter auf das Computernetz [...] erbeutet worden [...]. [Neue Zürcher Zeitung, 17.04.2015]
- Im März 2008 hatte Breininger seiner Schwester per **E-Mail** mitgeteilt, er sei in den pakistanischen Bergen und wolle weiter nach Afghanistan. [Spiegel, 03.12.2009 (online)]
- Schon jetzt kostet eine **E-Mail** erheblich weniger als ein Brief. [C't, 1998, Nr. 18]

2. **Computer- oder Webanwendung, die den elektronischen Austausch von (briefähnlichen) Nachrichten über ein Computernetzwerk ermöglicht**



Worthäufigkeit



Tokens: 1088377367, Hits: 2141

Ältere Wörterbücher

- Grimmsches Wörterbuch (¹DWB
- Wörterbuch der deutschen Gegenwartssprache (WDG) (0)

Korpustreffer

Referenzkorpora
- Kernkorpus Basis (3338)
- DWDS-Kernkorpus (86)
- DWDS-Kernkorpus 21 (3740)
- DTA+DWDS (86)
- Deutsches Textarchiv (0)

Zeitungskorpora
- Alle Zeitungen (86852)
- BamS (1916)
- BILD (3324)
- Berliner Zeitung (3332)
- FAZ (7344)

Gute-Belege-Extraktor im DWDS

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Mengenproblem / Arbeitsaufwand

- Häufig: mehrere Hundert oder Tausend Belege pro Stichwort im Korpus
- Es sollen (pro Lesart) 3-5 typische und gute Belege ausgewählt werden
- Bei einer Menge von > 20 000 Einträgen muss die Zahl der Belege drastisch reduziert werden
- Den Bearbeitern sollen dabei nur die besten Belege präsentiert werden

Gute-Belege-Extraktor im DWDS

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Ziel: Sortierung der Belege für eine Stichwort nach Gütekriterien (Orientierung an Kilgarriff‘s (2008) Ansatz) - Regelbasiert

Gütekriterien sind verschiedene linguistische Kriterien (Satzlänge, Anzahl Pronomen im Satz…)

Externe Kriterien (Zeit des Belegs, Quelle) spielen ebenfalls eine Rolle (Ausgewogenheit)

Erfahrung: es werden immer noch zu viele Belege ausgewählt, die die Lexikographen verwerfen

Gute-Belege-Extraktor im DWDS

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Fazit

Das Konzept „guter Beleg“ ist zu vage, um ihn so weite zu operationalisieren, dass die Ergebnisse optimal die Bedürfnisse der Lexikographen (und der Nutzer des Wörterbuchs) treffen

Erwartung: das Trainieren eines maschinellen Lerners mit diesen klassifizierten Daten (aufnehmen / verwerfen) kann helfen, in Zukunft die suboptimalen Vorschläge des regelbasierten Extraktors zu verbessern

Erweiterung mit ML-Techniken

Die ML-Komponente setzt an der Ausgabe des regelbasierten Verfahrens an (Ausgabe in der Regel zwischen 10 und 20 Beispielen pro Stichwort)

Handklassifizierte Daten: 13.200 Beispiele für 1.050 Stichwörter; Zwei Klassen: Aufnehmen? Ja/ Nein

Diese Daten wurden in zwei in etwa gleichgroße Partitionen geteilt: Trainings- und Testdaten

Erweiterung mit ML-Techniken

Ansatz: Support Vector Machines, Rapid Minder Software

Merkmale des Kontexts:

- wortbasiert (Bag of Words)
- Mit Wortarten (Parts of Speech)
- Satzstruktur (Parse Trees)

Evaluierung

Beste Merkmalskombination ist die Berücksichtigung aller drei Merkmale, der „Gewinn“ gegenüber dem einfachen wortbasierten Merkmal ist aber gering.
Testdaten:

ha ml	Accept	Dismiss	Total
Accept	603	487	1090
Dismiss	1,774	3,880	5,604
Total	2,377	4,317	6,694

Evaluierung

- Recall gute Beispiele = 603 / 2,377 = **25.3 %** (d.h. von den 2.377 durch den Lexikographen als “gut” bewerteten Belegen wurde 603 auf durch den ML so bewertet)
- Precision gute Beispiele: 603 / 1,090 = **55.3 %** (d.h. von den 1090 vom ML als “gut” klassifizierten Belegen wurden 603 auch vom Lexikographen so bewertet)
- Recall schlechte Beispiele = 3,830 / 4,317 = **88.7 %**
- Precision schlechte Beispiele: 3,830 / 5,604 = **68.3 %**.

Evaluierung

- F-score for gute Beispiele beträgt 0.34
- F-score für schlechte Beispiele beträgt 0.76
- Akkuratheit beträgt 0.66

(Akkuratheit = Anzahl der korrekt klassifizierten Beispeile dividiert durch die Gesamtzahl der Beispiele)

.

Fazit

Eine maschineller Lerner als zusätzliche Komponente, optimiert auf das Erkennen „schlechter“ Beispiele

1. reduziert deutlich die Liste der Belege, die ein Lexikograph ansehen sollte (auf ca. 16 %) (erwünscht)

2. filtert bei ca. 25 % der Stichwörter ALLE guten Belege raus (nicht erwünscht)

Lösung für (2):

- Der regelbasierte Extraktor liefert mehr Belege
- Die Performanz des Lerners wird verbessert durch weitere Merkmale, die „gute Belege“ von schlechten unterscheiden

Beteiligte

- Jörg Didakowski - Gute-Belege-Extraktor
- Alexander Geyken – Koordination
- Lothar Lemnitzer - Klassifikation, Evaluation
- Christian Pölitz - ML Experimente

Danke…

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

für Ihr Interesse!

Kontakt:

{didakowski,lemnitzer,geyken}@bbaw.de

poelitz@uni-dortmund.de

www.dwds.de; zwei.dwds.de (beta)