Induktive Topikmodellierung und extrinsische Topicdomänen

**Abstract**

Wir machen Topik von zwei Seiten, einmal von oben und einmal von unten und gucken dann, wie das zusammen passt. Genauer gesagt gucken wir, inwiefern sich extrinsisch motivierte Domänen mit ML-Verfahren aus induzierten Topiks lernen lassen. Für die Topikmodllierung nehmen wir nicht LDA, wie alle anderen, sondern LSI, das ist zwar alt, geht aber auf unseren Datensätzen besser. A propos Daten, die kommen aus COW (ein gecrawltes Webkorpus von den Autoren selbstgemacht) und Korpora mit vorwiegend Zeitungstexten (DeReKo, vom IDSs selbtsgemacht). Wir probieren viel mit verschiedenen Parametern herum und am Ende kommt heraus, dass es alles so mittelgut funktioniert.

# Motivation

Es ist ja kein Geheimnis, dass es mit „Topiks“ immer so eine Sache ist, eine schwere: sei es in der Informationsstruktur (um die es hier zwar gar nicht geht, die aber trotzdem kurz erwähnt sei, nicht zuletzt, weil man sonst vielleicht denken könnte, dass es doch darum geht), sei es in den digitalen Geisteswissenschaften, wo mit „Topik“ ja nun wieder etwas ganz anderes gemeint ist, und darum (also um letzeres) dreht es sich hier. Topik ist da (also in den DH) das Thema (jetzt wieder nicht informationsstrukturell gemeint, aufgepasst!) von einem Text, also das, worüber der Text spricht. Der vorliegende Text spricht beispielsweise von „Topiks“, sein Topik (oder Thema) ist also „Topiks“. Ganz so einfach ist das aber nun wieder auch nicht, und damit wären wir schon bei der Schwierigkeit: Ist „Topik“ denn überhaupt ein richtiges Topik, oder sollte man nicht eher sagen, das Topik (i.S.v. Thema) dieses Textes ist „Topikmodellierung“ (also ein bisschen genauer als nur „Topik“), oder aber so etwas wie „Textklassifikation“ (also ein bisschen ungenauer als „Topik“, dafür aber vielleicht für manche Sachen nützlicher). Das eigentliche Problem ist ja: so eine richtige, verbindliche, allgemein anerkannte Taxonomie von Themen gibt es ja gar nicht (da kann man auch mal bei den EAGLES guidelines gucken, die geben an der Stelle auch auf) und es kann so etwas wohl auch gar nicht geben. Wenn es aber eh in gewissem Maße beliebig ist, ob man jetzt das eine oder das etwas andere Klassifikationsschema nimmt, dann könnte man ja auch gleich das nehmen, das sich besonders gut aufgrund lexikalischer Verteilungen etablieren lässt. Man will dann also thematische Kategorien so haben, dass sie einerseits hinreichend gut interpretiert werden können, die sich andererseits aber auch hinreichend aus den Daten heraus rechtfertigen lassen, sich also für eine automatische Klassifikation gut eignen und gleichsam in den Daten wurzeln. Automatisch muss die Klassifikation sowieso sein, bei den Korpusgrößen wie man sie heute hat geht es ja gar nicht anders, und man möchte dann aber die Kategorien so wählen, dass man einen guten Kompromiss hat zwischen automatischer Klassifizierbarkeit und semantischer Interpretierbarkeit. Wir präsentieren jetzt hier also eine kleine Studie, bei der wir gucken, inwieweit sich induktiv ermittelte thematische Charakterisierungen von Texten (also ohne Rückgriff auf ein Inventar thematischer Kategorien) durch automatische Verfahren abbilden lassen auf ein gegebenes Klassifikationsschema für thematische Großbereiche. Solche Großbereiche nennen wir „Topik-Domänen“, sie sind zu verstehen als eine relativ abstrakte Ebene in einer thematischen Taxonomie (beispielsweise lassen sich einzelne Topiks wie „Fussball“, „Reiten“, „Eishockey“ zur Topikdomäne „Sport“ zusammenfassen).

# Daten

Stichproben aus DeReKo (...) und DECOW14 (...)

# Topiks und Topikdomänen

Das entspricht dem Top-down Ansatz:

* 2-3 Beispiele für solche Klassifikationsschemen
* CowCat
* Top down Ansatz (manuelle Annotation): Verteilung über Topikdomänen in DeReKo und COW
* Plot

# Topikmodellierung

Topikmodellierung bezeichnet eine Reihe von Verfahren zur automatischen thematischen Erschließung größerer Textmengen, die in Teilen der digitalen Geisteswissenschaften seit Jahren etabliert sind und deren Nutzen als heuristisches Werkzeug in zahlreichen Fallstudien dokumentiert wurde (paar Beispiel). Die am meisten verbreiteten Verfahren dieser Art stützen sich ausschließlich auf (Ko-) Okkurrenzen von Wörtern in Texten und erfassen durch verschiedene numerische Techniken Ähnlichkeiten zwischen den Texten. Da über ein bestimmtes Thema in der Regel mit einem bestimmten Vokabular gesprochen bzw. geschrieben wird, ergeben sich somit semantische Ähnlichkeiten. (Daneben gibt es Erweiterungen, bei denen auch andere Informationen Eingang in die Berechnung finden: ... ein oder zwei Beispiele). Ein sehr bekanntes Verfahren ist dabei *Latent Semantic Indexing* (LSI, auch bekannt als *Latent Semantic Analysis*; Landauer und Dumais 1997). Ein anderes, etwas rezenteres Verfahrenist *Latent Dirichlet Allocation* (LDA, REF). Bei beiden handelt es sich um unüberwachte , bottom-up Verfahren, bei denen nicht nach vorgegebenen Kategorien klassifiziert wird, sondern die heuristisch semantische Strukturen in Textsammlungen aufdecken. Im Zusammenhang dieses Papiers ist es wichtig zu betonen, dass weder LSI noch LDA *Bezeichnungen* für die induzierten Topiks liefern. Die Ausgabe beschränkt sich vielmehr auf eine Liste besonders charakteristischer Wörter für ein Topik, sowie eine gewichtete Zuordnung einzelner Texte zu den einzelnen Topiks. Inwieweit auf diese Weise induzierte Topiks-ohne-Namen automatisch zu vorgebenen Topikdomänen zusammengefasst werden können, ist Gegenstand dieser Studie. Wir verwenden *Latent Semantic Indexing* für die Topikmodellierung,

# Experiment #1: Topikmodellierung für Stichproben aus DeReKo und DECOW

- mit schönem Vergleichsplot vom Poster, wenn er lesbar reinpasst.

# Experiment #2: Automatisches Lernen vorgegebener Topikdomänen aus induzierten Topiks

- mit Beschreibung und allen Hotzenplots vom Klassifikationsexperiment

# Diskussion und Ausblick