Tagging von Online-Blogs

Gertrud Faaß (vertreten durch Josef Ruppenhofer)

STTS tagset and tagging: special corpora

24. September 2012

Korpuslinguistische studentische Projekte am IwiSt

SoSe2012

- Studentische Arbeit im interdisziplinären Seminar: *Korpuslinguistisches Experimentieren mit authentischen Texten* (IKÜ, Fr. Dr. Bedijs und IwiSt, Gertrud Faaß/PhD University of Pretoria). Bachelor, 4. Semester
- Aufgabenstellung: Eigenes Spezialkorpus aus dem Web sammeln, aufbereiten und untersuchen in Bezug auf Differenzen zur Allgemeinsprache
- Raissa Khattab: Modeblogs

Probleme bei der Vergabe solcher Aufgaben an BA-Studierende an der Uni Hi

- Kaum/keine Kenntnisse über Computer als MSV-Werkzeuge (Einsatz eines Crawlers ausgeschlossen)
- Kaum/keine Linux-Kenntnisse (Verwendung vorgefertigter Skripte ist möglich)
- Geringe Kenntnisse (nur aus dem Seminar) über
  - Korpusformate
  - Vorgehen beim Sammeln entsprechender Daten
- Kein CL-Problembewußtsein

Ein Korpus mit Modeblog-Beiträgen

Vorgehen der Studentin

- Manueller Download (copy/paste) von Beiträgen aus 10 online-Modeblogs und Erstellen einer .xls Datei
- (SGML)-Annotation von Metadaten: Source (url/Datum)
- Anonymisierung von Autorinnen
  (url in der CWB-Fassung des Korpus nicht mit enkodiert)

Ein Korpus mit Modeblog-Beiträgen

Gesammelte Daten

Korpusgröße: ca. 57.000 tokens (370 Beiträge)
Format:

<text>
<date>2012-06-11</date>
<title>New In.</title>
<text>Leider komme ich momentan nicht zum Outfit fotografieren. Ich hoffe, dass ändert sich im Laufe der Woche. [...] </text>

Ein Korpus mit Modeblog-Beiträgen

Weiteres Vorgehen

- Tokenisierung, Tagging (STTS)
- Enkodierung als CQP-Korpus
- Untersuchung von
  - Adjektiven, die Kleidungsstücke näher erläutern
  - Diminutiven

Symbols

Vorschlag: Kategorie/Unterkategorie **SYMB.emot**

- Häufige Verwendung von Smileys und Symbolen
  $\rightarrow$ Tokenizing-Probleme
- Beispiele:

;), ;)*	WINKSYMB	:), :)*	SMILESYMB
<3	HEARTSYMB	:D*	BIGSMILESYMB
:-*	KISSSYMB	:P*	TONGUESYMB

- Lösung: Umschreibung (per ssh-skript im Text geändert)
- Wie taggen? Einführung eines “emoticon” tags?
  Idee: Einfügung von **SYMB** mit Unterkategorie **emot**
- Wie wird das sonst in Chat-/Twitterkorpora gehandhabt?

Korpus MODEBLOGS

Vorbild: EAGLES / RF-Tagger

- EAGLES fordert dazu auf, je nach Nutzung genauere Annotation im Bedarfsfall einzuführen (EAGLES (1996)).
- RF-Tagger nutzt dieses Prinzip für die Annotation von morpho-syntaktischen Kategorien.
- Problem: Gleiche Ebene = Gleiches Merkmal (Schmid und Laws (2008))
- Regel-/lexikonbasiertes Ergänzen nach dem statistischen Taggen mit Tag-Erweiterungen: mehr Möglichkeiten.

Korpus MODEBLOGS

Vorschlag: Unterkategorie **NN.dim**

- Untersuchungsobjekt: Diminutive (*Herzchen, Nagelstäbchen, Schleifchen* sind signifikant häufig (Vergleich mit SDEWAC)
- Vorschlag: **NN.dim**
- Weitere: **ADJA/D.pos**, **ADJA/D.cmp**, **ADJA/D.sup**?

Korpus MODEBLOGS

Vorschlag: Unterkategorie **NN/NE.len**

- Untersuchungsobjekt: englische *loanwords* sind signifikant häufig (im Vergleich mit SDEWAC)
- Weitere Unterkategorie für NN (evtl. auch NE?): “LEN” (loanword english), würde auch “LFR” ermöglichen (*Trottoir*)
- Eventuell ist diese Unterkategorie auch für weitere Wortarten interessant (*downloaden*).

Referenzen

EAGLES (1996). *Evaluation of Natural Language Processing Systems*, EAGLES document EAG-EWG-PR.2: Final Report.

Khattab; Raissa (2012). Seminararbeit: Sprachgebrauch des Deutschen und Terminologie in Modeblogs. *SoSe2012: korpuslinguistisches Experimentieren mit authentischen Texten*, Dr. K. Bedijs (IÜF) und G. Faaß, PhD/University of Pretoria (IwiSt)

Schmid, Helmut; Laws, Florian Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging in *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008) pp. 777-784 Coling 2008 Organizing Committee, Manchester, UK.