Text Mining
Aplicaciones

Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física
UNC, Córdoba (Argentina)
http://www.cs.famaf.unc.edu.ar/~laura

SADIO
13 de Marzo de 2008

contenidos

-

bootstrapping

self-training:

1. un pequeño conjunto de ejemplos etiquetados (E) + un gran conjunto de ejemplos no etiquetados (N)
2. se entrena un clasificador en E, y se lo aplica a N
3. los ejemplos de N que han sido clasificados con mayor confianza pasan a formar parte de E
4. si quedan ejemplos sin clasificar en N, se vuelve al punto 2.

co-training:

1. se aprenden dos clasificadores independientes sobre los mismos datos etiquetados
2. si los dos clasificadores predicen la misma clase y con buena confianza, se toma la clase como buena

bootstraping para desambiguar sentidos

Dan Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. ACL'95.

determinar cuáles ocurrencias de la palabra "*plant*" tienen el significado de "*planta industrial*" y cuáles de "*ser vivo*"

bootstraping para desambiguar sentidos

Dan Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. ACL'95.

- un pequeño número de ejemplos etiquetados
- un gran número de ejemplos no etiquetados
- los ejemplos se caracterizan por
  - palabras en la oración
  - documento en el que ocurre

bootstraping para desambiguar sentidos

Dan Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. ACL'95.

procedimiento (*one sense per collocation*):

1. se obtiene una lista de decisión de los ejemplos etiquetados
   - reglas de asociación por clase
   - ordenadas por fiabilidad
   - basadas en las palabras de la oración
2. se clasifican los ejemplos sin etiquetar
3. los ejemplos que se pueden clasificar pasan a los etiquetados
4. se itera hasta que no se clasifiquen nuevos ejemplos

si quedan ejemplos sin etiquetar (*one sense per discourse*):

- se asigna a todos los ejemplos del documento el mismo sentido $\longrightarrow$ autocorrección
- si quedan ejemplos sin etiquetar, se vuelve a aplicar el procedimiento

descubrimiento de sentidos mediante clustering

Pantel, P. and Lin, D. 2002. Discovering Word Senses from Text. KDD-02

-

descubrimiento de sentidos mediante grafos

J. Véronis. 2004. HyperLex: Lexical Cartography for Information Retrieval. Computer, Speech and Language, 18 (3)

-

traducción automática estadística

Statistical MT Handbook by Kevin Knight

-

adquisición de paráfrasis

Regina Barzilay and Kathy McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. ACL.

-

asociación de pares pregunta–respuesta

Abdessamad Echihabi and Daniel Marcu (2003). A Noisy-Channel Approach to Question Answering. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), July 7-12, Sapporo, Japan.

-

inducción de estructura

Unsupervised learning of natural languages. Zach Solan, David Horn, Eytan Ruppin, Shimon Edelman, 2005. PNAS.

-

inducción de plantillas

Cross-component Clustering for Template Induction. Zvika Mark, Ido Dagan and Eli Shamir. 2002. Workshop on Text Learning (TextML 2002)

-

minado de opiniones

Overview of the TREC 2006 Blog Track

-

reconocimiento no supervisado de entidades con nombre

Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity Nadeau, D., Turney, P., Matwin, S, 2006 . 19th Canadian Conference on Artificial Intelligence.

-

reconocimiento de relaciones entre entidades

Discovering Relations among Named Entities from Large Corpora
Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman, 2004. ACL

-

evaluación

-