Kielitieteellisten aineistojen käsittely

1 Johdanto

Kielitieteellisten aineistojen kokoaminen ja etenkin kommentointi ja metadatalla varustaminen on suuri työ. Kommentit ja metadata voivat olla hyvin monenlaisia, mutta niiden luominen on työlästä. Yksinkertainen kieliopillinen luokittelu voidaan mahdollisesti tehdä automaattisestikin, mutta monimutkaisempi kielen jäsentely vaatii yleensä asiantuntijatyötä.

Jotta tällainen kieliaineistojen käsittely olisi kannattavaa, pitäisi aineistojen hyödyntämisen olla tehokasta ja monipuolista. Usein kieliaineistojen käyttö on rajoittunut erilaisiin frekvenssianalyyseihin ja kielen tyypillisten piireiden kartoittamiseen. Monimutkaisemmat analyysit vaativat pohjakseen yksityiskohtaisemmin jäsenneltyjä aineistoja, jotka on vielä jäsennelty ja kommentoitu juuri käsillä oleva tutkimusongelma huomioon ottaen.

Aineistoja voidaan käyttää paitsi kielitieteellisten ongelmien tutkimiseen ja selvittämiseen myös koneoppimisen opetusaineistona, tietämyksen muodostamisen pohja-aineistona jne.

2 Aineistojen kommentointi, metadatan tyypit

- erilaiset metadatan/kommentoinnin tasot (kieliopillinen, morfologinen, lauseenjäsennys,jne.)
- erilaiset metadatan/kommentoinnin tekniikat
- Lähteitä:
    - Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C. & Liberman, M. 2000. ATLAS: A Flexible an Extensible Architecture for Linguistic Annotation. Proceedings of the Second International Conference on Language Resources an Evaluation, European Language Resources Association, Paris, 1699-1706.
    - Bird, S. & Liberman, M. 2000. A Formal Framework for Linguistic Annotation. Speech Communication.
    - Ide, N. & Romary, L. A Registry of Standard Data Categories for Linguistic Annotation.
    - Ide, N. & Suderman, K. 2007. GrAF: a graph-based format for linguistic annotations. ACL Workshops. Proceedings of the Linguistic Annotation Workshop, 1 - 8.

- Katz, G. & Arosio, F. 2001. The annotation of temporal information in natural language sentences. Annual Meeting of th ACL. Proceedings of the workshop on Temporal an spatial information processing. Volume 13, Article No. 15.
- Xiaoyi, M., Lee, H., Bird, S. & Maeda, K. 2002. Models and Tools for Collaborative Annotation. Proceedings of the Third International Conference on Language Resources an Evaluation, Paris: European Language Resources Association, 2002.

3 Aineistojen käsittely

3.1 Rakenteisten kieliaineistojen kyselykielet

- Lause- ja virketason kommentit ovat luonteeltaan rakenteisia.
- Aineiston tutkimista varten on kehitetty useita erilaisia kyselykieliä.
- Esimerkkejä: LPath ja LPath$^+$ (XPathin kielitieteellisiä laajennoksia), MQL
- Graafisesti mallinnetut kommentit ja niiden kyselykielet (Emu)
- Lähteitä:
  - Cassidy, S. & Bird, S. 2000. Querying Databases of Annotated Speech.Database Technologies: Proceedings of the Eleventh Australasian Database Conference. IEEE Computer Society, 12-20.
  - Lai, C. & Bird, S. 2009.Querying linguistics trees. Journal of Logic. Language and information 19 (1), 53-73.
  - Petersen, U. 2004. Emdros: a test database engine for analyzed or annotated text. International Conference On Computational Linguistics. Proceedings of the 20th international conference on Computational Linguistics. Geneva, Switzerland.

3.2 Tiedonlouhinta monitasoisesti jäsennellyistä kieliaineistoista

- Pohjana on jäsennelty kieliaineisto.
- Aineisto täytyy kuvata uudestaan, abstrahoida, nimenomaan tutkittava ongelma huomioon ottaen
- Apuna voidaan käyttää rakennefragmentteja - fuzzy tree fragments, FTF

- Koneoppimisen algoritmit, joissa käytetään päättelysääntöjä (itsenäiset päättelysäännöt, päättelyrakenteet, järjestetyt säännöt) tai hahmontunnistusta (pattern matching)
- Lähteitä:
    - Aarts, B., Nelson, G. & Wallis, S. 2007. Using Fuzzy Tree Fragments to explore English grammar. English Today (2007), 23: 27-31. Cambridge University Press.
    - Cupit, J. & Shadbolt, N. 1996. Knowledge discovery in databases: Exploiting knowledge-level redescription. Advances in Knowledge Acquisition 1076: 245-261. Springer-Verlag, Berlin.
    - Halstead, P. & Rose, TG. Extracting conceptual knowledge from text using explicit relation markers. Advances in Knowledge Acquisition 1076: 147-162. Springer-Verlag, Berlin.
    - Wallis, S. & Nelson, G. 1997. Syntactic parsing as a knowledge acquisition problem. Knowledge Acquisition, modeling ja management 1319: 285- 300. Springe-Verlag, Berlin.
    - Wallis, S. & Nelson, G. 2001. Knowledge Discovery in Grammatically Analysed Corpora. Data Mining and Knowledge Discovery. Volume 5, Issue 4 (October 2001), 305 - 335.
    - Wallis, S. & Nelson, G. 1997. Syntactic parsing as a knowledge acquisition problem. Knowledge acquisition, modeling and management 1319: 285-300.

4 Yhteenveto

Riippumatta siitä, millä menetelmällä aineistoa on tarkoitus käyttää hyväksi, tärkeää on se, että aineiston kommentointi on tehty nimenomaan tutkittavaa ongelmaa varten. Tällöin ns. yleiskäyttöinen kieliaineisto vaatii yleensä vielä jonkinlaisen abstrahoinnin ja lisäkuvauksen, jotta kyselyt tai tiedonhaut tuottaisivat tuloksen.

Lähteet

BDG00 Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C. & Liberman, M. 2000. ATLAS: A Flexible an Extensible Architecture for Linguistic Annotation. Proceedings of the Second International Conference on Language Resources an Evaluation, European Language Resources Association, Paris, 1699-1706.

BiL00 Bird, S. & Liberman, M. 2000. A Formal Framework for Linguistic Annotation. Speech Communication.

IdeRo Ide, N. & Romary, L. A Registry of Standard Data Categories for Linguistic Annotation.