SYSTEM OF SEMANTIC TYPES IN PDEV

Outline of the presentation
--------------------------------

Prague, Dec 6, 2010

by Martin Holub and Lenka Smejkalova

Institute of Formal and Applied Linguistics
Charles University in Prague

1) The Prague-&-Brno PDEV team, the goals

* The team members:
  - Silvie Cinkova, Martin Holub, Lenka Smejkalova = Prague team
  - Adam Rambousek, Pavel Rychly = Brno team, infrastructure
  - Patrick Hanks = the CPA author, lexicographer, advisor

* PDEV as an NLP applicable source?
  - for NLP application the PDEV data should
    - be consistent as much as possible
    - make at least a representative sample (in statistical sense, we need corpus coverage)
    - be clear enough at least for humans (to test it we measure inter-annotator agreement)

* Two basic NLP tasks:
  - pattern recognition and pattern discovering
  - from the machine learning point of view:
    - the first task is a (standard) classification task, while
    - the second task is a clustering task
  - strategic application at UFAL: machine translation
  - fundamental assumption: patterns imply meaning, the task is semantically oriented
*** 2) Basic PDEV structure

* Three main components
  - pattern database
  - manually tagged reference samples attached to each PDEV entry
  - system of semantic types, corpus-driven, linguistically oriented

* What is a "good PDEV ontology"???
  - our view (if PDEV is used for NLP): "good ontology" means a system of semantic types that helps to automatically recognize patterns well

*** 3) Terminology: Semantic Types vs. Lexical Sets

* Terms
  - semantic types = "labels" used in pattern definitions
  - lexical sets = "groups of paradigmatically related words that may fill the argument positions in a pattern"

* Needs
  - humans need clear and consistent definitions of semantic types
  - on the other hand, for machine learning we do not need to define semantic types, because computers cannot understand human definitions; for machine learning purposes we need consistent (training) data - the greater volume, the better
  - lexical sets should be extracted from a large corpus and optimized by computer so that they serve to pattern recognition
  - to extract the whole set of nouns for a given semantic type we need the union of all relevant lexical sets
4) **Unclear semantic types can be a cause of inconsistencies in PDEV data**

- there is no documentation of the system of semantic types used in PDEV -- neither definitions, nor relations
- possible inconsistencies in using semantic types have not been explored/mapped yet

- consistent using and interpretation of semantic types requires their definitions:
  - we need good/clear definitions of semantic types in order to keep pattern database consistent: so that different lexicographers can use the established set of semantic types consistently
  - definitions of semantic types are also important for interpretation:
    - for lexicographers who browse the dictionary
      - for annotators (to make manually tagged data of good quality) and
    - for "normal" PDEV users

5) **The existing data about semantic types in the current PDEV**

* Extracting lexical sets from manually tagged sentences
  - the data used (about 200K manually tagged sentences)
  - verb arguments extraction using an automatic parser
  - the tools to browse the data:
    - filtering and sorting according to frequency and PMI
    - displaying relevant sentences

* Manually tagged data
  - almost 9000 pairs (ST, noun) tagged by Patrick, tagset={'T','C','M'}
  - randomly selected from the whole set extracted from tagged sentences
  - we obtained a small samples for some semantic types
  - machine learning still unsuccessful as the feature set used does not provide enough information

6) **Conclusion: what we need in the nearest future**

- semantic types definitions, guidelines for their use/interpretation
- more consistently annotated data for lexical sets extraction