Terminological variation and term candidate extraction

Ulrich Heid

Universität Hildesheim

TermVar Workshop, Hildesheim, 20/21-7.2017:
Introduction to discussion session 2
Term variation and term candidate extraction

General aspects

- Finding terms in texts:
  - Term recognition:
    Finding *known* terms in textual documents
    e.g. to identify example sentences
  - Term extraction:
    Finding *term candidates* in textual documents
    e.g. to find new terms

- Terminological variation
  - Relevant for both procedures
  - Identification of variants is typically a separate second process

- In this workshop:
  emphasis on term (candidate) extraction
Term candidate extraction

Who needs this technology? – Who needs data about variation?

- Translators
  - Term lists with explicit statements about variants
  
  See discussion this afternoon

- Knowledge Engineering
  - Terms as linguistic objects related with concepts
  - Variants related to the same concept:
    Different expressions for a given concept

- Information Retrieval
  - Terms as search items
  - Variants used to achieve more recall

- Natural Language Processing
  - Terms integratable into language resources
  - Variants used e.g. for coreference resolution
Term extraction procedures
Basics: different approaches to monolingual extraction

• Size of components searched:
  – Sub-word-based: by morphemes or letter sequences
  – Word-based: by words/lemmas, based on frequency
  – Word-sequence-based:
    * by word sequences and their frequency
    * by pos-shapes or other syntactic patterns

• Techniques used
  – Statistics based on word (sequence) frequency
    within specialized texts or by comparison with ”general language”
  – Statistics based on association (measures)
  – Symbolic patterns
  – Hybrid approaches: patterns plus statistics
Term extraction procedures and variation

Sub-word-based approaches

- Search for domain-relevant morphemes
  - e.g. neoclassical morphemes
  - typically: items from a database used as seeds

- Search for letter sequences
  - 4-tuples with high recurrence in domain texts, based on “informative words” (Vergne 2003)

- Treatment of variants:
  - The approaches find orthographically/morphologically related items, e.g. \textit{techn-} : \textit{-ique}, \textit{-ology}, \textit{-ical}, ...
  - Relationship between items found remains unclear: Need for additional categorization
Term extraction procedures and variation

Word- and word sequence-based approaches

- Statistical word-based approaches will extract variants, but not identify relations between them: again need for additional procedures

- Search for morphologically unrelated synonyms: Via distributional semantics: only in (very) large corpora

- Word-sequence-based approaches:
  - Morpho-syntactic patterns allow for an explicit description of relationships between variants: $N_1 N_2 \leftrightarrow N_2$ of $N_1$: energy production $\leftrightarrow$ production of energy
  - This approach can be combined with morphological analysis: DE Energieproduktion $\leftrightarrow$ Produktion von Energie
  - Still no information about status of variants: Heuristic assumption: most frequent variant is preferred
Richness of texts wrt variants

Observations from past experiments

- Texts produced in technical writing:
  The more controlled, the less variants
  cf. guidelines

- Technical texts from different sources:
  expectably more variation than from single source

- User-generated content:
  - Tendentially more variants than in expert text
  - Jargon:
    Abbreviations: Tischkreissäge – TKS
    Ad hoc short forms: BMW 730i – 730er
  - More story-like texts: more hypernym-like variants:
    ... the circular saw ... . This saw ...

- Experience from the TTC project:
  EU, 2010-2012
  More variation in Romance languages than in Germanic languages
Questions for discussion

- Variation at different levels of analysis:
  - Words (e.g. synonyms)
  - Multiword terms, possibly related with word formation products
    ▶ Which types can be extracted, with which quality?

- Relations between variants:
  Which extraction quality can be achieved?
    ▶ Semantic relations between variants, e.g. synonyms/hypernyms?
    ▶ Pragmatic relations: preferred variants, jargon, ...?

- Which and which amount of language resources are needed?
  - Patterns at POS-level?
  - Lexical resources, e.g. for neoclassical morphemes?
  - Deeper syntactic and/or morphological analysis,
    e.g. parsing, word formation analysis, ...
    ▶ Effort/Investment ↔ gain in quality?