Mining Maximal Frequent Sequences for Multilingual Multi-Word Expression Extraction

Antoine Doucet
University of La Rochelle, France

L3i Informatique Image Interaction Univ. La Rochelle

UNIVERSITÉ La Rochelle

Mining MFS for Multilingual MWE Extraction

- Maximal. . . Frequent. . . Sequences
- We can extract "former President Clinton" from the fragments
  - "the former President of the United States Bill Clinton"
  - "former President William Jefferson Clinton"
- The approach can handle
  - Collections of any size (parallel processing)
  - Collections in **any language**, of any document genre
- **To-Do:** Hybridization
  - As a prior, to avoid to process text that cannot contain MWEs
  - With linguistic knowledge in the loop (e.g., POS tagging)
  - With more integrated approaches, to be defined on a case-by-case basis

Mining MFS for Multilingual MWE Extraction

- Maximal. . . Frequent. . . Sequences
- We can extract "former President Clinton" from the fragments
  - "the former President of the United States Bill Clinton"
  - "former President William Jefferson Clinton"
- The approach can handle
  - Collections of any size (parallel processing)
  - Collections in **any language**, of any document genre
- **To-Do:** Hybridization
  - As a prior, to avoid to process text that cannot contain MWEs
  - With linguistic knowledge in the loop (e.g., POS tagging)
  - With more integrated approaches, to be defined on a case-by-case basis

Mining MFS for Multilingual MWE Extraction

- Maximal. . . Frequent. . . Sequences
- We can extract "former President Clinton" from the fragments
  - "the former President of the United States Bill Clinton"
  - "former President William Jefferson Clinton"
- The approach can handle
  - Collections of any size (parallel processing)
  - Collections in **any language**, of any document genre
- **To-Do:** Hybridization
  - As a prior, to avoid to process text that cannot contain MWEs
  - With linguistic knowledge in the loop (e.g., POS tagging)
  - With more integrated approaches, to be defined on a case-by-case basis

Mining MFS for Multilingual MWE Extraction

- Maximal. . . Frequent. . . Sequences
- We can extract “former President Clinton” from the fragments
  - “the former President of the United States Bill Clinton”
  - “former President William Jefferson Clinton”
- The approach can handle
  - Collections of any size (parallel processing)
  - Collections in **any language**, of any document genre
- **To-Do:** Hybridization
  - As a prior, to avoid to process text that cannot contain MWEs
  - With linguistic knowledge in the loop (e.g., POS tagging)
  - With more integrated approaches, to be defined on a case-by-case basis

Mining MFS for Multilingual MWE Extraction

- Maximal. . . Frequent. . . Sequences
- We can extract “former President Clinton” from the fragments
  - “the former President of the United States Bill Clinton”
  - “former President William Jefferson Clinton”
- The approach can handle
  - Collections of any size (parallel processing)
  - Collections in **any language**, of any document genre
- **To-Do:** Hybridization
  - As a prior, to avoid to process text that cannot contain MWEs
  - With linguistic knowledge in the loop (e.g., POS tagging)
  - With more integrated approaches, to be defined on a case-by-case basis

Mining MFS for Multilingual MWE Extraction

- Maximal. . . Frequent. . . Sequences
- We can extract “former President Clinton” from the fragments
  - “the former President of the United States Bill Clinton”
  - “former President William Jefferson Clinton”
- The approach can handle
  - Collections of any size (parallel processing)
  - Collections in **any language**, of any document genre
- **To-Do:** Hybridization
  - As a prior, to avoid to process text that cannot contain MWEs
  - With linguistic knowledge in the loop (e.g., POS tagging)
  - With more integrated approaches, to be defined on a case-by-case basis