Language Understanding Systems
Mid-Term Project: *FST & GRM Tools for SLU*

Evgeny A. Stepanov

SISL, DISI, UniTN
`evgeny.stepanov@unitn.it`

Objective

Develop **Spoken Language Understanding** (SLU) Module
for Movie Domain using NL-SPARQL Data Set

who plays luke on star wars new hope

Concept Tagging

```
who     O
plays   O
luke    B-character.name
on      O
star    B-movie.name
wars    I-movie.name
new     I-movie.name
hope    I-movie.name
```

IOB Notation

The notation is used to label *multi-word* spans in token-per-line format. Both, prefix and suffix notations are commons: `B-NP` vs. `NP-B`

- **B** for **B**eginning of span
- **I** for **I**nside of span
- **O** for **O**utside of span
- Sometimes **E-** for **E**nd of span

```
who      O
plays    O
luke     B-character.name
on       O
star     B-movie.name
wars     I-movie.name
new      I-movie.name
hope     I-movie.name
```

Tools

Develop **Spoken Language Understanding** (SLU) Module for Movie Domain using NL-SPARQL Data Set

Sequence Labeling
- OpenFST
- OpenGRM

Data Set

NL-SPARQL Data Set
See `readme.txt` in `data.zip`

Sequence Labeling

- Token-per-line
  - words (tokens)
  - concept tags (IOB-format)
- Additional Features (optional to use)
  - POS-tags (automatic)
  - Lemmas (automatic)

Tasks: Minimum

Train *Concept Tagger*

Sequence Labeling

- FST&GRM
  - Train WFST & LM
  - Experiment with different Language Model parameters
    - ngram size
    - smoothing
  - Take care of **unknown** words
    - e.g. lexicon frequency cut-off
- Evaluate with `conlleval.pl`

Expected Performance: $F_1 \approx 76$

Tasks: Improvements

Language Model trained on tags only: the majority is 'O'

- Train LM using the **words** & **concept tags**
- Implement the tagging pipeline to make use of that
- Evaluate and compare

Expected Performance: $F_1 \approx 82$

To Submit

REPORT (≈4 pages) that includes:

- Data Analysis
  - Data size
  - Distribution of concepts (not IOB-tags)
  - etc.
- Evaluation (with Baseline)
- Comparison of different training parameters and settings
- Error Analysis
- Discussion

CODE with readme (e.g. GitHub link)

Report template is provided