Language Understanding Systems

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

Evgeny A. Stepanov

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

Objective

Develop Concept Tagging Module for Movie Domain using NL-SPARQL Data Set

who plays luke on star wars new hope

Objective

Develop Concept Tagging 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

Signals & Interactive Systems

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 Concept Tagging Module for Movie Domain using NL-SPARQL Data Set

Sequence Labeling

- OpenFST
- OpenGRM

Data Set

NL-SPARQL Data Set
`https://github.com/esrel/NL2SparQL4NLU`

Sequence Labeling

- Token-per-line
  - words (tokens)
  - concept tags (IOB-format)

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` or `conll.py`
  - `https://github.com/esrel/LUS/` (src)

Expected Performance: $F_1 \approx 76$

Tasks: Improvements (2018)

Issue

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: $F_1 \approx 83$ *(Gobbi et al., 2018)*

Tasks: Improvements (2019)

Issue

- *Sparsity* for certain classes.
- *Overlap* for certain classes (e.g. 'actor' & 'director')

Entity Recognition + Concept Tagging

- Explore Normalization for NLU (using e.g. RegEx, NER)
- Generalize words (phrases) to entities prior to concept tagging (add a transducer)
- External resources like lookup tables are allowed to be used
- experiment on different entity sets

Expected Performance: inconclusive

Tasks: Improvements (2019)

Example

```
how many woody allen   movies starred diane keaton
how many person.name   movies starred person.name
O    O    director.name O      O       actor.name
```

Tasks: Improvements (2020)

Ideas

- Joint modeling with SCLM
- Comparison to NTLK HMM
- Sequential segmentation and labeling

Tasks: Improvements (2021)

Ideas

- Do not use additional features! (POS, LEMMA)
- Compare models on another data set (requires conversion to CoNLL format)
- Compare to published results

- `https://github.com/howl-anderson/ATIS_dataset`
- `https://github.com/sebischair/NLU-Evaluation-Corpora`
- `https://github.com/sonos/nlu-benchmark` (results)

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)