Site Update
Rio de Janeiro

Alexandre Rademaker

Jun, 25th 2023

summary

Jan 2023 - Feb 2023. I have Francis and Dan at FGV/EMAp.

Working with

- Portuguese Grammar - PorGram
- LTDB - Linguistic Type Data-Base
- GlossTag corpus - https://github.com/own-pt/glosstag

PorGram I

Hosted in https://github.com/LR-POR/PorGram

Fixed the ACE config to compile it.

Convert the text files into profiles and finished a first treebank called 'core'.

We started to consider next texts. 'The Speckled Band' from Arthur Conan Doyle translation here.

Dan added some initial punctuation support.

PorGram II

PorGram tries to reuse as much as possible MorphoBr (full forms morphological dictionary) and UD Portuguese Corpora.

How to handle morphology? Consider the verb 'pagar' (pay, infinitive) in the past imperfective form 'pagávamos'. The stressed syllable changed. The **HPSG morphology** can't handle it.

PorGram III

Options:

- External tool via YY format? Freeling? Foma?
- External tool via API
- Explore the `irregs.tab` (the exception list)

MorphoBr has 11.181.124 pairs of entries summing up 500MB.
Compiled into `morphobr.bin` which has 1.2MB.

PorGram IV

Next:

- test the `irregs.tab` idea
- more experiments on using corpora to extract valence information of verbs

Glosstag I

Princeton Wordnet 3.0 glosses (definitions and examples) annotated with senses. Unfinished work from Princeton team.

In 2019, we implemened a Emacs mode for annotation. Reusing data prepatation (tokenization, PoS tag and lemmatization). Goal is to finish the annotation (206K tokens, 31%). Paper at GWC 2019.

GWC 2023 paper about Glosstag 2022 release. ERG grammar for processing the sentences. Initial evaluation of ERG results. UKB for automatically complete the annotation.

How to map the surface MWE annotations (tokens) with the ERG predicates (MRS)? How to combine with Dan's work with ERG lexicon expansion from Wordnet forms?

MRS Semantics I

At IBM, we have some few opportunities for ‘deep’ linguistic processing!

Exploring applications for Logical reasoning from formulas obtained from sentences: questions from a risk assessement questionnaire.

Ex: “What is the water solubility of the compound?”

MRS Semantics II

We implemented a Python library for simple type language and KB integration. Called ULKB. Thanks Guilherme Lima (IBM).

We implemented a Python library for MRS transformation into FOL formulas. Uses: Utool (scope resolution), PyDelphin (MRS parsing).

How to test the library? SICK dataset of text entailment. Challenges for logical text entailments. Does LLM do better?

What is next? Previous works on "Dependent Type Semantics" (ncatlab). No work from MRS to other logical forms besides FOL. Coq and Lean provers.