Thematic fit evaluation: an aspect of selectional preferences

Asad Sayeed Clayton Greenberg Vera Demberg
Saarland University

Motivations and goals

**Thematic fit**: another way to see selectional preferences – how a processor reacts to a candidate filler given a verb and role.

(“cut”, instrument, “finger”) → 1 out of 7

(“cut”, instrument, “knife”) → 6 out of 7

A window into human semantic processing.

Applications: psycholinguistic modeling, dialogue systems, testing of coreference with previous discourse items

Future evaluation aims

More balanced datasets
- Need more judgement data designed to contrast particular semantic features.

Compositionality
- Existing evaluation datasets: verb-role-noun triplets – need new datasets with other slots filled.
- e.g. if agent is “chef”, then $\theta$-Fit(“whip”, patient, “cream”) > $\theta$-Fit(“whip”, patient, “horse”)

Perceptuomotor knowledge
- Existing thematic fit models – distributional.
- Can semantic knowledge fully be captured by distributional stats?
- Rating scheme to distinguish different semantic “knowledges”?

Standard evaluation procedure

Areas for improvement:

- thematic fit scoring process
- evaluation objective

Visualization from Roleo (Sayeed et al., 2016 ACL demos)

$\theta$-Fit(“cut”, patient, “paper”): excerpts of plots of top candidate vectors with “paper” vector.

Existing datasets and correlations

**Descriptions of datasets**

- **MSTNN**: 1444 agents & patients, items drawn from many experiments (see table on right)
- **P07**: 414 agents & patients, systematic selection frequent predicates and arguments
- **F-Loc**: 274 locations, **F-Inst**: 248 instruments, these roles tend to be less stable
- **GDS-mono**: 240 patients with monosemous verbs, verbs matched for freq. with GDS-poly
- **GDS-poly**: 240 patients with polysemous verbs. Each noun had a freq. and infreq. variant

**Dataset**  | **Centroid Greenberg-Clusters SDS2015-average SDS2015-swap**
--- | ---
**P07** | 59 | 55 | 59 | 48
**MSTNN** | 34 | 38 | 34 | 25
**F-Loc** | 23 | 29 | 21 | 19
**F-Inst** | 36 | 42 | 39 | 45
**GDS-mono** | 66 | 68 | - | -
**GDS-poly** | 43 | 47 | - | -

- Embeddings seem to perform more poorly when used on their own (Baroni et al., 2014).
- Greenberg clusters systematically improve correlations on datasets with wide freq. range.
- Spaces built on semantic vs. syntactic links capture complementary information.

Spearman’s $\rho \times 100$ correlation with human judgements

{asayeed,claytong,vera}@coli.uni-saarland.de

RepEval 2016 – Berlin

12 August 2016