Introduction

- Hindi (and Urdu) has a complex inventory of case markers and postpositions for marking semantic relations between words
  - Case markers: small set of markers that indicate core arguments to a verb, other basic relations
  - Focus markers: discourse relations, emphasis
  - Postpositions: large class of markers for narrow semantic relations, verbal adjuncts
- Understanding these relations is a difficult task for NLP
- Potential upstream benefits: semantic role labelling, translation

We created a Hindi corpus annotated with coarse semantic labels from the SNACS formalism and attempted automatic labelling with language models!

Automatic tagging

- We extended the lexical semantic recognition (Liu et al., 2021) task to Hindi: automatic tagging of coarse supersenses on case/adpositions
- Data processing
  - Convert labels to BIO-tagging scheme
  - Split into 80/10/10 train/dev/test set, check performance improvements on dev to stop training
- Architecture: contextual language model → biLSTM → CRF to output tags
  - We tested various BERT-like masked language models for Hindi, with a 2-layer biLSTM with dropout of 0.3, then to a CRF
  - Hyperparameters: [30, 60] epochs, [0.0001, 0.0002, 0.0005, 0.001] learning rate, [64, 128, 256, 512] LSTM layer size
  - Found to be better to use biLSTM+CRF than just biLSTM or Transformers
- Results: IndicTransformers BERT is the best, with comparable numbers to past work on English (two on left below)

Corpus analysis

- Breakdown of target and annotation types shown above
- Agreement: Cohen’s $\kappa$ for doubly-annotated targets was 0.78 on scene roles, 0.85 on functions, and 0.73 on construals (both together)
  - Apparently syntax is easier to categorise than semantics
- Semantic diversity: We estimated the entropy of the distribution over scene roles for each token type in the corpus. Found that case markers have very high entropy: highly semantically diverse.

SNACS

- SNACS is the Semantic Network of Adposition and Case Supersenses, already applied on English (L1 and L2), Korean, Mandarin, and German (Schneider et al., 2018a, 2020)
- Related to linguistic theories of argument structure and theta roles: Agent, Theme, Recipient, Causer, ...
- Construal system: attempts to separate syntax from semantics
  - Experiencer→Agent: a subject marked with the ergative (strongly agentive) with an experiencer predicate (e.g. feel)
  - Scene Role→Function
- Linguistic issues in annotating Hindi
  - Syntactic function of some case markers is hard to label
  - Non-nominative/ergative subjects
  - Causative constructions: are animate Instruments a thing?
  - Empathic particles

References