Benchmarking the GF translator

Prasanth Kolachina

LTRC, IIIT-Hyderabad
India

August 21^{th}, 2013

What I (we) have seen!

- How to write GF grammars in a functional programming style;
  - Resource grammars
  - Application grammars
- Distinction between Resource and Application grammars
  - Semantics vs Syntax!
- Good practises of GF grammar writing
- GF-parser and Data-driven parsing
- Web-applications/Tools from GF to support development of multilingual grammars needed for MT

Motivation

- Translation from Hindi to English
  - Cooking recipes?
- Translation from English to Hindi
  - Spatial reasoning CNL useful to humans?
- Translation across languages from different families

Motivation

- Translation across languages from different families
  - Translation between any languages is theoretically possible.
  - In practise;
    - $Difficulty_{Molto}(Lang_{src}, Lang_{tgt}) \approx \frac{1}{\sum_{l \in L_{Molto}} Maximum(Speakers(Lang_{src}\&l), Speakers(Lang_{tgt}\&l))}$
  - Translation between Swe-Eng, Fre-Eng ... YES!
  - Translation between Zulu-Eng? Greek-Eng? Hin-Eng? ... COULD BE
  - Translation between Tamil-Mandarin, or esoteric languages? ...

Motivation

- Translation using gate-keepers for each family
  - Only possible in the case of loss-less MT
  - Lang$_{family_1}$ $\rightarrow$ Gate-Keeper$_{family_1}$ $\rightarrow$ Gate-keeper$_{family_2}$ $\rightarrow$ Lang$_{family_2}$
  - Reduces the dependency on cross-lingual experts (rare in the case of esoteric languages)
  - Similar(?) to Pivot-based translation in SMT

What I would like to do?

- Gate-Keeper for Indian languages?
- Resource grammars for English and Hindi
  - Existing grammar in GF sufficient?
  - Superior lexicon exists already
  - Concrete syntax for Hindi in the resource grammar?
- Application grammars for different domains

Proposal

- Evaluation of existing GF robust parser for translation from English→Hindi
  - Are the resulting translations fluent?
  - Is the word-order reflective of actual Hindi syntax?
- Proper realization of semantic concepts
- Adequacy of the existing lexicon with respect to domain-specific terminology
- Agreement constraints to be satisfied
  - Inherent capability of the GF framework!

Parallel corpora for English-Hindi

- Multiple domain corpora in small sizes
  - Health, Tourism, Legal, Social sciences
- 500-700 sentences in each domain
- Multiple application grammars with one underlying resource grammar!

Expectations!

- "Holes" in the GF resource grammar with respect to cross-linguistic differences in syntax between English and Hindi
- Need for "English-independent" syntactic information in the Hindi core-grammar
- Realization of semantic concepts with frequent items in the vocabulary
  - Adds to the fluency!