COLING 2020 Tutorial

Cross-lingual Semantic Representation for NLP with UCCA: Cross-linguistic Studies

Omri Abend

Cross-linguistic Portability

- A scheme is portable if the same categories/guidelines can be consistently applied across languages
- Advantages: uniformity, allows cross-linguistic comparison, transfer learning across languages, multi-lingual parsing
- Typological theory is a key element for achieving this
  - Universal Dependencies is one recent example of the strength of such an approach

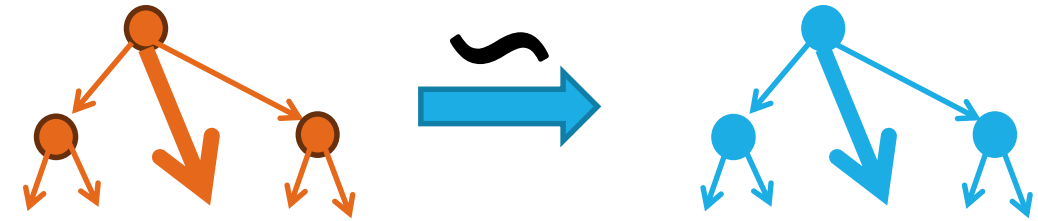
Cross-linguistic Portability

- In practice, this is a difficult criterion to meet, even for the most coarse-grained categories
- For example: What defines a verb?
  - Inflects for tense, person and number? (but what about when they don't?)
  - Refers to actions or movement? (but what about all the verbs that don't?)
  - …
- Semantics seeks to uncover a more uniform layer of representation, and can thus be used for cross-linguistically applicable definitions
  - Several schemes (e.g., UD and Unimorph) leverage this insight

Cross-linguistic Stability

- For semantic representation, we can aspire to something more than portability

- Stability: a representation scheme is (cross-linguistically) *stable* if the annotation of the text is preserved in literal translation



Cross-linguistic Stability: From Basic Linguistic Theory to UCCA

- Possession in BLT is defined based on semantic criteria (ownership, part/whole, kinship)
- Realization is discussed separately (e.g., NP-internally, as a verb, morphologically)

- This allows a uniform treatment of phenomena like:

  English: **I have** a pig

  Jakaltek: ay no' hin txitam (exist [CL:ANIMAL my pig];
  lit. ***my*** *pig* ***exists***)

  Hebrew: yesh li xazir (exist to-me pig; lit. ***there is*** *a pig* ***for me***)

Cross-linguistic Stability: From Basic Linguistic Theory to UCCA

- In UCCA, this translates to stability:

  English: **I**$_A$ **have**$_S$ [a pig]$_A$

  Jakaltek: ay$_S$ no' hin$_A$ txitam$_A$

  (exist [CL:ANIMAL my pig]; lit. ***my*** *pig* ***exists***)

  Hebrew: yesh$_S$ li$_A$ xazir$_A$

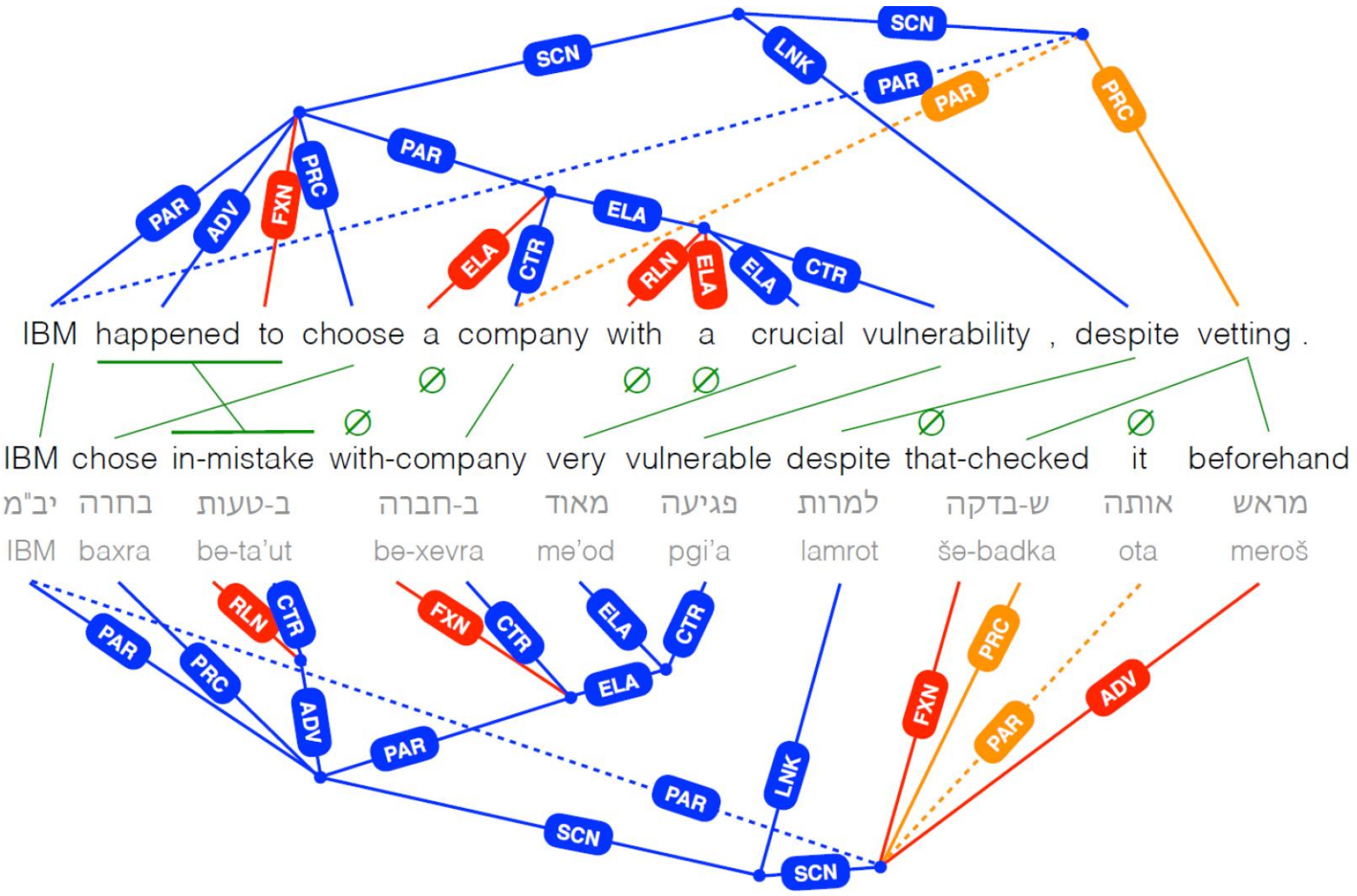
  (exist to-me pig; lit. ***there is*** *a pig* ***for me***)

The Appeal of Cross-linguistic Stability

- Invariance to meaning-preserving variation – one of the goals of meaning representations
- Potentially useful for applications:
  - Machine translation (an invariant over translations)
  - Multi-lingual parsing (e.g., zero-shot learning)
  - Annotation projection

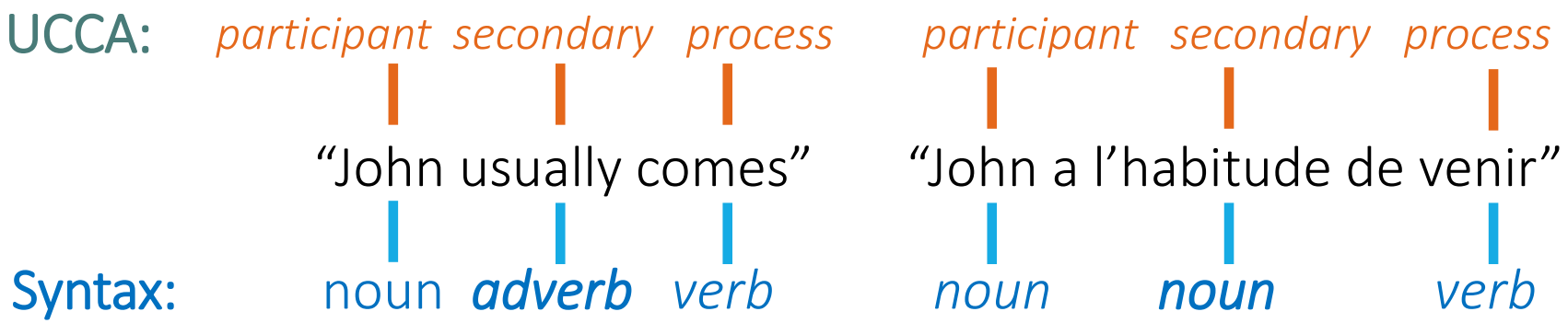
- Full stability is a holy grail, but
  - Instead, stability up to some pre-specified variation
  - Increasing stability is feasible!

Empirically Assessing Translation Divergences



A Corpus Study of Translation Stability

- How well does UCCA *preserve* structure across English-French translations? (Sulem et al., 2015)



Cross-linguistic Stability: Results

- Scene divergences:

  → 92% of the English Scenes and 95% of the French Scenes have a correspondent on the other side

- Comparison to syntax: stability of the number of Scenes/paragraph vs. the number of clauses/paragraph

  → UCCA is more stable than PTB-style syntactic trees

UCCA: Cross-linguistic Stability

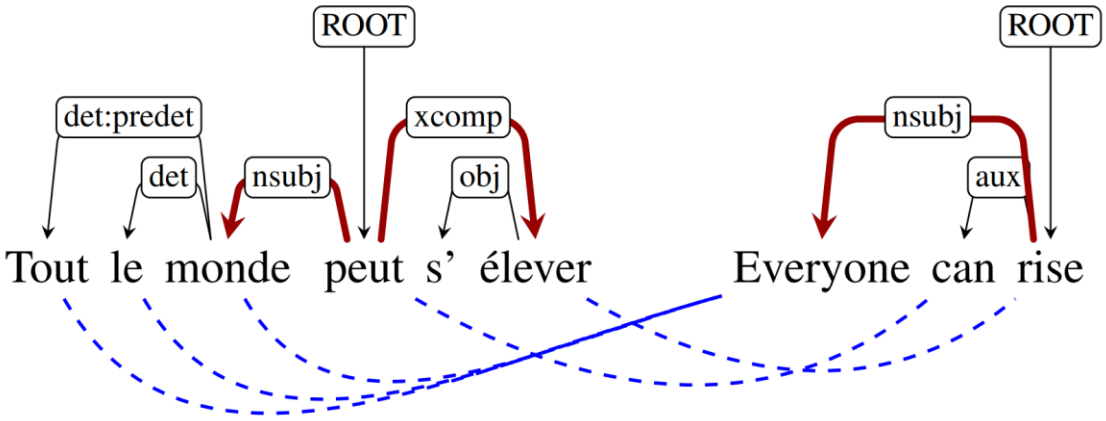
Nevertheless, occasional UCCA divergences are found:

Officers were probing the increasing gloom with their night glasses

Les officiers **armés$_S$ de leur lorgnette de nuit** fouillaient l'obscurité croissante

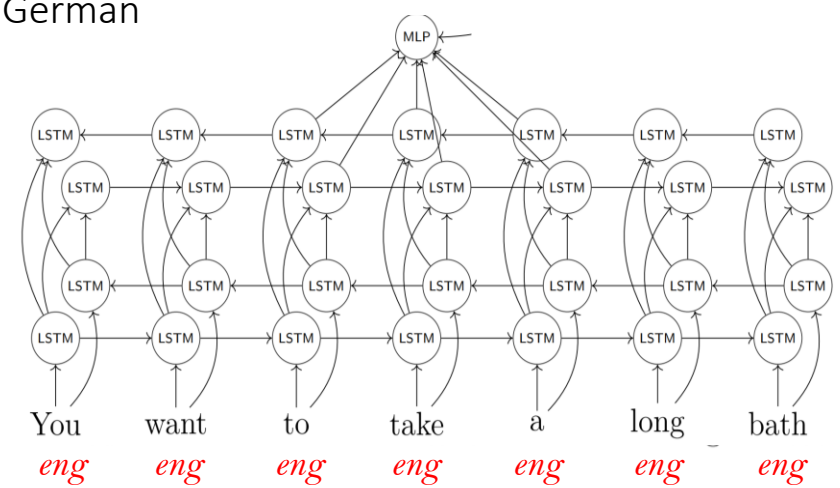
Empirically Assessing Translation Divergences

- However, we have yet to conduct a more comprehensive comparison of UCCA structures across languages
- We did, however, assess divergences with Universal Dependencies (Nikolaev et al., 2020)
  - We found that some recurrent divergences with UD are bridged in UCCA.



SemEval 2019 Shared Task on UCCA Parsing

- SemEval 2019 hosted a shared task on UCCA parsing
  - English, German and French (15 training sentences)
- In all three languages, notable improvement over TUPA
  - In French, results of 75-80 F-score (vs. 50 with supervised TUPA)
  - All systems used projection/transfer techniques
  - For instance, merging the training sets for the three languages + language embeddings
    - Helped in parsing French and German



Intermediate Summary

- UCCA adopts some of the distinctions in BLT to a semantic annotation scheme
  - Abstracts away from much syntactic variation
  - Applicable to a number of languages
- The field should revisit the issue of translation divergences
  - Advance our understanding on one of the core questions of language
  - Much applicative potential: multi-lingual transfer, projection, structure-aware machine translation
- UCCA parsing can benefit from cross-lingual transfer
  - Current research explores the relation between stability and transferability of annotation schemes