Approaches

Current Arabic Parsers:
- Segmentation/Tagging Oracle
- A Factored Model
- A Joint Model
- An Integrated Model
- A Generative Model

Our Model

The Lattice
- A sequence of surface words
- A morphological analyzer maps sentences to lattices
- Each token is mapped to a lattice representing its morphological analyses.
- Our lattice is a concatenation of the different word-frames.
- All segmentation possibilities are represented as lattice paths.
- Each arc in the lattice corresponds to a tagged segment.
- We assume a lexeme-based lexicon consisting of tagged lexemes.
- We assume all lattice paths are a-priori equally likely.

The Grammar
- A probabilistic lexeme-based context-free grammar read off of the Modern Hebrew Treebank (Simaan et al., 2001).
- Three types of rules:
  - Syntactic rules: $S \rightarrow NP \ VP$
  - Pre-Terminal rules: $NP \rightarrow$ non-terminal
  - Lexical rules: $VP \rightarrow$ (fd, Verb)

This work:
- A Joint Model

The Naive Solution:
- Pipeline
- The Input: A sequence of surface words
- A morphological analyzer maps sentences to lattices
- Each token is mapped to a lattice representing its morphological analyses.
- Our lattice is a concatenation of the different word-frames.
- All segmentation possibilities are represented as lattice paths.
- Each arc in the lattice corresponds to a tagged segment.
- We assume a lexeme-based lexicon consisting of tagged lexemes.
- We assume all lattice paths are a-priori equally likely.

The Main Point
- When modelling the different lexeme probabilities, we do not treat inter-token lexeme sequences as complex tags, and do not take linear context into account.
- Instead, the different lexemes are generated independently based on their corresponding POS tags.
- The context is modeled via the PCFG subdistribution resulting in the different lexemes.
- For example, we model the probability of the event $f_a g$ resulting in the morpheme sequence $fREL \rightarrow (fREL) x M_1 J_3 J_3$

References

- Cohen and Smith 2007

To Sum Up
- Better grammars yield better results on all tasks (in line with Cohen and Smith, 2007).
- We tested our system with increasingly complex grammars.
- All lexically pruned models outperform S&C non-oracle results.
- Our best lexically-pruned model outperforms S&C oracle results.
- We propose a single, clean generative model that outperforms previous models on the joint task.

Acknowledgments

We thank Meni Adler and Michael Elhadad,⊊FRegressor,隍Khe latin KPC(G), for feedback and discussion.

The work of R. Tsarfaty and Y. Goldberg was supported by the Lyn and William Frankel Foundation for Computer Science, and The Netherlands Organization for Scientific Research (NWO) VENI grant 639.63.011.

Graphs designed courtesy of Yoav Goldberg.