Measuring the morpho-syntactic ambiguity using the edit distance functions

SAMI 2021

Peter Bednár

Motivation (1)

- Dependency parsing is an important task in NLP:
  - Information extraction
  - Question answering and dialog systems
  - Automatic translation
- Dependency parser must perform:
  - Tokenization/sentence segmentation
  - Lemmatization
  - Morphological tagging
  - Dependency parsing

Motivation (2)

1. We would like to measure how much are the sequences of morphological features ambiguous with regards of the dependency parsing
2. We would like to estimate how much information for parsing can be gained solely from the order of tokens and their morphological categories

Edit distances

1. **Sequence edit distance**
   - Takes into the consideration linear order of the tokens and their morphological categories
2. **Tree edit distance**
   - Takes into the consideration dependency relations between the tokens/words

Sequence edit distance

- Extension of *Levenshtein* distance
- Counts number of operations required to transform one sentence sequences to another one
- Operations:
  - Insertion of token
  - Deletion of token
  - Substitution of single token/toke property

Tree edit distance

- Counts number of operations required to transform one sentence dependency tree to another one
- Operations:
  - Insertion of token
  - Deletion of token
  - Substitution of single token/toke property

Experiments and Results (1)

- Data from Universal Dependencies corpora
- Primary focus on Slovak language

- **Experiment 1**
  - How good can be Tree edit distance approximated by the Sequence edit distance
  - Difference is 0.043 average per sentence (i.e., under 5%)
  - Quartiles for nearest neighbors – 0, 2, 25
  - Exact match 35 % of sentences

Experiments and Results (2)

- **Experiment 2**
  - Focus on morphological ambiguities
  - Compare distances based solely on the morphological features with the dependency structure
  - Difference is 0.8 average per sentence
  - Quartiles 0.12, 1.22 and 3.41
  - Exact match 46% of sentences

Experiments and Results (3)

- **Experiment 3**
  - Measure language similarities (Slavic + English as reference)

Language	# of tokens	Dist./ Sent.	Dist./ token	Norm. dist.
Belarusian	6 383	5.6272	0.5925	0.4642
Bulgarian	124 336	4.3949	0.4627	0.3293
Czech	1 173 282	3.6833	0.3878	0.2679
English	204 585	4.8498	0.5106	0.3637
Croatian	152 857	4.6685	0.4916	0.3583
Upper Sorbian	460	6.8975	0.7262	0.5864
Old Russian	118 630	5.4960	0.5786	0.4305
Polish	281 736	4.511	0.4749	0.3472
Russian	870 479	3.9593	0.4168	0.2956
Slovenian	112 530	4.3234	0.4552	0.3284
Serbian	74 259	5.0665	0.5334	0.4010
Ukrainian	92 401	4.4874	0.4724	0.3387

Conclusion and Discussion

- Proposed Sequence and Tree edit distances for morpho-syntactic structures
- Good approximation of Tree edit distance with Sequence edit distance
- Low ambiguity of morphological features
- In the future, can be used to:
  - Semi-supervised learning with alignment of unlabeled data
  - Transfer learning from one language to another