Multilingual Parsing from Raw Text to Universal Dependencies

Joakim Nivre

Uppsala University
Linguistics and Philology
Overall Plan

1. Basic notions of dependency grammar and dependency parsing
2. Graph-based and transition-based dependency parsing
3. Advanced graph-based parsing techniques
4. Advanced transition-based parsing techniques
5. Neural network techniques in dependency parsing
6. Multilingual parsing from raw text to universal dependencies
Parsing vs. Grammar

- Dependency parsing models are linguistically atheoretical
  - No assumptions about head-dependent criteria (structure)
  - No assumptions about linguistic categories (labels)
  - May be limited to some formal class of dependency trees

- Is this good or bad?
  - Practically useful when dealing with heterogeneous data sets
  - Prevents full exploitation of linguistic annotation

- Linguistics for parsers: $h \xrightarrow{?} d$
Universal Dependencies

- Dependency annotation is consistent across languages
  - Allows meaningful comparisons across languages
  - Facilitates cross-lingual approaches to parsing
- Dependency annotation is based on a linguistic theory
  - Grammatical relations between content words
  - Function words specify content words
  - Special relations for coordination, MWEs, etc.
- Parsers can learn to handle more than one relation
Universal Dependencies

The dog was chased by the cat

Hunden jagades av katten

Pes byl honěn kočkou
Dependency Parsing Shared Tasks

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
Dependency Parsing Shared Tasks

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
Dependency Parsing Shared Tasks

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
Dependency Parsing Shared Tasks

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
- SPMRL 2013 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- SPMRL 2014 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
Dependency Parsing Shared Tasks

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
- SPMRL 2013 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- SPMRL 2014 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- VarDial 2017 (cross-lingual: cs-sk, sl-hr, da/sv-no)
Dependency Parsing Shared Tasks

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
- SPMRL 2013 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- SPMRL 2014 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- VarDial 2017 (cross-lingual: cs-sk, sl-hr, da/sv-no)
- CoNLL 2017 (45 languages + surprise + end-to-end parsing)
What is new?

- Cross-linguistically consistent annotation
  - Facilitates comparison across languages
  - Enables cross-lingual and universal approaches
  - Surprise languages at test time
- From raw text to universal dependencies
  - No gold segmentation or annotation
Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free

- New parallel test set (DFKI, Google and others):
  - 15–20 languages
Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
- Arabic NYUAD: not available free of charge
Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
- Arabic NYUAD: not available free of charge
- At least 10K test words ⇒
  - Exclude: Belarusian, Coptic, Lithuanian, Sanskrit, Tamil
  - Include but small training: French ParTUT, Galician TreeGal, Irish, Kazakh, Latin, Slovenian SST, Ukrainian, Uyghur

A few surprise languages at test time:
- New: Buryat, Kurmanji, North S´ami, Upper Sorbian

New parallel test set (DFKI, Google and others):
- 15–20 languages
Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
- Arabic NYUAD: not available free of charge
- At least 10K test words ⇒
  - Exclude: Belarusian, Coptic, Lithuanian, Sanskrit, Tamil
  - Include but small training: French ParTUT, Galician TreeGal, Irish, Kazakh, Latin, Slovenian SST, Ukrainian, Uyghur
- A few surprise languages at test time:
Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
- Arabic NYUAD: not available free of charge
- At least 10K test words ⇒
  - Exclude: Belarusian, Coptic, Lithuanian, Sanskrit, Tamil
  - Include but small training: French ParTUT, Galician TreeGal, Irish, Kazakh, Latin, Slovenian SST, Ukrainian, Uyghur
- A few surprise languages at test time:
  - New: Buryat, Kurmanji, North Sámi, Upper Sorbian
Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free

- Arabic NYUAD: not available free of charge

- At least 10K test words ⇒
  - Exclude: Belarusian, Coptic, Lithuanian, Sanskrit, Tamil
  - Include but small training: French ParTUT, Galician TreeGal, Irish, Kazakh, Latin, Slovenian SST, Ukrainian, Uyghur

- A few surprise languages at test time:
  - New: Buryat, Kurmanji, North Sámi, Upper Sorbian

- New parallel test set (DFKI, Google and others):
  - 15–20 languages
Additional Data

- Just one “closed” track
- Registered participants were asked for suggestions
- CommonCrawl + word embeddings
- Word Atlas of Language Structures (WALS)
- Wikipedia Dumps
  - Wikipedia word vectors (90 languages) by Facebook
- Opus Parallel Corpora
- WMT 2016 Parallel + Monolingual Data
- Apertium + Giellatekno Morphological Analyzers
- French Treebank UD v2 conversion
Multi-Language and Multi-Domain

- English language
  - **UD English** (*Web Treebank*): blog, social, reviews
    - 205K train, 25K dev, 25K test
  - **UD English LinES**: fiction, nonfiction (sw localization), spoken
    - 50K train, 17K dev, 16K test
  - **UD English ParTUT**: legal, news, wiki
    - 26K train, 12K dev, 12K test
  - **UD English DGPT**: nonfiction/legal (EuroParl), news, wiki
    - roughly 20K test only!

- You can train one model for all if you want
- But they are different domains!

- Main system score:
  - Macro-average LAS across all test sets (not languages)
End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data
End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data
- Sentence segmentation
End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data
- Sentence segmentation
- Tokenization
- Word segmentation (multi-word tokens)
End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data
- Sentence segmentation
- Tokenization
- Word segmentation (multi-word tokens)
- Morphological analysis
  - If your parser needs it
  - Exception: predicted morphology for surprise languages
End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data
- Sentence segmentation
- Tokenization
- Word segmentation (multi-word tokens)
- Morphological analysis
  - If your parser needs it
  - Exception: predicted morphology for surprise languages
- Parsing
Baseline Models

- UDPipe (ÚFAL): trained segmenter, tagger+lemmatizer, parser
- Pre-processed test data (except syntax) directly available
- Just use that if you don’t have anything better

- SyntaxNet / ParseySaurus (Google)

- No interest in surprise languages?
- Use simple delexicalized parser
Evaluation Metrics

- Align system-output tokens to gold tokens

Al-Zaman: American forces killed Shaikh Abdullah al-Ani, the preacher at the mosque in the town of Qaim, near the Syrian border.

GOLD: Al-Zaman: American forces killed Shaikh
OFFSET: 0-1 2 3-7 9 11-18 20-25 27-32 34-39

- All characters except for whitespace match => easy align!

SYSTEM: Al-Zaman: American forces killed Shaikh
OFFSET: 0-7 9 11-18 20-25 27-32 34-39
Evaluation Metrics

- Align system-output tokens to gold tokens

*Die Kosten sind definitiv auch im Rahmen.*

**GOLD:** Die Kosten sind definitiv auch im Rahmen .

**SPLIT:** Die Kosten sind definitiv auch in dem Rahmen .

**OFFSET:** 0-2 4-9 11-14 16-24 26-29 31-32 34-39 40

- Corresponding but not identical spans?

- Find longest common subsequence

**SYSTEM:** Kosten sind definitiv auch im Rahmen .

**SPLIT:** Kosten sind de finitiv auch im Rahmen .

**OFFSET:** 4-9 11-14 16-24 26-29 31-32 34-39 40
Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  - $head_{System}(i) = head_{Gold}(i)$
  - (Comparing just integers here.)
Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  - $head_{System}(i) = head_{Gold}(i)$
  - (Comparing just integers here.)
- Compare aligned nodes, if alignment is found
  - $node : Integer \rightarrow Node$
  - $align : SystemNode \rightarrow GoldNode$
  - $align(head_{System}(node_i)) = head_{Gold}(align(node_i))$
  - (Comparing node objects.)
Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  - \( \text{head}_{\text{System}}(i) = \text{head}_{\text{Gold}}(i) \)
  - (Comparing just integers here.)
- Compare aligned nodes, if alignment is found
  - \( \text{node} : \text{Integer} \rightarrow \text{Node} \)
  - \( \text{align} : \text{SystemNode} \rightarrow \text{GoldNode} \)
  - \( \text{align}(\text{head}_{\text{System}}(\text{node}_i)) = \text{head}_{\text{Gold}}(\text{align}(\text{node}_i)) \)
  - (Comparing node objects.)
- Cannot align? No point for attachment!
Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  
  \[ \text{head}_{\text{System}}(i) = \text{head}_{\text{Gold}}(i) \]
  
  (Comparing just integers here.)
- Compare aligned nodes, if alignment is found
  
  \[ \text{node} : \text{Integer} \rightarrow \text{Node} \]
  
  \[ \text{align} : \text{SystemNode} \rightarrow \text{GoldNode} \]
  
  \[ \text{align} (\text{head}_{\text{System}}(\text{node}_i)) = \text{head}_{\text{Gold}}(\text{align}(\text{node}_i)) \]
  
  (Comparing node objects.)
- Cannot align? No point for attachment!
- Wrong sentence boundary?
  
  - one or more wrong relations
Labeled Attachment Score

- Correct relation...alignment of parent equals parent of alignment, and the universal prefix of dependency relation types match on both sides.

- Precision: \( P = \frac{\text{#correctRelations}}{\text{#systemNodes}} \)

- Recall: \( R = \frac{\text{#correctRelations}}{\text{#goldNodes}} \)

- LAF (labeled attachment \( F_1 \)-score): \( LAF = \frac{2PR}{P+R} \)
UD-specific Weighted Metric (Experimental)

- Relations between content words are more important cross-linguistically
- Attachment of function word = morphology in other languages
- Weighted scoring of correct relations:
  - **Weight = 1** for root, nsubj, obj, iobj, csubj, ccomp, xcomp, obl, vocative, expl, dislocated, advcl, advmod, discourse, nmod, appos, nummod, acl, amod, conj, fixed, flat, compound, list, parataxis, orphan, goeswith, reparandum, dep
  - **Weight = 0** for aux, case, cc, clf, cop, det, mark
  - **Weight = 0** for punct
More information at:

http://universaldependencies.org/conll17/