Universal Dependency Parsing

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
The Grand Vision
The Grand Vision

The problem

• Why 90% parsing accuracy for English but only 80% for Finnish?
• Are some languages intrinsically harder to parse?
• Not just morphological richness – many typological parameters
The Grand Vision

The problem

• Why 90% parsing accuracy for English but only 80% for Finnish?
• Are some languages intrinsically harder to parse?
• Not just morphological richness – many typological parameters

An ideal solution

• A universal parser for all languages
• Linguistic universals are hard-coded
• Typological parameters are learned from data
The Grand Vision

The problem

• Why 90% parsing accuracy for English but only 80% for Finnish?
• Are some languages intrinsically harder to parse?
• Not just morphological richness – many typological parameters

An ideal solution

• A universal parser for all languages
• Linguistic universals are hard-coded
• Typological parameters are learned from data
Pieces of the Puzzle

1. Morphosyntactic disambiguation
2. Universal dependency annotation
3. Parsing with universal dependencies
Part I

Morphosyntactic Disambiguation

Joint work with Bernd Bohnet, Igor Boguslavsky, Richárd Farkas, Filip Ginter and Jan Hajič
Background

Parsing accuracy for morphologically rich languages tends to be lower than for languages like English (Nivre et al., 2007)

Hypothesized explanations (Tsarfaty et al. 2010, 2013):

- Strict separation morphology-syntax
- Data sparsity due to high type-token ratio

Suggested remedies

- Joint morphological and syntactic analysis (Lee et al., 2011)
- Lexical resources (Hajič, 2000; Goldberg and Elhadad, 2013)
This Study

Parsing techniques:
• Transition-based model for joint morphological and syntactic analysis
• Lexical resources integrated as hard or soft constraints

Evaluation on five morphologically rich languages:
• Czech, Finnish, German, Hungarian, Russian
• New state of the art in dependency parsing for all languages

Representations
Representations

For each word of a sentence $w_1, \ldots, w_n$: 

Ein Haus hat er in Ulm gebaut.
Representations

For each word of a sentence $w_1, \ldots, w_n$:

- A part-of-speech tag $p \in P$
For each word of a sentence $w_1, \ldots, w_n$:

- A part-of-speech tag $p \in P$
- A morphological feature bundle $m \in M$

Ein Haus hat er in Ulm gebaut.
Representations

For each word of a sentence $w_1, \ldots, w_n$:

- A part-of-speech tag $p \in P$
- A morphological feature bundle $m \in M$
- A lemma $l \in Z^*$
For each word of a sentence $w_1, \ldots, w_n$:

- A part-of-speech tag $p \in P$
- A morphological feature bundle $m \in M$
- A lemma $l \in Z^*$
- A syntactic head $h \in \{0, w_i, \ldots, w_n\}$
Representations

For each word of a sentence $w_1, \ldots, w_n$:

- A part-of-speech tag $p \in P$
- A morphological feature bundle $m \in M$
- A lemma $l \in Z^*$
- A syntactic head $h \in \{0, w_i, \ldots, w_n\}$
- A dependency label $d \in D$
Parsing Framework

Transition system (Bohnet and Nivre, 2012):
- Arc-standard system with online reordering (Nivre, 2009)
- Select morphology when shifting words to the stack

Beam search and structured learning (Zhang and Clark, 2008)

Preprocessing (at learning and parsing time)
- Tagger assigns $k$ best tags and feature bundles
- Parser can only select analyses licensed by preprocessor
Experiment 1

**Pipeline**

Preprocessor assigns a single tag and feature bundle per word
Beam = 40 distinct trees

**Joint**

Preprocessor assigns up to 2 tags and feature bundles per word
Beam = 40 distinct trees + 8 tag variants + 8 feature variants

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>P</th>
<th>M</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs PDT (Hajič et al., 2001)</td>
<td>1249K</td>
<td>88K</td>
<td>70K</td>
<td>12</td>
<td>1851</td>
<td>49</td>
</tr>
<tr>
<td>de Tiger (Brants et al., 2002)</td>
<td>648K</td>
<td>32K</td>
<td>32K</td>
<td>54</td>
<td>257</td>
<td>43</td>
</tr>
<tr>
<td>fi TDT (Haverinen et al., 2013)</td>
<td>183K</td>
<td></td>
<td>21K</td>
<td>12</td>
<td>1917</td>
<td>47</td>
</tr>
<tr>
<td>hu Szeged (Farkas et al., 2012)</td>
<td>1101K</td>
<td>210K</td>
<td>171K</td>
<td>22</td>
<td>1105</td>
<td>33</td>
</tr>
<tr>
<td>ru SynTagRus (Boguslavsky et al., 2000)</td>
<td>575K</td>
<td>73K</td>
<td>72K</td>
<td>14</td>
<td>454</td>
<td>78</td>
</tr>
</tbody>
</table>
Morphological Accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>93.0</td>
<td>93.7</td>
</tr>
<tr>
<td>de</td>
<td>89.1</td>
<td>90.0</td>
</tr>
<tr>
<td>fi</td>
<td>88.8</td>
<td>89.1</td>
</tr>
<tr>
<td>hu</td>
<td>96.1</td>
<td>96.2</td>
</tr>
<tr>
<td>ru</td>
<td>92.6</td>
<td>92.8</td>
</tr>
</tbody>
</table>
Morphological Accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>93.0</td>
<td>93.7</td>
</tr>
<tr>
<td>de</td>
<td>89.1</td>
<td>90.0</td>
</tr>
<tr>
<td>fi</td>
<td>88.8</td>
<td>89.1</td>
</tr>
<tr>
<td>hu</td>
<td>96.1</td>
<td>96.2</td>
</tr>
<tr>
<td>ru</td>
<td>92.6</td>
<td>92.8</td>
</tr>
</tbody>
</table>
Syntactic Accuracy

Pipeline
Joint

<table>
<thead>
<tr>
<th>Language</th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>83.1</td>
<td>83.3</td>
</tr>
<tr>
<td>de</td>
<td>91.8</td>
<td>92.0</td>
</tr>
<tr>
<td>fi</td>
<td>79.9</td>
<td>80.6</td>
</tr>
<tr>
<td>hu</td>
<td>88.4</td>
<td>88.9</td>
</tr>
<tr>
<td>ru</td>
<td>87.4</td>
<td>87.6</td>
</tr>
</tbody>
</table>
Syntactic Accuracy

<table>
<thead>
<tr>
<th>Language</th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>83.1</td>
<td>83.3</td>
</tr>
<tr>
<td>de</td>
<td>91.8</td>
<td>92.0</td>
</tr>
<tr>
<td>fi</td>
<td>79.9</td>
<td>80.6</td>
</tr>
<tr>
<td>hu</td>
<td>88.4</td>
<td>88.9</td>
</tr>
<tr>
<td>ru</td>
<td>87.4</td>
<td>87.6</td>
</tr>
</tbody>
</table>
Experiment 2

Morphological lexicons

- Lookup of tag, feature bundle and lemma
- Added as hard or soft constraints to preprocessor and parser
- Lemma selected deterministically from form + tag + feature bundle

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Language</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>PDT</td>
<td>(Hajič and Hladká, 1998)</td>
</tr>
<tr>
<td>de</td>
<td>SMOR</td>
<td>(Schmid et al., 2004)</td>
</tr>
<tr>
<td>fi</td>
<td>OMorFi</td>
<td>(Pirinen, 2011)</td>
</tr>
<tr>
<td>hu</td>
<td>morphdb.hu</td>
<td>(Trón et al., 2006)</td>
</tr>
<tr>
<td>ru</td>
<td>ETAP-3</td>
<td>(Apresian et al., 2003)</td>
</tr>
</tbody>
</table>
Syntactic Accuracy

- **Joint**
- **LexHard**
- **LexSoft**

<table>
<thead>
<tr>
<th>Language</th>
<th>Joint</th>
<th>LexHard</th>
<th>LexSoft</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>83.3</td>
<td>83.4</td>
<td>83.5</td>
</tr>
<tr>
<td>de</td>
<td>92.0</td>
<td>92.0</td>
<td>92.1</td>
</tr>
<tr>
<td>fi</td>
<td>80.6</td>
<td>82.5</td>
<td>82.3</td>
</tr>
<tr>
<td>hu</td>
<td>88.9</td>
<td>89.1</td>
<td>89.1</td>
</tr>
<tr>
<td>ru</td>
<td>87.6</td>
<td>88.0</td>
<td>87.7</td>
</tr>
</tbody>
</table>
Discussion

Joint inference benefits both morphology and syntax
  • Strongest effect on morphology for Czech and German – syncretism?
  • Strongest effect on syntax for Finnish and Hungarian – why?

Lexical resources mitigate data sparseness
  • Strongest effect on morphology – soft constraints always best?
  • Weaker effect on syntax – relevant errors fixed by joint inference?
  • Strong effect on syntax for Finnish – sparse data?
Discussion

Joint inference benefits both morphology and syntax

- Strongest effect on morphology for Czech and German – syncretism?
- Strongest effect on syntax for Finnish and Hungarian – why?

Lexical resources mitigate data sparseness

- Strongest effect on morphology – soft constraints always best?
- Weaker effect on syntax – relevant errors fixed by joint inference?
- Strong effect on syntax for Finnish – sparse data?

Nice! But why only 82% for Finnish?

Can we even compare the numbers?
Part 2
Universal Dependency Annotation

Joint work with Ryan McDonald, Slav Petrov, Chris Manning, Marie de Marneffe, Jinho Choi, Filip Ginter, Yoav Goldberg, Jan Hajič and Reut Tsarfaty
Apples and Oranges
Apples and Oranges

Treebank annotation schemes vary across languages

- Hard to compare parsing results across languages (Nivre et al., 2007)
- Hard to evaluate cross-lingual learning (McDonald et al., 2013)
- Hard to make progress towards a universal parser?
Treebank annotation schemes vary across languages

- Hard to compare parsing results across languages (Nivre et al., 2007)
- Hard to evaluate cross-lingual learning (McDonald et al., 2013)
- Hard to make progress towards a universal parser?

Recent initiatives:

- **HamleDT**: Conversion of 29 existing treebanks to a PDT-like annotation scheme (Zeman et al., 2012)
- **Universal Dependency Treebank Project**: New annotation, conversion and harmonization to Google universal PoS tags and Stanford dependencies (McDonald et al., 2013)
Case Study

Delexicalized transfer parsing with universal PoS tags

<table>
<thead>
<tr>
<th>Source Training Language</th>
<th>Target Test Language</th>
<th>Unlabeled Attachment Score (UAS)</th>
<th>Labeled Attachment Score (LAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Germanic</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE</td>
<td>EN</td>
</tr>
<tr>
<td>DE</td>
<td>74.86</td>
<td>55.05</td>
<td>65.89</td>
</tr>
<tr>
<td>EN</td>
<td>58.50</td>
<td>83.33</td>
<td>70.56</td>
</tr>
<tr>
<td>SV</td>
<td>61.25</td>
<td>61.20</td>
<td>80.01</td>
</tr>
<tr>
<td>ES</td>
<td>55.39</td>
<td>58.56</td>
<td>66.84</td>
</tr>
<tr>
<td>FR</td>
<td>55.05</td>
<td>59.02</td>
<td>65.05</td>
</tr>
<tr>
<td>KO</td>
<td>33.04</td>
<td>32.20</td>
<td>27.62</td>
</tr>
</tbody>
</table>

Case Study

Delexicalized transfer parsing with universal PoS tags

First ever evaluation of labeled accuracy

<table>
<thead>
<tr>
<th>Source Training Language</th>
<th>Target Test Language</th>
<th>Unlabeled Attachment Score (UAS)</th>
<th>Labeled Attachment Score (LAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Germanic</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE</td>
<td>EN</td>
</tr>
<tr>
<td>DE</td>
<td>74.86</td>
<td>55.05</td>
<td>65.89</td>
</tr>
<tr>
<td>EN</td>
<td>58.50</td>
<td>83.33</td>
<td><strong>70.56</strong></td>
</tr>
<tr>
<td>SV</td>
<td><strong>61.25</strong></td>
<td><strong>61.20</strong></td>
<td>80.01</td>
</tr>
<tr>
<td>ES</td>
<td>55.39</td>
<td>58.56</td>
<td>66.84</td>
</tr>
<tr>
<td>FR</td>
<td>55.05</td>
<td>59.02</td>
<td>65.05</td>
</tr>
<tr>
<td>KO</td>
<td>33.04</td>
<td>32.20</td>
<td>27.62</td>
</tr>
</tbody>
</table>

Case Study

Delexicalized transfer parsing with universal PoS tags

<table>
<thead>
<tr>
<th>Source Training Language</th>
<th>Target Test Language</th>
<th>Unlabeled Attachment Score (UAS)</th>
<th>Labeled Attachment Score (LAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Germanic</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE</td>
<td>EN</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td>74.86</td>
<td>55.05</td>
</tr>
<tr>
<td>EN</td>
<td></td>
<td>58.50</td>
<td>83.33</td>
</tr>
<tr>
<td>SV</td>
<td></td>
<td><strong>61.25</strong></td>
<td>61.20</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>55.39</td>
<td>58.56</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>55.05</td>
<td>59.02</td>
</tr>
<tr>
<td>KO</td>
<td></td>
<td>33.04</td>
<td>32.20</td>
</tr>
</tbody>
</table>

Case Study

Delexicalized transfer parsing with universal PoS tags

Results make typological sense

<table>
<thead>
<tr>
<th>Source Training Language</th>
<th>Target Test Language</th>
<th>Unlabeled Attachment Score (UAS)</th>
<th>Labeled Attachment Score (LAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Germanic</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE</td>
<td>EN</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td>74.86</td>
<td>55.05</td>
</tr>
<tr>
<td>EN</td>
<td></td>
<td>58.50</td>
<td>83.33</td>
</tr>
<tr>
<td>SV</td>
<td></td>
<td><strong>61.25</strong></td>
<td><strong>61.20</strong></td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>55.39</td>
<td>58.56</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>55.05</td>
<td>59.02</td>
</tr>
<tr>
<td>KO</td>
<td></td>
<td>33.04</td>
<td>32.20</td>
</tr>
</tbody>
</table>

Case Study

Delexicalized transfer parsing with universal PoS tags

<table>
<thead>
<tr>
<th>Source Training Language</th>
<th>Target Test Language</th>
<th>Unlabeled Attachment Score (UAS)</th>
<th>Labeled Attachment Score (LAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Germanic</td>
<td>Romance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DE</td>
<td>EN</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td>74.86</td>
<td>55.05</td>
</tr>
<tr>
<td>EN</td>
<td></td>
<td>58.50</td>
<td>83.33</td>
</tr>
<tr>
<td>SV</td>
<td></td>
<td><strong>61.25</strong></td>
<td><strong>61.20</strong></td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>55.39</td>
<td>58.56</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td>55.05</td>
<td>59.02</td>
</tr>
<tr>
<td>KO</td>
<td></td>
<td>33.04</td>
<td>32.20</td>
</tr>
</tbody>
</table>

Too Many Standards?
Too Many Standards?

English SD
de Marneffe et al. (2006)
Too Many Standards?

- **English SD**
  - de Marneffe et al. (2006)

- **Google SD**
  - McDonald et al. (2013)
Too Many Standards?

- English SD
  - de Marneffe et al. (2006)

- Google SD
  - McDonald et al. (2013)

- Finnish SD
  - Haverinen et al. (2013)

- Hebrew SD
  - Tsarfaty (2013)
Too Many Standards?

CLEAR
Choi (2012)

English SD
de Marneffe et al. (2006)

Google SD
McDonald et al. (2013)

HamleDT
Zeman et al. (2012)

Finnish SD
Haverinen et al. (2013)

Hebrew SD
Tsarfaty (2013)
Too Many Standards?

Universal Dependencies

Universal Dependencies

http://universaldependencies.github.io/docs/
Universal Dependencies

http://universaldependencies.github.io/docs/

In dem Restaurant isst Maria den Fisch.

ADP DET NOUN VERB PNOUN DET NOUN
CASE=DAT CASE=DAT TENSE=PRES CASE=NOM CASE=ACC CASE=ACC
NUM=SIN NUM=SIN GEN=NEU GEN=NEU NUM=SIN NUM=SIN
GEN=MAS GEN=MAS

Google Universal Part-of-Speech Tags + Morphological Features
Universal Dependencies

http://universaldependencies.github.io/docs/
Guiding Principles
Guiding Principles

Maximize parallelism

• Don’t annotate the same thing in different ways
• Don’t make different things look the same
Guiding Principles

Maximize parallelism

• Don’t annotate the same thing in different ways
• Don’t make different things look the same

But don’t overdo it

• Don’t annotate things that are not there
• Languages select from a universal pool of categories
• Allow language-specific extensions
Keeping content words as heads promotes parallelism

• Function words often correlate with morphology
• Keeping content words as heads promotes parallelism
• Function words often correlate with morphology
Keeping content words as heads promotes parallelism

Function words often correlate with morphology
### Dependency Relations

- **Taxonomy of 42 universal grammatical relations**, broadly supported across many languages in language typology
- **Language specific subtypes** can be added
The lexicalist hypothesis

- Grammatical relations hold between words (including clitics)
- Morphological categories are properties of words

Morphological annotation

- Revised Google Universal Part-of-Speech Tags (Petrov et al., 2012)
- Universal inventory of morphological features (under construction)
Tokens and Words

Principle of recoverability

- Clitics and contractions are split to allow meaningful annotation
- Mapping from basic tokenization is explicitly represented
- Heuristic mapping of annotation to basic tokens is provided

Encoded in revised CoNLL format (CoNLL-U)

Vamos

1.1
Vámonos

1.2

nos

a

el

mar

.
Tokens and Words

Principle of recoverability

- Clitics and contractions are split to allow meaningful annotation
- Mapping from basic tokenization is explicitly represented
- Heuristic mapping of annotation to basic tokens is provided

Encoded in revised CoNLL format (CoNLL-U)
Tokens and Words

Principle of recoverability

- Clitics and contractions are split to allow meaningful annotation
- Mapping from basic tokenization is explicitly represented
- Heuristic mapping of annotation to basic tokens is provided

Encoded in revised CoNLL format (CoNLL-U)
Tokens and Words

Principle of recoverability

• Clitics and contractions are split to allow meaningful annotation
• Mapping from basic tokenization is explicitly represented
• Heuristic mapping of annotation to basic tokens is provided

Encoded in revised CoNLL format (CoNLL-U)
Work in Progress

Current time plan:
• Stable annotation guidelines by end of September
• First release of data sets before the end of 2014

Follow our progress and give feedback:
• Universal Dependencies: http://universaldependencies.github.io/docs/

Check out old releases:
• Uni-Dep-TB: https://code.google.com/p/uni-dep-tb/

Let’s work together!
Current time plan:

- Stable annotation guidelines by end of September
- First release of data sets before the end of 2014

Follow our progress and give feedback:

- Universal Dependencies: http://universaldependencies.org/docs/

Check out old releases:

- Uni-Dep-TB: https://code.google.com/p/uni-dep-tb/

Let’s work together!
Part 3

Parsing with Universal Dependencies

Confessions of a converted dependency parser
Taking Stock
Taking Stock

Motivation for universal dependencies

• Improve comparability of parsing results across languages
• Facilitate development of multilingual systems
• Enable typological studies of syntactic structure
Taking Stock

Motivation for universal dependencies

• Improve comparability of parsing results across languages
• Facilitate development of multilingual systems
• Enable typological studies of syntactic structure

What about parsing?

• Not likely to improve parsing accuracy with existing parsers
• Parsers tend to prefer function words as heads (Schwarz et al., 2012)
• We risk bringing English down to 80% instead of Finnish up to 90%
What’s the Problem?
What's the Problem?

- Dependency parsers know only one syntactic relation
What’s the Problem?

- Dependency parsers know only one syntactic relation
- They do not interpret dependency labels
What’s the Problem?

- Dependency parsers know only one syntactic relation
- They do not interpret dependency labels
- They represent a construction primarily by its head
A Simple Case

The dog was chased by the *black cat*. 
A Simple Case

The dog was chased by the *black* *cat*. 

- All criteria point to *cat* being the head
A Simple Case

The dog was chased by the black cat.

- All criteria point to *cat* being the head
- Little (syntactic) information is lost by dropping *black*
A Tricky Case

The dog *was chased* by the black cat.
A Tricky Case

The dog *was* *chased* by the black cat.

• Some criteria point to *was*, others to *chased* as the head
A Tricky Case

The dog *chased* by the black cat.

• Some criteria point to *was*, others to *chased* as the head

• Neither word can represent the whole
A Tricky Case

The dog was chased *by* the black *cat*.

- Some criteria point to *was*, others to *chased* as the head
- Neither word can represent the whole
- The same problem arises with *by* (the black *cat*)
Back to the Roots

- Three fundamental syntactic relations (Tesnière, 1959)
- Tesnière-style dependency treebank (Sangati and Mazza, 2009)
Universal Dependencies
The UD representation is a simple dependency tree
• The UD representation is a simple dependency tree
• But labels are universal and can be interpreted
Universal Dependencies

The UD representation is a simple dependency tree

**was** chased

Fido by **Kitty—and—Tiger**

- The UD representation is a simple dependency tree
- But labels are universal and can be interpreted
- Therefore, we can put more knowledge into the parser
Final Remarks
Final Remarks

- Constituency parsing – largely driven by PTB
  - Perhaps too much emphasis on English (until recently)
  - But deep analysis of categories and representations
Final Remarks

• Constituency parsing – largely driven by PTB
  • Perhaps too much emphasis on English (until recently)
  • But deep analysis of categories and representations

• Dependency parsing – largely driven by CoNLL data
  • More attention to typological diversity
  • But ignorance and blind faith with respect to representations
Final Remarks

- Constituency parsing – largely driven by PTB
  - Perhaps too much emphasis on English (until recently)
  - But deep analysis of categories and representations

- Dependency parsing – largely driven by CoNLL data
  - More attention to typological diversity
  - But ignorance and blind faith with respect to representations

- Can UD give us the best of both worlds?
Thanks for Your Attention!
Questions?

http://stp.lingfil.uu.se/~nivre/