Dependency Grammar

Introduction

Dependency Grammar and Dependency Structure

Dependency syntax postulates that syntactic structure consists of relations between lexical items, normally binary asymmetric relations ("arrows") called dependencies.

The arrows are commonly typed with the name of grammatical relations (subject, prepositional object, apposition, etc.)

The arrow connects a head (governor, superior, regent) with a dependent (modifier, inferior, subordinate)

Usually, dependencies form a tree (connected, acyclic, single-head)

• Some people draw the arrows one way; some the other way!
• Tesnière had them point from head to dependent...
• Usually add a fake ROOT so every word is a dependent of precisely 1 other node

Dependency Grammar and Dependency Structure

ROOT Discussion of the outstanding issues was completed.

• The idea of dependency structure goes back a long way
  • To Pāṇini's grammar (c. 5th century BCE)
  • Basic approach of 1st millennium Arabic grammarians
• Constituency is a new-fangled invention
  • 20th century invention (R.S. Wells, 1947)
• Modern dependency work often linked to work of L. Tesnière (1959)
  • Was dominant approach in "East" (Russia, China, …)
• Good for free-er word order languages
• Among the earliest kinds of parsers in NLP, even in the US:
  • David Hays, one of the founders of U.S. computational linguistics, built early (1st?) dependency parser (Hays 1962)
Christopher Manning

**Relation between phrase structure and dependency structure**

- A dependency grammar has a notion of a head. Officially, CFGs don’t.
- But modern linguistic theory and all modern statistical parsers (Charniak, Collins, Stanford, …) do, via hand-written phrasal “head rules”:
  - The head of a Noun Phrase is a noun/number/adjective...
  - The head of a Verb Phrase is a verb/adj...

- The head rules can be used to extract a dependency parse from a CFG parse

- The closure of dependencies gives constituency from a dependency tree
- But the dependents of a word must be at the same level (i.e., “flat”) – there can be no VP!

**Dependency Parsing**

- A sentence is parsed by choosing for each word what other word (including ROOT) that it is a dependent of.
  - Usually some constraints:
    - Only one word is a dependent of ROOT
    - Don’t want cycles A → B, B → A
    - This makes the dependencies a tree
    - Final issue is whether arrows can cross (non-projective) or not

**Methods of Dependency Parsing**

1. Dynamic programming (like in the CKY algorithm)
   You can do it similarly to lexicalized PCFG parsing: an O(n^3) algorithm
   Eisner (1996) gives a clever algorithm that reduces the complexity to O(n^2), by producing parse items with heads at the ends rather than in the middle
2. Graph algorithms
   You create a Minimum Spanning Tree for a sentence
   McDonald et al.’s (2005) MSTParser scores dependencies independently using a ML classifier (he uses MIRA, for online learning, but it can be something else)
3. Constraint Satisfaction
   Edges are eliminated that don’t satisfy hard constraints. Karlsson (1990), etc.
4. “Deterministic parsing”
   Greedy choice of attachments guided by good machine learning classifiers
   MaltParser (Nivre et al. 2008)

**Dependency Conditioning Preferences**

What are the sources of information for dependency parsing?

1. Bilexical affinities 
   "issues → the" is plausible
2. Dependency distance 
   modify with nearby words
3. Intervening material
   Dependencies rarely span intervening verbs or punctuation
4. Valency of heads
   How many dependents on which side are usual for a head?

**Probabilistic dependency grammar: generative model**

1. Start with left wall $S$
2. Generate root $w_1$
3. Generate left children $w_2, \ldots, w_j$ from the FSA $\lambda w_i$
4. Generate right children $w_j, \ldots, w_r$ from the FSA $\rho w_i$
5. Recurse on each $w_i$ for $i$ in $\{j, \ldots, 1\}$, sampling from $P(\alpha | \{\lambda w_i, \rho w_i\})$
6. Return $\alpha_1, \ldots, \alpha_j, \ldots, \alpha_r$
Naive Recognition/Parsing

$O(n^2)$ combinations

Dependency Grammar Cubic Recognition/Parsing (Eisner & Satta, 1999)

- **Triangles**: span over words, where tall side of triangle is the head, other side is dependent, and no non-head words expecting more dependents
- **Trapezoids**: span over words, where larger side is head, smaller side is dependent, and smaller side is still looking for dependents on its side of the trapezoid

Graph Algorithms: MSTs
Christopher Manning

This is the common “arc-eager” variant: a head can immediately take a right dependent, before its dependents are found.

MaltParser

A simple form of greedy discriminative dependency parser

The parser does a sequence of bottom up actions

Roughly like “shift” or “reduce” in a shift-reduce parser, but the “reduce” actions are specialized to create dependencies with head on left or right

The parser has:

- a stack σ, written with top to the right
- which starts with the ROOT symbol
- a buffer β, written with top to the left
- which starts with the input sentence
- a set of dependency arcs A
- which starts off empty
- a set of actions

McDonald et al. (2005 ACL):

Online Large-Margin Training of Dependency Parsers

- Edge scoring is via a discriminative classifier
  - Can condition on rich features in that context
  - Each dependency is a linear function of features times weights
- Feature weights were learned by MIRA, an online large-margin algorithm
  - But you could use an SVM, maxent, or a perceptron
- Features cover:
  - Head and dependent word and POS separately
  - Head and dependent word and POS bigram features
  - Words between head and dependent
  - Length and direction of dependency

MaltParser

[Nivre et al. 2008]

- A simple form of greedy discriminative dependency parser
- The parser does a sequence of bottom up actions
- Roughly like “shift” or “reduce” in a shift-reduce parser, but the “reduce” actions are specialized to create dependencies with head on left or right
- The parser has:
  - a stack σ, written with top to the right
  - which starts with the ROOT symbol
  - a buffer β, written with top to the left
  - which starts with the input sentence
  - a set of dependency arcs A
  - which starts off empty
  - a set of actions

Greedy Transition-Based Parsing

Basic transition-based dependency parser

Start: σ = [ROOT], β = w₁,..., wₙ, A = ∅
1. Shift σ, w₁|β, A → σ|w₁, β, A
2. Left-Arc σ|w₁, w₂|β, A → σ|w₁, w₂|β, A U {σ|w₁, w₂}
3. Reduce σ|w₁, w₂|β, A → σ, β, A
4. Shift σ, w₁|β, A → σ|w₁, β, A
Finish: β = ∅

Notes:

- Unlike the regular presentation of the CFG reduce step, dependencies combine one thing from each of stack and buffer

Example

Happy children like to play with their friends.

| [ROOT] | Happy, children,... | ∅ |
| [ROOT, Happy] | children, like,... | ∅ |
| [ROOT] | like, to,... | (amod(children, happy)) = A₁ |
| [ROOT, children] | like, to,... | A₁ |
| [ROOT, like] | to, play,... | A₂ |
| [ROOT, like] | to, play,... | A₂ |
| [ROOT] | aux, (play, to) = A₃ |
| [ROOT, like] | with, their,... | A₄ |
| [ROOT, like] | with, their,... | A₄ |
| [ROOT, like] | with, their,... | A₄ |
Happy children like to play with their friends.

RA_comp  [ROOT, like, play]  [with their, ...]  \( A_1 \cup \{ \text{like}, \text{play} \} = A_1 \)
RA_adv  [ROOT, like, play, with]  [their, friends, ...]  \( A_1 \cup \{ \text{preps}, \text{play, with} \} = A_2 \)
Shift  [ROOT, like, play, with, their]  [friends, ...]  \( A_3 \)
LA_adv  [ROOT, like, play, with, friends]  [.]  \( A_4 \)
Reduce  [ROOT, like, play, with]  [.]  \( A_4 \)
Reduce  [ROOT, like]  [.]  \( A_4 \)
RA_adv  [ROOT, like, ...]  [.]  \( A_4 \cup \{ \text{punc}, \text{like}, ... \} = A_4 \)

You terminate as soon as the buffer is empty. Dependencies = \( A_4 \)
Example

1. Left-Arc: \( \sigma(w, \beta, \alpha, \gamma) \in \mathcal{E}(A, w, \alpha) \cap \mathcal{E}(A, \alpha) \)
   \( \text{Precondition: } (w, \beta, \alpha, \gamma) \in Q \)
2. Right-Arc: \( \sigma(w, \beta, \alpha, \gamma) \in \mathcal{E}(A, \beta, \gamma) \cap \mathcal{E}(A, \gamma) \)
   \( \text{Precondition: } (w, \beta, \alpha, \gamma) \in Q \)
3. Reduce: \( \delta(w, \beta, \alpha, \gamma) \in \mathcal{E}(A, \beta, \gamma) \cap \mathcal{E}(A, \gamma) \)
   \( \text{Precondition: } (w, \beta, \alpha, \gamma) \in Q \)
4. Shift: \( \epsilon(w, \beta, \alpha, \gamma) \in \mathcal{E}(A, \beta, \gamma) \cap \mathcal{E}(A, \gamma) \)
   \( \text{Precondition: } (w, \beta, \alpha, \gamma) \in Q \)

**Shift**

\[ \text{ROOT: Red figures on the screen indicated falling stocks} \]

**Left-arc**

\[ \text{ROOT: Red figures on the screen indicated falling stocks} \]

**Right-arc**

\[ \text{ROOT: Red figures on the screen indicated falling stocks} \]

**Reduce**

\[ \text{ROOT: Red figures on the screen indicated falling stocks} \]

**Reduce**

\[ \text{ROOT: Red figures on the screen indicated falling stocks} \]

**Left-arc**

\[ \text{ROOT: Red figures on the screen indicated falling stocks} \]
Example 39

1. Left-Arc: σ, w' ∈ S, A ∈ V
   Precondition: (w, w') ∈ Σ
   Residue: σ, w'/A → σ, w/A
   Result: (w, w') ∈ A

2. Right-Arc: σ, w' ∈ S, A ∈ V
   Precondition: (w, w') ∈ Σ
   Residue: σ, w/A → σ, w'/A
   Result: (w, w') ∈ A

3. Reduce: σ, w' ∈ S, A ∈ V
   Precondition: (w, w') ∈ Σ
   Residue: σ, w/A → σ
   Result: (w, w') ∈ Σ

4. Shift: σ, w' ∈ S, A ∈ V
   Precondition: (w, w') ∈ Σ
   Residue: σ, w/A → σ
   Result: (w, w') ∈ Σ

Left-arc

Right-arc

Shift

Reduce
MaltParser
[Nivre et al. 2008]

- We have left to explain how we choose the next action
- Each action is predicted by a discriminative classifier (often SVM, can be perceptron, maxent classifier) over each legal move
  - Max of 4 untyped choices, max of |R| = 2 + 2 when typed
  - Features: top of stack word, POS; first in buffer word, POS; etc.
- There is NO search (in the simplest and usual form)
  - But you could do some kind of beam search if you wish
  - It provides very fast linear time parsing
- The model’s accuracy is slightly below the best Lexicalized PCFGs (evaluated on dependencies), but
  - It provides close to state of the art parsing performance

Evaluation of Dependency Parsing:
(labeled) dependency accuracy

Evaluation of Dependency Parsing:
(labeled) dependency accuracy

Who did Bill buy the coffee from yesterday?

Representative performance numbers

- The CoNLL-X (2006) shared task provides evaluation numbers for various dependency parsing approaches over 13 languages
  - MALT: LAS scores from 65–92%, depending greatly on language/treebank
- Here we give a few UAS numbers for English to allow some comparison to constituency parsing

<table>
<thead>
<tr>
<th>Parser</th>
<th>UAS%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sagae and Lavie (2006) ensemble of dependency parsers</td>
<td>92.7</td>
</tr>
<tr>
<td>Charniak (2000) generative, constituency, as dependencies</td>
<td>92.2</td>
</tr>
<tr>
<td>Collins (1999) generative, constituency, as dependencies</td>
<td>91.7</td>
</tr>
<tr>
<td>McDonald and Pereira (2005) – MST graph-based dependency</td>
<td>91.5</td>
</tr>
<tr>
<td>Yamada and Matsumoto (2003) – transition-dependency</td>
<td>90.8</td>
</tr>
</tbody>
</table>

Projectivity

- Dependencies from a CFG tree using heads, must be projective
  - There must not be any crossing dependency arcs when the words are laid out in their linear order, with all arcs above the words.
- But dependency theory normally does allow non-projective structures to account for displaced constituents
  - You can’t easily get the semantics of certain constructions right without these nonprojective dependencies

Handling non-projectivity

- The arc-eager algorithm we presented only builds projective dependency trees
- Possible directions to head:
  1. Just declare defeat on nonprojective arcs
  2. Use a dependency formalism which only admits projective representations (a CFG doesn’t represent such structures…)
  3. Use a postprocessor to a projective dependency parsing algorithm to identify and resolve nonprojective links
  4. Add extra types of transitions that can model at least most non-projective structures
  5. Move to a parsing mechanism that does not use or require any constraints on projectivity (e.g., the graph-based MSTParser)

Dependencies encode relational structure

Relation Extraction
with Stanford Dependencies
**Dependency paths identify relations like protein interaction**

[Erkan et al. EMNLP 07, Fundel et al. 2007]

- Demonstrated results that KaiC interacts rhythmically with SasA
- KaiC nsubj interacts prep_with SasA
- KaiC nsubj interacts prep_with SasA conj_and KaiA
- KaiC nsubj interacts prep_with SasA conj_and KaiB

**Stanford Dependencies**

[de Marneffe et al. LREC 2006]

- The basic dependency representation is projective
- It can be generated by postprocessing headed phrase structure parses (Penn Treebank syntax)
- It is generated directly by dependency parsers, such as MaltParser, or the Easy-First Parser

**BioNLP 2009/2011 relation extraction shared tasks**

[Björne et al. 2009]

- Many relationships become short distance!

- Dependency distance
- Linear distance

**Graph modification to facilitate semantic analysis**

Bell, based in LA, makes and distributes electronic and computer products.

- nsubj Bell
- partmod based in LA
- nsubj makes conj_and distributes
- products amod electronic and computer
- nsubj makes conj_and distributes
- products amod electronic and computer

- nsubj Bell
- partmod based in LA
- nsubj makes conj_and distributes
- products amod electronic and computer