Natural Language Processing

Introduction to dependency parsing

Marco Kuhlmann Department of Computer and Information Science



This work is licensed under a <u>Creative Commons Attribution 4.0 International License</u>.

Dependency parsing

- **Syntactic parsing** is the task of mapping a sentence to a formal representation of its syntactic structure.
- We focus on representations in the form of **dependency trees**.



A syntactic dependency is an asymmetric relation between a head and a dependent.

universaldependencies.org

Ċ

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Akkadian 1 1K 1K Afro-Asiatic, Semitic Albanian 1 <1K W IE, Albanian Amharic 1 10K Afro-Asiatic, Semitic Ambaric 1 10K Afro-Asiatic, Semitic Ancient Greek 2 416K Afro-Asiatic, Semitic Arabic 3 1,042K Afro-Asiatic, Semitic Arabic 3 1,042K Afro-Asiatic, Semitic Armenian 1 52K Afro-Asiatic, Semitic Barpara 1 13K Afro-Asiatic, Semitic Basque 1 12K Basque Basque Basque 1 12K Basque Basque Basque 1 13K Afro-Maintic, Semitic Basque 1 10K Afro-Maintic, Semitic Catalan 1 10K Afro-Maintic, Semitic Castalan		Afrikaans	1	49K	<0	IE, Germanic
Albanian 1 <1K		Akkadian	1	1K		Afro-Asiatic, Semitic
Amharic 1 10K AFro-Asiatic, Semitic Ancient Greek 2 416K AFro-Asiatic, Semitic Arabic 3 1,042K BW Afro-Asiatic, Semitic Armenian 1 52K BW Afro-Asiatic, Semitic Assyrian 1 <1K	-	Albanian	1	<1K	W	IE, Albanian
Ancient Greek 2 416K 460 IE, Greek Arabic 3 1,042K IV Afro-Asiatic, Semitic Armenian 1 52K IE, Armenian IE, Armenian Assyrian 1 41K IIIIO Afro-Asiatic, Semitic Bambara 1 13K IIIIO Afro-Asiatic, Semitic Basque 1 121K IIIIO Basque Belarusian 1 13K IIIIO Basque Bibipuri 2 4K IIIIO IE, Islavic Bulgarian 1 10K IIIIIOK IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		Amharic	1	10K		Afro-Asiatic, Semitic
Image: Second Secon		Ancient Greek	2	416K		IE, Greek
Armenian 1 52K market Armenian N Assyrian 1 <1K		Arabic	3	1,042K		Afro-Asiatic, Semitic
> X Assyrian 1 <1K		Armenian	1	52K		IE, Armenian
▶ Bambara 1 13K Implementation Mande ▶ Basque 1 121K Implementation Basque ▶ Belarusian 1 13K Implementation Implementation Implementation ▶ Belarusian 1 13K Implementation Implementation Implementation Implementation ▶ Belarusian 1 13K Implementation Implementation <thimpleme< td=""><td></td><td>Assyrian</td><td>1</td><td><1K</td><td></td><td>Afro-Asiatic, Semitic</td></thimpleme<>		Assyrian	1	<1K		Afro-Asiatic, Semitic
Image: Second		Bambara	1	13K		Mande
> Belarusian 1 13K # 1 13K # 1 16K 15K 15K <td></td> <td>Basque</td> <td>1</td> <td>121K</td> <td></td> <td>Basque</td>		Basque	1	121K		Basque
> Bhojpuri 2 4K IE, Indic > Breton 1 10K 2/IIIO/W IE, Celtic > Bulgarian 1 156K 2/IIIO/W IE, Slavic > Buryat 1 10K 2/IIIO/W IE, Slavic > Survat 1 10K 2/IIIO Mongolic > IC cantonese 1 13K 9 Sino-Tibetan IC catalan 1 531K III F, Romance IC catalan 1 531K III Sino-Tibetan IC cataian 1 74K III Sino-Tibetan IC cassical Chinese 1 74K IIII Sino-Tibetan IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		Belarusian	1	13K		IE, Slavic
> Breton 1 10K #/************************************		Bhojpuri	2	4K		IE, Indic
Bulgarian 1 156K Image: Construction of the second o		Breton	1	10K		IE, Celtic
Buryat 1 10K #/FE Mongolic Cantonese 1 13K Sino-Tibetan Catalan 1 531K IE, Romance Catalan 1 531K IE, Romance Chinese 5 285K ZED/W Sino-Tibetan Classical Chinese 1 74K 6 Sino-Tibetan Coptic 1 40K IE Sino-Tibetan Mongolic IE, Germanic IE, Slavic IE Danish 2 100K IE Germanic Dutch 2 306K IE Germanic IE English 9 620K IE Germanic IE Faroese 1 10K W IE, Germanic IE Faroese 1 10K W IE, Germanic IE French 8		Bulgarian	1	156K		IE, Slavic
Cantonese 1 13K Sino-Tibetan Catalan 1 531K IE, Romance Chinese 5 285K Sino-Tibetan Classical Chinese 1 74K 6 Sino-Tibetan Coptic 1 40K IE, Romance Sino-Tibetan Coptic 1 40K IE, Sino-Tibetan Sino-Tibetan Coptic 1 40K IE, Gery IE, Slavic Coptic 1 40K IE, Slavic IE, Slavic Cocech 5 2,222K IE, Germanic IE, Germanic Danish 2 100K IE, Germanic IE, Germanic Dutch 2 306K IE, Germanic IE, Germanic English 9 620K IE, Germanic IE, Germanic Erzya 1 15K IIII Uralic, Mordvin Estonian 2 465K IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		Buryat	1	10K		Mongolic
▶ ■ Catalan 1 531K □ IE, Romance ▶ ■ Chinese 5 285K □ □ Sino-Tibetan ▶ ■ Classical Chinese 1 74K 6 Sino-Tibetan ▶ ■ Coptic 1 40K ● Afro-Asiatic, Egyptian ▶ ■ Coptic 1 40K ● Afro-Asiatic, Egyptian ▶ ■ Croatian 1 199K □ W IE, Slavic ▶ ■ Croatian 1 199K □ W IE, Slavic ▶ ■ Czech 5 2,222K ■ ▲ /> IE, Germanic ▶ ■ Danish 2 100K □ IE, Germanic IE, Germanic ■ ■ Dutch 2 306K □ Uralic, Mordvin ■ ■ Erzya 1 15K □ Uralic, Finnic ■ ■ ■ 10K W Uralic, Finnic IE, Germanic	*	Cantonese	1	13K	P	Sino-Tibetan
▶ Chinese 5 285K Image: Constraint of the second s		Catalan	1	531K		IE, Romance
▶ Classical Chinese 1 74K Image: Sino-Tibetan ▶ Coptic 1 40K Image: Sino-Tibetan ▶ Coptic 1 40K Image: Sino-Tibetan ▶ Croatian 1 199K Image: Sino-Tibetan ▶ Croatian 1 199K Image: Sino-Tibetan ▶ Croatian 1 199K Image: Sino-Tibetan > Croatian 1 199K Image: Sino-Tibetan > Croatian 1 199K Image: Sino-Tibetan > Danish 2 2222K Image: Sino-Tibetan > Danish 2 100K Image: Sino-Tibetan > Danish 2 100K Image: Sino-Tibetan > English 9 620K Image: Sino-Tibetan > English 9 620K Image: Sino-Tibetan > Estonian 2 465K Image: Sino-Ow Image: Sino-Tibetan > Estonian 2 465K Image: Sino-Ow Image: Sino-Tibetan <t< td=""><td>•</td><td>Chinese</td><td>5</td><td>285K</td><td>CEOW</td><td>Sino-Tibetan</td></t<>	•	Chinese	5	285K	C EOW	Sino-Tibetan
Image: Solution of the system of the sys		Classical Chinese	1	74K	0	Sino-Tibetan
▶ Croatian 1 199K □QW IE, Slavic ▶ Czech 5 2,222K □A, IE, Slavic ▶ Danish 2 100K □QO IE, Germanic ▶ Danish 2 306K □W IE, Germanic ▶ Dutch 2 306K □W IE, Germanic ▶ English 9 620K ☞ □Composition IE, Germanic ▶ English 9 620K ☞ □Composition IE, Germanic ▶ Erzya 1 15K □ Uralic, Mordvin ▶ Estonian 2 465K □Die@@ Uralic, Finnic ▶ Estonian 2 465K □Die@@ Uralic, Finnic ▶ Faroese 1 10K W Uralic, Finnic ▶ Finnish 3 377K □Die@ W Uralic, Finnic ▶ French 8 1,157K □A IE, Romance IE, Romance ▶ Galician 2 164K		Coptic	1	40K		Afro-Asiatic, Egyptian
▶ Czech 5 2,222K ▷ ▷ ▷ □ <td< td=""><td></td><td>Croatian</td><td>1</td><td>199K</td><td></td><td>IE, Slavic</td></td<>		Croatian	1	199K		IE, Slavic
> Danish 2 100K Image: Second		Czech	5	2,222K		IE, Slavic
▶ Dutch 2 306K IE, Germanic ▶ English 9 620K IE, Germanic ▶ Erzya 1 15K Uralic, Mordvin ▶ Estonian 2 465K Uralic, Finnic ▶ Faroese 1 10K W IE, Germanic ▶ Faroese 1 10K W IE, Germanic ▶ Finnish 3 377K IIII (IIII) IE, Romance ▶ French 8 1,157K IIIIIIII IE, Romance ▶ Galician 2 164K IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		Danish	2	100K		IE, Germanic
Image: Second		Dutch	2	306K		IE, Germanic
Erzya 1 15K Uralic, Mordvin Estonian 2 465K Uralic, Finnic Faroese 1 10K Uralic, Finnic Finnish 3 377K Uralic, Finnic French 8 1,157K Value Uralic, Finnic Galician 2 164K IE, Romance IE, Romance German 4 3,753K IEOCOW IE, Germanic Gothic 1 55K IE, Germanic IE, Germanic Greek 1 63K IEOW IE, Greek Hebrew 1 161K IE Afro-Asiatic, Semitic		English	9	620K	춛∰▇▋▓₢ጲ▣❶ᠿእ₽ঔ₩	IE, Germanic
Estonian 2 465K Image: Second and a second and and a second and a second and a second and and and a second and		Erzya	1	15K		Uralic, Mordvin
Faroese 1 10K W IE, Germanic Finnish 3 377K Image: W Uralic, Finnic French 8 1,157K Image: W IE, Romance Source Galician 2 164K Age: W IE, Romance Source German 4 3,753K Image: W IE, Germanic Source Gothic 1 55K IE, Germanic IE, Germanic Source 1 63K Image: W IE, Greek Hebrew 1 161K Image: W Afro-Asiatic, Semitic		Estonian	2	465K		Uralic, Finnic
Finnish 3 377K ∰ ♥ ▲ ♥ ♥ Uralic, Finnic French 8 1,157K ● ▲ ♥ ♥ △ ♥ ♥ △ ♥ ○ ○ ♥ IE, Romance Image: Second		Faroese	1	10K	W	IE, Germanic
French 8 1,157K A and a an		Finnish	3	377K	Hext	Uralic, Finnic
Image: Second system Calician 2 164K ★ / I I I I I I I I I I I I I I I I I I		French	8	1,157K		IE, Romance
German 4 3,753K COMPANY IE, Germanic Gothic 1 55K IE, Germanic IE, Germanic Greek 1 63K COMPANY IE, Greek Hebrew 1 161K Company Afro-Asiatic Semitic		Galician	2	164K		IE, Romance
Gothic 1 55K IE, Germanic Greek 1 63K IE, Greek Hebrew 1 161K Afro-Asiatic, Semitic		German	4	3,753K	œ₿₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽₽	IE, Germanic
Greek 1 63K C K IE, Greek Hebrew 1 161K Afro-Asiatic, Semitic		Gothic	1	55K	•	IE, Germanic
Hebrew 1 161K E Afro-Asiatic Semitic		Greek	1	63K	ECW	IE, Greek
		Hebrew	1	161K		Afro-Asiatic, Semitic



Dependency trees

- A dependency tree for a sentence x is a digraph G = (V, A)where $V = \{1, ..., |x|\}$ and where there exists a $r \in V$ such that every $v \in V$ is reachable from r via exactly one directed path.
- The vertex *r* is called the **root** of *G*.
- The arcs of a dependency tree may be labelled to indicate the type of the syntactic relation that holds between the two elements. Universal Dependencies v2 uses 37 universal syntactic relations (list).

Representation of dependency trees



6

Two parsing paradigms

Graph-based dependency parsing

Cast parsing as a combinatorial optimisation problem over a (possibly restricted) set of dependency trees.

Transition-based dependency parsing

Cast parsing as a sequence of local classification problems: at each point in time, predict one of several parser actions.

Graph-based dependency parsing

Given a sentence x and a set Y(x) of candidate dependency trees for *x*, we want to find a highest-scoring tree $\hat{y} \in Y(x)$:

$$\hat{y} = \underset{y \in Y(x)}{\operatorname{arg\,max\,score}}(x, y)$$

The computational complexity of this problem depends on the choice of the set Y(x) and the scoring function.

The arc-factored model

Under the **arc-factored model**, the score of a dependency tree is expressed as the sum of the scores of its arcs:

$$\hat{y} = \underset{y \in Y(x)}{\operatorname{arg\,max}} \sum_{a \in y} \operatorname{score}(x, a)$$
 ——

The score of a single arc can be computed by means of a neural network that receives the head and the dependent as input. for example, a simple linear layer: $score(x, h \rightarrow d) = [h; d] \cdot w + b$

— head–dependent arc

Computational complexity

- Under the arc-factored model, the highest-scoring dependency tree can be found in $O(n^3)$ time (n = sentence length). Chu–Liu/Edmonds algorithm; <u>McDonald et al. (2005)</u>
- Even seemingly minor extensions of the arc-factored model entail intractable parsing.

McDonald and Satta (2007)

For some of these extensions, polynomial-time parsing is possible for restricted classes of dependency trees.

Transition-based dependency parsing

- We cast parsing as a sequence of local classification problems such that solving these problems builds a dependency tree.
- In most approaches, the number of classifications required for this is linear in the length of the sentence.

Transition-based dependency parsing

- The parser starts in the **initial configuration**. empty dependency graph
- It then calls a classifier, which predicts the **transition** that the parser should make to move to a next configuration. extend the partial dependency tree
- This process is repeated until the parser reaches a **terminal** configuration.

complete dependency tree

Training transition-based dependency parsers

- To train a transition-based dependency parser, we need a treebank with gold-standard dependency trees.
- In addition to that, we need an algorithm that tells us the gold-standard transition sequence for a tree in that treebank.
- Such an algorithm is conventionally called an **oracle**.

Comparison of the two parsing paradigms

Graph-based parsing

slow (in practice, cubic in the length of the sentence)

restricted feature models (in practice, arc-factored)

features and weights directly defined on target structures

Transition-based parsing

fast (quasi-linear in the length) of the sentence)

rich feature models defined on configurations

indirection – features and weights defined on transitions