Deriving Multi-Headed Planar Dependency Parses from Link Grammar Parses

Juneki Hong — Carnegie Mellon University

Jason Eisner — Johns Hopkins University

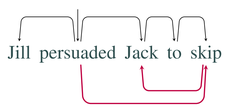
Summary

- Multi-headed dependency corpora would allow for the development of richer syntactic formalisms.
- Link Grammar can produce projective multi-headed corpora, but Link Grammar parses are undirected.
- We want to recover this "missing information" by consistently directionalizing Link Grammar parses, subject to constraints such as acyclicity and reachability.
- Starting with a corpus of LG parses, we utilize ILP to find a minimum set of directionality assignments subject to these constraints.[3]
- The resulting parses differ in style from CoNLL-style parses of the same sentences.

Multi-Headed Dependency Parsing

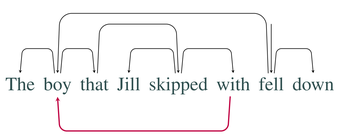
Relaxing single-headed constraints common in dependency parsing would allow for constructions such as Control, Relativization, and Conjunction.

Control



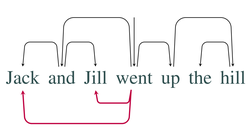
Jack is the object of one verb and the subject of another

Relativization



The boy is the object of *with* as well as the subject of *fell*.

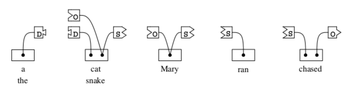
Conjunction



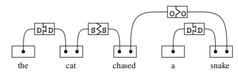
Jack and Jill serve as the two arguments to *and*, but are also subjects of *went*.

Link Grammars

- Grammar-based formalism for projective dependency parsing with undirected links.
- A label on the link describes the relationship between two words.
- The original formalism and English Link Grammar was created by Davy Temperley, Daniel Sleator, and John Lafferty[4].



Example visualization of a link grammar, taken from their original paper. Words have "links" that attach to others. These link attachments can either be optional or required.



A parse where all the link attachments have been satisfied. Link attachments must be projective. A parse cannot be completed until all words have a complete set of link attachments.

Integer Linear Programming

- ILP is an optimization problem where the objective function and constraints are linear, while some or all of the variables are integers.
- In general, it's NP-Hard, but good solvers exist that often work well.
- Our ILP is encoded as a ZIMPL program and solved with SCIP Optimization Suite[2, 1]

ILP Link Orientation Variables

For each sentence, for each edge i, j, where $i < j$

Orientation of each link can be represented by variables that can either be oriented left or oriented right:

$$x_{ij}, x_{ji} \in \mathbb{Z} \geq 0$$

$$x_{ij} + x_{ji} = 1$$

ILP Constraints

Acyclicity

Given that node u is the parent of v

n_v: length of the sentence containing node v

$d_v \in [0, n_v]$: depth of the node from the root of the sentence

We enforce that the depth of a child is greater than that of the parent:

$$(\forall_u)\ d_v + (1 + n_v) \cdot (1 - x_{uv}) \geq 1 + d_u \tag{1}$$

Connectedness

To ensure that every word is reachable from a root, a word must have at least one parent. Together with acyclicity, this enforces reachability.

$$\sum_u x_{uv} \geq 1 \tag{2}$$

Consistency of Directionalized Links

Links with same label type are encouraged to be oriented in the same way. We introduce variables to represent whether links with label L are allowed to go left or right.

$$r_L, \ell_L \in \{0, 1\}$$

We introduce slack variables s_{ij} to allow some links to go in disallowed directions with a penalty.

$s_{ij} \in \mathbb{R} \geq 0$

N_L: number of link tokens with label L

$$x_{ij} \leq r_L + s_{ij} \qquad x_{ji} \leq \ell_L + s_{ij} \tag{3}$$

$$objective = \min \left(\sum_L r_L + \ell_L \right) \frac{N_L}{4} + \sum_{ij} s_{ij} \tag{4}$$

Data Sets

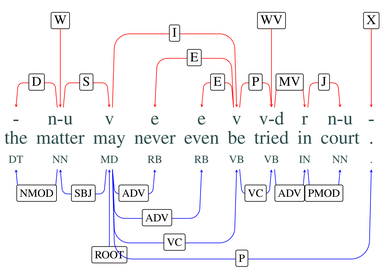
Data Sets taken from:

CoNLL 2007 Shared Task (English)

ACL 2013 Shared Task of Machine Translation (Russian)

	Input Sentences	Output Connected Parses
English	18,577	10,960
Russian	18,577	4,913

1 Experiments and Results



Bottom half from CoNLL 2007 shared task. Top half is our directionalized link parse.

Multiheadedness

On the English data Set, the link data has 8% additional edges over the CoNLL. (average about 2 multiheaded words per sentence)

CoNLL Matches

52% of links match CoNLL arcs

57% of CoNLL arcs match links

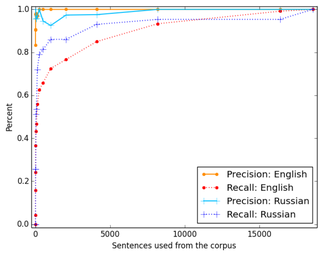
Directionality

6.19% of link types allowed both directions

2.07% of link tokens required disallowed direction via slack

Stability of Results

To see whether the recovered direction mapping might be unstable and sensitive to the input corpus, we compared results of increasing runs of sentences.



The direction mappings obtained on small datasets have high *precision* relative to the one obtained on the largest dataset. Their *recall* grows as more link types are seen and directionalized.

Backwards Subject-Verb Links

In our directionalized corpus subjects point to verbs instead of verbs pointing to subjects. This is due to a possible inconsistency of the Link Grammar discovered by our method.

- Link Grammar seems to be inconsistent about whether the auxiliary verb or the main verb is the head of a clause.
- Governing verb links to either auxilliary or main, depending on the clause type, but governing verbs usually link to subject when there is one. This makes subject a consistent choice to make head of a clause. To fix, we can edit the link grammar, link parses, or the ILP.

Conclusions

- Link Grammar parses can be oriented into connected DAGs
- A new corpus for building multi-headed dependency parsers.
- ILP can be used to help annotate missing data in corpora.

References

[1] Tobias Achterberg. SCIP: Solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.

[2] Thorsten Koch. *Rapid Mathematical Programming*. PhD thesis, Technische Universität Berlin, 2004. ZIB-Report 04-58.

[3] Sujith Ravi and Kevin Knight. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of ACL-IJCNLP*, pages 504–512. Association for Computational Linguistics, 2009.

[4] Daniel Sleator and Davy Temperley. Parsing English with a link grammar. Computer Science Technical Report CMU-CS-91-196, Carnegie Mellon University, October 1991.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1423276. The work was mainly conducted while the first author was at Johns Hopkins University.