

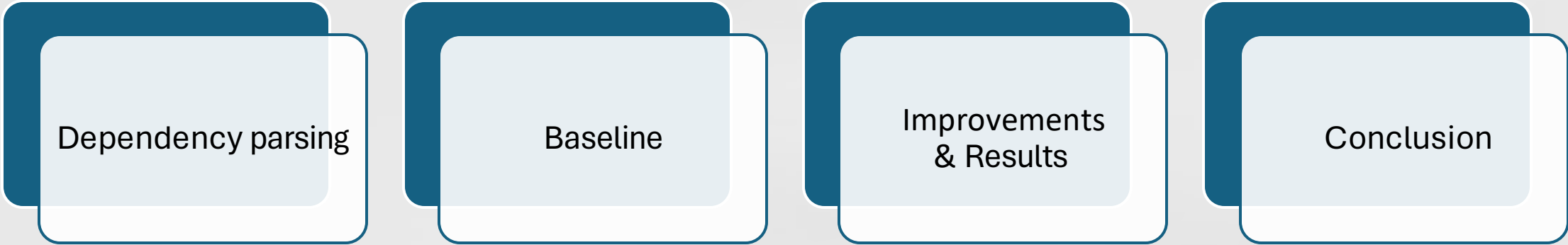


Beams are all you need

Evaluating Beam Search for Dependency Parsing

G15

Agenda



Dependency parsing

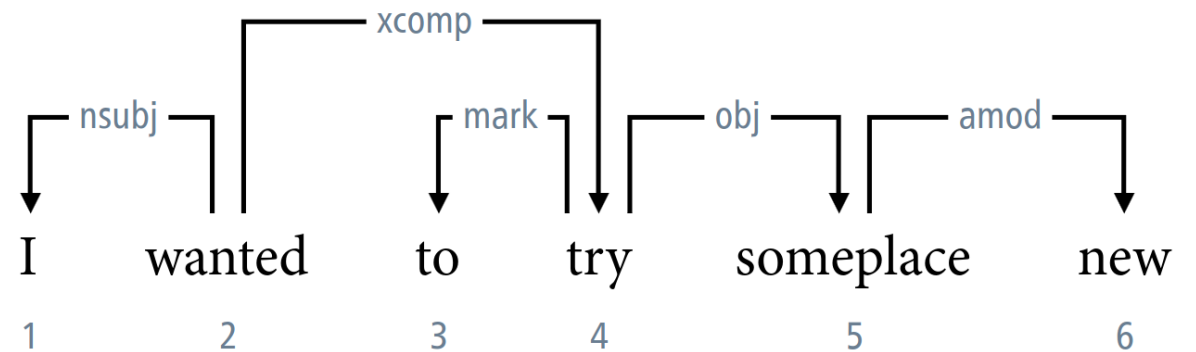
Baseline

Improvements
& Results

Conclusion

Dependency parsing

- What is it?
- What is it used for?



word position	1	2	3	4	5	6
head position	2	0	4	2	4	5
dependency relation	nsubj	root	mark	xcomp	obj	amod

Baseline



Tagger

Assigns tags to words, ex. Noun, verb

Fixed window model



Parser

Determines syntactic structure

Arc-standard algorithm

Fixed window model



Dataset – Universal Dependencies

English Web Treebank

Swedish_LinES

Projectivize

Improvements

Implemented

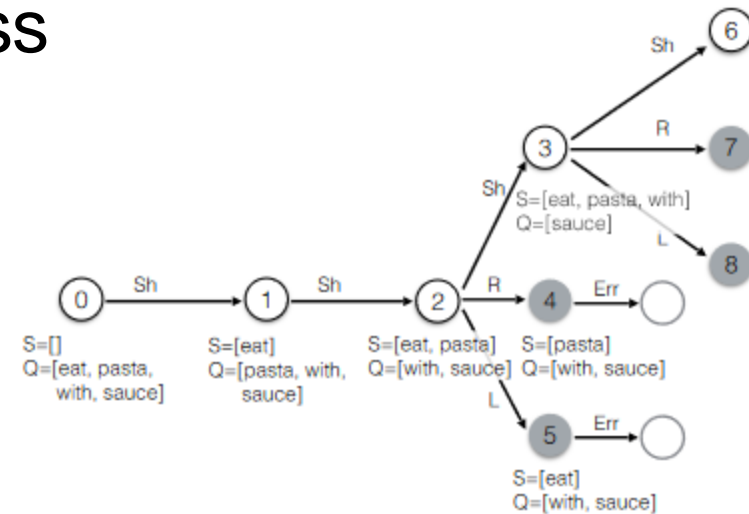
- Beam search
 - Small improvement
 - Slow
- Error states

Not implemented

- Best-first beam search
- Globalized model

Error states

- Beam search suffers from locality in the scoring, how do scores from one step relate to the next?
- Vaswani et al (2016) suggest introducing error states during training
- Main idea is to occupy probability mass for features with incorrect heads.
- Error state not used during prediction



Results vaswani et al (2016)

Best-first beam search

Meister et al 2020

Based on A*

Priority queue of beams

- Expand beam with highest priority, not timestep
- Priority is highest scoring hypothesis
- Prioritize promising beams

Same result

10x faster

Local VS. Global model

- Daniel Andor et al suggest a globalized model

Method	WSJ		Union-News		Union-Web		Union-QTB	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
Martins et al. (2013)*	92.89	90.55	93.10	91.13	88.23	85.04	94.21	91.54
Zhang and McDonald (2014)*	93.22	91.02	93.32	91.48	88.65	85.59	93.37	90.69
Weiss et al. (2015)	93.99	92.05	93.91	92.25	89.29	86.44	94.17	92.06
Alberti et al. (2015)	94.23	92.36	94.10	92.55	89.55	86.85	94.74	93.04
Our Local (B=1)	92.95	91.02	93.11	91.46	88.42	85.58	92.49	90.38
Our Local (B=32)	93.59	91.70	93.65	92.03	88.96	86.17	93.22	91.17
Our Global (B=32)	94.61	92.79	94.44	92.93	90.17	87.54	95.40	93.64
Parsey McParseface (B=8)	-	-	94.15	92.51	89.08	86.29	94.77	93.17

Results

Baseline UAS score: 0.6993 for eng dataset, 0.7283 for swe dataset (Parser)

- **Beam-16:** Improved scores: 0.7055 & 0.7399
- **Error state:** Lowered scores: 0.6977 & 0.7383

Baseline UAS score: 0.6569 for eng dataset, 0.6683 for swe dataset (Parser and Tagger)

- **Beam-16:** Improved scores: 0.6639 & 0.6760
- **Error state:** Lowered scores: 0.6542 & 0.6755

Beam search: ca 1% improvement from baseline

- **State errors:** no improvement

Research Literature: UAS score of 0.7696 for baseline and 0.8135 with beam search

Future Improvements

Analysis of result

- Our baseline was not the exact same as paper.
- We did not have the exact same dataset, model and baseline accuracy as reference papers.
- Could have done more testing on different datasets to further validate results.

Further analysis and conclusions

- Project shows that beam search can improve UAS accuracy over greedy searches.
- Locality in predictions should be handled when using beam search
- Error states might be more suitable for models with many features, ours has 6.

System	UAS
Local-14-rand (beam 1)	90.96
Local-14-rand (beam 4)	91.21 (+0.25)
ErrSt-14-rand (beam 1)	90.83
ErrSt-14-rand (beam 4)	91.98 (+1.15)
ErrSt-25-rand	92.29



Thank you for your attention

Questions?