Supervised sentiment analysis: Stanford Sentiment Treebank

Christopher Potts

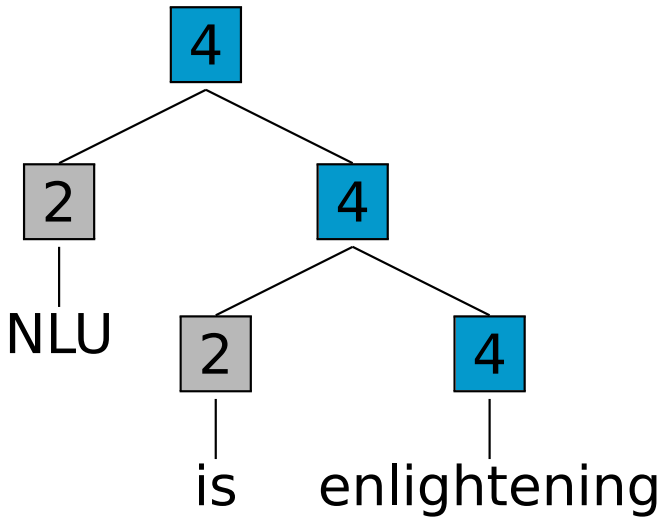
Stanford Linguistics

CS224u: Natural language understanding

SST project overview

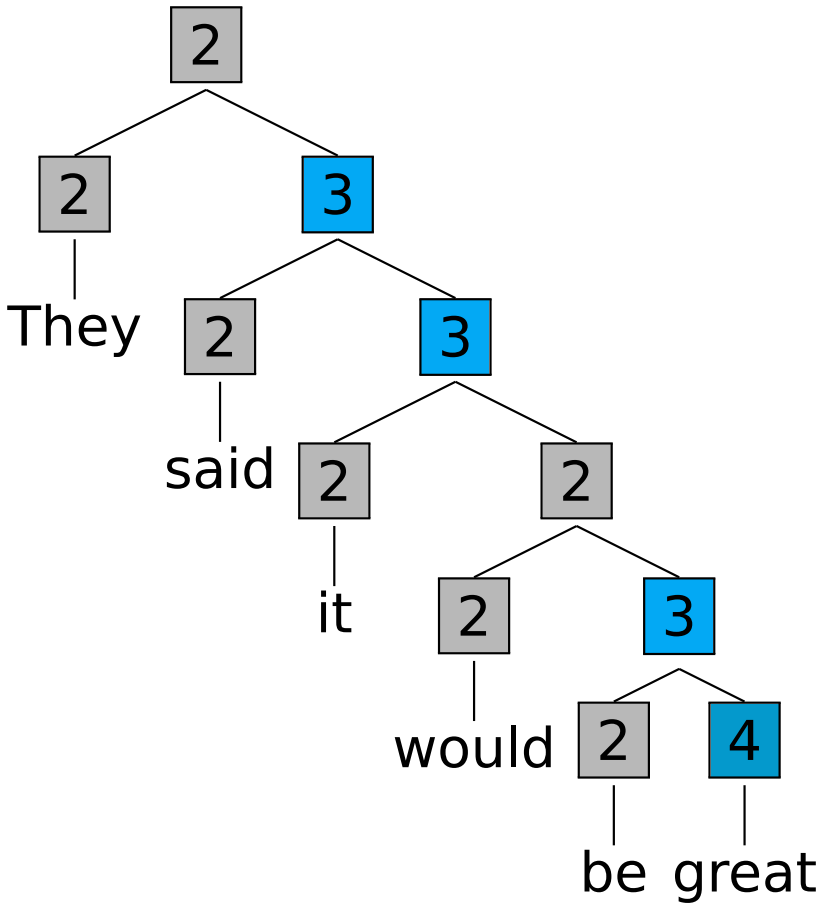
1. Socher et al. (2013)
2. Full code and data release: https://nlp.stanford.edu/sentiment/
3. Sentence-level corpus (10,662 sentences)
4. Original data from Rotten Tomatoes (Pang and Lee 2005)
5. Fully-labeled trees (crowdsourced labels)
6. The 5-way labels were extracted from workers' slider responses.

Fully labeled trees



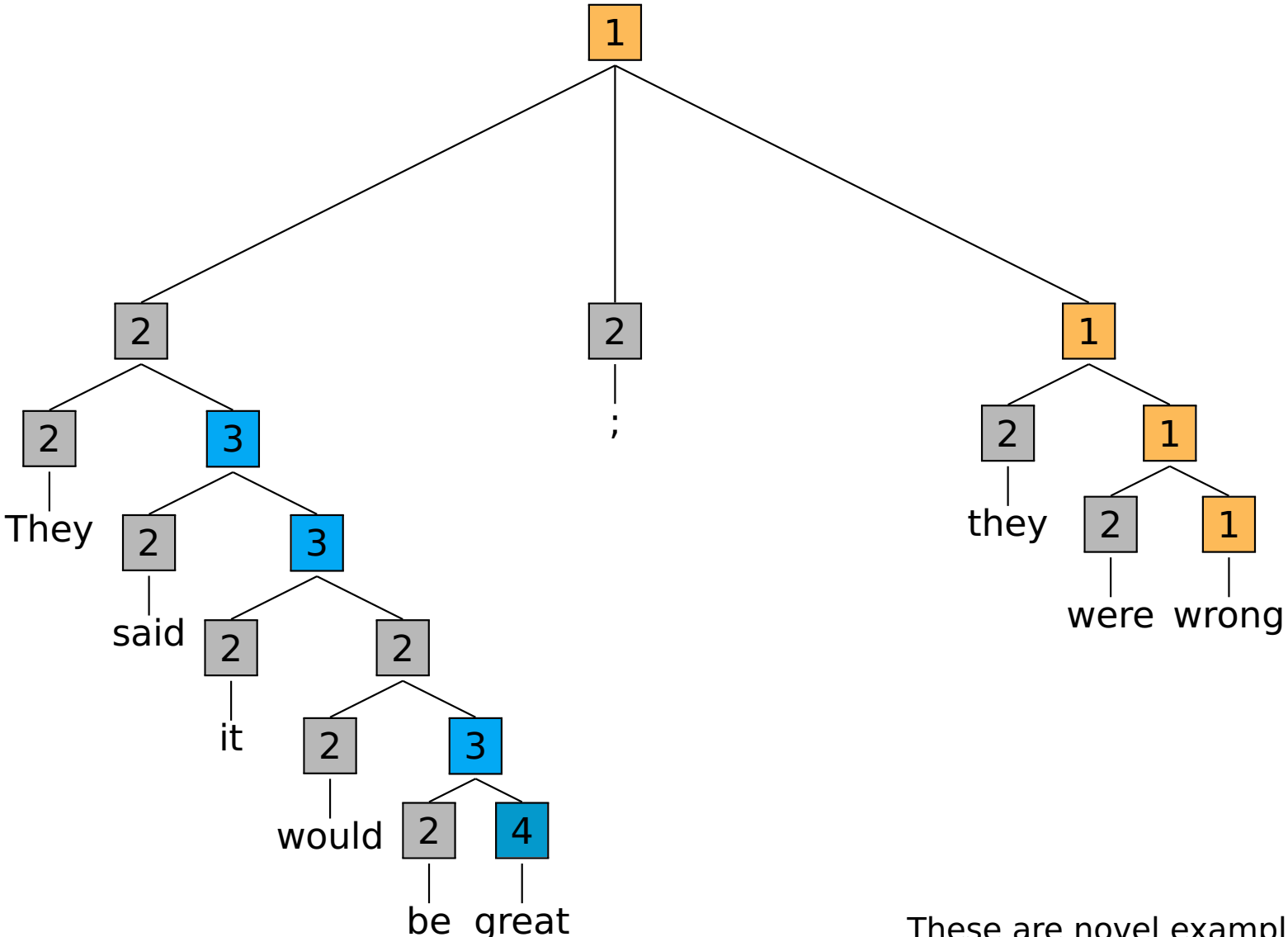
These are novel examples,
and the labels are actual output from
https://nlp.stanford.edu/sentiment/

Fully labeled trees



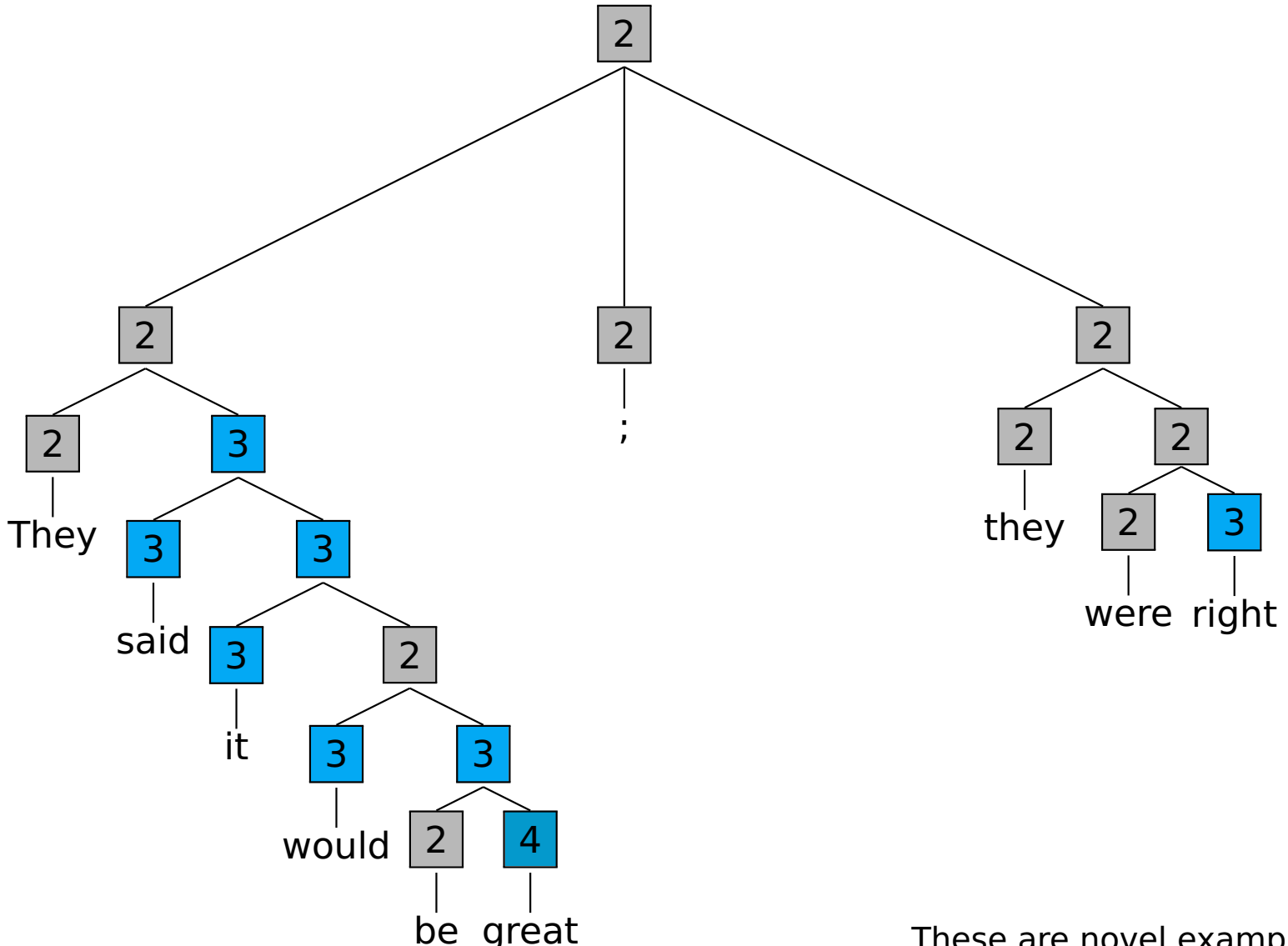
These are novel examples,
and the labels are actual output from
https://nlp.stanford.edu/sentiment/

Fully labeled trees



These are novel examples,
and the labels are actual output from
https://nlp.stanford.edu/sentiment/

Fully labeled trees



These are novel examples,
and the labels are actual output from
https://nlp.stanford.edu/sentiment/

Root-level tasks

Five-way problem

Label	Meaning	Train	Dev
0	very negative	1,092	139
1	negative	2,218	289
2	neutral	1,624	229
3	positive	2,322	279
4	very positive	1,288	165
		8,544	1,101

Note: 4 > 3 (more positive) but 0 > 1 (more negative)

Root-level tasks

Five-way problem

Label	Meaning	Train	Dev
0	very negative	1,092	139
1	negative	2,218	289
2	neutral	1,624	229
3	positive	2,322	279
4	very positive	1,288	165
		8,544	1,101

Note: 4 > 3 (more positive) but 0 > 1 (more negative)

Ternary problem

Label	Meaning	Train	Dev
0, 1	negative	3,310	428
2	neutral	1,624	229
3, 4	positive	3,610	444
		8,544	1,101

Root-level tasks

Five-way problem

Label	Meaning	Train	Dev
0	very negative	1,092	139
1	negative	2,218	289
2	neutral	1,624	229
3	positive	2,322	279
4	very positive	1,288	165
		8,544	1,101

Note: 4 > 3 (more positive) but 0 > 1 (more negative)

Binary problem (neutral data simply excluded)

Label	Meaning	Train	Dev
0, 1	negative	3,310	428
3, 4	positive	3,610	444
		6,920	872

All-nodes tasks

Five-way problem

Label	Meaning	Train	Dev
0	very negative	40,774	5,217
1	negative	82,854	10,757
2	neutral	58,398	8,227
3	positive	89,308	11,001
4	very positive	47,248	6,245
		318,582	41,447

Note: 4 > 3 (more positive) but 0 > 1 (more negative)

All-nodes tasks

Five-way problem

Label	Meaning	Train	Dev
0	very negative	40,774	5,217
1	negative	82,854	10,757
2	neutral	58,398	8,227
3	positive	89,308	11,001
4	very positive	47,248	6,245
		318,582	41,447

Note: 4 > 3 (more positive) but 0 > 1 (more negative)

Ternary problem

Label	Meaning	Train	Dev
0, 1	negative	123,628	15,974
2	neutral	58,398	8,227
3, 4	positive	136,556	17,246
		318,582	41,447

All-nodes tasks

Five-way problem

Label	Meaning	Train	Dev
0	very negative	40,774	5,217
1	negative	82,854	10,757
2	neutral	58,398	8,227
3	positive	89,308	11,001
4	very positive	47,248	6,245
		318,582	41,447

Note: 4 > 3 (more positive) but 0 > 1 (more negative)

Binary problem (neutral data simply excluded)

Label	Meaning	Train	Dev
0, 1	negative	123,628	15,974
3, 4	positive	136,556	17,246
		260,184	33,220

Train/dev/test scenarios

Train

Full examples and/or subphrases with or without repeats:

NLU is enlightening	positive
is enlightening	positive
enlightening	positive
is	neutral
NLU	neutral
Not enlightening	negative
enlightening	positive
Not	negative

Dev/test

Full sentences only.

Additional details and analyses

https://github.com/cgpotts/cs224u/blob/master/sst_01_overview.ipynb

References I

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, MI. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA. Association for Computational Linguistics.