A probabilistic generative model for an intermediate constituency-dependency representation

Federico Sangati (f.sangati@uva.nl)
ILLC, University of Amsterdam
Institute for Logic, Language and Computation
Netherlands Organisation for Scientific Research

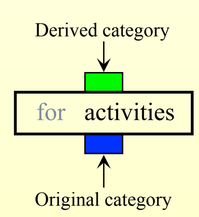
Abstract: We present a probabilistic model extension to the Tesnire Dependency Structure (TDS) framework formulated in (Sangati and Mazza, 2009). This representation incorporates aspects from both constituency and dependency theory. In addition, it makes use of junction structures to handle coordination constructions. We test our model on parsing the English Penn WSJ treebank using a re-ranking framework. This technique allows us to efficiently test our model without needing a specialized parser, and to use the standard evaluation metric on the original Phrase Structure version of the treebank.

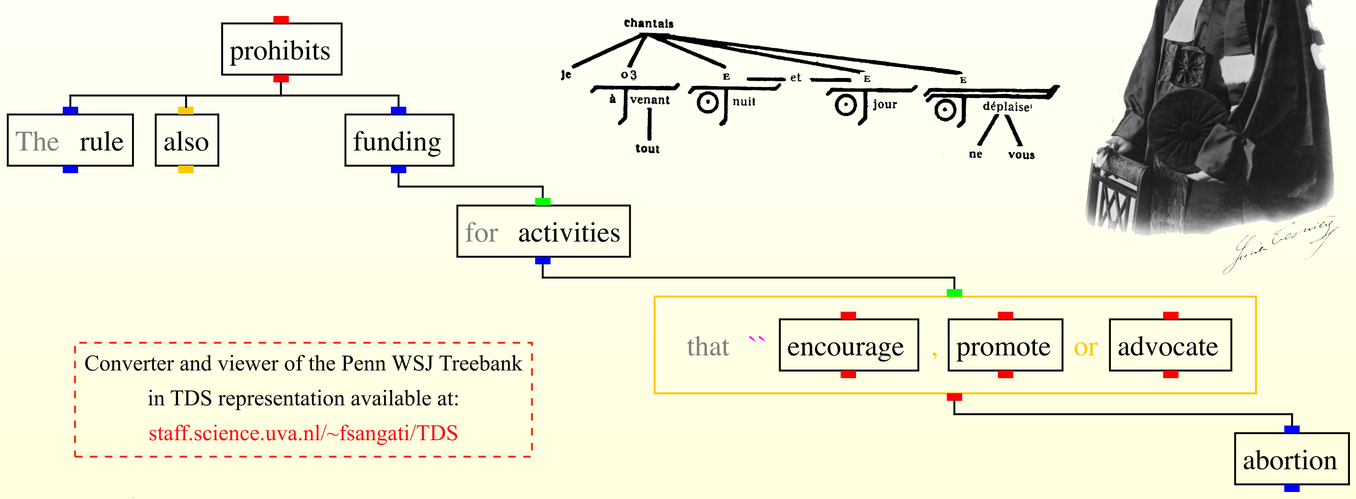
The TDS representation: Tesnière Dependency Structures

Words:
- Content words (activities, encourage, ...)
- Functional words (the, that, for, ...)
- Punctuation words (, `` " ...)

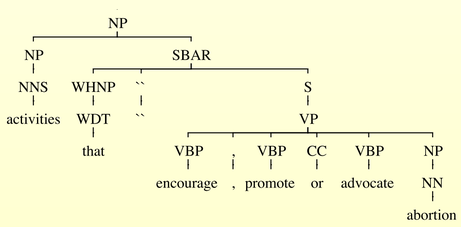
Blocks:
- Standard blocks
- Junction blocks

Categories:
- Verbs
- Adverbs
- Nouns
- Adjectives

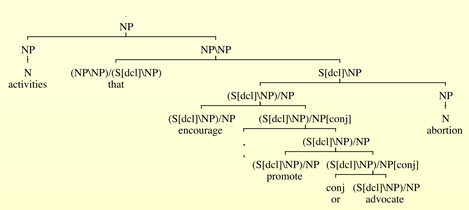




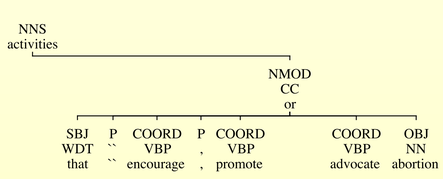
Converter and viewer of the Penn WSJ Treebank in TDS representation available at: staff.science.uva.nl/~fsangati/TDS



Original Phrase Structure (PS)



CCG Bank Representation



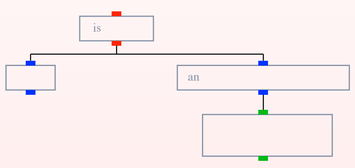
Dependency Structure (CONLL07)

A Probabilistic Generative Model for TDS

The Model: 3 Generative Processes + PoS-Tagging & Chunking

1. Block Generation

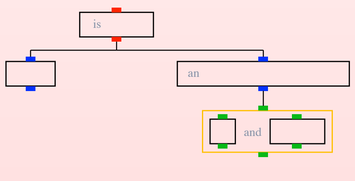
$$P_{BGM}(S) = \prod_{B \in dependentBlocks(S)} P(B|parent(B), direction(B), leftSibling(B))$$



	B	Parent	Direction	Left Sibling
	VER, VER, is	-	-	-
	NOU, NOU	VER	Left	-
	STOP	VER	Left	NOU, NOU
	STOP	VER	Inner	-
	NOU, NOU, an	VER	Right	-
	STOP	VER	Right	NOU, NOU, an
	STOP	NOU	Left	-
	STOP	NOU	Right	-
	STOP	NOU	Inner	-
	STOP	NOU	Left	-
	ADJ, ADJ	NOU	Inner	-
	STOP	NOU	Inner	ADJ, ADJ
	STOP	NOU	Right	-
3×	STOP	ADJ	Left	-
3×	STOP	ADJ	Inner	-
3×	STOP	ADJ	Right	-

2. Block Expansion

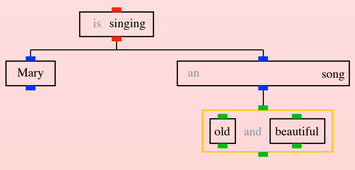
$$P_{BEM}(S) = \prod_{B \in blocks(S)} P(elements(B)|derivedCat(B))$$



Elements(B)	Derived Cat
VER	VER
NOU	NOU
an NOU	NOU
ADJ and ADJ	ADJ
ADJ	ADJ
ADJ	ADJ

3. Words Filling

$$P_{WFM}(S) = \prod_{B \in standardBlocks(S)} P(cw(B)|cw(parent(B)), cats(B), fw(B), dir(B))$$



cw(B)	cw(parent(B)	cats(B)	fw(B)	direction
VBG	-	VER	is	-
NNP	VBG	NOU	-	Left
NN	VBG	NOU	an	Right
JJ	NN	ADJ	-	Left
JJ	NN	ADJ	-	Left

4. PoS-Tagging & Chunking

$$P_{T\&C}(S) = \prod_{i=1}^{i=|sentenceLength|} P(tag(i)|word_i, word_{i-1}, word_{i-2})$$

Words	Mary	is	singing	an	old	and	beautiful	song
PoS-tags	NNP	AUX	VBG	DT	JJ	CC	JJ	NN
Chunk tags	N	N	I	N	-N	C	N	+I

PoS	$word_i$	pos_{i-1}	pos_{i-2}
NNP	Mary	-	-
AUX	is	NNP	-
VBG	singing	AUX	NNP
DT	an	VBG	AUX
JJ	old	DT	VBG
⋮	⋮	⋮	⋮

Chunk	pos_i	pos_{i-1}	pos_{i-2}
N	NNP	-	-
N	AUX	NNP	-
I	VBG	AUX	NNP
N	DT	VBG	AUX
-N	JJ	DT	VBG
⋮	⋮	⋮	⋮

$$P(S) = P_{BGM}(S) \cdot P_{BEM}(S) \cdot P_{WFM}(S) \cdot P_{T\&C}(S)$$

Parsing through re-ranking

Training Phase:
- Decompose each TDS structure in the training corpus into events.
- Keep track of the frequency of each event and conditioning context (history).

Testing Phase:
- Get n-best PS candidates for each test sentence from an other parser.
- Convert each candidate into TDS and decompose it into events.
- Compute the probability of each candidate, and output the one with maximum probability.

Results

Treebank: Penn WSJ

Training/Test: sec 02-21 / sec 22

n-best candidates: Charniak's Max-Ent parser

New Evaluation metrics
- Block Detection Score (BDS): the accuracy of detecting the correct boundaries of the blocks in the structure.
- Block Attachment Score (BAS): the accuracy of detecting the correct governing block of each block in the structure.
- Junction Detection Score (JDS): the accuracy of detecting the correct list of content-words composing each junction block in the structure.

		F-Score	UAS	BDS	BAS	JDS
Charniak	$(n = 1)$	89.4	92.5	95.0	89.5	77.6
PCFG-reranker	$(n = 5)$	89.0	92.4	**95.1**	89.2	77.5
PCFG-reranker	$(n = 1000)$	83.5	88.4	92.9	83.6	71.8
TDS-reranker	$(n = 5)$	**89.6**	92.5	**95.2**	89.5	77.6
TDS-reranker	$(n = 10)$	89.1	92.3	95.1	89.0	76.7
TDS-reranker	$(n = 100)$	86.9	91.0	94.4	87.0	73.1
TDS-reranker	$(n = 1000)$	84.8	89.3	93.5	84.9	69.7