A List of heuristic rules

Table 1 lists the methods used to identify and locate the signal tokens based on the annotated signal labels and the gold lexical and syntactic parse provided in Penn Treebank (Marcus et al., 1993). Certain signal labels are supplemented with optional *meta comment* that specifies the signal in the particular relation.

Signals of the relation are identified from the *elementary discourse units* (EDUs) related by the relation. As a result, a word-level annotation of the discourse signals is obtained. For example:

Tokenized text

In the first year , the bank eliminated 800 jobs .
Now it says it will trim more in the next year .

RST Treebank

Relation Index	Nucleus	Satellite	Sense
R1	word 1-11	word 12-23	Temporal
...	...	...	...

RST Signaling Corpus

Signal Index	Relation Index	Signal label	Meta comment
S1	R1	(13) discourse marker	now
S2	R1	(12) tense	
S3	R1	(25) lexical chain	first year – next year
...	...	...	...

Using the heuristics in Table 1, the signal tokens are identified (the underlined words below):

In the <u>first year</u> , the bank <u>eliminated</u> 800 jobs.
<u>Now</u> it <u>says</u> it <u>will trim</u> more in the <u>next year</u> .

Each of the signal tokens is tagged by the corresponding signal index, leading to below word-level annotation:
- - 3 3 - - - 2 - - - 1 - 2 - 2 2 - - - 3 3 -
and the relation boundaries and senses can be retrieved with reference to the RST Signaling Corpus, and, in turn, the RST Discourse Treebank.

The converted annotation is available on http://www.coli.uni-saarland.de/~frances/contents/rstsignal/rstsignal.html.

References

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Category	Signal label		Method
			identify POS/syntactic patterns:
	1	relative clause	onset of SBAR, e.g. *that*
	2	infinitive clause	POS patterns: TO VB
	3	present participial clause	first VBG in VP
	4	past participial clause	first VBN in VP
	5	imperative clause	manually identify imperative verb
Syntactic	6	nominal modifier	POS patterns: TO VB
	7	adjective modifier	POS patterns: TO VB
	8*	parrallel syntactic construction	–
	9	subject-auxillary inversion	manually identify subject-auxillary inversion
	10*	interrupted matrix	–
	11	reported speech	identify a list of verbs, e.g. *said,according to*
Morphological	12	tense	identify verbs in different tenses
Discoure marker	13	one of 201 defined markers	identify labelled discourse marker e.g. *when,in addition to*
			identify tokens/phrases specified in meta comment:
Reference	14	personal reference	e.g. *Johnson – she*
	15	comparative reference	e.g. *equal*
	16	demonstrative reference	e.g. *these*
	17	propositional reference	e.g. *naming a candidate – it*
			identify tokens/phrases specified in meta comment:
Lexical	18	alternative expression	e.g. *what's more*
	19	indicative word	e.g. *compared with*
			identify tokens/phrases specified in meta comment:
	20	synonomy	e.g. *United Airlines – UAL*
	21	antonymy	e.g. *short–long*
	22	meronymy	e.g. *Californians–Johnson*
Semantic	23	repetition	the word / phrase that is repeated
	24	indicative word pair	e.g. *asked–replied*
	25	lexical chain	e.g. *company–sold–shares–holdings* unspecific comments, e.g. *a few lexical chains*, are excluded
	26	general word	e.g. *issues*
Numerical	27	same count	identify CD
	28	colon	identify *:*
	29	semi-colon	identify *;*
Graphical	30	dash	identify - or –
	31	comma	identify *,*
	32	parentheses	identify *-LRB-*,*-RRB-*, *-LSB-*,*-RSB-*,*-LCB-*,*-RCB-*
	33*	items in sequence	–
	34*	newspaper style attribution	
	35*	newspaper style contrast	
Genre	36*	newspaper style elaboration	–
	37*	newspaper layout	
	38*	newspaper style definition	
	39*	inverted pyramid cheme	
Unsure	40*	no signals identified	–

Table 1: The heuristics used to identify and locate the signal tokens based on the annotated signal labels. Signals marked with * are excluded in analysis.