Chapter 22:
Discourse Coherence

Miyamae Yuichi

Table of Contents

Table of Contents

0. Introduction

- Discourse is a coherent structured group of sentences.
  - chapters on SLP
  - news articles
  - conversations
  - threads on social media

- Coherence is the relationship between sentences that makes discourses different than random assemblages.
  - local coherence
  - global coherence

0. Introduction: local coherence

1. Sentences or clauses in real discourses are related to nearby sentences in systematic ways.

John took a train from Paris to Istanbul. He likes spinach.

?

John took a train from Paris to Istanbul. He had to attend a conference.

REASON (coherence relations)

0. Introduction: local coherence

2. entity-based coherence: some entities are salient, and the focus doesn't swing between multiple entities.

   John wanted to buy a piano for his living room.
   Jenny also wanted to buy a piano.
   He went to the piano store.
   It was nearby.
   The living room was on the second floor.
   She didn't find anything she liked.
   The piano he bought was hard to get up to that floor.

- Centering Theory
- the entity grid model

0. Introduction: local coherence

3. topically coherent: nearby sentences are about the same topic and use same or similar vocabulary.

- Topically coherent discourses exhibit lexical cohesion: the sharing of identical words in nearby sentences

Before winter I built a **chimney**, and shingled the sides of my **house**...

I have thus a tight shingled and plastered **house**... with a **garret** and a **closet**, a large **window** on each side....

0. Introduction: global coherence

- Many genres of text are associated with particular conventional discourse structures.

genre	conventional discourse structures
academic articles	sections about Methodology or Results
stories	conventional plotlines or motifs
essays	a particular claim a structured set of premises

0. Introduction

- Coherence detection can measure the quality of a text.
  - essay grading / essay quality measurement
  - summarization
  - symptoms of schizophrenia (mental health)

Table of Contents

1. Coherence Relations

<u>John took a train from Paris to Istanbul</u>. <u>He likes spinach</u>.

?

<u>John took a train from Paris to Istanbul</u>. <u>He had to attend a conference</u>.

REASON (coherence relations)

- Rhetorical Structure Theory (RST)
- Penn Discourse TreeBank (PDTB)

1-1. Rhetorical Structure Theory (RST)

- RST relations are defined between two spans of text, generally a nucleus and a satellite.

Reason

- The nucleus is an action carried by an agent
- The satellite is the reason for the nucleus

[$_{NUC}$ Jane took a train from Paris to Istanbul.] [$_{SAT}$ She had to attend a conference.]

1-1. Rhetorical Structure Theory (RST)

Elaboration

- The satellite gives additional information or detail about the nucleus

[$_{NUC}$ Dorothy was from Kansas.] [$_{SAT}$ She lived in the midst of the great Kansas prairies.]

Evidence

- The satellite gives additional information or detail about the nucleus to convince the reader to accept them

[$_{NUC}$ Kevin must be here.] [$_{SAT}$ His car is parked outside.]

1-1. Rhetorical Structure Theory (RST)

Attribution

- The satellite gives the source of attribution for the nucleus

[$_{SAT}$ Analysts estimated] [$_{NUC}$ that sales at U.S. stores declined in the quarter, too]

List

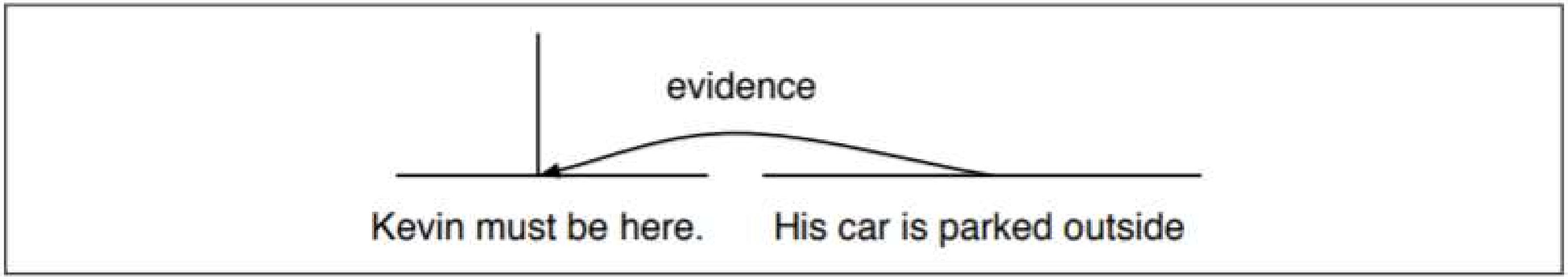
- A series of nuclei is given without contrast or comparison

[$_{NUC}$ Billy Bones was the mate;] [$_{NUC}$ Long John, he was quartermaster]

1-1. Rhetorical Structure Theory (RST)

- RST relations are traditionally represented graphically.

[$_{\text{NUC}}$ Kevin must be here.] [$_{\text{SAT}}$ His car is parked outside.]



1-1. Rhetorical Structure Theory (RST)

- The coherence of a larger text can be captured by the hierarchical structure between coherence relations.

With its distant orbit–50 percent farther from the sun than Earth–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure.

1-1. Rhetorical Structure Theory (RST)

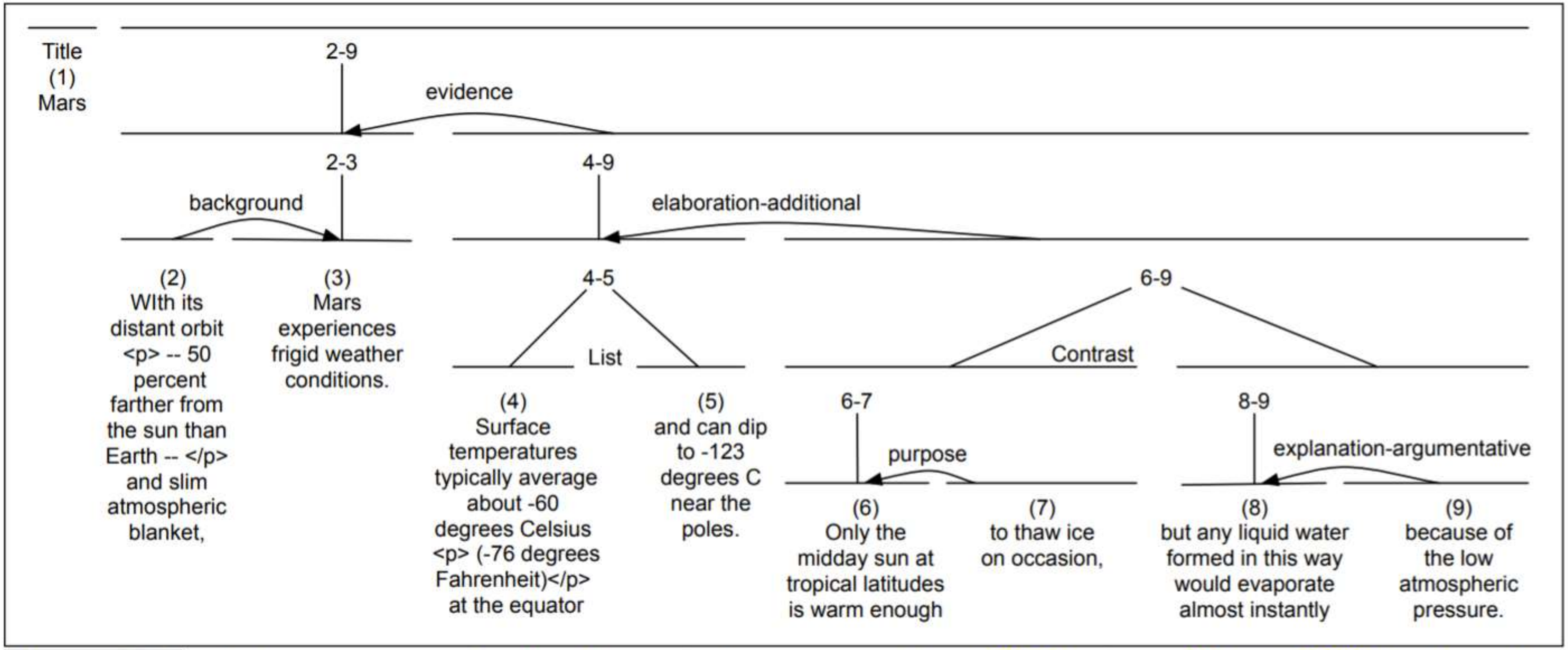


Figure 22.1 A discourse tree for the *Scientific American* text in (22.12), from Marcu (2000a). Note that asymmetric relations are represented with a curved arrow from the satellite to the nucleus.

1-1. Rhetorical Structure Theory (RST)

With its distant orbit–50 percent farther from the sun than Earth–and slim atmospheric blanket, | Mars experiences frigid weather conditions. | Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator | and can dip to -123 degrees C near the poles. | Only the midday sun at tropical latitudes is warm enough | to thaw ice on occasion, | but any liquid water formed in this way would evaporate almost instantly | because of the low atmospheric pressure.

- elementary discourse units (EDUs)
  - text spans of a sentence, clause or phrase
  - also called as discourse segments

1-2. Penn Discourse TreeBank (PDTB)

- PDTB labeling is lexically grounded.
  - Annotators were given words that signal discourse relations
  - discourse connectives: because, when, since, as a result, etc.

Arg1

Jewelry displays in department stores were often cluttered and uninspired. And the merchandise was, well, fake. **As a result, marketers of faux gems steadily lost space in department stores to more fashionable rivals—cosmetics makers.**

Arg2

1-2. Penn Discourse TreeBank (PDTB)

In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. (implicit=as a result) **By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.**

- "as a result" signals a CAUSAL sense.
- "since" signals a CAUSAL or a TEMPORAL sense.

Connective	Senses
since	reason (94), succession (78), succession-reason (10), other (2)

1-2. Penn Discourse TreeBank (PDTB)

Class	Type	Example
TEMPORAL	SYNCHRONOUS	The parishioners of St. Michael and All Angels stop to chat at the church door, as members here always have. (Implicit while) In the tower, five men and women pull rhythmically on ropes attached to the same five bells that first sounded here in 1614.
CONTINGENCY	REASON	Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda. (implicit=because) **As a former White House aide who worked closely with Congress, he is savvy in the ways of Washington.**
COMPARISON	CONTRAST	The U.S. wants the removal of what it perceives as barriers to investment; Japan denies there are real barriers.
EXPANSION	CONJUNCTION	Not only do the actors stand outside their characters and make it clear they are at odds with them, but they often literally stand on their heads.

Figure 22.2 The four high-level semantic distinctions in the PDTB sense hierarchy

Temporal
- Asynchronous
- Synchronous (Precedence, Succession)

Comparison
- Contrast (Juxtaposition, Opposition)
- *Pragmatic Contrast (Juxtaposition, Opposition)*
- Concession (Expectation, Contra-expectation)
- *Pragmatic Concession*

Contingency
- Cause (Reason, Result)
- Pragmatic Cause (Justification)
- *Condition (Hypothetical, General, Unreal Present/Past, Factual Present/Past)*
- *Pragmatic Condition (Relevance, Implicit Assertion)*

Expansion
- *Exception*
- Instantiation
- Restatement (Specification, Equivalence, Generalization)
- Alternative (Conjunction, Disjunction, Chosen Alternative)
- List

Figure 22.3 The PDTB sense hierarchy. There are four top-level classes, 16 types, and 23 subtypes (not all types have subtypes). 11 of the 16 types are commonly used for implicit argument classification; the 5 types in italics are too rare in implicit labeling to be used.

1-2. Penn Discourse TreeBank (PDTB)

- PDTB does not annotate anything above the span level
  - It does not commit to higher-level discourse structure
  - In contrast, RST integrates these pairwise coherence relations into a global tree structure

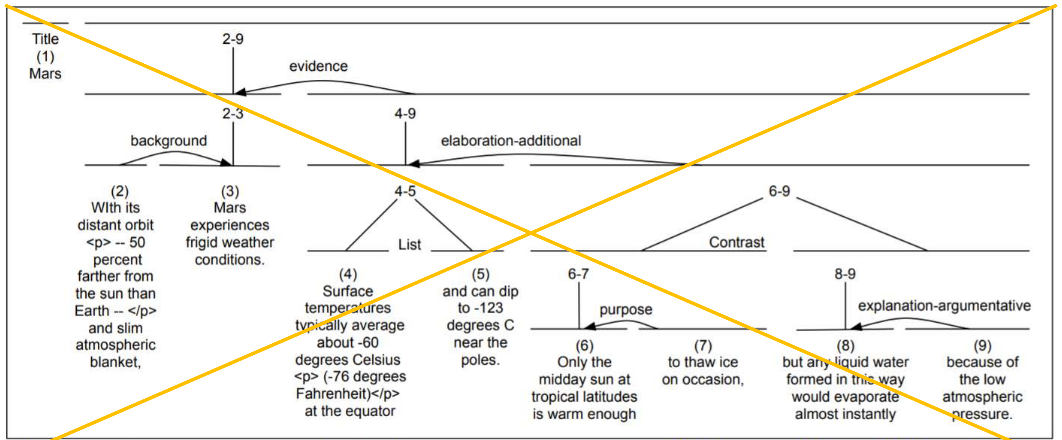


Figure 22.1 A discourse tree for the *Scientific American* text in (22.12), from Marcu (2000a). Note that asymmetric relations are represented with a curved arrow from the satellite to the nucleus.

Table of Contents

2. Discourse Structure Parsing

- discourse parsing: automatically determine the coherence relations given a sequence of sentences

2-1. EDU segmentation for RST parsing

[Mr. Rambo says]$_{e1}$ [that a 3.2-acre property]$_{e2}$ [overlooking the San Fernando Valley]$_{e3}$ [is priced at $4 million]$_{e4}$ [because the late actor Erroll Flynn once lived there.]$_{e5}$

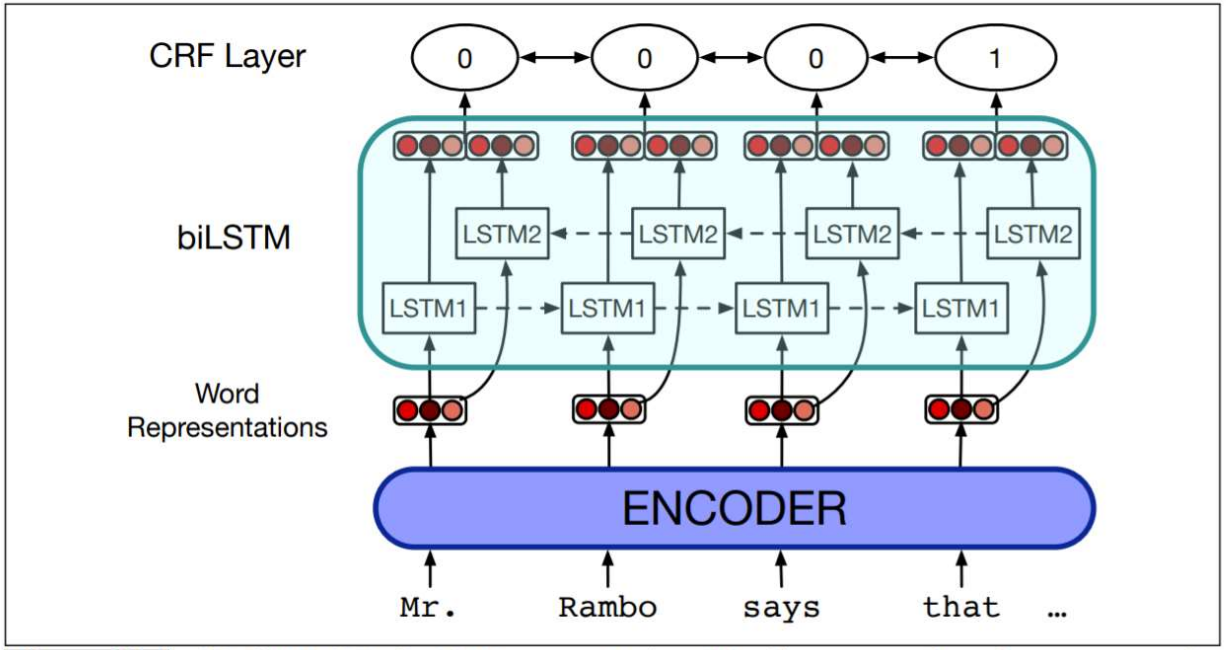


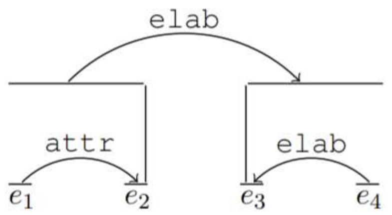
Figure 22.4 biLSTM-CRF for EDU segmentation. Word inputs can draw from any encoder for contextual embeddings like BERT.

2-2. RST parsing

- shift-reduce parser
  - Which consists of a stack, a queue and actions

- Actions include:
  - shift: pushes the first EDU in the queue onto the stack
  - reduce(l, d): merges the top two subtrees on the stack
    - l: the coherence relation label (e.g. attribution, elaboration)
    - d: the nuclearity direction ($d \in \{NN, NS, SN\}$)
  - pop root: removes the final tree from the stack

2-2. RST parsing



e_1: American Telephone & Telegraph Co. said it
e_2: will lay off 75 to 85 technicians here , effective Nov. 1.
e_3: The workers install , maintain and repair its private branch exchanges,
e_4: which are large intracompany telephone networks.

Figure 22.5 Example RST discourse tree, showing four EDUs. Figure from Yu et al. (2018).

Step	Stack	Queue	Action	Relation
1	$\varnothing$	e_1, e_2, e_3, e_4	SH	$\varnothing$
2	e_1	e_2, e_3, e_4	SH	$\varnothing$
3	e_1, e_2	e_3, e_4	RD(attr,SN)	$\varnothing$
4	$e_{1:\mathbf{2}}$	e_3, e_4	SH	$\widehat{e_1\mathbf{e_2}}$
5	$e_{1:\mathbf{2}}$, e_3	e_4	SH	$\widehat{e_1\mathbf{e_2}}$
6	$e_{1:\mathbf{2}}$, e_3, e_4	$\varnothing$	RD(elab,NS)	$\widehat{e_1\mathbf{e_2}}$
7	$e_{1:\mathbf{2}}$, $e_{\mathbf{3}:4}$	$\varnothing$	RD(elab,SN)	$\widehat{e_1\mathbf{e_2}}, \widehat{\mathbf{e_3}e_4}$
8	$e_{1:4}$	$\varnothing$	PR	$\widehat{e_1\mathbf{e_2}}, \widehat{\mathbf{e_3}e_4}, \widehat{e_{1:2}\mathbf{e_{3:4}}}$

Figure 22.6 Parsing the example of Fig. 22.5 using a shift-reduce parser. Figure from Yu et al. (2018).

2-2. RST parsing

encoder

- input: a sentence $w_1, w_2, \dots, w_m$

  ➡ word representation $x_1^w, x_2^w, \dots, x_m^w$

  ➡ $h_1^w, h_2^w, \dots, h_m^w = biLSTM(x_1^w, x_2^w, \dots, x_m^w)$

  ➡ $x^e = \dfrac{1}{t-s+1} \sum_{k=s}^{t} h_k^w \quad \begin{cases} w_s, w_{s+1}, \dots, w_t \\ \rightarrow x_s^w, x_{s+1}^w, \dots, x_t^w \\ \rightarrow h_s^w, h_{s+1}^w, \dots, h_t^w \end{cases}$

  ➡ $h_1^e, h_2^e, \dots, h_n^e = biLSTM(x_1^e, x_2^e, \dots, x_n^e)$

- output: EDU representations $h_1^e, h_2^e, \dots, h_n^e$

2-2. RST parsing

decoder

- input: a concatenation of the top 3 subtrees on the stack (s_0, s_1, s_2) plus the first EDU in the queue (q_0)

Step	Stack	Queue	Action
5	$\boxed{e_{1:2}, e_3}$	$\boxed{e_4}$	SH
6	$\boxed{e_{1:2}, e_3, e_4}$	$\boxed{\varnothing}$	RD (elab,NS)

$$o = W\left(h^t_{s0}, h^t_{s1}, h^t_{s2}, h^e_{q0}\right)$$

(W is a feedforward network)

$$\left(h^t_s = \frac{1}{j-i+1}\sum_{k=i}^{j} h^e_k\right) \quad \left(e_{1:2}\text{: } h^t_s = \frac{1}{2}\left(h^e_1 + h^e_2\right)\right)$$

- output: an action o

2-2. RST parsing

- The standard cross-entropy loss (with l_2 regularization) is used to train the system

$$p_a = \frac{\exp(o_a)}{\sum_{a' \; in \; A} \exp(o_{a'})}$$

$$L_{CE}(\,) = -\log(p_a) + \frac{\lambda}{2}||\Theta||^2$$

2-3. PDTB discourse parsing

- This is sometimes called shallow discourse parsing
  - Because the task just involves flat relationships between spans

1. Find the discourse connectives (disambiguation)
2. Find the two spans for each connective
3. Label the relationship between these spans
4. Assign a relation between every adjacent pair of sentences

2-3. PDTB discourse parsing

1. Find the discourse connectives (disambiguation)
   - disambiguate connectives from their non-discourse use

○ Selling picked up as previous buyers bailed out of their positions <u>and</u> aggressive short sellers—anticipating further declines—moved in.

× My favorite colors are blue <u>and</u> green.

- Recent work performs the task end-to-end from word inputs using a biLSTM-CRF with BIO outputs (B-CONN, I-CONN, O)

2-3. PDTB discourse parsing

2. Find the two spans for each connective
   - use the same sequence models used to find RST EDUs

2-1. EDU segmentation for RST parsing

[Mr. Rambo says]$_{e1}$ [that a 3.2-acre property]$_{e2}$ [overlooking the San Fernando Valley]$_{e3}$ [is priced at $4 million]$_{e4}$ [because the late actor Erroll Flynn once lived there.]$_{e5}$

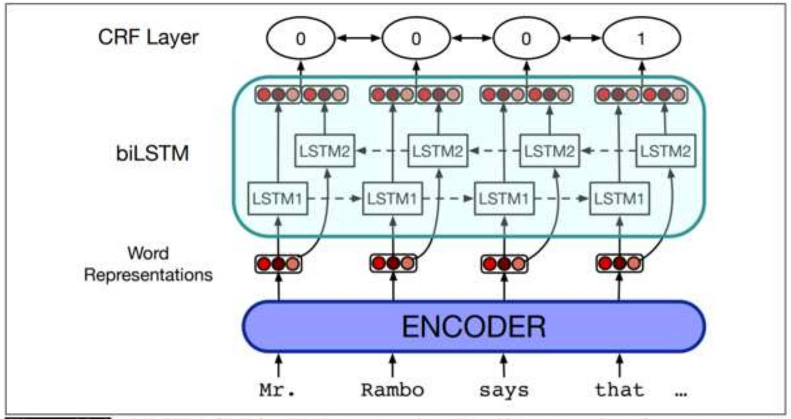


Figure 22.4 biLSTM-CRF for EDU segmentation. Word inputs can draw from any encoder for contextual embeddings like BERT.

2-3. PDTB discourse parsing

4. Assign a relation between every adjacent pair of sentences
    - input: a pair of adjacent sentences
    - output: a coherence relation sense label (one of the 11 second-level PDTB tags or none)

1. Represent each of the two spans by BERT embeddings and take the last layer hidden state corresponding to [CLS]
2. Pass this through a single layer tanh feedforward network and then a softmax for sense classification

Temporal
- Asynchronous
- Synchronous (Precedence, Succession)

Comparison
- Contrast (Juxtaposition, Opposition)
- *Pragmatic Contrast (Juxtaposition, Opposition)*
- Concession (Expectation, Contra-expectation)
- *Pragmatic Concession*

Contingency
- Cause (Reason, Result)
- Pragmatic Cause (Justification)
- *Condition (Hypothetical, General, Unreal Present/Past, Factual Present/Past)*
- *Pragmatic Condition (Relevance, Implicit Assertion)*

Expansion
- *Exception*
- Instantiation
- Restatement (Specification, Equivalence, Generalization)
- Alternative (Conjunction, Disjunction, Chosen Alternative)
- List

Table of Contents

3. Centering and Entity-Based Coherence

- entity-based coherence
  - Centering Theory
  - the entity grid model

3-1. Centering

- One of the entities in the discourse is salient.
- Discourses maintaining the same salient entity, e.g.

  a. John went to his favorite music store to buy a piano.
  b. He had frequented the store for many years.
  c. He was excited that he could finally buy a piano.
  d. He arrived just as the store was closing for the day.

  are more coherent than those don't, e.g.

  a. John went to his favorite music store to buy a piano.
  b. It was a store John had frequented for many years.
  c. He was excited that he could finally buy a piano.
  d. It was closing just as John arrived.

3-1. Centering

- backward-looking center $C_b(U_n)$
  - the current salient entity
  - $C_b(U_n) = C_p(U_{n-1})$

- forward-looking centers $C_f(U_n)$
  - a set of potential future salient entities

    a. John went to his favorite music store to buy a piano.
  - ranked according to factors like discourse salience and grammatical role (e.g. subjects > objects > others)
  - $C_p(U_n)$: prefferred center, the highest-ranked C_f

3-1. Centering

	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Figure 22.7 Centering Transitions for Rule 2 from Brennan et al. (1987).

Rule 1: If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.

Rule 2: Transition states are ordered. Continue is preferred to Retain is preferred to Smooth-Shift is preferred to Rough-Shift.

→ Continue > Retain > Smooth-Shift > Rough-Shift

3-1. Centering

	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

U_1: John went to his favorite music store to buy a piano.

$C_f(U_1)$ = {John, store, piano}; $C_p(U_1)$ = John; $C_b(U_1)$ = undefined

U_2: He was excited that he could finally buy a piano.

$C_f(U_2)$ = {John, piano}; $C_p(U_2)$ = John; $C_b(U_2)$ = John

Continue ($C_b(U_2) = C_p(U_2)$; $C_b(U_1)$ = undefined)

U_3: He arrived just as the store was closing for the day.

$C_f(U_3)$ = {John, store}; $C_p(U_3)$ = John; $C_b(U_3)$ = John

Continue ($C_b(U_3) = C_p(U_3)$; $C_b(U_3) = C_b(U_2)$)

U_4: It was closing just as John arrived.

$C_f(U_4)$ = {store, John}; $C_p(U_4)$ = store; $C_b(U_4)$ = John

Retain ($C_b(U_4) \neq C_p(U_4)$; $C_b(U_4) = C_b(U_3)$)

3-1. Centering

	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

U_1: <u>John went to his favorite music store to buy a piano.</u>

$C_f(U_1)$ = {John, store, piano}; $C_p(U_1)$ = John; $C_b(U_1)$ = undefined

U_2: <u>It was a store John had frequented for many years.</u>

$C_f(U_2)$ = {store, John}; $C_p(U_2)$ = store; $C_b(U_2)$ = John

Retain ($C_b(U_2) \neq C_p(U_2)$; $C_b(U_1)$ = undefined)

U_3: <u>He was excited that he could finally buy a piano.</u>

$C_f(U_3)$ = {John, piano}; $C_p(U_3)$ = John; $C_b(U_3)$ = undefined

Rough-Shift ($C_b(U_3) \neq C_p(U_3)$; $C_b(U_3) \neq C_b(U_2)$)

U_4: <u>It was closing just as John arrived.</u>

$C_f(U_4)$ = {store, John}; $C_p(U_4)$ = store; $C_b(U_4)$ = John

Retain ($C_b(U_4) \neq C_p(U_4)$; $C_b(U_3)$ = undefined)

3-2. Entity Grid model

- The entity grid model uses machine learning to induce the patterns of entity that make a discourse more coherent.

1 [The Justice Department]$_s$ is conducting an [anti-trust trial]$_o$ against [Microsoft Corp.]$_x$ with [evidence]$_x$ that [the company]$_s$ is increasingly attempting to crush [competitors]$_o$.
2 [Microsoft]$_o$ is accused of trying to forcefully buy into [markets]$_x$ where [its own products]$_s$ are not competitive enough to unseat [established brands]$_o$.
3 [The case]$_s$ revolves around [evidence]$_o$ of [Microsoft]$_s$ aggressively pressuring [Netscape]$_o$ into merging [browser software]$_o$.
4 [Microsoft]$_s$ claims [its tactics]$_s$ are commonplace and good economically.
5 [The government]$_s$ may file [a civil suit]$_o$ ruling that [conspiracy]$_s$ to curb [competition]$_o$ through [collusion]$_x$ is [a violation of the Sherman Act]$_o$.
6 [Microsoft]$_s$ continues to show [increased earnings]$_o$ despite [the trial]$_x$.

Figure 22.9 A discourse with the entities marked and annotated with grammatical functions. Figure from Barzilay and Lapata (2008).

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	s	o	s	x	o	–	–	–	–	–	–	–	–	–	–	1
2	–	–	o	–	–	x	s	o	–	–	–	–	–	–	–	2
3	–	–	s	o	–	–	–	–	s	o	o	–	–	–	–	3
4	–	–	s	–	–	–	–	–	–	–	–	s	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	s	o	–	5
6	–	x	s	–	–	–	–	–	–	–	–	–	–	–	o	6

s (subject) o (object)
x (neither) - (absent)

3-2. Entity Grid model

- Coherence is measured by patterns of local entity transition
  - For example, the transition of “Department” from sentence 1 to sentence 2 is [s -]
  - The transition [s -] occurs 6 times out of 75 transitions (0.08%)

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

3-2. Entity Grid model

- Coherence is measured by patterns of local entity transition

	S S	S O	S X	S –	O S	O O	O X	O –	X S	X O	X X	X –	– S	– O	– X	– –
d_1	.01	.01	0	.08	.01	0	0	.09	0	0	0	.03	.05	.07	.03	.59
d_2	.02	.01	.01	.02	0	.07	0	.02	.14	.14	.06	.04	.03	.07	0.1	.36
d_3	.02	0	0	.03	.09	0	.09	.06	0	0	0	.05	.03	.07	.17	.39

Figure 22.10 A feature vector for representing documents using all transitions of length 2. Document d_1 is the text in Fig. 22.9. Figure from Barzilay and Lapata (2008).

  - These can be used as features for a machine learning model.
  - This model can be a text classifier trained with human-labeled scores (e.g. coherent or incoherent), which is expensive.

3-3. Evaluating Neural and Entity-based coherence

- Models are evaluated in one of two ways.
  1. We can have humans rate the coherence of a document and train a classifier to predict these human ratings.
  2. self-supervision: we pair up a natural discourse with a pseudo-document created by changing the ordering.

- Self-supervision has been implemented in 3 ways.
  1. sentence order discrimination task
     - compare a document to a random permutation
     - k documents → kn pairs of (original, permuted)

3-3. Evaluating Neural and Entity-based coherence

- Self-supervision has been implemented in 3 ways.

2. sentence insertion task

   - remove one of n sentences in a document
     → create $n-1$ copies
     → insert the removed sentence into each position
   - decide which document has the original ordering

3. sentence order reconstruction task

   - put the randomized sentences back in the correct order
   - k documents → kn pairs of (original, permuted)

Table of Contents

4. Representation learning models for local coherence

- The field was pioneered by a series of unsupervised models that made use of lexical cohesion
  1. lexical chains of words that occurred through a discourse
     - such as {pine, bush trees, trunk}
     - which were related in Roget's Thesaurus
     - the number and density of chain correlated with the topic

  2. TextTiling
     - compute the cosine between neighboring text spans (represented as vectors of raw word counts)
     - sentences or paragraph in a subtopic have high cosine

4. Representation learning models for local coherence

3. LSA Coherence method

- model the coherence between two sentences as the cosine between their LSA sentence embedding vectors

$$
\begin{aligned}
\mathrm{sim}(s,t) &= \cos(\mathbf{s},\mathbf{t}) \\
&= \cos\left(\sum_{w\in s}\mathbf{w}, \sum_{w\in t}\mathbf{w}\right)
\end{aligned}
$$

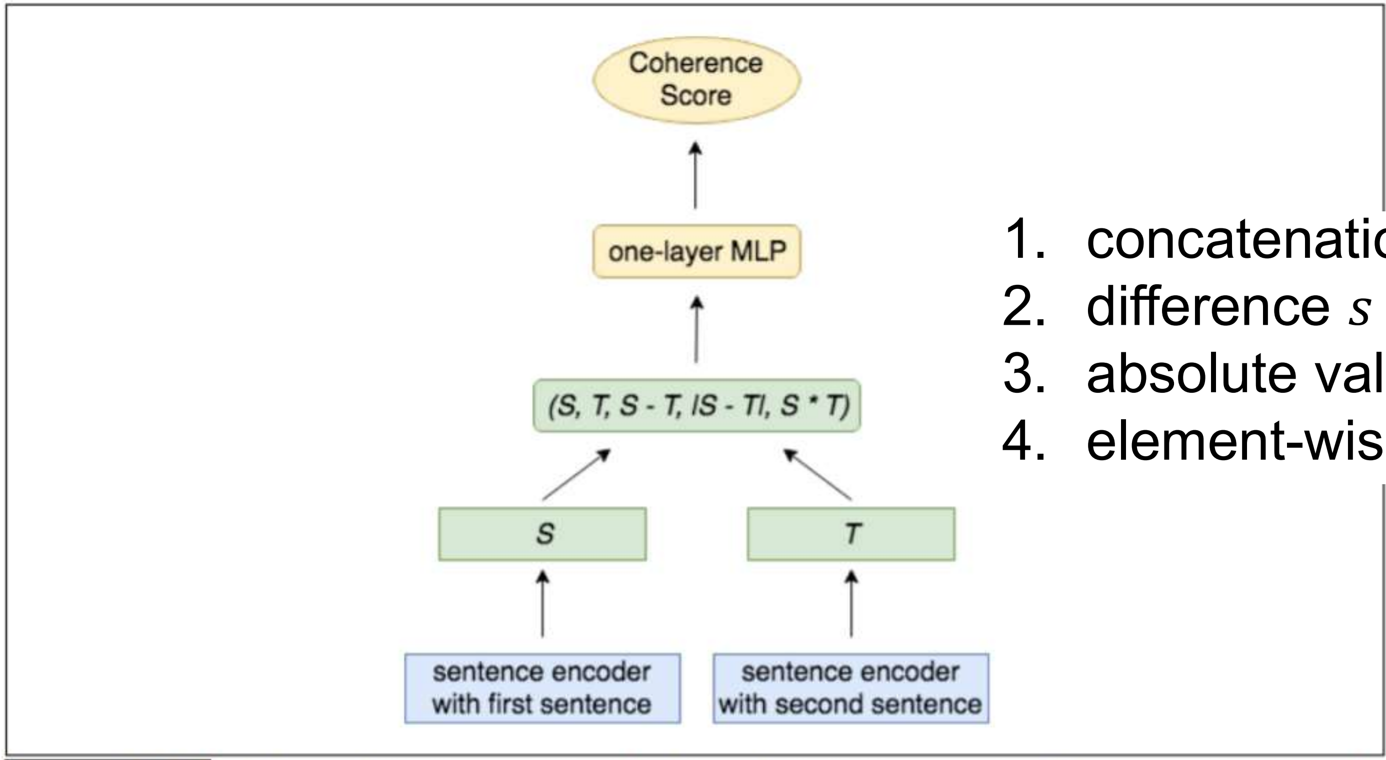
$$
\mathrm{coherence}(T) = \frac{1}{n-1}\sum_{i=1}^{n-1}\cos(s_i, s_{i+1})
$$

4. Representation learning models for local coherence

- local coherence discriminator (LCD)
  - compute the coherence of a text as the average of coherence scores between consecutive pairs of sentences
  - self-supervised model trained to discriminate consecutive sentence pairs (s_i, s_{i+1}) from incoherent pairs (s_i, s')

4. Representation learning models for local coherence

- local coherence discriminator



1. concatenation of s and t
2. difference $s - t$
3. absolute value of difference $|s - t|$
4. element-wise product $s \odot t$

Figure 22.11 The architecture of the LCD model of document coherence, showing the computation of the score for a pair of sentences s and t. Figure from Xu et al. (2019).

4. Representation learning models for local coherence

- local coherence discriminator

$$L_\theta = \sum_{d \in C} \sum_{s_i \in d} \mathop{\mathbb{E}}_{p(s'|s_i)} [L(f_\theta(s_i, s_{i+1}), f_\theta(s_i, s'))]$$

$$L(f^+, f^-) = \max(0, \eta - f^+ + f^-)$$

$$f^+ = f_\theta(s_i, s_{i+1})$$

$$f^- = f_\theta(s_i, s')$$

$$\eta = margin\ hyperparameter$$

Table of Contents

5. Global Coherence

- Propp models the discourse structure of Russian folktales
  - a set of character categories called dramatis personae
    e.g. Hero, Villain, Donor, Helper
  - a set of events called functions
    e.g. “Villain commits kidnapping”, “Donor tests Hero”, “Hero is pursued”

- Different tales choose different subsets of functions, but always in the same order.

5. Global Coherence

1. the structure of arguments
   - the way people attempt to convince each other in persuasive essays by offering claims and supporting premises

2. the structure of scientific papers
   - the way authors present their goals, results, and relationship to prior work in their papers

5-1. Argumentation Structure

- Analyzing people’s argumentation computationally is often called argumentation mining.

- Aristotle described three components of a good argument
  - pathos: appealing to the emotions of the listener
  - ethos: appealing to the speaker’s personal character
  - logos: the logical structure of the argument

- Most of the discourse structure studies focus on logos.
  - build and train on annotated datasets of persuasive essays etc.

5-1. Argumentation Structure

- Corpora often include annotations of argumentative components
  - claims: the central component of the argument
  - premises: the reasons to persuade the reader
  - argumentative relations: e.g. SUPPORT, ATTACK

"(1) **Museums and art galleries provide a better understanding about arts than Internet.** (2) In most museums and art galleries, detailed descriptions in terms of the background, history and author are provided. (3) Seeing an artwork online is not the same as watching it with our own eyes, as (4) the picture online does not show the texture or three-dimensional structure of the art, which is important to study."

  - SUPPORT(2, 1), SUPPORT(3, 1), SUPPORT(4, 3)

5-1. Argumentation Structure

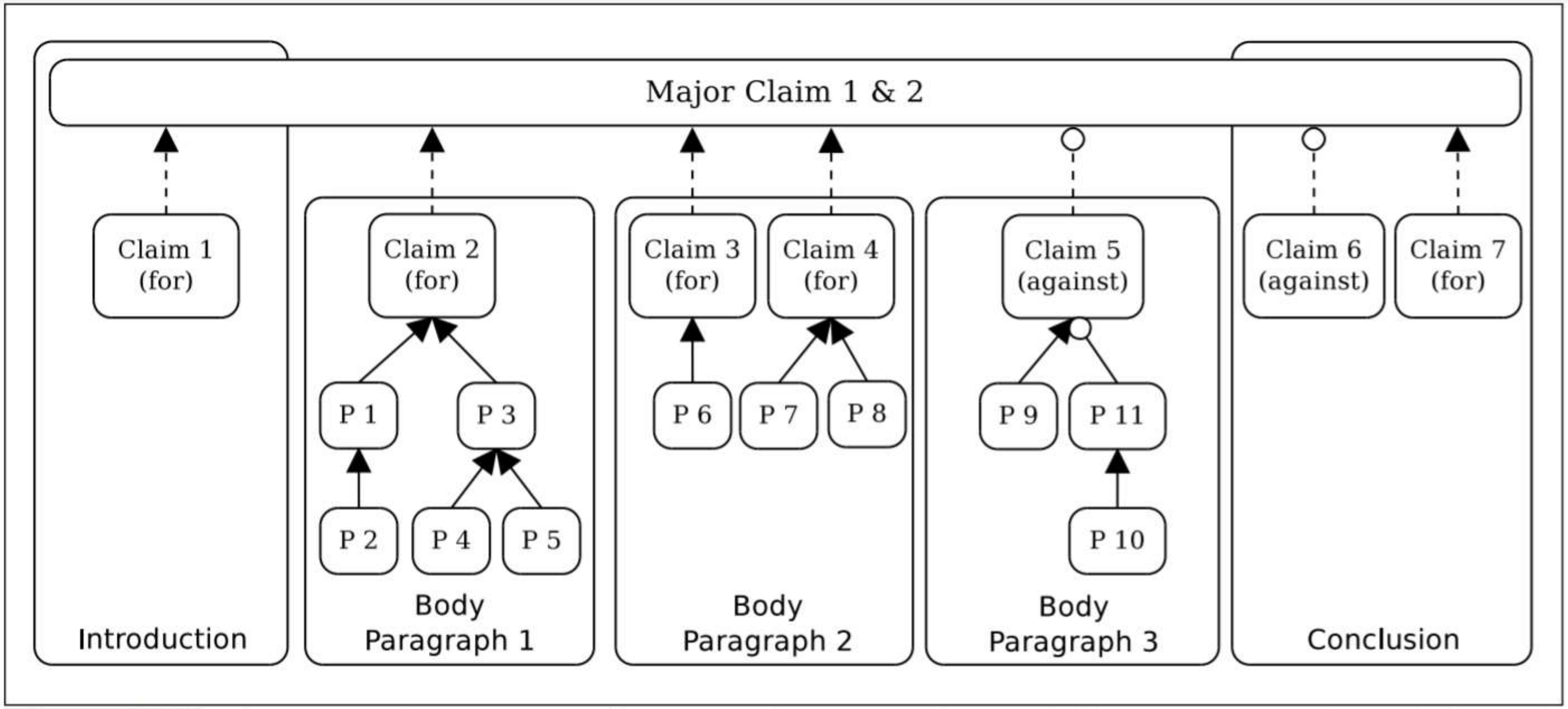


Figure 22.12 Argumentation structure of a persuasive essay. Arrows indicate argumentation relations, either of SUPPORT (with arrowheads) or ATTACK (with circleheads); P denotes premises. Figure from Stab and Gurevych (2017).

5-1. Argumentation Structure

- Algorithms for detecting argumentation structure include
  - classifiers for distinguishing claims, premises, or non-argumentation
  - classifiers for deciding if two spans have the SUPPORT, ATTACK, or neither relation

- Preliminary efforts on detecting argumentation schemes
  - argument from example
  - argument from cause to effect
  - argument from consequences

5-1. Argumentation Structure

- Aristotle’s ethos and pathos techniques are relevant in the detection of persuasion

- For example scholars have investigated the linguistic realization of features studied by social scientists like
  - reciprocity: people return favors
  - social proof: people follow others’ choices
  - authority: people are influenced by those with power
  - scarcity: people value things that are scarce

5-2. The structure of scientific discourse

- One popular annotation scheme is the argumentative zoning
  - Each scientific paper tries to make a knowledge claim about a new piece of knowledge being added to the field.

Category	Description	Example
AIM	Statement of specific research goal, or hypothesis of current paper	"The aim of this process is to examine the role that training plays in the tagging process"
OWN_METHOD	New Knowledge claim, own work: methods	"In order for it to be useful for our purposes, the following extensions must be made:"
OWN_RESULTS	Measurable/objective outcome of own work	"All the curves have a generally upward trend but always lie far below backoff (51% error rate)"
USE	Other work is used in own work	"We use the framework for the allocation and transfer of control of Whittaker...."
GAP_WEAK	Lack of solution in field, problem with other solutions	"Here, we will produce experimental evidence suggesting that this simple model leads to serious overestimates"
SUPPORT	Other work supports current work or is supported by current work	"Work similar to that described here has been carried out by Merialdo (1994), with broadly similar conclusions."
ANTISUPPORT	Clash with other's results or theory; superiority of own work	"This result challenges the claims of..."

Figure 22.13 Examples for 7 of the 15 labels from the Argumentative Zoning labelset (Teufel et al., 2009).

Table of Contents

6. Summary

In this chapter we introduced local and global models for discourse **coherence**.

- Discourses are not arbitrary collections of sentences; they must be *coherent*. Among the factors that make a discourse coherent are coherence relations between the sentences, entity-based coherence, and topical coherence.
- Various sets of **coherence relations** and **rhetorical relations** have been proposed. The relations in Rhetorical Structure Theory (**RST**) hold between spans of text and are structured into a tree. Because of this, shift-reduce and other parsing algorithms are generally used to assign these structures. The Penn Discourse Treebank (**PDTB**) labels only relations between pairs of spans, and the labels are generally assigned by sequence models.
- **Entity-based coherence** captures the intuition that discourses are **about** an entity, and continue mentioning the entity from sentence to sentence. **Centering Theory** is a family of models describing how salience is modeled for discourse entities, and hence how coherence is achieved by virtue of keeping the same discourse entities salient over the discourse. The **entity grid** model gives a more bottom-up way to compute which entity realization transitions lead to coherence.
- Many different genres have different types of **global coherence**. Persuasive essays have claims and premises that are extracted in the field of **argument mining**, scientific articles have structure related to aims, methods, results, and comparisons.