CHAPTER

22 Discourse Coherence

> And even in our wildest and most wandering reveries, nay in our very dreams, we shall find, if we reflect, that the imagination ran not altogether at adventures, but that there was still a connection upheld among the different ideas, which succeeded each other. Were the loosest and freest conversation to be transcribed, there would immediately be transcribed, there would immediately be observed something which connected it in all its transitions.
>
> David Hume, *An enquiry concerning human understanding*, 1748

Orson Welles' movie *Citizen Kane* was groundbreaking in many ways, perhaps most notably in its structure. The story of the life of fictional media magnate Charles Foster Kane, the movie does not proceed in chronological order through Kane's life. Instead, the film begins with Kane's death (famously murmuring *"Rosebud"*) and is structured around flashbacks to his life inserted among scenes of a reporter investigating his death. The novel idea that the structure of a movie does not have to linearly follow the structure of the real timeline made apparent for 20th century cinematography the infinite possibilities and impact of different kinds of coherent narrative structures.

But coherent structure is not just a fact about movies or works of art. Like movies, language does not normally consist of isolated, unrelated sentences, but instead of collocated, structured, **coherent** groups of sentences. We refer to such a coherent structured group of sentences as a **discourse**, and we use the word **coherence** to refer to the relationship between sentences that makes real discourses different than just random assemblages of sentences. The chapter you are now reading is an example of a discourse, as is a news article, a conversation, a thread on social media, a Wikipedia page, and your favorite novel.

What makes a discourse coherent? If you created a text by taking random sentences each from many different sources and pasted them together, would that be a coherent discourse? Almost certainly not. Real discourses exhibit both **local coherence** and **global coherence**. Let's consider three ways in which real discourses are locally coherent;

First, sentences or clauses in real discourses are related to nearby sentences in systematic ways. Consider this example from Hobbs (1979):

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

This sequence is incoherent because it is unclear to a reader why the second sentence follows the first; what does liking spinach have to do with train trips? In fact, a reader might go to some effort to try to figure out how the discourse could be coherent; perhaps there is a French spinach shortage? The very fact that hearers try to identify such connections suggests that human discourse comprehension involves the need to establish this kind of coherence.

By contrast, in the following coherent example:

(22.2)   Jane took a train from Paris to Istanbul. She had to attend a conference.

the second sentence gives a REASON for Jane's action in the first sentence. Structured relationships like REASON that hold between text units are called **coherence relations**, and coherent discourses are structured by many such coherence relations. Coherence relations are introduced in Section 22.1.

coherence relations

A second way a discourse can be locally coherent is by virtue of being "about" someone or something. In a coherent discourse some entities are **salient**, and the discourse focuses on them and doesn't go back and forth between multiple entities. This is called **entity-based coherence**. Consider the following incoherent passage, in which the salient entity seems to wildly swing from John to Jenny to the piano store to the living room, back to Jenny, then the piano again:

(22.3) 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.

Entity-based coherence models measure this kind of coherence by tracking salient entities across a discourse. For example **Centering Theory** (Grosz et al., 1995), the most influential theory of entity-based coherence, keeps track of which entities in the discourse model are salient at any point (salient entities are more likely to be pronominalized or to appear in prominent syntactic positions like subject or object). In Centering Theory, transitions between sentences that maintain the same salient entity are considered more coherent than ones that repeatedly shift between entities. The **entity grid** model of coherence (Barzilay and Lapata, 2008) is a commonly used model that realizes some of the intuitions of the Centering Theory framework. Entity-based coherence is introduced in Section 22.3.

Centering Theory

entity grid

Finally, discourses can be locally coherent by being **topically coherent**: nearby sentences are generally about the same topic and use the same or similar vocabulary to discuss these topics. Because topically coherent discourses draw from a single semantic field or topic, they tend to exhibit the surface property known as **lexical cohesion** (Halliday and Hasan, 1976): the sharing of identical or semantically related words in nearby sentences. For example, the fact that the words *house*, *chimney*, *garret*, *closet*, and *window*— all of which belong to the same semantic field— appear in the two sentences in (22.4), or that they share the identical word *shingled*, is a cue that the two are tied together as a discourse:

topically coherent

lexical cohesion

(22.4) 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....

In addition to the local coherence between adjacent or nearby sentences, discourses also exhibit **global coherence**. Many genres of text are associated with particular conventional discourse structures. Academic articles might have sections describing the Methodology or Results. Stories might follow conventional plotlines or motifs. Persuasive essays have a particular claim they are trying to argue for, and an essay might express this claim together with a structured set of premises that support the argument and demolish potential counterarguments. We'll introduce versions of each of these kinds of global coherence.

Why do we care about the local or global coherence of a discourse? Since coherence is a property of a well-written text, coherence detection plays a part in any

task that requires measuring the **quality** of a text. For example coherence can help in pedagogical tasks like essay grading or essay quality measurement that are trying to grade how well-written a human essay is (Somasundaran et al. 2014, Feng et al. 2014, Lai and Tetreault 2018). Coherence can also help for summarization; knowing the coherence relationship between sentences can help know how to select information from them. Finally, detecting incoherent text may even play a role in mental health tasks like measuring symptoms of schizophrenia or other kinds of disordered language (Ditman and Kuperberg 2010, Elvevåg et al. 2007, Bedi et al. 2015, Iter et al. 2018).

22.1 Coherence Relations

Recall from the introduction the difference between passages (22.5) and (22.6).

(22.5) Jane took a train from Paris to Istanbul. She likes spinach.

(22.6) Jane took a train from Paris to Istanbul. She had to attend a conference.

The reason (22.6) is more coherent is that the reader can form a connection between the two sentences, in which the second sentence provides a potential REASON for the first sentences. This link is harder to form for (22.5). These connections between text spans in a discourse can be specified as a set of **coherence relations**. The next two sections describe two commonly used models of coherence relations and associated corpora: Rhetorical Structure Theory (RST), and the Penn Discourse TreeBank (PDTB).

coherence relation

22.1.1 Rhetorical Structure Theory

The most commonly used model of discourse organization is **Rhetorical Structure Theory** (**RST**) (Mann and Thompson, 1987). In RST relations are defined between two spans of text, generally a **nucleus** and a **satellite**. The nucleus is the unit that is more central to the writer's purpose and that is interpretable independently; the satellite is less central and generally is only interpretable with respect to the nucleus. Some symmetric relations, however, hold between two nuclei.

RST
nucleus
satellite

Below are a few examples of RST coherence relations, with definitions adapted from the RST Treebank Manual (Carlson and Marcu, 2001).

Reason: The nucleus is an action carried out by an animate agent and the satellite is the reason for the nucleus.

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

Elaboration: The satellite gives additional information or detail about the situation presented in the nucleus.

(22.8) [$_{NUC}$ 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 situation presented in the nucleus. The information is presented with the goal of convince the reader to accept the information presented in the nucleus.

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

Attribution: The satellite gives the source of attribution for an instance of reported speech in the nucleus.

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

List: In this multinuclear relation, a series of nuclei is given, without contrast or explicit comparison:

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

RST relations are traditionally represented graphically; the asymmetric Nucleus-Satellite relation is represented with an arrow from the satellite to the nucleus:
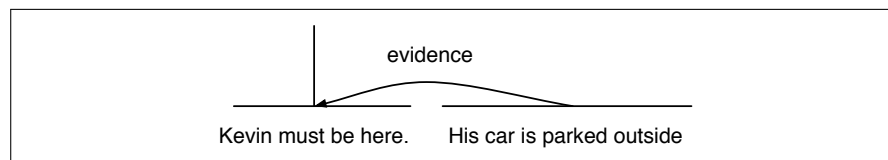
We can also talk about the coherence of a larger text by considering the hierarchical structure between coherence relations. Figure 22.1 shows the rhetorical structure of a paragraph from Marcu (2000a) for the text in (22.12) from the *Scientific American* magazine.

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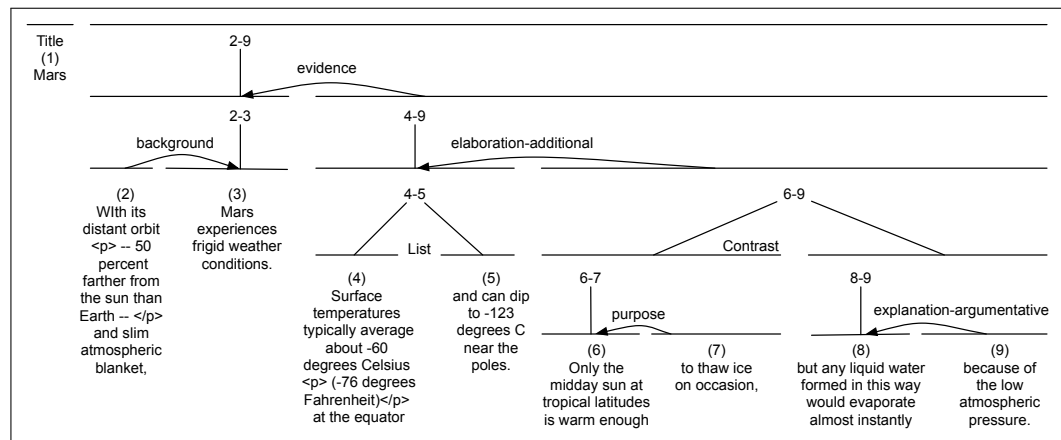
Figure 22.1 A discourse tree for the *Scientific American* text in (22.12), from Marcu (2000a). Note that asymmetric relations are represented with a curved arrow from the satellite to the nucleus.

The leaves in the Fig. 22.1 tree correspond to text spans of a sentence, clause or phrase that are called **elementary discourse units** or **EDU**s in RST; these units can also be referred to as **discourse segments**. Because these units may correspond to arbitrary spans of text, determining the boundaries of an EDU is an important task for extracting coherence relations. Roughly speaking, one can think of discourse

segments as being analogous to constituents in sentence syntax, and indeed as we'll see in Section 22.2 we generally draw on parsing algorithms to infer discourse structure.

There are corpora for many discourse coherence models; the RST Discourse TreeBank (Carlson et al., 2001) is the largest available discourse corpus. It consists of 385 English language documents selected from the Penn Treebank, with full RST parses for each one, using a large set of 78 distinct relations, grouped into 16 classes. RST treebanks exist also for Spanish, German, Basque, Dutch and Brazilian Portuguese (Braud et al., 2017).

Now that we've seen examples of coherence, we can see more clearly how a coherence relation can play a role in summarization or information extraction. For example, the nuclei of a text presumably express more important information than the satellites, which might be dropped in a summary.

22.1.2 Penn Discourse TreeBank (PDTB)

PDTB

The **Penn Discourse TreeBank** (**PDTB**) is a second commonly used dataset that embodies another model of coherence relations (Miltsakaki et al. 2004, Prasad et al. 2008, Prasad et al. 2014). PDTB labeling is *lexically grounded*. Instead of asking annotators to directly tag the coherence relation between text spans, they were given a list of **discourse connectives**, words that signal discourse relations, like *because*, *although*, *when*, *since*, or *as a result*. In a part of a text where these words marked a coherence relation between two text spans, the connective and the spans were then annotated, as in Fig. 22.13, where the phrase as a result signals a causal relationship between what PDTB calls *Arg1* (the first two sentences, here in italics) and **Arg2** (the third sentence, here in bold).

discourse connectives

(22.13) *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.**

(22.14) *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.**

Not all coherence relations are marked by an explicit discourse connective, and so the PDTB also annotates pairs of neighboring sentences with no explicit signal, like (22.14). The annotator first chooses the word or phrase that could have been its signal (in this case **as a result**), and then labels its sense. For example for the ambiguous discourse connective *since* annotators marked whether it is using a CAUSAL or a TEMPORAL sense.

The final dataset contains roughly 18,000 explicit relations and 16,000 implicit relations. Fig. 22.2 shows examples from each of the 4 major semantic classes, while Fig. 22.3 shows the full tagset.

Unlike the RST Discourse Treebank, which integrates these pairwise coherence relations into a global tree structure spanning an entire discourse, the PDTB does not annotate anything above the span-pair level, making no commitment with respect to higher-level discourse structure.

There are also treebanks using similar methods for other languages; (22.15) shows an example from the Chinese Discourse TreeBank (Zhou and Xue, 2015). Because Chinese has a smaller percentage of explicit discourse connectives than English (only 22% of all discourse relations are marked with explicit connectives,

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 <u>while</u>) 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=<u>because</u>) **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	<u>Not only</u> do the actors stand outside their characters and make it clear they are at odds with them, <u>but</u> 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.

compared to 47% in English), annotators labeled this corpus by directly mapping pairs of sentences to 11 sense tags, without starting with a lexical discourse connector.

(22.15) [**Conn** 为] [**Arg2** 推动图们江地区开发]，[Arg1 韩国捐款一百万美元设立了图们江发展基金]

"[In order to] [**Arg2** promote the development of the Tumen River region], [**Arg1** South Korea donated one million dollars to establish the Tumen River Development Fund]."

These discourse treebanks have been used for shared tasks on multilingual discourse parsing (Xue et al., 2016).

22.2 Discourse Structure Parsing

discourse parsing Given a sequence of sentences, how can we automatically determine the coherence relations between them? This task is often called **discourse parsing** (even though for PDTB we are only assigning labels to leaf spans and not building a full parse

tree as we do for RST).

22.2.1 EDU segmentation for RST parsing

RST parsing is generally done in two stages. The first stage, **EDU segmentation**, extracts the start and end of each EDU. The output of this stage would be a labeling like the following:

(22.16) [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.]$_{e5}$

Since EDUs roughly correspond to clauses, early models of EDU segmentation first ran a syntactic parser, and then post-processed the output. Modern systems generally use neural sequence models supervised by the gold EDU segmentation in datasets like the RST Discourse Treebank. Fig. 22.4 shows an example architecture simplified from the algorithm of Lukasik et al. (2020) that predicts for each token whether or not it is a break. Here the input sentence is passed through an encoder and then passed through a linear layer and a softmax to produce a sequence of 0s and 1, where 1 indicates the start of an EDU.
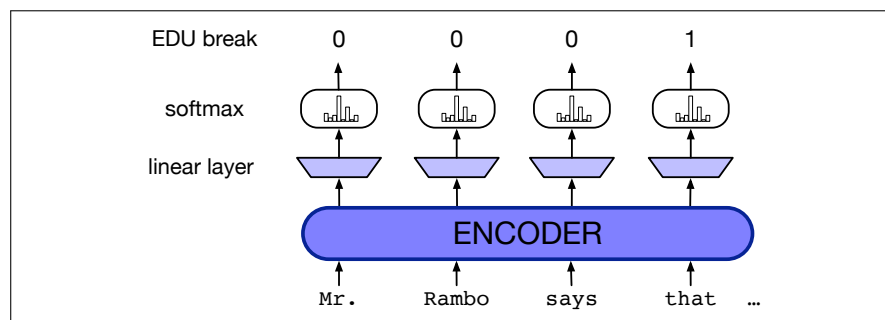
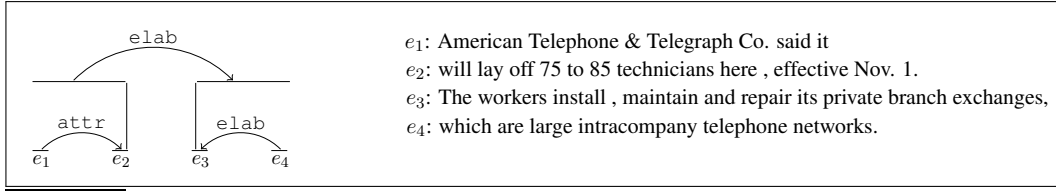**Figure 22.4**   Predicting EDU segment beginnings from encoded text.

22.2.2 RST parsing

Tools for building RST coherence structure for a discourse have long been based on syntactic parsing algorithms like shift-reduce parsing (Marcu, 1999). Many modern RST parsers since Ji and Eisenstein (2014) draw on the neural syntactic parsers we saw in Chapter 14, using representation learning to build representations for each span, and training a parser to choose the correct shift and reduce actions based on the gold parses in the training set.

We'll describe the shift-reduce parser of Yu et al. (2018). The parser state consists of a stack and a queue, and produces this structure by taking a series of actions on the states. Actions include:

- **shift**: pushes the first EDU in the queue onto the stack creating a single-node subtree.
- **reduce**(l,d): merges the top two subtrees on the stack, where l is the coherence relation label, and d is the nuclearity direction, $d \in \{NN, NS, SN\}$.

As well as the **pop root** operation, to remove the final tree from the stack.

Fig. 22.6 shows the actions the parser takes to build the structure in Fig. 22.5.

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	∅	e_1, e_2, e_3, e_4	SH	∅
2	e_1	e_2, e_3, e_4	SH	∅
3	e_1, e_2	e_3, e_4	RD(attr,SN)	∅
4	$e_{1:2}$	e_3, e_4	SH	$\widehat{e_1 e_2}$
5	$e_{1:2}, e_3$	e_4	SH	$\widehat{e_1 e_2}$
6	$e_{1:2}, e_3, e_4$	∅	RD(elab,NS)	$\widehat{e_1 e_2}$
7	$e_{1:2}, e_{3:4}$	∅	RD(elab,SN)	$\widehat{e_1 e_2}, \widehat{e_3 e_4}$
8	$e_{1:4}$	∅	PR	$\widehat{e_1 e_2}, \widehat{e_3 e_4}, \widehat{e_{1:2} e_{3:4}}$

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

The Yu et al. (2018) uses an encoder-decoder architecture, where the encoder represents the input span of words and EDUs using a hierarchical biLSTM. The first biLSTM layer represents the words inside an EDU, and the second represents the EDU sequence. Given an input sentence $w_1, w_2, ..., w_m$, the words can be represented as usual (by static embeddings, combinations with character embeddings or tags, or contextual embeddings) resulting in an input word representation sequence $\mathbf{x}_1^w, \mathbf{x}_2^w, ..., \mathbf{x}_m^w$. The result of the word-level biLSTM is then a sequence of $\mathbf{h^w}$ values:

$$\mathbf{h}_1^w, \mathbf{h}_2^w, ..., \mathbf{h}_m^w = \text{biLSTM}(\mathbf{x}_1^w, \mathbf{x}_2^w, ..., \mathbf{x}_m^w) \tag{22.17}$$

An EDU of span $w_s, w_{s+1}, ..., w_t$ then has biLSTM output representation $\mathbf{h}_s^w, \mathbf{h}_{s+1}^w, ..., \mathbf{h}_t^w$, and is represented by average pooling:

$$\mathbf{x}^e = \frac{1}{t-s+1} \sum_{k=s}^{t} \mathbf{h}_k^w \tag{22.18}$$

The second layer uses this input to compute a final representation of the sequence of EDU representations $\mathbf{h}^e$:

$$\mathbf{h}_1^e, \mathbf{h}_2^e, ..., \mathbf{h}_n^e = \text{biLSTM}(\mathbf{x}_1^e, \mathbf{x}_2^e, ..., \mathbf{x}_n^e) \tag{22.19}$$

The decoder is then a feedforward network $\mathbf{W}$ that outputs an action o based on a concatenation of the top three subtrees on the stack (s_o, s_1, s_2) plus the first EDU in the queue (q_0):

$$\mathbf{o} = \mathbf{W}(\mathbf{h}_{s0}^t, \mathbf{h}_{s1}^t, \mathbf{h}_{s2}^t, \mathbf{h}_{q0}^e) \tag{22.20}$$

where the representation of the EDU on the queue $\mathbf{h}_{q0}^e$ comes directly from the encoder, and the three hidden vectors representing partial trees are computed by average pooling over the encoder output for the EDUs in those trees:

$$\mathbf{h}_s^t = \frac{1}{j-i+1} \sum_{k=i}^{j} \mathbf{h}_k^e \tag{22.21}$$

Training first maps each RST gold parse tree into a sequence of oracle actions, and then uses the standard cross-entropy loss (with l_2 regularization) to train the system to take such actions. Give a state S and oracle action a, we first compute the decoder output using Eq. 22.20, apply a softmax to get probabilities:

$$p_a = \frac{\exp(\mathbf{o}_a)}{\sum_{a' \in A} \exp(\mathbf{o}_{a'})} \tag{22.22}$$

and then computing the cross-entropy loss:

$$L_{CE}() = -\log(p_a) + \frac{\lambda}{2}||\Theta||^2 \tag{22.23}$$

RST discourse parsers are evaluated on the test section of the RST Discourse Treebank, either with gold EDUs or end-to-end, using the RST-Pareval metrics (Marcu, 2000b). It is standard to first transform the gold RST trees into right-branching binary trees, and to report four metrics: trees with no labels (S for Span), labeled with nuclei (N), with relations (R), or both (F for Full), for each metric computing micro-averaged F_1 over all spans from all documents (Marcu 2000b, Morey et al. 2017).

22.2.3 PDTB discourse parsing

shallow discourse parsing

PDTB discourse parsing, the task of detecting PDTB coherence relations between spans, is sometimes called **shallow discourse parsing** because the task just involves flat relationships between text spans, rather than the full trees of RST parsing.

The set of four subtasks for PDTB discourse parsing was laid out by Lin et al. (2014) in the first complete system, with separate tasks for explicit (tasks 1-3) and implicit (task 4) connectives:

1. Find the discourse connectives (disambiguating them from non-discourse uses)
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

Many systems have been proposed for Task 4: taking a pair of adjacent sentences as input and assign a coherence relation sense label as output. The setup often follows Lin et al. (2009) in assuming gold sentence span boundaries and assigning each adjacent span one of the 11 second-level PDTB tags or none (removing the 5 very rare tags of the 16 shown in italics in Fig. 22.3).

A simple but very strong algorithm for Task 4 is to represent each of the two spans by BERT embeddings and take the last layer hidden state corresponding to the position of the [CLS] token, pass this through a single layer tanh feedforward network and then a softmax for sense classification (Nie et al., 2019).

Each of the other tasks also have been addressed. Task 1 is to disambiguating discourse connectives from their non-discourse use. For example as Pitler and Nenkova (2009) point out, the word *and* is a discourse connective linking the two clauses by an elaboration/expansion relation in (22.24) while it's a non-discourse NP conjunction in (22.25):

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

(22.25) My favorite colors are blue <u>and</u> green.

Similarly, *once* is a discourse connective indicating a temporal relation in (22.26), but simply a non-discourse adverb meaning 'formerly' and modifying *used* in (22.27):

(22.26) The asbestos fiber, crocidolite, is unusually resilient <u>once</u> it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.

(22.27) A form of asbestos <u>once</u> used to make Kent cigarette filters has caused a high percentage of cancer deaths among a group of workers exposed to it more than 30 years ago, researchers reported.

Determining whether a word is a discourse connective is thus a special case of word sense disambiguation. Early work on disambiguation showed that the 4 PDTB high-level sense classes could be disambiguated with high (94%) accuracy used syntactic features from gold parse trees (Pitler and Nenkova, 2009). Recent work performs the task end-to-end from word inputs using a biLSTM-CRF with BIO outputs (B-CONN, I-CONN, O) (Yu et al., 2019).

For task 2, PDTB spans can be identified with the same sequence models used to find RST EDUs: a biLSTM sequence model with pretrained contextual embedding (BERT) inputs (Muller et al., 2019). Simple heuristics also do pretty well as a baseline at finding spans, since 93% of relations are either completely within a single sentence or span two adjacent sentences, with one argument in each sentence (Biran and McKeown, 2015).

22.3   Centering and Entity-Based Coherence

A second way a discourse can be coherent is by virtue of being "about" some entity. This idea that at each point in the discourse some entity is salient, and a discourse is coherent by continuing to discuss the same entity, appears early in functional linguistics and the psychology of discourse (Chafe 1976, Kintsch and Van Dijk 1978), and soon made its way to computational models. In this section we introduce two **entity-based** models of this kind of **entity-based coherence**: **Centering Theory** (Grosz et al., 1995), and the **entity grid** model of Barzilay and Lapata (2008).

22.3.1   Centering

Centering Theory (Grosz et al., 1995) is a theory of both discourse salience and discourse coherence. As a model of discourse salience, Centering proposes that at any given point in the discourse one of the entities in the discourse model is salient: it is being "centered" on. As a model of discourse coherence, Centering proposes that discourses in which adjacent sentences CONTINUE to maintain the same salient entity are more coherent than those which SHIFT back and forth between multiple entities (we will see that CONTINUE and SHIFT are technical terms in the theory).

The following two texts from Grosz et al. (1995) which have exactly the same propositional content but different saliences, can help in understanding the main Centering intuition.

(22.28)   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.

(22.29) 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.

While these two texts differ only in how the two entities (John and the store) are realized in the sentences, the discourse in (22.28) is intuitively more coherent than the one in (22.29). As Grosz et al. (1995) point out, this is because the discourse in (22.28) is clearly about one individual, John, describing his actions and feelings. The discourse in (22.29), by contrast, focuses first on John, then the store, then back to John, then to the store again. It lacks the "aboutness" of the first discourse.

backward-looking center

forward-looking center

Centering Theory realizes this intuition by maintaining two representations for each utterance U_n. The **backward-looking center** of U_n, denoted as $C_b(U_n)$, represents the current salient entity, the one being focused on in the discourse after U_n is interpreted. The **forward-looking centers** of U_n, denoted as $C_f(U_n)$, are a set of potential future salient entities, the discourse entities evoked by U_n any of which could serve as C_b (the salient entity) of the following utterance, i.e. $C_b(U_{n+1})$.

The set of forward-looking centers $C_f(U_n)$ are ranked according to factors like discourse salience and grammatical role (for example subjects are higher ranked than objects, which are higher ranked than all other grammatical roles). We call the highest-ranked forward-looking center C_p (for "preferred center"). C_p is a kind of prediction about what entity will be talked about next. Sometimes the next utterance indeed talks about this entity, but sometimes another entity becomes salient instead.

We'll use here the algorithm for centering presented in Brennan et al. (1987), which defines four intersentential relationships between a pair of utterances U_n and U_{n+1} that depend on the relationship between $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$; these are shown in Fig. 22.7.

	$C_b(U_{n+1}) = C_b(U_n)$ 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).

The following rules are used by the algorithm:

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.

Rule 1 captures the intuition that pronominalization (including zero-anaphora) is a common way to mark discourse salience. If there are multiple pronouns in an utterance realizing entities from the previous utterance, one of these pronouns must realize the backward center C_b; if there is only one pronoun, it must be C_b.

Rule 2 captures the intuition that discourses that continue to center the same entity are more coherent than ones that repeatedly shift to other centers. The transition table is based on two factors: whether the backward-looking center C_b is the same from U_n to U_{n+1} and whether this discourse entity is the one that was preferred (C_p) from U_n. If both of these hold, a CONTINUE relation, the speaker has been talking about the same entity and is going to continue talking about that entity. In a RETAIN

relation, the speaker intends to SHIFT to a new entity in a future utterance and meanwhile places the current entity in a lower rank C_f. In a SHIFT relation, the speaker is shifting to a new salient entity.

Let's walk though the start of (22.28) again, repeated as (22.30), showing the representations after each utterance is processed.

(22.30)   John went to his favorite music store to buy a piano. (U_1)

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

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

It was closing just as John arrived (U_4)

Using the grammatical role hierarchy to order the C_f, for sentence U_1 we get:

$C_f(U_1)$: {John, music store, piano}
$C_p(U_1)$: John
$C_b(U_1)$: undefined

and then for sentence U_2:

$C_f(U_2)$: {John, piano}
$C_p(U_2)$: John
$C_b(U_2)$: John
Result: Continue   ($C_p(U_2)=C_b(U_2)$; $C_b(U_1)$ undefined)

The transition from U_1 to U_2 is thus a CONTINUE. Completing this example is left as exercise (1) for the reader

22.3.2   Entity Grid model

Centering embodies a particular theory of how entity mentioning leads to coherence: that salient entities appear in subject position or are pronominalized, and that discourses are salient by means of continuing to mention the same entity in such ways.

The **entity grid** model of Barzilay and Lapata (2008) is an alternative way to capture entity-based coherence: instead of having a top-down theory, the entity-grid model using machine learning to induce the patterns of entity mentioning that make a discourse more coherent.

The model is based around an **entity grid**, a two-dimensional array that represents the distribution of entity mentions across sentences. The rows represent sentences, and the columns represent discourse entities (most versions of the entity grid model focus just on nominal mentions). Each cell represents the possible appearance of an entity in a sentence, and the values represent whether the entity appears and its grammatical role. Grammatical roles are subject (S), object (O), neither (X), or absent (–); in the implementation of Barzilay and Lapata (2008), subjects of passives are represented with O, leading to a representation with some of the characteristics of thematic roles.

Fig. 22.8 from Barzilay and Lapata (2008) shows a grid for the text shown in Fig. 22.9. There is one row for each of the six sentences. The second column, for the entity 'trial', is O − − − X, showing that the trial appears in the first sentence as direct object, in the last sentence as an oblique, and does not appear in the middle sentences. The third column, for the entity Microsoft, shows that it appears as subject in sentence 1 (it also appears as the object of the preposition *against*, but entities that appear multiple times are recorded with their highest-ranked grammatical function). Computing the entity grids requires extracting entities and doing coreference

entity grid

```
                  D                                      G
                  e                C                     o
                  p        E       o   M                 v
                  a        v       m   a   P             e
                  r    M   i       p   r   r             r
                  t    i   d       e   k   o   B     N   S n   S
                  m    c   e       t   e   d   r   C e   o m   u   E
                  e T  r   n       i   t   u   a   a t   f e   a E a
                  n r  o   c       t   s   c   n   s s   t n   c r r
                  t i  s   e       o       t   d   e c   w t   t n t
                  t a  o   e       r   s   s   s   e a   a a   i m i
                    l  f   r       s       o       e p   r c   c e o
                                   s                 e   e s   s s n
                                                         s
             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
```

Figure 22.8 Part of the entity grid for the text in Fig. 22.9. Entities are listed by their head noun; each cell represents whether an entity appears as subject (S), object (O), neither (X), or is absent (–). Figure from Barzilay and Lapata (2008).

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).

resolution to cluster them into discourse entities (Chapter 22) as well as parsing the sentences to get grammatical roles.

In the resulting grid, columns that are dense (like the column for Microsoft) indicate entities that are mentioned often in the texts; sparse columns (like the column for earnings) indicate entities that are mentioned rarely.

In the entity grid model, coherence is measured by patterns of **local entity transition**. For example, Department is a subject in sentence 1, and then not mentioned in sentence 2; this is the transition [S –]. The transitions are thus sequences $\{$S,O X, $-\}^n$ which can be extracted as continuous cells from each column. Each transition has a probability; the probability of [S –] in the grid from Fig. 22.8 is 0.08 (it occurs 6 times out of the 75 total transitions of length two). Fig. 22.10 shows the distribution over transitions of length 2 for the text of Fig. 22.9 (shown as the first row d_1), and 2 other documents.

	SS	SO	SX	S–	OS	OO	OX	O–	XS	XO	XX	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).

The transitions and their probabilities can then be used as features for a machine learning model. This model can be a text classifier trained to produce human-labeled coherence scores (for example from humans labeling each text as coherent or incoherent). But such data is expensive to gather. Barzilay and Lapata (2005) introduced a simplifying innovation: coherence models can be trained by **self-supervision**: trained to distinguish the natural original order of sentences in a discourse from

a modified order (such as a randomized order). We turn to these evaluations in the next section.

22.3.3 Evaluating Neural and Entity-based coherence

Entity-based coherence models, as well as the neural models we introduce in the next section, are generally evaluated in one of two ways.

First, we can have humans rate the coherence of a document and train a classifier to predict these human ratings, which can be categorial (high/low, or high/mid/low) or continuous. This is the best evaluation to use if we have some end task in mind, like essay grading, where human raters are the correct definition of the final label.

Alternatively, since it's very expensive to get human labels, and we might not yet have an end-task in mind, we can use natural texts to do self-supervision. In self-supervision we pair up a natural discourse with a pseudo-document created by changing the ordering. Since naturally-ordered discourses are more coherent than random permutation (Lin et al., 2011), a successful coherence algorithm should prefer the original ordering.

Self-supervision has been implemented in 3 ways. In the **sentence order discrimination** task (Barzilay and Lapata, 2005), we compare a document to a random permutation of its sentence. A model is considered correct for an (original, permuted) test pair if it ranks the original document higher. Given k documents, we can compute n permutations, resulting in kn pairs each with one original document and one permutation, to use in training and testing.

In the **sentence insertion** task (Chen et al., 2007) we take a document, remove one of the n sentences s, and create $n-1$ copies of the document with s inserted into each position. The task is to decide which of the n documents is the one with the original ordering, distinguishing the original position for s from all other positions. Insertion is harder than discrimination since we are comparing documents that differ by only one sentence.

Finally, in the **sentence order reconstruction** task (Lapata, 2003), we take a document, randomize the sentences, and train the model to put them back in the correct order. Again given k documents, we can compute n permutations, resulting in kn pairs each with one original document and one permutation, to use in training and testing. Reordering is of course a much harder task than simple classification.

22.4 Representation learning models for local coherence

The third kind of local coherence is topical or semantic field coherence. Discourses cohere by talking about the same topics and subtopics, and drawing on the same semantic fields in doing so.

The field was pioneered by a series of unsupervised models in the 1990s of this kind of coherence that made use of **lexical cohesion** (Halliday and Hasan, 1976): the sharing of identical or semantically related words in nearby sentences. Morris and Hirst (1991) computed **lexical chains** of words (like *pine*, *bush trees*, *trunk*) that occurred through a discourse and that were related in Roget's Thesaurus (by being in the same category, or linked categories). They showed that the number and density of chain correlated with the topic structure. The **TextTiling** algorithm of Hearst (1997) computed the cosine between neighboring text spans (the normalized dot product of vectors of raw word counts), again showing that sentences or paragraph in

lexical cohesion

TextTiling

a subtopic have high cosine with each other, but not with sentences in a neighboring subtopic.

A third early model, the LSA Coherence method of Foltz et al. (1998) was the first to use embeddings, modeling the coherence between two sentences as the cosine between their LSA sentence embedding vectors[1], computing embeddings for a sentence s by summing the embeddings of its words w:

$$\begin{aligned} \text{sim}(s,t) &= \cos(\mathbf{s}, \mathbf{t}) \\ &= \cos(\sum_{w \in s} \mathbf{w}, \sum_{w \in t} \mathbf{w}) \end{aligned} \tag{22.31}$$

and defining the overall coherence of a text as the average similarity over all pairs of adjacent sentences s_i and s_{i+1}:

$$\text{coherence}(T) = \frac{1}{n-1} \sum_{i=1}^{n-1} \cos(s_i, s_{i+1}) \tag{22.32}$$

Modern neural representation-learning coherence models, beginning with Li et al. (2014), draw on the intuitions of these early unsupervised models for learning sentence representations and measuring how they change between neighboring sentences. But the new models also draw on the idea pioneered by Barzilay and Lapata (2005) of self-supervision. That is, unlike say coherence relation models, which train on hand-labeled representations for RST or PDTB, these models are trained to distinguish natural discourses from unnatural discourses formed by scrambling the order of sentences, thus using representation learning to discover the features that matter for at least the ordering aspect of coherence.

Here we present one such model, the local coherence discriminator (LCD) (Xu et al., 2019). Like early models, LCD computes the coherence of a text as the average of coherence scores between consecutive pairs of sentences. But unlike the early unsupervised models, LCD is a self-supervised model trained to discriminate consecutive sentence pairs (s_i, s_{i+1}) in the training documents (assumed to be coherent) from (constructed) incoherent pairs (s_i, s'). All consecutive pairs are positive examples, and the negative (incoherent) partner for a sentence s_i is another sentence uniformly sampled from the same document as s_i.

Fig. 22.11 describes the architecture of the model f_θ, which takes a sentence pair and returns a score, higher scores for more coherent pairs. Given an input sentence pair s and t, the model computes sentence embeddings $\mathbf{s}$ and $\mathbf{t}$ (using any sentence embeddings algorithm), and then concatenates four features of the pair: (1) the concatenation of the two vectors (2) their difference $\mathbf{s} - \mathbf{t}$; (3) the absolute value of their difference $|\mathbf{s} - \mathbf{t}|$; (4) their element-wise product $\mathbf{s} \odot \mathbf{t}$. These are passed through a one-layer feedforward network to output the coherence score.

The model is trained to make this coherence score higher for real pairs than for negative pairs. More formally, the training objective for a corpus C of documents d, each of which consists of a list of sentences s_i, is:

$$L_\theta = \sum_{d \in C} \sum_{s_i \in d} \mathbb{E}_{p(s'|s_i)} [L(f_\theta(s_i, s_{i+1}), f_\theta(s_i, s'))] \tag{22.33}$$

$\mathbb{E}_{p(s'|s_i)}$ is the expectation with respect to the negative sampling distribution conditioned on s_i: given a sentence s_i the algorithms samples a negative sentence s'

---

[1] See Chapter 6 for more on LSA embeddings; they are computed by applying SVD to the term-document matrix (each cell weighted by log frequency and normalized by entropy), and then the first 300 dimensions are used as the embedding.
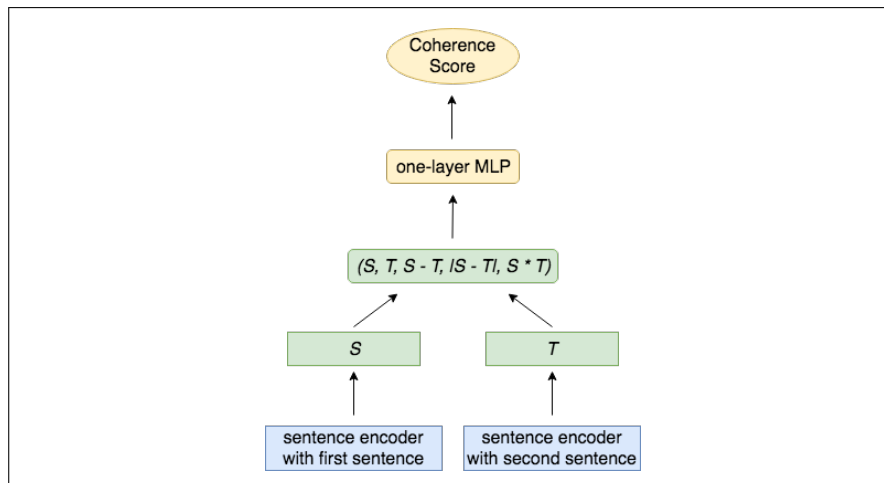
Figure 22.11 The architecture of the LCD model of document coherence, showing the computation of the score for a pair of sentences *s* and *t*. Figure from Xu et al. (2019).

uniformly over the other sentences in the same document. *L* is a loss function that takes two scores, one for a positive pair and one for a negative pair, with the goal of encouraging $f^+ = f_\theta(s_i, s_{i+1})$ to be high and $f^- = f_\theta(s_i, s')$ to be low. Fig. 22.11 use the margin loss $l(f^+, f^-) = \max(0, \eta - f^+ + f^-)$ where η is the margin hyperparameter.

Xu et al. (2019) also give a useful baseline algorithm that itself has quite high performance in measuring perplexity: train an RNN language model on the data, and compute the log likelihood of sentence s_i in two ways, once given the preceding context (conditional log likelihood) and once with no context (marginal log likelihood). The difference between these values tells us how much the preceding context improved the predictability of s_i, a predictability measure of coherence.

Training models to predict longer contexts than just consecutive pairs of sentences can result in even stronger discourse representations. For example a Transformer language model trained with a contrastive sentence objective to predict text up to a distance of ± 2 sentences improves performance on various discourse coherence tasks (Iter et al., 2020).

Language-model style models are generally evaluated by the methods of Section 22.3.3, although they can also be evaluated on the RST and PDTB coherence relation tasks.

22.5 Global Coherence

A discourse must also cohere globally rather than just at the level of pairs of sentences. Consider stories, for example. The narrative structure of stories is one of the oldest kinds of global coherence to be studied. In his influential *Morphology of the Folktale*, Propp (1968) models the discourse structure of Russian folktales via a kind of plot grammar. His model includes a set of character categories he called **dramatis personae**, like Hero, Villain, Donor, or Helper, and a set of events he called **functions** (like "Villain commits kidnapping", "Donor tests Hero", or "Hero is pursued") that have to occur in particular order, along with other components. Propp shows that the plots of each of the fairy tales he studies can be represented as

a sequence of these functions, different tales choosing different subsets of functions, but always in the same order. Indeed Lakoff (1972) showed that Propp's model amounted to a discourse grammar of stories, and in recent computational work Finlayson (2016) demonstrates that some of these Proppian functions could be induced from corpora of folktale texts by detecting events that have similar actions across stories. Bamman et al. (2013) showed that generalizations over dramatis personae could be induced from movie plot summaries on Wikipedia. Their model induced latent personae from features like the actions the character takes (e.g., Villains strangle), the actions done to them (e.g., Villains are foiled and arrested) or the descriptive words used of them (Villains are evil).

In this section we introduce two kinds of such global discourse structure that have been widely studied computationally. The first is the structure of arguments: the way people attempt to convince each other in persuasive essays by offering claims and supporting premises. The second is somewhat related: the structure of scientific papers, and the way authors present their goals, results, and relationship to prior work in their papers.

22.5.1   Argumentation Structure

The first type of global discourse structure is the structure of **arguments**. Analyzing people's argumentation computationally is often called **argumentation mining**.

argumentation mining

The study of arguments dates back to Aristotle, who in his Rhetorics described three components of a good argument: **pathos** (appealing to the emotions of the listener), **ethos** (appealing to the speaker's personal character), and **logos** (the logical structure of the argument).

pathos
ethos
logos

Most of the discourse structure studies of argumentation have focused on **logos**, particularly via building and training on annotated datasets of persuasive essays or other arguments (Reed et al. 2008, Stab and Gurevych 2014a, Peldszus and Stede 2016, Habernal and Gurevych 2017, Musi et al. 2018). Such corpora, for example, often include annotations of argumentative components like **claims** (the central component of the argument that is controversial and needs support) and **premises** (the reasons given by the author to persuade the reader by supporting or attacking the claim or other premises), as well as the **argumentative relations** between them like SUPPORT and ATTACK.

claims
premises
argumentative relations

Consider the following example of a persuasive essay from Stab and Gurevych (2014b). The first sentence (1) presents a claim (in bold). (2) and (3) present two premises supporting the claim. (4) gives a premise supporting premise (3).

> "(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."

Thus this example has three argumentative relations: SUPPORT(2,1), SUPPORT(3,1) and SUPPORT(4,3). Fig. 22.12 shows the structure of a much more complex argument.

While argumentation mining is clearly related to rhetorical structure and other kinds of coherence relations, arguments tend to be much less local; often a persuasive essay will have only a single main claim, with premises spread throughout the text, without the local coherence we see in coherence relations.
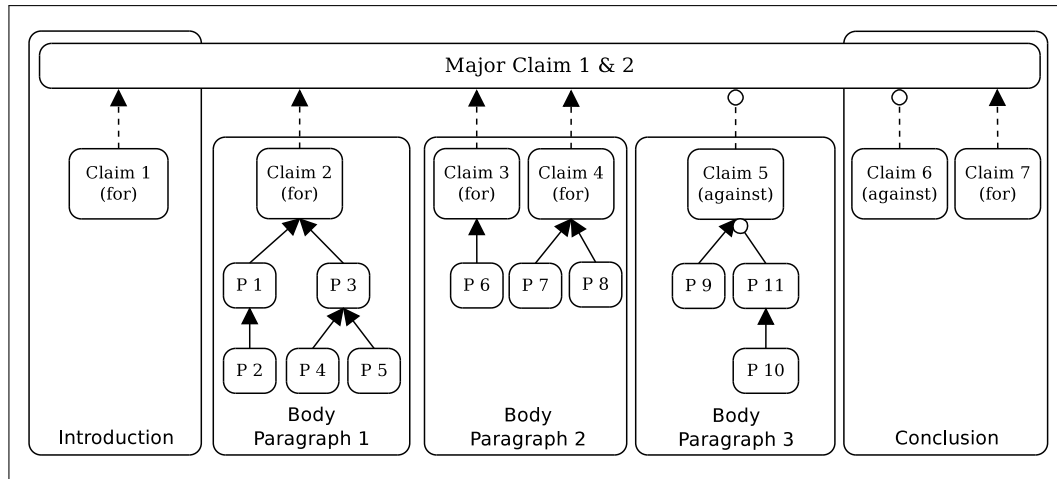
Figure 22.12   Argumentation structure of a persuasive essay. Arrows indicate argumentation relations, either of SUPPORT (with arrowheads) or ATTACK (with circleheads); P denotes premises. Figure from Stab and Gurevych (2017).

Algorithms for detecting argumentation structure often include classifiers for distinguishing claims, premises, or non-argumentation, together with relation classifiers for deciding if two spans have the SUPPORT, ATTACK, or neither relation (Peldszus and Stede, 2013). While these are the main focus of much computational work, there is also preliminary efforts on annotating and detecting richer semantic relationships (Park and Cardie 2014, Hidey et al. 2017) such as detecting **argumentation schemes**, larger-scale structures for argument like **argument from example**, or **argument from cause to effect**, or **argument from consequences** (Feng and Hirst, 2011).

argumentation schemes

Another important line of research is studying how these argument structure (or other features) are associated with the success or persuasiveness of an argument (Habernal and Gurevych 2016, Tan et al. 2016, Hidey et al. 2017. Indeed, while it is Aristotle's logos that is most related to discourse structure, Aristotle's ethos and pathos techniques are particularly relevant in the detection of mechanisms of this sort 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), and **scarcity** (people value things that are scarce), all of which can be brought up in a persuasive argument (Cialdini, 1984). Rosenthal and McKeown (2017) showed that these features could be combined with argumentation structure to predict who influences whom on social media, Althoff et al. (2014) found that linguistic models of reciprocity and authority predicted success in online requests, while the semisupervised model of Yang et al. (2019) detected mentions of scarcity, commitment, and social identity to predict the success of peer-to-peer lending platforms.

persuasion

See Stede and Schneider (2018) for a comprehensive survey of argument mining.

22.5.2   The structure of scientific discourse

Scientific papers have a very specific global structure: somewhere in the course of the paper the authors must indicate a scientific goal, develop a method for a solution, provide evidence for the solution, and compare to prior work. One popular

annotation scheme for modeling these rhetorical goals is the **argumentative zoning** model of Teufel et al. (1999) and Teufel et al. (2009), which is informed by the idea that each scientific paper tries to make a **knowledge claim** about a new piece of knowledge being added to the repository of the field (Myers, 1992). Sentences in a scientific paper can be assigned one of 15 tags; Fig. 22.13 shows 7 (shortened) examples of labeled sentences.

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).

Teufel et al. (1999) and Teufel et al. (2009) develop labeled corpora of scientific articles from computational linguistics and chemistry, which can be used as supervision for training standard sentence-classification architecture to assign the 15 labels.

22.6 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.

Bibliographical and Historical Notes

Coherence relations arose from the independent development of a number of scholars, including Hobbs (1979) idea that coherence relations play an inferential role for the hearer, and the investigations by Mann and Thompson (1987) of the discourse structure of large texts. Other approaches to coherence relations and their extraction include Segmented Discourse Representation Theory (**SDRT**) (Asher and Lascarides 2003, Baldridge et al. 2007) and the Linguistic Discourse Model (Polanyi 1988, Scha and Polanyi 1988, Polanyi et al. 2004). Wolf and Gibson (2005) argue that coherence structure includes crossed bracketings, which make it impossible to represent as a tree, and propose a graph representation instead. A compendium of over 350 relations that have been proposed in the literature can be found in Hovy (1990).

RST parsing was first proposed by Marcu (1997), and early work was rule-based, focused on discourse markers (Marcu, 2000a). The creation of the RST Discourse TreeBank (Carlson et al. 2001, Carlson and Marcu 2001) enabled a wide variety of machine learning algorithms, beginning with the shift-reduce parser of Marcu (1999) that used decision trees to choose actions, and continuing with a wide variety of machine learned parsing methods (Soricut and Marcu 2003, Sagae 2009, Hernault et al. 2010, Feng and Hirst 2014, Surdeanu et al. 2015, Joty et al. 2015) and chunkers (Sporleder and Lapata, 2005). Subba and Di Eugenio (2009) integrated sophisticated semantic information into RST parsing. Ji and Eisenstein (2014) first applied neural models to RST parsing neural models, leading to the modern set of neural RST models (Li et al. 2014, Li et al. 2016, Braud et al. 2017, Yu et al. 2018, inter alia) as well as neural segmenters (Wang et al. 2018). and neural PDTB parsing models (Ji and Eisenstein 2015, Qin et al. 2016, Qin et al. 2017).

Barzilay and Lapata (2005) pioneered the idea of self-supervision for coherence: training a coherence model to distinguish true orderings of sentences from random permutations. Li et al. (2014) first applied this paradigm to neural sentence-representation, and many neural self-supervised models followed (Li and Jurafsky 2017, Logeswaran et al. 2018, Lai and Tetreault 2018, Xu et al. 2019, Iter et al. 2020)

Another aspect of global coherence is the global topic structure of a text, the way the topics shift over the course of the document. Barzilay and Lee (2004) introduced an HMM model for capturing topics for coherence, and later work expanded this intuition (Soricut and Marcu 2006, Elsner et al. 2007, Louis and Nenkova 2012, Li and Jurafsky 2017).

The relationship between explicit and implicit discourse connectives has been a fruitful one for research. Marcu and Echihabi (2002) first proposed to use sentences with explicit relations to help provide training data for implicit relations, by removing the explicit relations and trying to re-predict them as a way of improving performance on implicit connectives; this idea was refined by Sporleder and Lascarides (2005), (Pitler et al., 2009), and Rutherford and Xue (2015). This rela-

tionship can also be used as a way to create discourse-aware representations. The DisSent algorithm (Nie et al., 2019) creates the task of predicting explicit discourse markers between two sentences. They show that representations learned to be good at this task also function as powerful sentence representations for other discourse tasks.

The idea of entity-based coherence seems to have arisen in multiple fields in the mid-1970s, in functional linguistics (Chafe, 1976), in the psychology of discourse processing (Kintsch and Van Dijk, 1978), and in the roughly contemporaneous work of Grosz, Sidner, Joshi, and their colleagues. Grosz (1977) addressed the focus of attention that conversational participants maintain as the discourse unfolds. She defined two levels of focus; entities relevant to the entire discourse were said to be in *global* focus, whereas entities that are locally in focus (i.e., most central to a particular utterance) were said to be in *immediate* focus. Sidner 1979; 1983 described a method for tracking (immediate) discourse foci and their use in resolving pronouns and demonstrative noun phrases. She made a distinction between the current discourse focus and potential foci, which are the predecessors to the backward- and forward-looking centers of Centering theory, respectively. The name and further roots of the centering approach lie in papers by Joshi and Kuhn (1979) and Joshi and Weinstein (1981), who addressed the relationship between immediate focus and the inferences required to integrate the current utterance into the discourse model. Grosz et al. (1983) integrated this work with the prior work of Sidner and Grosz. This led to a manuscript on centering which, while widely circulated since 1986, remained unpublished until Grosz et al. (1995). A collection of centering papers appears in Walker et al. (1998). See Karamanis et al. (2004) and Poesio et al. (2004) for a deeper exploration of centering and its parameterizations, and the History section of Chapter 22 for more on the use of centering on coreference.

The grid model of entity-based coherence was first proposed by Barzilay and Lapata (2005) drawing on earlier work by Lapata (2003) and Barzilay, and then extended by them Barzilay and Lapata (2008) and others with additional features (Elsner and Charniak 2008, 2011, Feng et al. 2014, Lin et al. 2011) a model that projects entities into a global graph for the discourse (Guinaudeau and Strube 2013, Mesgar and Strube 2016), and a convolutional model to capture longer-range entity dependencies (Nguyen and Joty, 2017).

Theories of discourse coherence have also been used in algorithms for interpreting discourse-level linguistic phenomena, including verb phrase ellipsis and gapping (Asher 1993, Kehler 1993), and tense interpretation (Lascarides and Asher 1993, Kehler 1994, Kehler 2000). An extensive investigation into the relationship between coherence relations and discourse connectives can be found in Knott and Dale (1994).

Useful surveys of discourse processing and structure include Stede (2011) and Webber et al. (2012).

Andy Kehler wrote the Discourse chapter for the 2000 first edition of this textbook, which we used as the starting point for the second-edition chapter, and there are some remnants of Andy's lovely prose still in this third-edition coherence chapter.

Exercises

22.1 Finish the Centering Theory processing of the last two utterances of (22.30), and show how (22.29) would be processed. Does the algorithm indeed mark (22.29) as less coherent?

22.2 Select an editorial column from your favorite newspaper, and determine the discourse structure for a 10–20 sentence portion. What problems did you encounter? Were you helped by superficial cues the speaker included (e.g., discourse connectives) in any places?

Althoff, T., C. Danescu-Niculescu-Mizil, and D. Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. *ICWSM 2014*.

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Studies in Linguistics and Philosophy (SLAP) 50, Kluwer.

Asher, N. and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Baldridge, J., N. Asher, and J. Hunter. 2007. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift für Sprachwissenschaft*, 26:213–239.

Bamman, D., B. O'Connor, and N. A. Smith. 2013. Learning latent personas of film characters. *ACL*.

Barzilay, R. and M. Lapata. 2005. Modeling local coherence: An entity-based approach. *ACL*.

Barzilay, R. and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Barzilay, R. and L. Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *HLT-NAACL*.

Bedi, G., F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1.

Biran, O. and K. McKeown. 2015. PDTB discourse parsing as a tagging task: The two taggers approach. *SIGDIAL*.

Braud, C., M. Coavoux, and A. Søgaard. 2017. Cross-lingual RST discourse parsing. *EACL*.

Brennan, S. E., M. W. Friedman, and C. Pollard. 1987. A centering approach to pronouns. *ACL*.

Carlson, L. and D. Marcu. 2001. Discourse tagging manual. Technical Report ISI-TR-545, ISI.

Carlson, L., D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *SIGDIAL*.

Chafe, W. L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li, editor, *Subject and Topic*, pages 25–55. Academic Press.

Chen, E., B. Snyder, and R. Barzilay. 2007. Incremental text structuring with online hierarchical ranking. *EMNLP/CoNLL*.

Cialdini, R. B. 1984. *Influence: The psychology of persuasion*. Morrow.

Ditman, T. and G. R. Kuperberg. 2010. Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia. *Journal of neurolinguistics*, 23(3):254–269.

Elsner, M., J. Austerweil, and E. Charniak. 2007. A unified local and global model for discourse coherence. *NAACL-HLT*.

Elsner, M. and E. Charniak. 2008. Coreference-inspired coherence modeling. *ACL*.

Elsner, M. and E. Charniak. 2011. Extending the entity grid with entity-specific features. *ACL*.

Elvevåg, B., P. W. Foltz, D. R. Weinberger, and T. E. Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Feng, V. W. and G. Hirst. 2011. Classifying arguments by scheme. *ACL*.

Feng, V. W. and G. Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. *ACL*.

Feng, V. W., Z. Lin, and G. Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. *COLING*.

Finlayson, M. A. 2016. Inferring Propp's functions from semantically annotated text. *The Journal of American Folklore*, 129(511):55–77.

Foltz, P. W., W. Kintsch, and T. K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Grosz, B. J. 1977. The representation and use of focus in a system for understanding dialogs. *IJCAI-77*. Morgan Kaufmann.

Grosz, B. J., A. K. Joshi, and S. Weinstein. 1983. Providing a unified account of definite noun phrases in English. *ACL*.

Grosz, B. J., A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Guinaudeau, C. and M. Strube. 2013. Graph-based local coherence modeling. *ACL*.

Habernal, I. and I. Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. *ACL*.

Habernal, I. and I. Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Halliday, M. A. K. and R. Hasan. 1976. *Cohesion in English*. Longman. English Language Series, Title No. 9.

Hearst, M. A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64.

Hernault, H., H. Prendinger, D. A. duVerle, and M. Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Hidey, C., E. Musi, A. Hwang, S. Muresan, and K. McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. *4th Workshop on Argument Mining*.

Hobbs, J. R. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

Hovy, E. H. 1990. Parsimonious and profligate approaches to the question of discourse structure relations. *Proceedings of the 5th International Workshop on Natural Language Generation*.

Iter, D., K. Guu, L. Lansing, and D. Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. *ACL*.

Iter, D., J. Yoon, and D. Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. *Fifth Workshop on Computational Linguistics and Clinical Psychology*.

Ji, Y. and J. Eisenstein. 2014. Representation learning for text-level discourse parsing. *ACL*.

Ji, Y. and J. Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *TACL*, 3:329–344.

Joshi, A. K. and S. Kuhn. 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. *IJCAI-79*.

Joshi, A. K. and S. Weinstein. 1981. Control of inference: Role of some aspects of discourse structure – centering. *IJCAI-81*.

Joty, S., G. Carenini, and R. T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Karamanis, N., M. Poesio, C. Mellish, and J. Oberlander. 2004. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. *ACL*.

Kehler, A. 1993. The effect of establishing coherence in ellipsis and anaphora resolution. *ACL*.

Kehler, A. 1994. Temporal relations: Reference or discourse coherence? *ACL*.

Kehler, A. 2000. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

Kintsch, W. and T. A. Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363–394.

Knott, A. and R. Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Lai, A. and J. Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. *SIGDIAL*.

Lakoff, G. 1972. Structural complexity in fairy tales. In *The Study of Man*, pages 128–50. School of Social Sciences, University of California, Irvine, CA.

Lapata, M. 2003. Probabilistic text structuring: Experiments with sentence ordering. *ACL*.

Lascarides, A. and N. Asher. 1993. Temporal interpretation, discourse relations, and common sense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Li, J. and D. Jurafsky. 2017. Neural net models of open-domain discourse coherence. *EMNLP*.

Li, J., R. Li, and E. H. Hovy. 2014. Recursive deep models for discourse parsing. *EMNLP*.

Li, Q., T. Li, and B. Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. *EMNLP*.

Lin, Z., M.-Y. Kan, and H. T. Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. *EMNLP*.

Lin, Z., H. T. Ng, and M.-Y. Kan. 2011. Automatically evaluating text coherence using discourse relations. *ACL*.

Lin, Z., H. T. Ng, and M.-Y. Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Logeswaran, L., H. Lee, and D. Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. *AAAI*.

Louis, A. and A. Nenkova. 2012. A coherence model based on syntactic patterns. *EMNLP*.

Lukasik, M., B. Dadachev, K. Papineni, and G. Simões. 2020. Text segmentation by cross segment attention. *EMNLP*.

Mann, W. C. and S. A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report RS-87-190, Information Sciences Institute.

Marcu, D. 1997. The rhetorical parsing of natural language texts. *ACL*.

Marcu, D. 1999. A decision-based approach to rhetorical parsing. *ACL*.

Marcu, D. 2000a. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Marcu, D., editor. 2000b. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Marcu, D. and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. *ACL*.

Mesgar, M. and M. Strube. 2016. Lexical coherence graph modeling using word embeddings. *ACL*.

Miltsakaki, E., R. Prasad, A. K. Joshi, and B. L. Webber. 2004. The Penn Discourse Treebank. *LREC*.

Morey, M., P. Muller, and N. Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the rst-dt. *EMNLP*.

Morris, J. and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Muller, P., C. Braud, and M. Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. *Workshop on Discourse Relation Parsing and Treebanking*.

Musi, E., M. Stede, L. Kriese, S. Muresan, and A. Rocci. 2018. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. *LREC*.

Myers, G. 1992. "in this paper we report...": Speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Nguyen, D. T. and S. Joty. 2017. A neural local coherence model. *ACL*.

Nie, A., E. Bennett, and N. Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. *ACL*.

Park, J. and C. Cardie. 2014. Identifying appropriate support for propositions in online user comments. *First workshop on argumentation mining*.

Peldszus, A. and M. Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Peldszus, A. and M. Stede. 2016. An annotated corpus of argumentative microtexts. *1st European Conference on Argumentation*.

Pitler, E., A. Louis, and A. Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. *ACL IJCNLP*.

Pitler, E. and A. Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. *ACL IJCNLP*.

Poesio, M., R. Stevenson, B. Di Eugenio, and J. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Polanyi, L. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12.

Polanyi, L., C. Culy, M. van den Berg, G. L. Thione, and D. Ahn. 2004. A rule based approach to discourse parsing. *Proceedings of SIGDIAL*.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. 2008. The Penn Discourse TreeBank 2.0. *LREC*.

Prasad, R., B. L. Webber, and A. Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Propp, V. 1968. *Morphology of the Folktale*, 2nd edition. University of Texas Press. Original Russian 1928. Translated by Laurence Scott.

Qin, L., Z. Zhang, and H. Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. *EMNLP*.

Qin, L., Z. Zhang, H. Zhao, Z. Hu, and E. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. *ACL*.

Reed, C., R. Mochales Palau, G. Rowe, and M.-F. Moens. 2008. Language resources for studying argument. *LREC*.

Rosenthal, S. and K. McKeown. 2017. Detecting influencers in multiple online genres. *ACM Transactions on Internet Technology (TOIT)*, 17(2).

Rutherford, A. and N. Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. *NAACL HLT*.

Sagae, K. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. *IWPT-09*.

Scha, R. and L. Polanyi. 1988. An augmented context free grammar for discourse. *COLING*.

Sidner, C. L. 1979. Towards a computational theory of definite anaphora comprehension in English discourse. Technical Report 537, MIT Artificial Intelligence Laboratory, Cambridge, MA.

Sidner, C. L. 1983. Focusing in the comprehension of definite anaphora. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse*, pages 267–330. MIT Press.

Somasundaran, S., J. Burstein, and M. Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. *COLING*.

Soricut, R. and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. *HLT-NAACL*.

Soricut, R. and D. Marcu. 2006. Discourse generation using utility-trained coherence models. *COLING/ACL*.

Sporleder, C. and A. Lascarides. 2005. Exploiting linguistic cues to classify rhetorical relations. *RANLP-05*.

Sporleder, C. and M. Lapata. 2005. Discourse chunking and its application to sentence compression. *EMNLP*.

Stab, C. and I. Gurevych. 2014a. Annotating argument components and relations in persuasive essays. *COLING*.

Stab, C. and I. Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. *EMNLP*.

Stab, C. and I. Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Stede, M. 2011. *Discourse processing*. Morgan & Claypool.

Stede, M. and J. Schneider. 2018. *Argumentation Mining*. Morgan & Claypool.

Subba, R. and B. Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. *NAACL HLT*.

Surdeanu, M., T. Hicks, and M. A. Valenzuela-Escarcega. 2015. Two practical rhetorical structure theory parsers. *NAACL HLT*.

Tan, C., V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *WWW-16*.

Teufel, S., J. Carletta, and M. Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. *EACL*.

Teufel, S., A. Siddharthan, and C. Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. *EMNLP*.

Walker, M. A., A. K. Joshi, and E. Prince, editors. 1998. *Centering in Discourse*. Oxford University Press.

Wang, Y., S. Li, and J. Yang. 2018. Toward fast and accurate neural discourse segmentation. *EMNLP*.

Webber, B. L., M. Egg, and V. Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Wolf, F. and E. Gibson. 2005. Representing discourse coherence: A corpus-based analysis. *Computational Linguistics*, 31(2):249–287.

Xu, P., H. Saghir, J. S. Kang, T. Long, A. J. Bose, Y. Cao, and J. C. K. Cheung. 2019. A cross-domain transferable neural coherence model. *ACL*.

Xue, N., H. T. Ng, S. Pradhan, A. Rutherford, B. L. Webber, C. Wang, and H. Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. *CoNLL-16 shared task*.

Yang, D., J. Chen, Z. Yang, D. Jurafsky, and E. H. Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. *NAACL HLT*.

Yu, N., M. Zhang, and G. Fu. 2018. Transition-based neural RST parsing with implicit syntax features. *COLING*.

Yu, Y., Y. Zhu, Y. Liu, Y. Liu, S. Peng, M. Gong, and A. Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. *Workshop on Discourse Relation Parsing and Treebanking 2019*.

Zhou, Y. and N. Xue. 2015. The Chinese Discourse Tree-Bank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.