A Rule Based Approach to Discourse Parsing

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, David Ahn[1]
FX Palo Alto Laboratory
3400 Hillview Avenue, Bldg 4
Palo Alto CA 94304
{polanyi|culy|vdberg|thione}@fxpal.com ; ahn@science.uva.nl

Abstract

In this paper we present an overview of recent developments in discourse theory and parsing under the Linguistic Discourse Model (LDM) framework, a semantic theory of discourse structure. We give a novel approach to the problem of discourse segmentation based on discourse semantics and sketch a limited but robust approach to symbolic discourse parsing based on syntactic, semantic and lexical rules. To demonstrate the utility of the system in a real application, we briefly describe the architecture of the PALSUMM system, a symbolic summarization system being developed at FX Palo Alto Laboratory that uses discourse structures constructed using the theory outlined to summarize written English prose texts.

1 Introduction

In this paper we present an overview of recent theoretical and computational developments in discourse theory and parsing under the Linguistic Discourse Model (LDM) framework, a semantic account of discourse structure. In Section 2, we present an overview of what we will term the Classical LDM (C-LDM) and identify critical problems encountered in implementing the model: the difficulty in segmenting complex sentences within a text and calculating the attachment site and relationship of an incoming unit to an appropriate node in a developing Discourse Tree. In Sections 3 and 4 we introduce the Unified Linguistic Discourse Model (U-LDM) that incorporates solutions to these problems. Specifically, in Section 3 we describe a novel approach to discourse segmentation based on the relationship of sentential syntax to discourse semantics. In Section 4, a limited but robust approach to symbolic discourse processing based on syntactic, semantic and lexical rules is given. In Section 5, we sketch the architecture of the PALSUMM system, a summarization system being developed at FX Palo Alto Laboratory that uses algorithms operating on discourse representations generated by a U-LDM parser to summarize written English prose texts. In Section 6 we present our conclusions and suggest directions for future research.

2 The Classical Linguistic Discourse Model (C-LDM)

Unlike the Discourse Structures Model (DSM) of Grosz and Sidner (1986), a *pragmatic and psychological theory* that aims to clarify the relationship between speakers' intentions and their focus of attention in discourse, or *the rhetorical model* of Rhetorical Structures Theory (Mann and Thompson, 1988) that is designed to identify the coherence relations between segments of text, the Linguistic Discourse Model (LDM) (Polanyi and Scha, 1984; Polanyi, 1988; Polanyi

[1] Current address:
Language and Inference Technology Group,
ILLC, University of Amsterdam,
Nieuwe Achtergracht 166,
1018 WV Amsterdam, The Netherlands

and van den Berg, 1996) is a *syntactically informed, semantically driven model* developed to provide proper semantic interpretation for every utterance in a discourse despite the apparent discontinuities that are present even in well structured written texts. In its focus on understanding discourse meaning, the LDM is close in spirit to Structured Discourse Representation Theory (S-DRT) (Asher, 1993). While S-DRT attempts to account for discourse structure purely semantically, the LDM framework is concerned to maintain a separation between discourse "syntactic" structure, on the one hand, and discourse interpretation on the other. Therefore, like DSM and RST, the LDM incorporates an explicit tree structured model of relationships between discourse segments as its model of discourse "syntax". In discourse parsing under the LDM, any attachment to the developing discourse tree of a textual unit is treated as an instruction to update an appropriate semantic representation. We construct dynamic semantic representations (DSRs), similar to the Discourse Representation Structures (Kamp, 1981; Kamp and Reyle, 1993) used in S-DRT as its model of discourse semantics. The DSRs correspond to the contexts relative to which subsequent segments can be interpreted.

The analysis of intra-sentential structure is done by sentential syntax which identifies the syntactic and semantic structures within the sentence and makes the resulting analysis available for discourse processing.

2.1 Overview of the Classic LDM

In the Linguistic Discourse Model (LDM) discourse is formed through the recursive combination of discourse constituent units (DCUs). The structure of a discourse is represented by an open right tree of DCUs. Basic discourse units (BDUs), resulting from a segmentation of the discourse according to rules of discourse segmentation, form the content of the leaves of the tree. Once a text has been segmented into BDUs, an open right tree representing the structure of the discourse is built up. The completed tree shows, for any given point in the discourse, which discourse units (DCUs) remain available for continuation and which DCUs are no longer available. Because discourse anaphora resolution is critically constrained by discourse structure, the tree representation makes clear the domain in which the antecedent for a given anaphoric referential expression is to be found. Antecedents must be available at a node along the right edge of the discourse tree. (Polanyi, 1985; Grosz and Sidner 1986; Webber, 1991)

The LDM posits three structural relations between discourse units:

1. **discourse coordination**
   a. Units related by bearing a similar relationship to an existing or newly formed common parent in the tree (lists, narratives).
   b. Available at the C-node is information common to all child nodes.
2. **discourse subordination**
   a. Units related by an *elaboration* relationship in which the subordinated unit provides more information about an entity or situation described in the subordinating unit.
   b. Units unrelated to existing units available on the right edge of the tree, viewed as intrusions or interruptions.
   c. Available at the S-node is information specific only to the subordinating or dominant constituent (usually the left child).
3. **n-ary constructions**
   a. Units related by logical or rhetorical, genre or interactional conventions specific to a given language.
   b. Preposed modifier, sentence initial adverbial, "cue word", (reported speech) attribution phrase.
   c. Available at N-nodes is information about each constituent and the relationship connecting them.

Although we believe that the general approach to discourse structure captured by the Classical LDM is essentially sound, there are three critical problems with the existing framework:

1. Segmenting the incoming text into BDUs
2. Determining the existing or new node at which to attach an incoming BDU
3. Determining the relationship between the incoming BDU and the attachment node

Although very difficult challenges associated with each of these discourse parsing tasks remain, in developing the Unified Linguistic Discourse Model (U-LDM) we have made significant progress recently on solving them. These are discussed in Sections 3 and 4 below.

3 Discourse Segmentation

The problem of segmenting discourse into the elementary units appropriate for building up the

structure of the discourse is an extremely difficult one. Each discourse theory must specify how "segments" should be identified in light of the questions the theory is set up to answer. Models based on Grosz and Sidner's 1986 work, especially those which form the basis of spoken language systems, define segments in terms of the intentions of the speaker: when the speaker's intention shifts, the segment associated with that intention ends and immediately following talk is included in new (or resumed) segments. While very useful in dealing with task oriented talk where speakers move between asking questions, informing others and giving commands, this model is less applicable to determining discourse segments within a sentence. The problem is an acute one for the analysis of written texts because often a subsequent, not necessarily adjacent, segment will continue the development of material introduced in a sub-sentential, often subordinate, constituent. Construction of the appropriate representation of the rhetorical or semantic structure of discourse must therefore keep sub-sentential units available for attachment at independent nodes on the tree along. The entire sentence or sentential main constituent must also be available to be continued after any continuation on sub-sentential units has been completed. As reported by Carlson et al. (2003), under RST[2], lexical and syntactic information used to segment discourse into *Elementary Discourse Units* (*EDUs*) is based on verbal constituents including clauses and infinitives.[3]

As we show below, the approach taken to segmentation under the U-LDM, while it includes as segments (and non-segments) many of the constructions currently used in RST, provides a rationalization for the choice of units. Rather than posit which syntactic objects function as discourse segments, we started by establishing the semantic basis for functioning as a segment and then identified which syntactic constructions carry the semantic information needed for discourse segment status. We then identified as Basic Discourse Units (BDUs) segments that have the potential to *independently* establish an anchor point for future continuation. We then drew a further distinction between BDUs as a class of syntactic structures with the potential to establish anchor points and the actual BDUs in a given sentence which can function as indexical anchor points in a specific discourse. We believe these distinctions, while cumbersome, are necessary for both theoretical and practical text analysis.

3.1 Discourse Segments under the U-LDM

As a semantic theory, the U-LDM must account for the interpretation of utterances. Specifically, we must account for the availability for update of appropriate discourse contexts or sub-contexts introduced in earlier text. In order to do so, we must be able to match incoming discourse utterances with their target contexts, some of which may have been introduced in syntactically subordinated positions within a sentence. *Therefore, in designing U-LDM discourse segmentation, we have identified the syntactic reflexes of the semantic content of the linguistic or paralinguistic phenomena making up discourse*.

Since elementary discourse units are needed to build up discourse structure recursively, we have identified as discourse segments the syntactic constructions that encode a minimum unit of **meaning** and/or discourse **function** interpretable relative to a set of contexts. We understand a minimum unit of *meaning* to communicate information about not more than one "event", "event-type" or state of affairs in a "possible world" of some type[4]. Clauses, and many other verb based structures, carry indexical information that ties the content to the context in which it is to be interpreted. Minimal *functional* units, on the other hand encode information about how previously occurring (or possibly subsequent) linguistic gestures relate structurally, semantically, interactionally or rhetorically to other units in the discourse or to information in the context in which the discourse takes place[5].

Examples of discourse segments are given in Table 1. Note that while discourse segments under the U-LDM are the syntactic reflex of a linguistically realized semantic "gesture" interpreted relative to context, they need not be contiguous, but may completely surround another segment (e.g. an appositive, or non-restrictive relative clause.) *Discontinuous seg-*

[2] Under S-DRT, no explicit structural tree is constructed and no explicit segmentation criteria have been proposed in the literature.

[3] Although some clauses are not treated as elementary units and "a small number of phrasal EDUs are allowed, provided that the phrase begins with a strong discourse marker."

[4] Roughly speaking an "elementary proposition", "event-type predicate" etc. In a Davidsonian style semantics, quantification over an event variable signals a separate unit of meaning.

[5] Greetings, discourse PUSH/POP markers and other "cue phrases", connectives etc. are all functional segments.

Segments	Common realizations	Examples
Content segments		
Eventualities (activities or states) and their participants.	Clauses: main, subordinate	[I heard the dog] [that was barking.]
	Predication	[California elected Schwarzenegger] [governor]
	Participial modifiers	[The donkey [braying next door] was annoying.]
	Infinitival modifiers	[We persuaded them] [to leave.] [They left] [to get the tickets.]
Interpolations	Parentheticals	[The show [(and what a show it was)] lasted 4 hours.]
	Appositives	[The building, [an example of the Mozarabic style,] was recently restored.]
	Interruptions	[They were [– Stop that! –] leaving at 8:00.]
Fragments	Section headings	[4. Discussion]
	List items	[e.g., [hydrogen,] [helium]]
	"Restarted" material	[My dog,] [no,] [my cat ran away.]
Operator Segments		
Conjunction	Conjunctions	[We arrived] [and] [got seats.]
Discourse operators	"scene-setting" preposed modifiers	[On Tuesday,] [we will see the sites.]
	"cue" words	[Anyway,] [we did get there on time.]

Table 1. Examples of Discourse Segments, Unlabelled bracketing is used to indicate segments.

ments occur when there is overt material on both sides of the intervening segment. With *fragmentary segment*s, the full interpretation remains unrecoverable from surrounding context. For example: a single word answer to a question is a complete segment, whereas the same word uttered but "left hanging" would be an uninterpretable fragment. (See Appendix for extensive example of a segmented text.)

3.2 Basic Discourse Units

An important contrast between the U-LDM and other approaches to segmentation concerns the distinction made in the U-LDM between discourse segments such as those we have identified above and Basic Discourse Units (BDUs). *While all BDUs under the U-LDM are segments, not all segments are BDUs. BDUs, under this model, are discourse segments of a type that can be independently continued*: operator segments are one example of non-BDU segments. Other verb bases constituents that might be expected to be segments are not because they do not establish an interpretation context independent of other segments that can be updated by subsequent units. In general, these "notable non-segments", summarized in Table 2, are heavily integrated into other nominal or verbal constructions and cannot be accessed for independent continuation.

Non-segments	Examples
Gerunds	[Singing is fun.]
Nominalizations	[Rationalization is useless.]
Auxiliary and modal verbs	[I might have succeeded.]
Clefts	[It was the tiger that we liked best.]

Table 2. Notable non-segments (underlined). [6]

In order to account for continuation in specific sentences, we further identify one class of instances of BDU: Active BDUs (A-BDUs) are BDUs on the right edge of a discourse tree. The main clause of any sentence will be an A-BDU and, depending on the deployment of BDU segments within a given sentence, other BDUs may also be accessible for continuation. (See Section 4 below.)

4 Discourse Parsing with the U-LDM

Ascertaining the relationship of a BDU to the discourse is a complex parsing process involving lexical, semantic, structural and syntactic information[7]. For the case of written prose we are concentrating on here, the unit of analysis is the sentence (or sentence fragment). Sentences are attached to the DPT of the text as a unit[8]. Discourse attachment of the sentence involves two decisions: where along the right edge to attach, and what is the relationship to the attachment point. The process, which includes constructing a BDU tree of the sentence, can be summarized as follows:

[6] In answer to a reviewer who asked if in "Singing is fun", *singing* should not be an independent segment, we would answer that this sentences concerns one eventuality (something being fun), not two. Since any noun can be referred to by a pronoun in the next sentence simply referring to the noun is not equivalent to referring to the eventuality in which the referent of the noun is a participant.

[7] Although the linguistic (and lexical) information we discuss could be augmented with processes relying upon high level world knowledge and inference, we believe that it is extremely significant to see how far one can get with discourse parsing without invoking non-linguistic information.

[8] See discussion of MBDU below.

- Identify potential BDUs within sentence using sentential syntax
- Construct a BDU-tree from the segments of the sentence, using sentential syntactic information and discourse rules to map segments and relationships among them. This BDU-tree is itself an Open Right Tree dominated by the node corresponding to the Main clause of sentence[9]. (This is the *Main BDU* or *MBDU*).
- Attach the BDU-tree as a unit to the Discourse Parse Tree by computing the relationship of MBDU and preposed modifiers, if any, to accessible DCUs aligned along the right edge of the tree using rules of discourse relations (See Section 4.1 below). Lexical information used for attachment decisions can come from anywhere in the BDU tree.
- Once the BDU-tree is attached, its terminal leaves are terminal nodes of the Discourse Parse Tree (DPT) and any terminal or intermediary nodes on the right edge of the BDU tree are DCUs on the DPT accessible for attachment in the next iteration of the process.

In order to determine which accessible DCUs are candidates for M-BDU attachment and what relationship obtains between the incoming unit and the selected DCU, a number of distinct types of evidence are used, including:

1) **lexical information**
   reuse somewhere in the BDU tree of the same lexeme, synonym/antonym, hypernym, or participation in the same lexical frame or "semantic field" as item in target node.
2) **syntactic information**
   parallel syntactic structure; topic/focus and centering information, syntactic status of reused lexemes, pre-posed adverbial constituents, etc.
3) **semantic information**
   realis status, genericity, tense, aspect, point of view etc. in the MBDU
4) **constituents of incomplete n-ary constructions on the right edge**
   Questions, initial greetings, genre-internal units like sections and sub-sections, etc.
5) **structure of both the local attachment point and the BDU-tree**

While we are still experimenting with understanding the complexities involved in attachment, we believe that different types of evidence have different weights[10] and that the combined weight of evidence determines the attachment point. We have noted, however even at this stage of our investigations, that the weight given to each type of information differs for attachment site selection and relationship determination. Lexical information, for example, is often very important in determining site, while semantic and syntactic information is most relevant in determining relationship. In the remainder of this section we will give a small set of robust rules for determining the attachment site and relationship of an incoming BDU-tree to the existing parse tree of the discourse.

4.1 Rules for Determining Discourse Attachment Site Candidates and Attachment Relations

Both the attachment site choice and the actual attachment process rely on partially ordered sets of hybrid rules, each of which are conditioned on a set of constraints. Constraints for rules used in attachment site selection are primarily lexical constraints, although other information is also relevant.

All types of evidence play a role in choosing the attachment relation. A rule is a pair: **Rule <C, O>** where **C** is the set of constraints that enable the rule and **O** is the associated operation. The operation associated with a rule can therefore be either the markup of a DCU as a possible attachment site, or an actual discourse relation, such as Subordination, Coordination or N-ary. A rule is enabled when all sub-conditions in ***C*** are satisfied and no other rules having priority are enabled. Rules may combine different sources of evidential information (semantic, syntactic, structural and lexical). If more than one rule is enabled at the same time, ambiguous parses are produced[11]. Some rules are listed in Table 3.

[9] This process is too complex to describe in detail here but it involves looking at both the F-structure of the sentential parsing information returned by the XLE and applying discourse rules to the BDUs identified. Soricut and Marcu (2003) also build up RST sentential trees to use in discourse parsing. Both the information and methods used to construct RST trees as well as the trees themselves differ from ours.

[10] We assign weights heuristically at this point.

[11] At this stage in our research, we rely only on a partial order among the rules. In future work, we will investigate (1) how evidence is weighed and combined in order to make better attachment deci-

Attachment Relation	Sub Conditions
Nary-Attachment	Frame(AP,MBDU) matches genre-specific construction Greetings, Argument, Question/Answer, Speech Event, Genre Meta Structure(Story, Technical Paper, Lecture, etc..) Reported speech/reporting clause
Subordination	M-BDU Realis status differs from Status of AP (MBDU is *Irrealis;* AP is R*ealis* OR MBDU is *Realis;* AP is *Irrealis*)
Nary-Attachment (intrasentential)	Tense(AP) = *past* Tense(MBDU) = *pluperfect* AP is time-reference for MBDU
Nary-Attachment (intrasentential)	VerbClass(AP)="SpeechAct" Type(MBDU) = ADJUNCT
Nary-Attachment (intrasentential)	Tense(AP) = *present* Tense(MBDU) = *past* AP is time-reference for MBDU
Coordinate	Parent(AP) is Coordination Parent(AP) would coordinate with MBDU AP would coordinate with MBDU
Subordination	Tense(AP) = *past* Genericty(AP) = *specific* Tense(MBDU) = *present* Genericty(MBDU) = *generic*
Subordination	M-BDU genericity status differs from Status of AP (MBDU is *specific;* AP is *generic* OR MBDU is *generic;* AP is *specific*)
Subordination	SUBJ(MBDU) = OBJ(AP)
Subordination	SUBJ(MBDU) = XCOMP(AP)
Subordination	MBDU/Lexeme is a subcase of AP/Lexeme Role(AP/Lexeme) = Role(MBDU/Lexeme)
Right Headed Subordination (intrasentential)	Type(AP) = ADJUNCT Type(MBDU) = S
Nary-Attach (intrasentential)	PRED(ADJUNCT(AP)) = "if" AP is *Irrealis* MBDU is *Realis*
Nary-Attachment (intrasentential)	AP and MBDU related by logical connective (cf Webber& Joshi, 1998; Forbes (2003)
Subordination	Tense(AP) = *past* Tense(MBDU) = *pluperfect*
Subordination	Tense(AP) = *present* Tense(MBDU) = *past*
Subordinate	AP is Bottom of DPT M-BDU is Footnote or Parenthetical
Coordinate	AP is Narrative(= *Specific*, *punctual* ,*event*) MBDU is Narrative
Coordinate	Tense(AP) = Tense(MBDU) Aspect(AP) = Aspect(MBDU)
Coordinate	MBDU/Lexeme is synonym or antonym of AP/Lexeme Role(AP/Lexeme) = Role(MBDU/Lexeme)
Subordinate	AP is Bottom of DPT

Table 3. Discourse Attachment Rules ordered to express priority of the rules. AP denotes (potential) attachment point.

The parsing process at the Discourse Parse Tree (DPT) level works as follows. When a BDU-Tree has been constructed and is ready to be attached to the right edge of the DPT, each DCU along the right edge is examined and the lexical information in the right-edge DPT nodes are compared with the lexical evidence retrieved from the incoming BDU-Tree. This process, guided by the set of discourse rules, produces an ordered set of active DCUs, representing the possible attachment points in order of likelihood. The set can then be pruned of its n lowest scoring constituents, according to an appropriate policy such as a threshold.

In a second stage, each attachment rule is checked against possible attachment sites. Rules that fire successfully attach the BDU-Tree to the DPT at the chosen site with the relationship specified by the rule. Local semantic, lexical and syntactic information is then percolated up to the

sions and (2) the extent to which discourse ambiguity generated in this fashion is legitimate and how to reduce grammar overgeneration by more efficient handling of interactions among rules and the weighing of the linguistic evidence.

DCU consisting of the parent of both attachment point and incoming MBDU according to constraints of the discourse relation selected. If multiple attachments at different sites are possible, ambiguous parses are generated; less preferred attachments are discarded and the remaining attachment choices generate valid parse trees.

5 PALSUMM Text Summarization

So far, we have described the U-LDM only as a theoretical approach to discourse parsing. We now turn briefly to describe a computational implementation of these methods. The PALSUMM Text Summarization System is a domain independent symbolic sentence extraction system that produces high level readable summaries that preserve the language and style of the original text and eliminate problems with unresolved or incorrect reference. Our system is currently used to summarize a corpus of 300 technical reports produced by our laboratory. [12]

The PALSUMM System relies on the Xerox Linguistic Environment (XLE) to parse the sentences of our source texts. The f-structure output of the XLE parser is segmented into units according to the criteria identified above. The segments are then combined into a BDU-tree. Using syntactic information about syntactic coordination and subordination relations, lexical ontological information taken from WordNet and a customized lexical domain ontology as well as discourse rules, the M-BDU of the sentence along with any other BDUs that must be accessible along the right edge of the discourse tree to accommodate possible continuations are identified, Both the site of attachment and the attachment relation are then computed using discourse attachment rules of the type presented above. Text summarization algorithms are then applied to the resulting tree.

Running in purely symbolic mode, the tree is pruned at a given level of embeddedness to produce a summary of a desired length or degree of summarization.[13] Because the resulting summaries may be longer than desired, alternatively we also use statistical methods to identify salient information (see discussion and references in Marcu 2003) and then construct a partial discourse tree that includes only information identified as most salient and the text at all nodes dominating that salient information.

6 Conclusions and Directions for future work

The U-LDM discussed in this paper represents a significant advance in the theoretical understanding of the nature of discourse structure. The explicit rules for discourse segmentation based on the syntactic reflexes of semantic structures allow analysts for the first time to relate the semantics underlying the syntactic structure of sentences to the discourse segments needed to account for continuity. In order to adapt the rules to other languages which may have different syntactic reflexes of semantic information, understanding the semantic justification for the choice of segments is important. In addition, the rules for discourse attachment for the first time make clear the principles of discourse continuity for "coherent" discourse. In the future, we plan to deepen our understanding of the rules for discourse attachment and, in particular, begin to apply machine learning techniques to increase our understanding of the complex interrelationship that obtain among them.

While full implementation of the principles of discourse organization outlined here are beyond the state of the art in some respects (i.e. determining that a sentence is generic in English is non-trivial in many instances although machine learning techniques might be useful in this regard), we believe that the PALSUMM System demonstrates the practicality of symbolic discourse parsing using the U-LDM Model. The infrastructure for this system has been successfully applied to the task of summarizing documents without a complex semantic component, extensive world knowledge and inference or a subjectively annotated corpus. We believe that the U-LDM parsing methods discussed here can be used for all other complex NLP tasks in which symbolic parsing is appropriate, especially

[12] For illustration purposes, we present in Appendix A a summary of a document that was hand coded using the rules given and then summarized automatically using the PALSUMM tree pruning algorithm. The PALSUMM Summaries were judged to be significantly more readable than summaries produced by MEAD in a small comparative study. In Appendix B, we present a diagram of the PALSUMM system.

[13] Although closely related to methods reported by Marcu (1999, 2000) for summarization using RST trees, our basic algorithm is essentially simpler because RST trees are dependency trees over a large set of different link types, whereas LDM trees are constituent trees over effectively two basic node types: subordinations and non-subordinations.

those involving high value document collections where precision is critical. In addition, the structures generated through symbolic parsing by the system will be invaluable for training statistical and probabilistic systems.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in English: A Philosophical Semantics for Natural Language Metaphysics.* Kluwer Academic Publishers.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. To appear. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue,* Jan van Kuppevelt and Ronnie Smith eds. Kluwer Academic Publishers.

Katherine M. Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi and Bonnie Webber. 2003. D-LTAG System - Discourse Parsing with a Lexicalized Tree-Adjoining Grammar, Journal of Language, Logic and Information, 12(3).

Barbara Grosz and Candace Sidner. 1986. Attention, Intention and the Structure of Discourse. *Computational Linguistics 12*:175-204..

Hans Kamp. 1981. A theory of truth and semantic representation." In *Formal Methods in the Study of Language.* Jeroen A. G. Groenendijk, Theo Janssen, and Martin Stokhof (eds.). Amsterdam: Mathematisch Centrum, 277-322.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory**:** Towards a Functional Theory of Text Organization. *Text* 8(3)243-281.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. In *Advances in Automatic Text Summarization*. I. Mani and Mark Maybury (Eds.), 123-136, The MIT Press.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press. Cambridge, MA.

Daniel Marcu. 2003. Automatic Abstracting, *Encyclopedia of Library and Information Science*, 245-256, 2003.

Livia Polanyi. 1988. A Formal Model of Discourse Structure. *Journal of Pragmatics* 12: 601-639.

Livia Polanyi and Martin van den Berg. 1996. Discourse Structure and Discourse Interpretation. In *Proceedings of the 10th Amsterdam Colloquium on Formal Semantics*. University of Amsterdam.

Livia Polanyi and Remko Scha. 1984. A syntactic approach to discourse semantics. In *Proceedings of COLING 6*. Stanford, CA. 413-419.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, and John Blitzer, Arda Çelebi, Elliott Drabek, Wai Lam, Danyu Liu, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel. 2003. "The MEAD Multidocument Summarizer". `http://www.summarization.com/mead/`

Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *Proceedings of HLT/NAACL'03*, May 27-June 1, Edmonton, Canada

Bonnie Webber.1991. Structure and Ostension in the Interpretation of discourse Deixis. In *Language and Cognitive Processes*, 6(2):107-135.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. *ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada.

Alec Wilkinson. 2003. *Talk of the Town, September 26*, 2003, New Yorker

APPENDIX A. PALSUMM Example

The text below, taken from a recent issue of *The New Yorker* magazine (Alec Wilkinson, 2003)[14], has been analyzed by hand using the segmentation and discourse structure construction rules given in Sections 3 and 4 above, resulting in the Discourse Parse Tree given in Figure 1. The summary of the text was automatically generated using the automatic summarization algorithm mentioned in Section 5 and a Genre Specific rule for stories in which stories are treated as consisting of an Orientation, Narrative and Coda. The first specific, non-habitual eventive clause closes the Orientation and begins the Narrative Section. The function of a Coda is to make the point of a story explicit. This is often done, as in the present case, by using an anaphor that refers to an entire section of text (Webber, 1991)[15].

(1) In the spring of 1947, (2)William Katavolos is the solitary occupant of the Ram's Head Inn, (3) on Ram Island,(4) off eastern Long Island. (5) Katavolos is twenty-three. (6) His father has leased the inn. (7) Katavolos has returned from the war (8) and (9) wants a place (10) where he can paint (11) and (12) be left alone. (13) The hotel is reached by a causeway from Shelter Island, (14) and the causeway sometimes floods, (15) leaving Katavolos as isolated as a lighthouse keeper. (16) To amuse himself one evening, (17) he puts some water in a glass, (18) covers the rim of the glass with waxed paper, (19) then presses the paper into the water (20) to create a vacuum. (21) He secures the paper to the glass with a rubber band, (22) then turns the glass upside down. (23) The water fills the vacuum, (24) preserving the dome (25) — it looks like the bottom of a wine bottle. (26) Then he begins to wonder (27) what would happen (28) if he repeated the experiment on a larger scale.(29) A few days later, (30) he throws a tarpaulin over a section of Gardiners Bay (31) He weights down the edges (32) so that no air can get beneath the tarpaulin, (33) then he swims underneath it. (34) Using two oars, (35) he raises the center of the tarpaulin. (36) The water fills the cavity (37) and he swims into it, (38) floating above sea level, (39) which, (40) he says later, (41) "fascinated the hell out of me." (42) This is the beginning of (43) what Katavolos will call hydronics, (44) the practice of making buildings from soft plastic forms (45) filled with water. (46) In 1949, (47) Katavolos gives up painting (48) to design furniture (49) — his chairs are in the collections of the Museum of Modern Art, the Metropolitan Museum, and the Louvre— (50) and, (51) in 1960, (52) he begins teaching architecture at the Pratt Institute, (54) in Brooklyn, (55) where he will become the co-director of the Center for Experimental Structures. (56) In 1970, (57) in a courtyard at Pratt, (58) he builds the first hydronic structure (59) — a plastic dome filled with water (60) and supported by a plastic cylinder, (61) also filled with water. (62) The plastic is like Saran Wrap, (63) only thicker. (64) Each year after that, (65) he builds a new structure (66) He calls the structures (67) liquid villas. (68) They consist of columns, arches, and vaults. (69) The elements, (70) that is (71) of classical architecture.

Summary 152/363 = 42%

In the spring of 1947, William Katavolos is the solitary occupant of the Ram's Head Inn, Katavolos is twenty-three. To amuse himself one evening, he puts some water in a glass, covers the rim of the glass with waxed paper, then presses the paper into the water. He secures the paper to the glass with a rubber band, then turns the glass upside down. A few days later, he throws a tarpaulin over a section of Gardiners Bay.

He weights down the edges then he swims underneath it. Using two oars, he raises the center of the tarpaulin. The water fills the cavity, This is the beginning of what Katavolos will call hydronics. In 1949, Katavolos gives up painting and in 1960 he begins teaching architecture at the Pratt Institute. In 1970, in a courtyard at Pratt, he builds the first hydronic structure. Each year after that, he builds a new structure.

[14] Alec Wilkinson. 2003. *Talk of the Town, September 26*, 2003, New Yorker

[15] Bonnie Webber.1991. Structure and Ostension in the Interpretation of discourse Deixis. In *Language and Cognitive Processes*, 6(2):107-135.

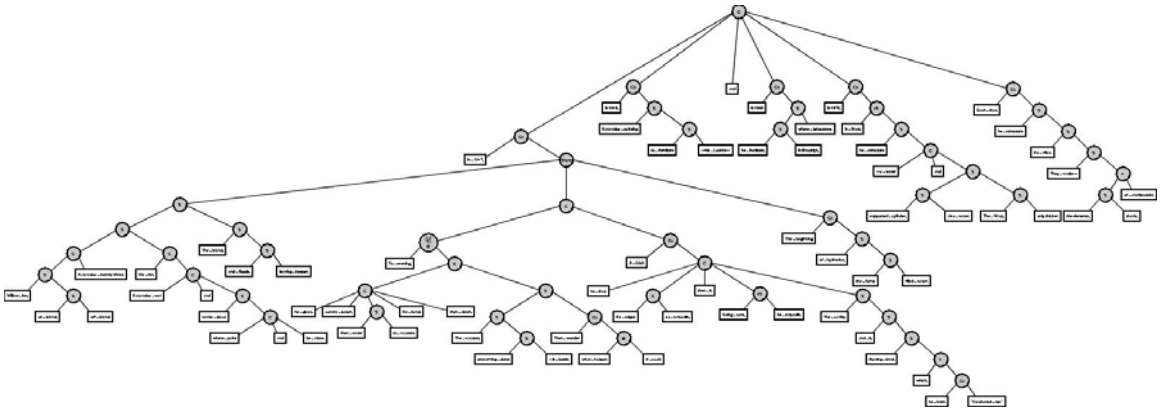

Figure 1. Discourse Parse Tree of the New Yorker text.

APPENDIX B. System Diagram

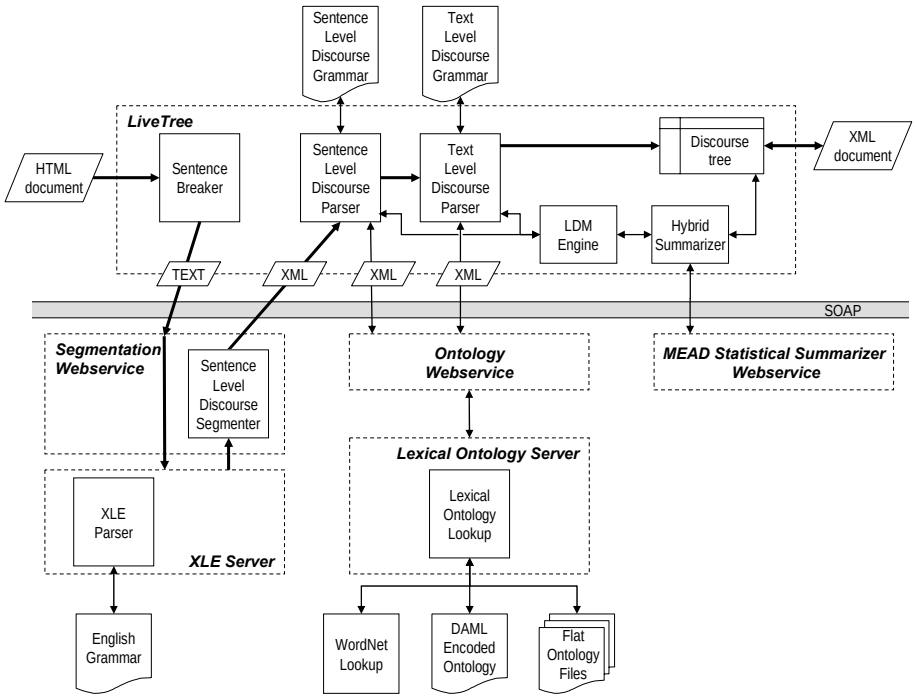

Figure 2. Diagram of the PALSUMM system, a symbolic summarization system currently being developed at FX Palo Alto Laboratory.