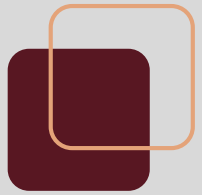


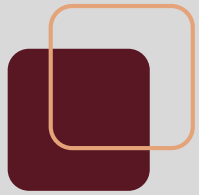
The Penn Discourse Tree Bank

Nikolaos Bampounis

20 May 2014

Seminar:
Recent Developments
in Computational Discourse Processing

What is the PDTB?

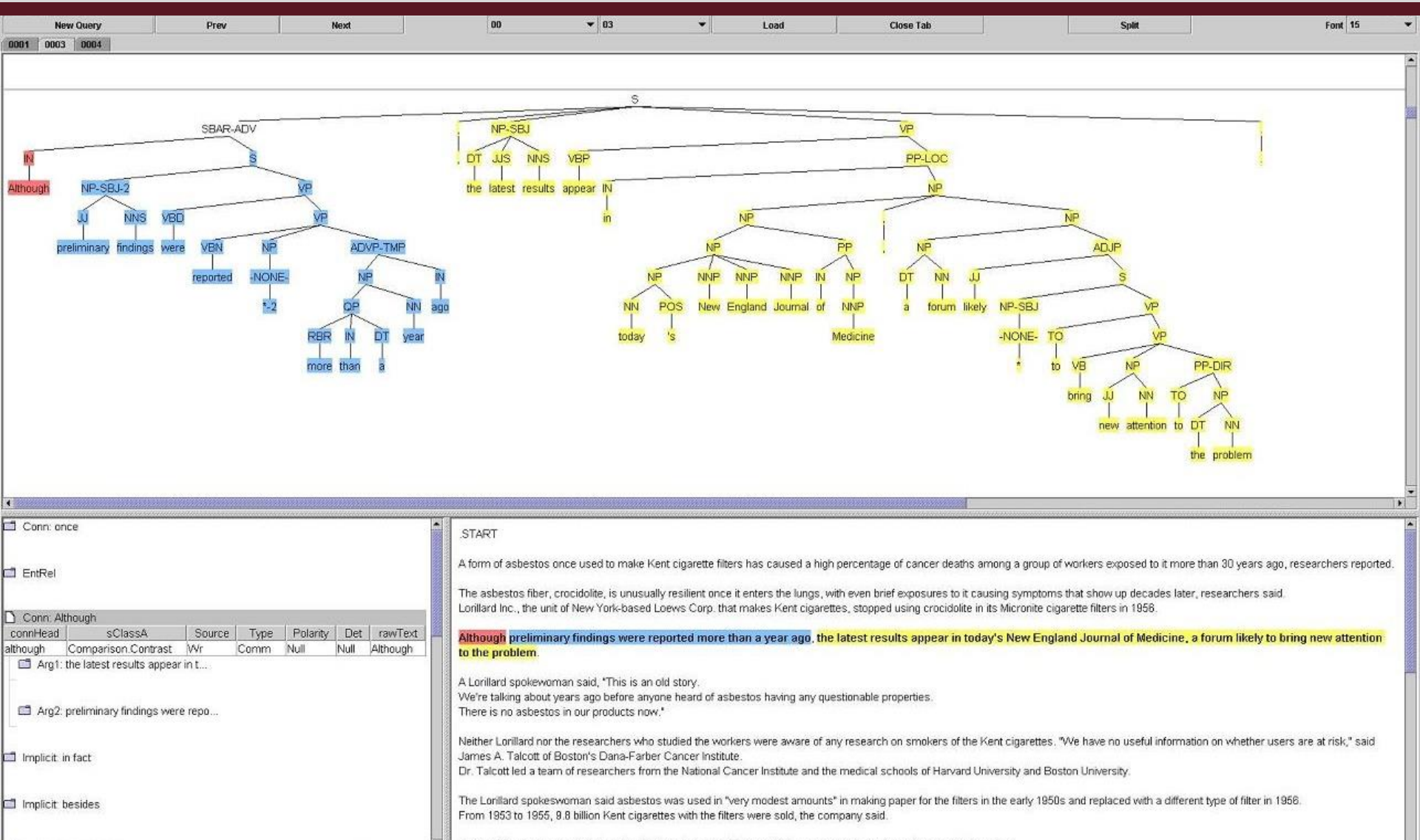


- Developed on the 1 million word WSJ corpus of Penn Tree Bank
- Enables access to syntactic, semantic and discourse information on the same corpus
- Lexically-grounded approach

Motivation

- Theory-neutral framework:
  - ❑No higher-level structures imposed
  - ❑Just the connectives and their arguments
- Validation of different views on higher level discourse structure
- Solid training and testing data for LT applications

How it looks


New Query
Prev
Next
00
03
Load
Close Tab
Split
Font 15
0001
0003
0004
S
SBAR-ADV
IN
Although
S
NP-SBJ-2
VP
JJ
NNS
VBD
VP
preliminary
findings
were
VBN
NP
ADVP-TMP
reported
-NONE-
NP
IN
*-2
QP
NN
ago
RBR
IN
DT
year
more
than
a
NP-SBJ
VP
DT
JJS
NNS
VBP
PP-LOC
the
latest
results
appear
IN
NP
in
NP
NP
NP
PP
NP
NNP
NNP
NNP
IN
NP
NN
POS
New
England
Journal
of
NNP
today
's
Medicine
NP
ADJP
DT
NN
JJ
S
a
forum
likely
NP-SBJ
VP
-NONE-
TO
VP
*
to
VB
NP
PP-DIR
bring
JJ
NN
TO
NP
new
attention
to
DT
NN
the
problem
Conn: once
EntRel
Conn: Although
connHead
sClassA
Source
Type
Polarity
Det
rawText
although
Comparison.Contrast
Wr
Comm
Null
Null
Although
Arg1: the latest results appear in t...
Arg2: preliminary findings were repo...
Implicit: in fact
Implicit: besides
.START
A form of asbestos once 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.
The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.
Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.
Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.
A Lorillard spokewoman said, "This is an old story.
We're talking about years ago before anyone heard of asbestos having any questionable properties.
There is no asbestos in our products now."
Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.
Dr. Talcott led a team of researchers from the National Cancer Institute and the medical schools of Harvard University and Boston University.
The Lorillard spokeswoman said asbestos was used in "very modest amounts" in making paper for the filters in the early 1950s and replaced with a different type of filter in 1956.
From 1953 to 1955, 9.8 billion Kent cigarettes with the filters were sold, the company said.

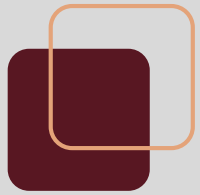
What is annotated

- Argument structure, type of discourse connective and attribution

  According to Mr. Salmore, the ad was "devastating" because it raised question about Mr. Counter's credibility. → **CAUSE**

- Connectives are treated as discourse level predicates with two abstract objects as arguments: because(Arg1, Arg2)
- Only paragraph-internal relations are considered

Connectives relations



- Explicit
- Implicit
- AltLex
- EntRel
- NoRel

Explicit connectives

- Straight-forward
- Belong to syntactically well-defined classes
  - Subordinate conjunctions: *as soon as, because, if etc.*
  - Coordinating conjunctions: *and, but, or etc.*
  - Adverbial connectives: *however, therefore, as a result etc.*

Explicit connectives

- Straight-forward
- Belong to syntactically well-defined classes

The federal government suspended sales of U.S. savings bonds because Congress hasn't lifted the ceiling on government debt.

Arguments

- Conventionally named Arg1 and Arg2

  The federal government suspended sales of U.S. savings bonds because Congress hasn't lifted the ceiling on government debt.

- The extent of arguments may range widely:
  - ❑A single clause, a single sentence, a sequence of clauses and/or sentences
  - ❑Nominal phrases or discourse deictics that express an event or state

Arguments

- Information supplementary to an argument may be labelled accordingly

  [Workers described "clouds of blue dust"] that hung over parts of the factory, even though exhaust fans ventilated the area.

Implicit connectives

- Absence of an explicit connective
- Relation between sentences is inferred
- Annotators were actually required to provide an explicit connective

Implicit connectives

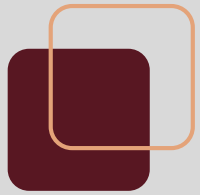
- Absence of an explicit connective
- Relation between sentences is inferred

The $6 billion that some 40 companies are looking to raise in the year ending March 31 compares with only $2.7 billion raised on the capital market in the previous fiscal year. [In contrast] In fiscal 1984 before Mr. Gandhi came to power, only $810 million was raised.

Implicit connectives

- But what if the annotators fail to provide a connective expression?

Implicit connectives



- But what if the annotators fail to provide a connective expression?

✓Three distinct labels are available:

- ❑AltLex
- ❑EntRel
- ❑NoRel

AltLex

- Insertion of a connective would lead to redundancy
- The relation is already **alt**ernatively **lex**icalized by a non-connective expression

After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%.
AltLex **The reason:** Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.

EntRel

- **Ent**ity-based coherence **rel**ation
- A certain entity is realized in both sentences

Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. EntRel **Mr. Milgrim** succeeds David Berman, who resigned last month.

NoRel

- **No** discourse or entity-based **rel**ation can be inferred
- Remember: Only adjacent sentences are taken into account

Jacobs is an international engineering and construction concern. NoRel Total capital investment at the site could be as much as $400 million, according to Intel.

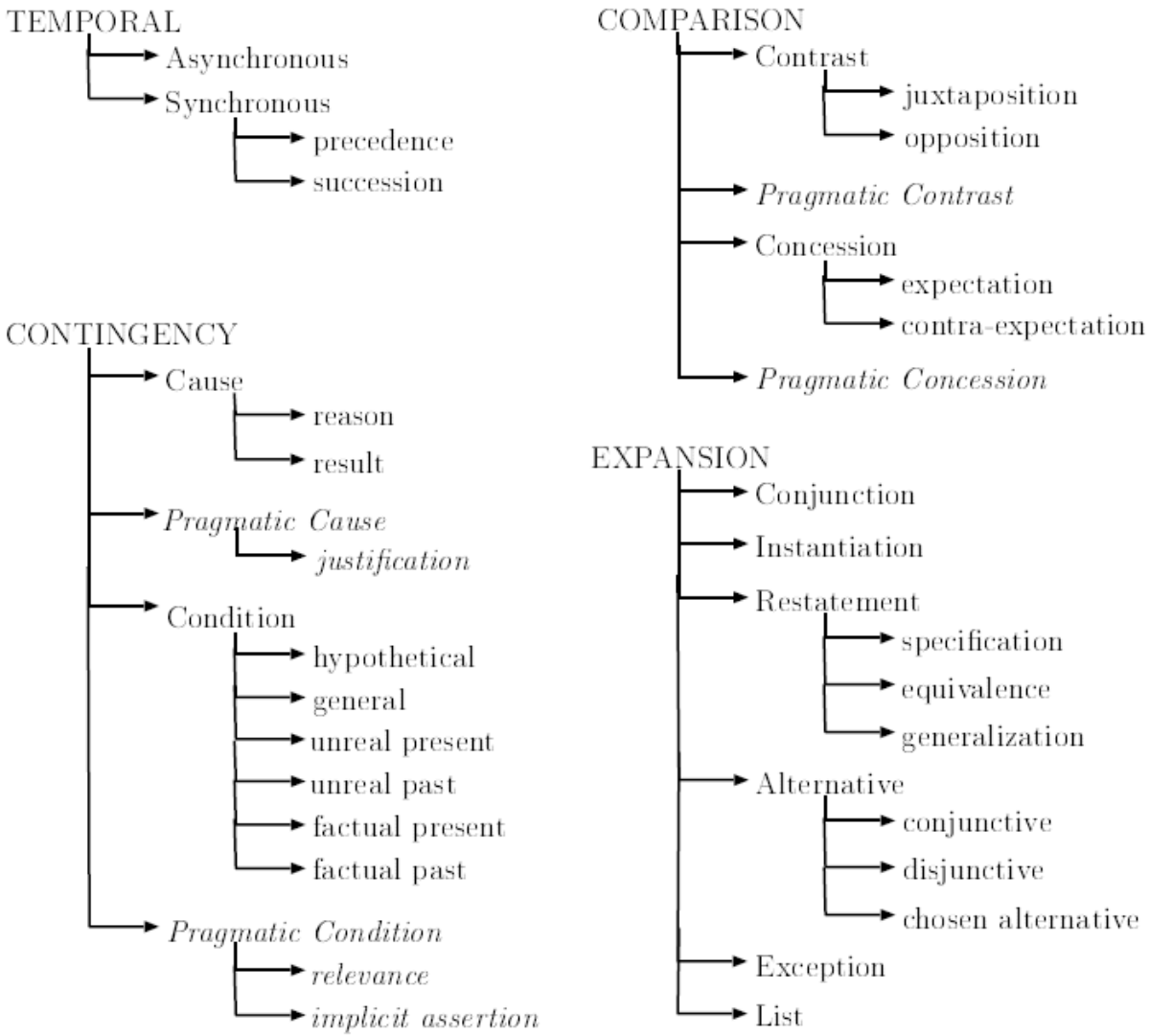
Senses

- Both explicit and inferred discourse relations (implicit and AltLex) were labelled for connective sense.

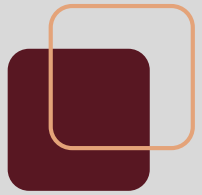
The Mountain View, Calif., company has been receiving 1,000 calls a day about the product since it was demonstrated at a computer publishing conference several weeks ago. → **TEMPORAL**

It was a far safer deal for lenders since NWA had a healthier cash flow. → **CAUSAL**

Hierarchy of sense tags



Attribution



- A relation of "ownership" between abstract objects and agents

  "The public is buying the market when in reality there is plenty of grain to be shipped,"

  said Bill Biedermann, Allendale Inc. director.

- Technically irrelevant, as it's not a relation between abstract objects

Attribution

- Is the attribution itself part of the relation?

When Mr. Green won a $240,000 verdict in a land condemnation case against the state in June 1983, he says Judge O'Kicki unexpectedly awarded him an additional $100,000.

Advocates said the 90-cent-an-hour rise, to $4.25 an hour, is too small for the working poor, while opponents argued that the increase will still hurt small business and cost many thousands of jobs.

Attribution

- Is the attribution itself part of the relation?
- Who are the relation and its arguments attributed to?
  - ❑the writer
  - ❑someone else than the writer
  - ❑different sources

Editions

- PDTB 1.0 released in 2006
- PDTB 2.0 released in 2008
  - ❑Annotation of the entire corpus
  - ❑More detailed classification of senses

Statistics

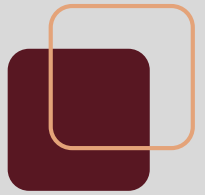
- Explicit: 18,459 tokens and 100 distinct connective types
- Implicit: 16,224 tokens and 102 distinct connective types
- AltLex: 624 tokens with 28 distinct senses
- EntRel: 5,210 tokens
- NoRel: 254 tokens

Let's practice!

- ✓ Annotate the text:
  - Explicit connectives
  - Implicit connectives
  - AltLex
  - EntRel
  - NoRel
  - Arg1/Arg2
  - Attribution
  - Sense of connectives

What about PDTB annotators?

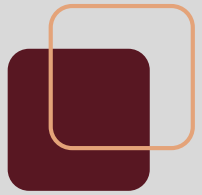
- Agreement on extent of arguments:
  - ❑90.2-94.4% for explicit connectives
  - ❑85.1-92.6% for implicit connectives
- Agreement on sense labelling:
  - ❑94% for Class
  - ❑84% for Type
  - ❑80% for Subtype



A PDTB-Styled End-to-End Discourse Parser

Lin et al., 2012

Discourse Analysis vs Discourse Parsing



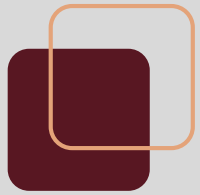
- **Discourse analysis:** the process of understanding the internal structure of a text
- **Discourse parsing:** the process of realizing the semantic relations between text units

The parser

- Performs parsing in the PDTB representation on unrestricted text
  - ✓Only Level 2 senses used (11 types out of 13)
- Combines all sub-tasks into a single pipeline of probabilistic classifiers[1]
- Data-driven

[1] OpenNLP maximum entropy package

The algorithm

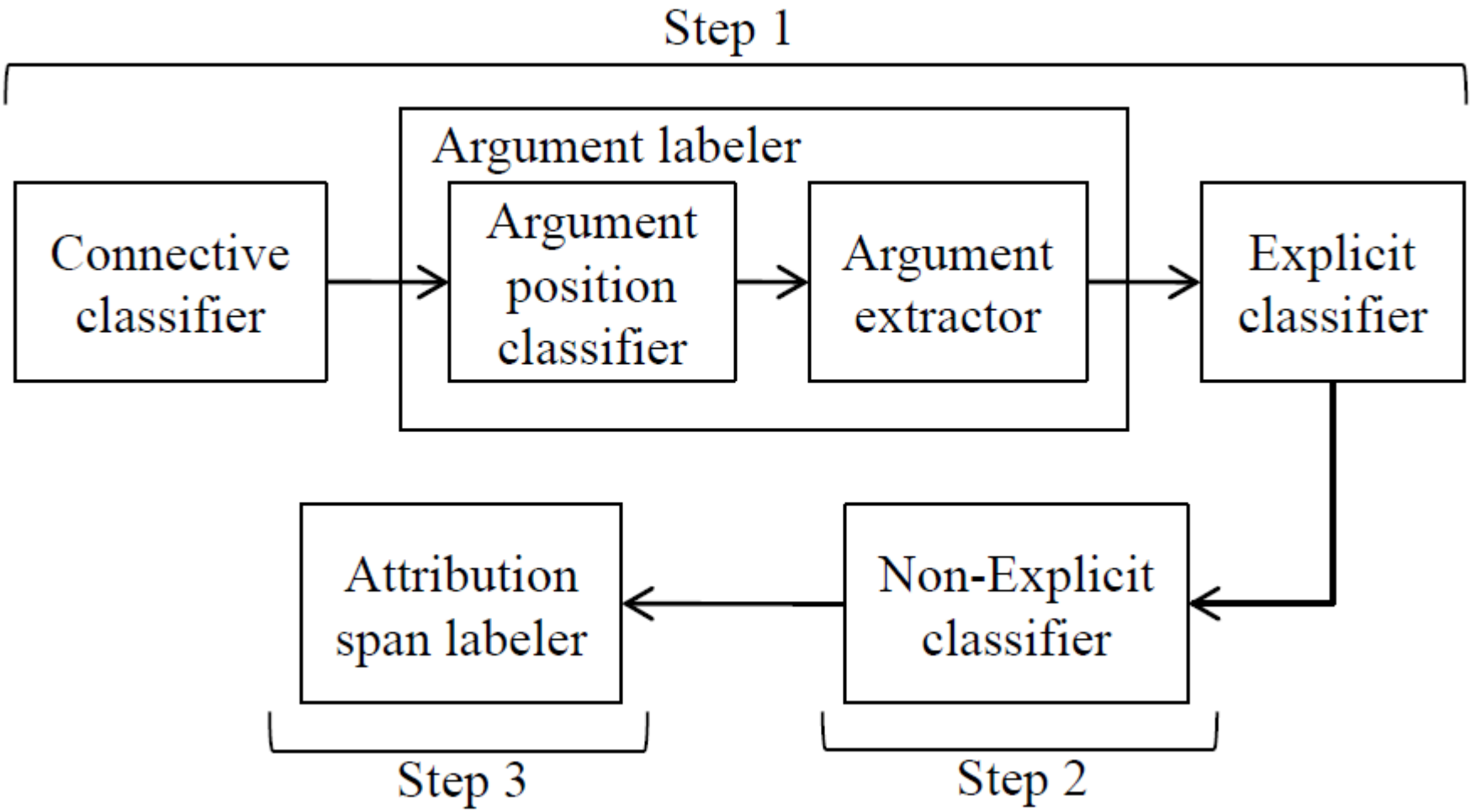


- Supposed to mimic the real annotation procedure

Input: free text T

Output: discourse structure of T

The system pipeline

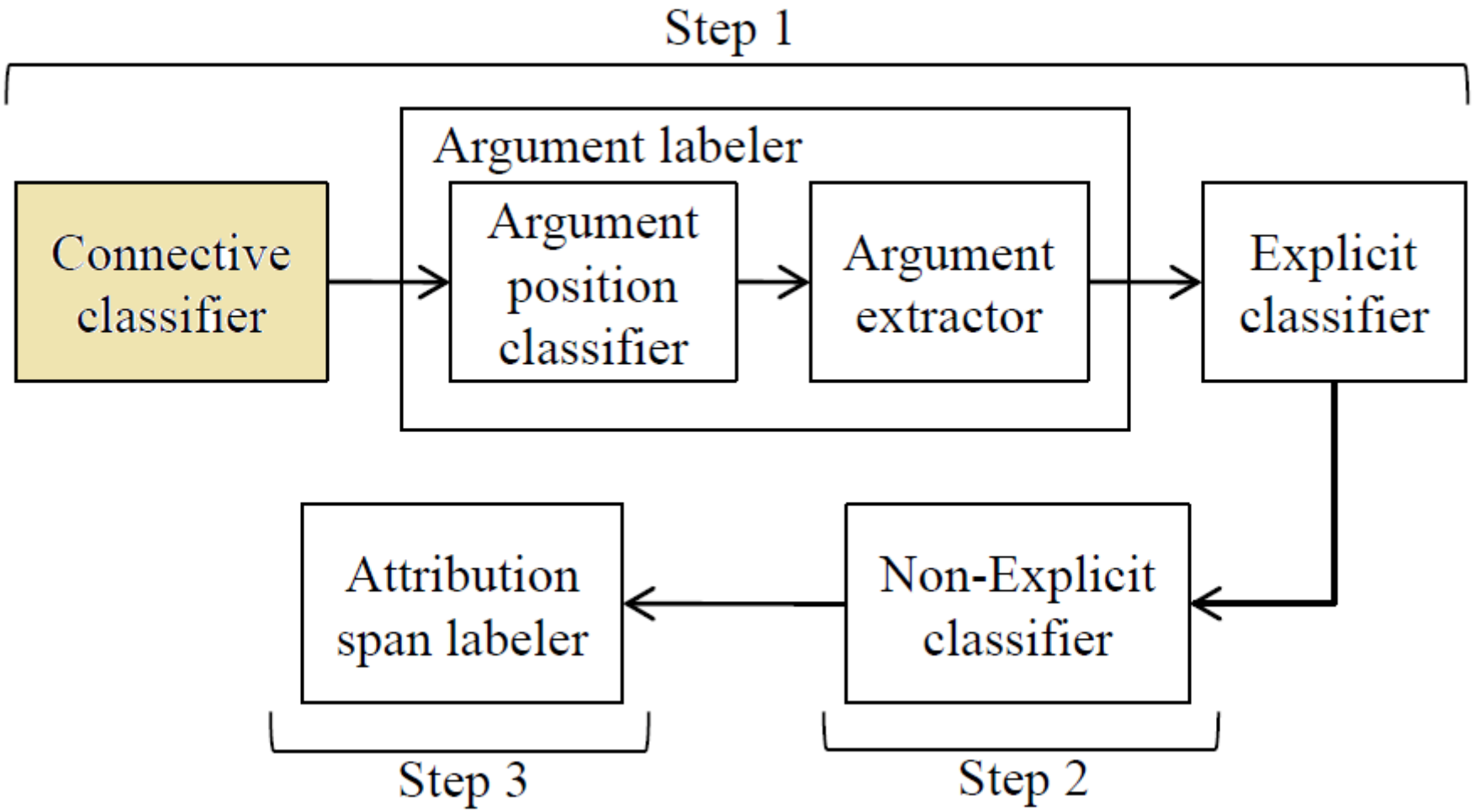


The evaluation method

- For the evaluation of the system, 3 experimental settings were used:
  - ❑GS without EP
  - ❑GS with EP
  - ❑Auto with EP

  GS: Gold standard parses and sentence boundaries
  EP: error propagation
  Auto: Automatic parsing and sentence splitting
- In the next slides, we will be referring to GS without EP

The system pipeline

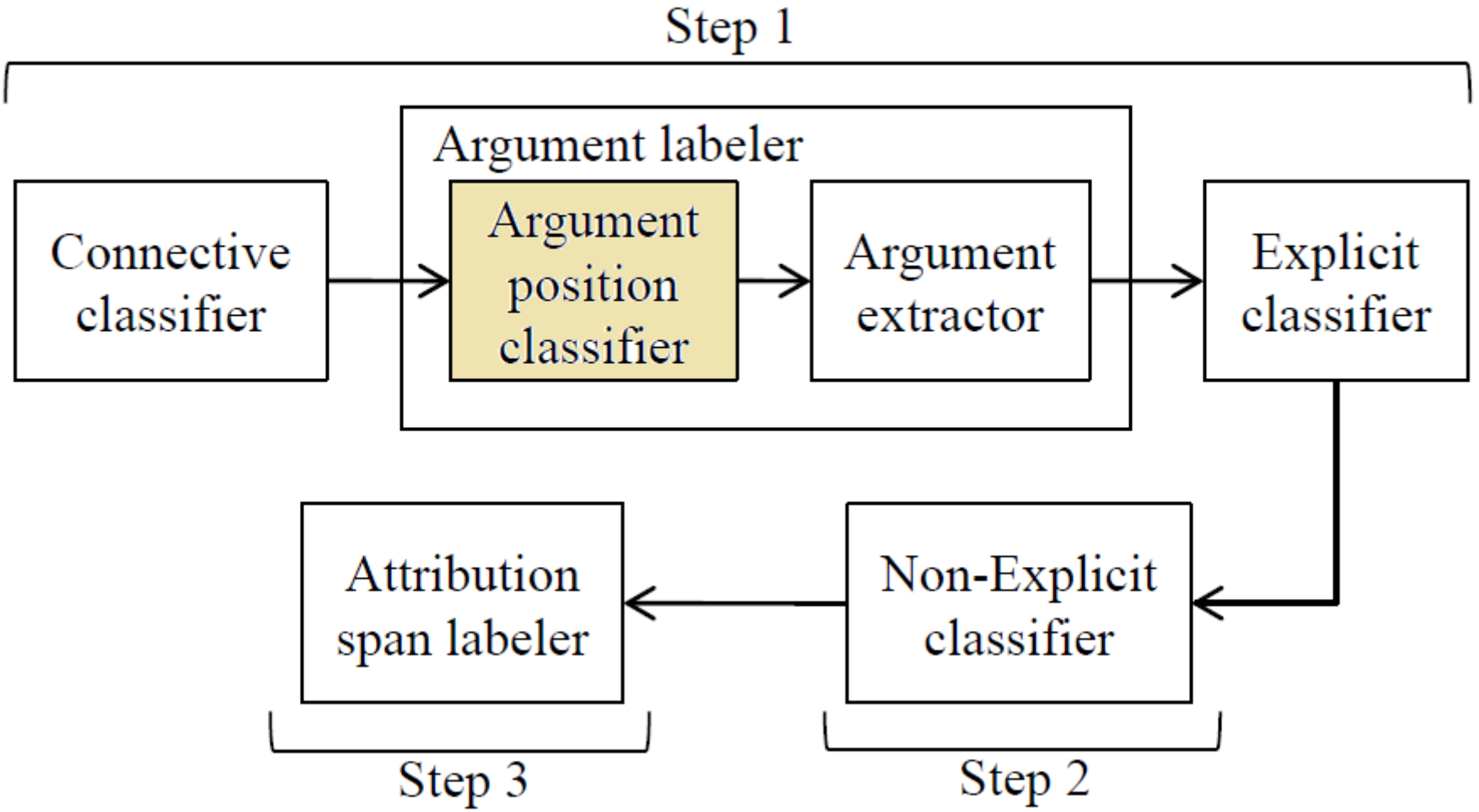


Connective classifier

- Finds all explicit connectives
- Labels them as being discourse connectives or not
  - ✓ Syntactic and lexico-syntactic features used

➢ F_1: 95.76%

System pipeline



Argument position classifier

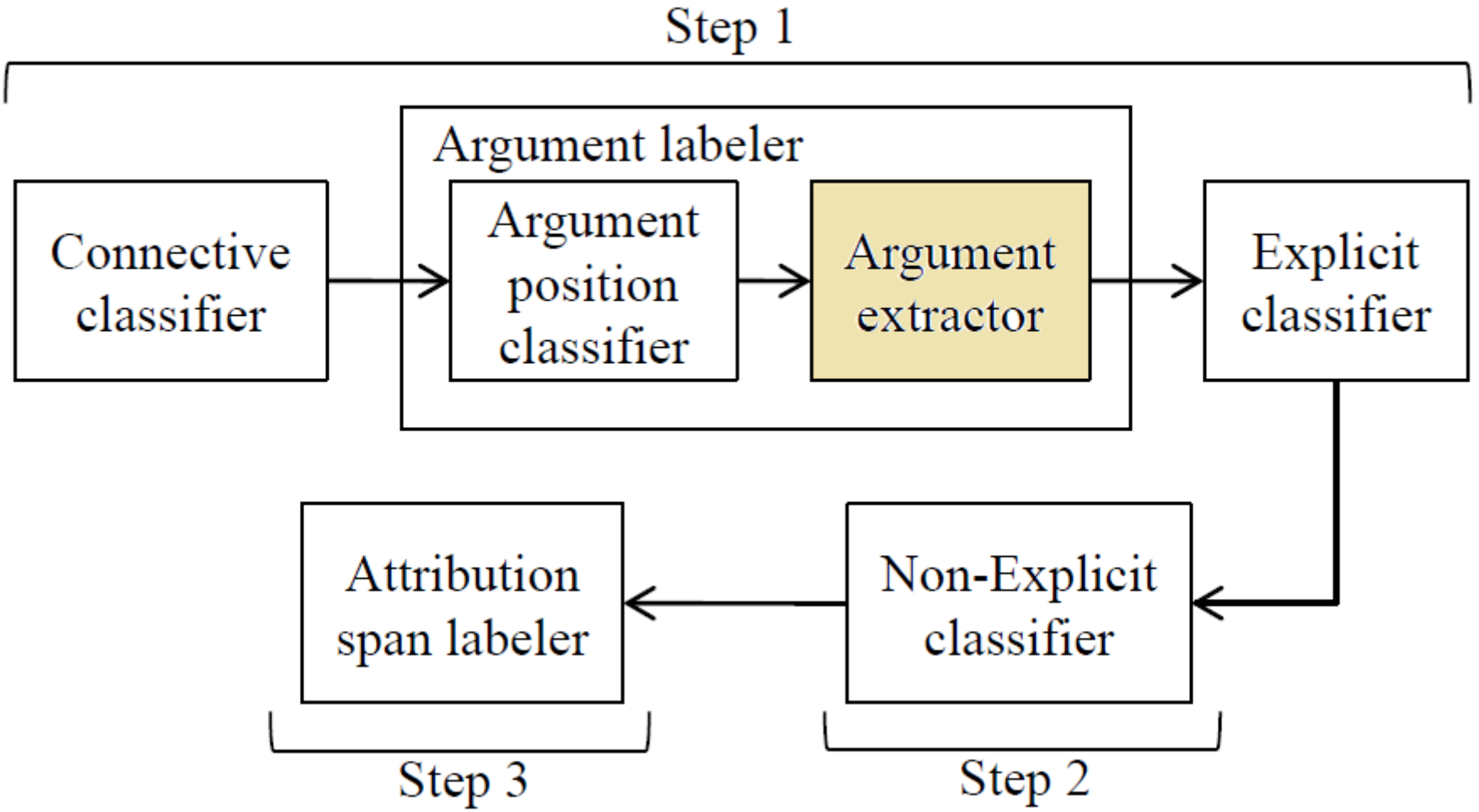
- For discourse connectives, Arg2 and relative position of Arg1 are identified

✓ The classifier (SS or PS) uses:

- ❑ position of connective itself
- ❑ contextual features

➢ Component F_1: 97.94%

System pipeline



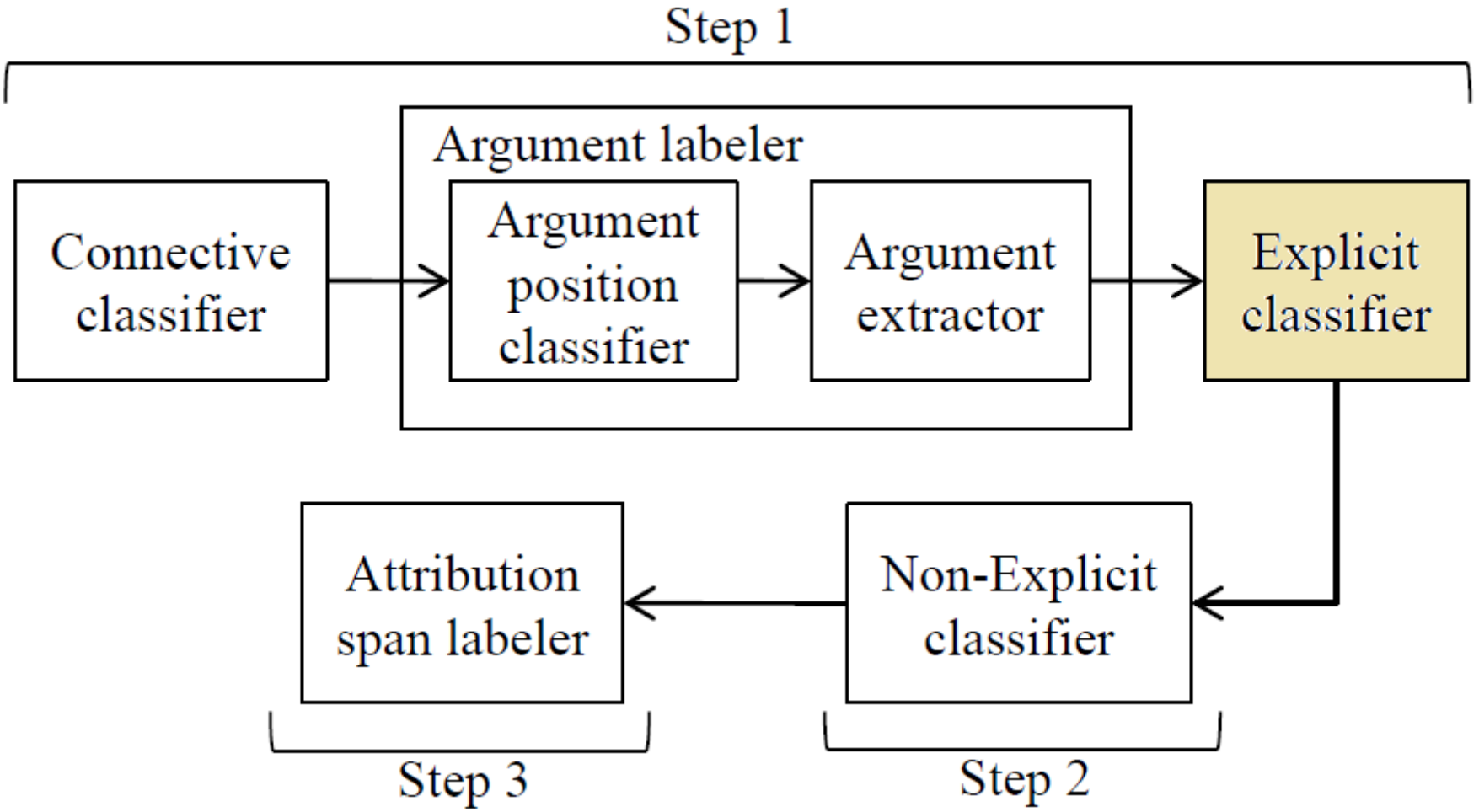
Argument extractor

- The span of the identified arguments is extracted
- When Arg1 and Arg2 are in the same sentence, extraction is not trivial
  - ❑Sentence is splitted into clauses
  - ❑Probabilities are assigned to each node

➢ Component F_1:

- 86.24% for partial matches
- 53.85% for exact matches

System pipeline

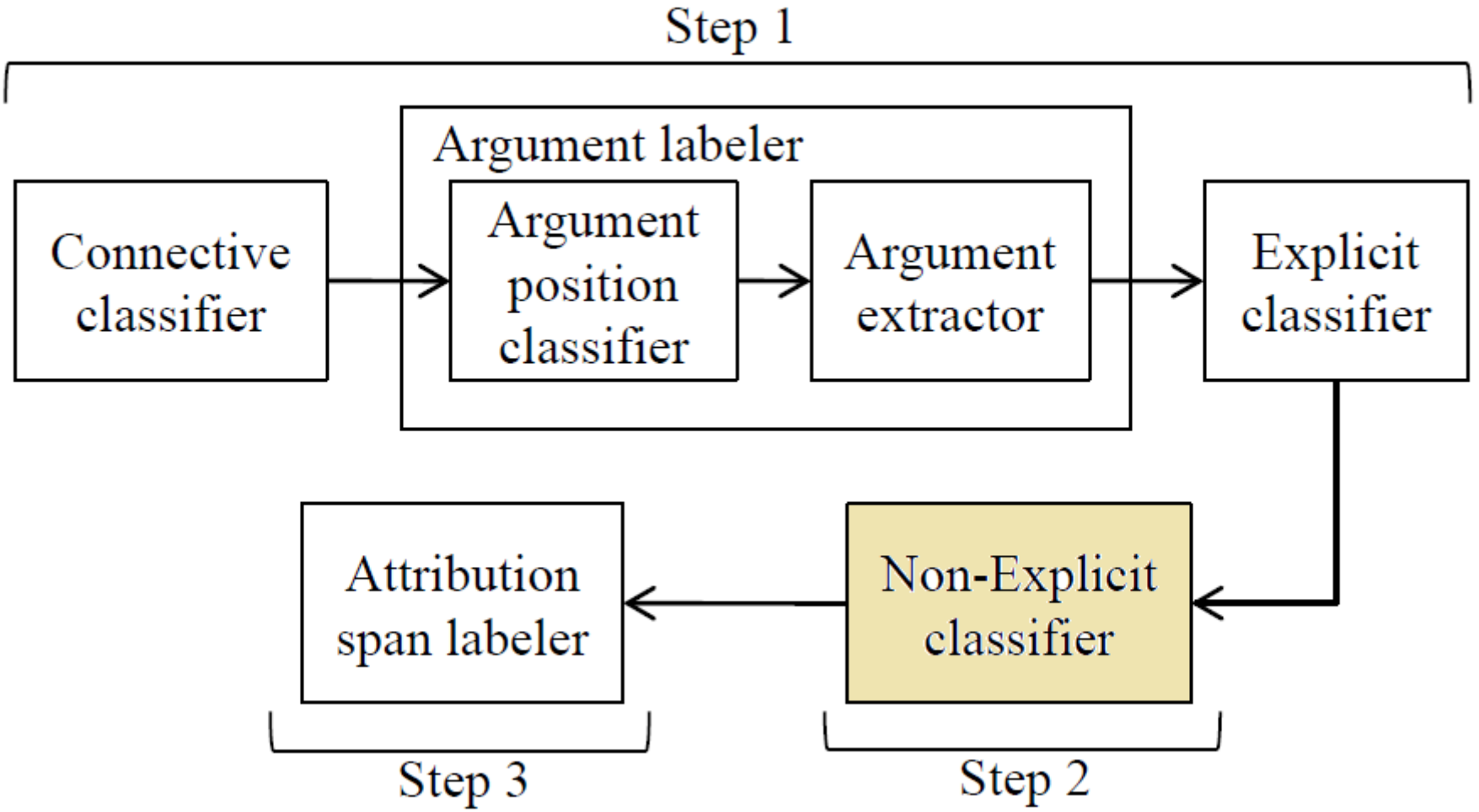


Explicit classifier

- Identifies the semantic type of the connective
- Features used by the classifier:
  - the connective
  - its POS
  - the previous word

➢ Component F_1: 86.77%

System pipeline



Non-Explicit classifier

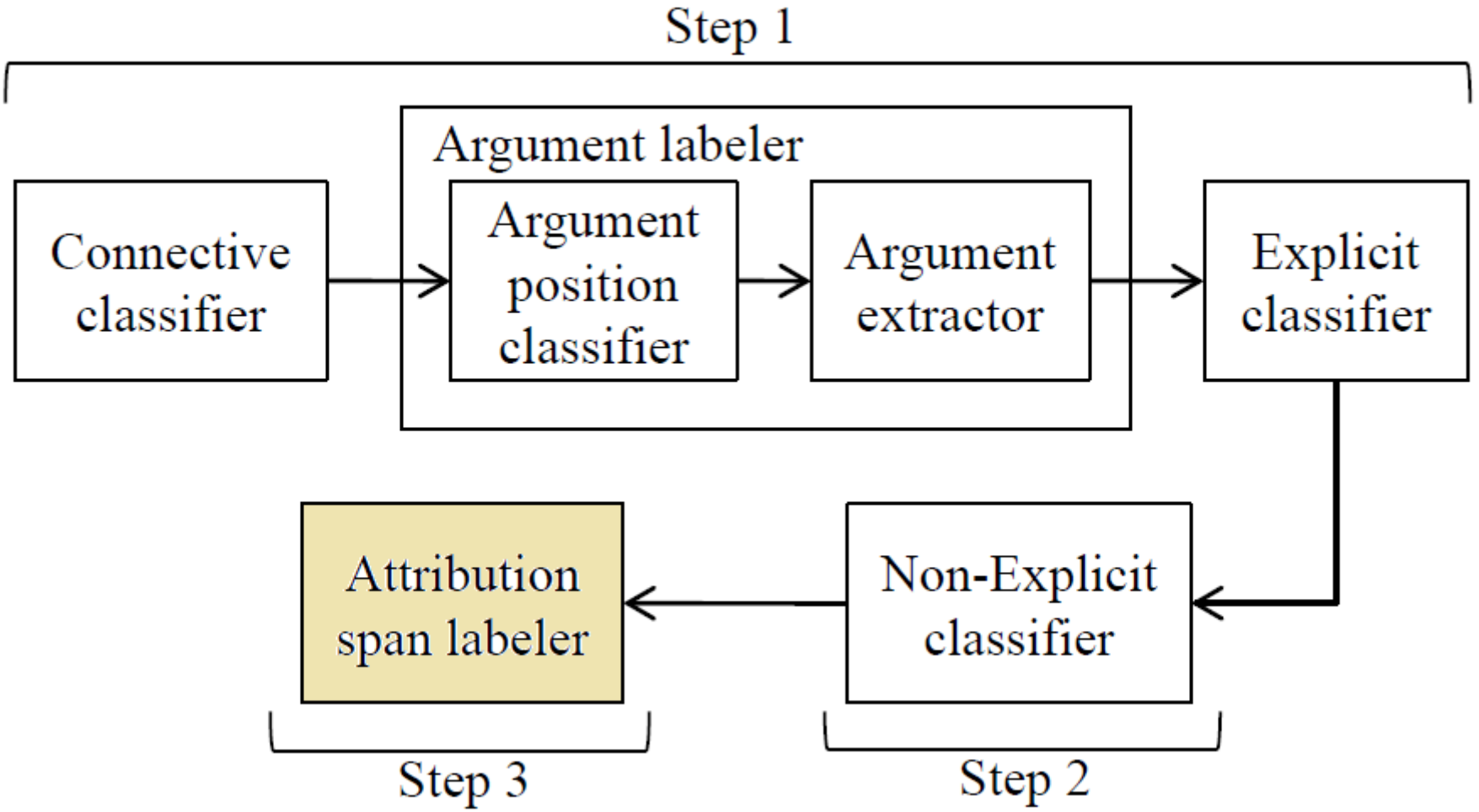
- For all adjacent sentences within a single paragraph (for which no explicit relation was identified), relation is classified as:
  - Implicit
  - AltLex
  - EntRel
  - NoRel
- Implicit and AltLex are also classified for sense type

Non-Explicit classifier

- Used for the classifier:
  - ❑Contextual features
  - ❑Constituent parse features
  - ❑Dependency parse features
  - ❑Word-pair features
    - The first three words of Arg2: used for indicating AltLex relations

➢ Component F_1: 39.63%

System pipeline



Attribution span labeler

- Breaks sentences into clauses
- For each clause, checks if it constitutes an attribution span
- The classifier uses features extracted from the current, the previous and the next clauses

➢ Component F_1:

  - 79.68% for partial matches
  - 65.95% for exact matches

So, how well does the system do?

- Considering the fully automated pipeline performance, the F_1 results are not that good:

	Partial match F_1	**Exact match F_1**
GS + EP	46.80%	33.00%
Auto + EP	38.18%	20.64%

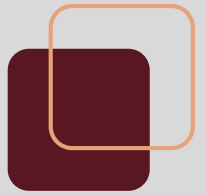
- Great part of these low figures is due to the low performance of the Non-explicit classifier

But still…

- Most of the components have a relatively good performance if fed with correct data
- It can provide useful aid for many LT tasks e.g. identifying redundancy in summarization tasks or answering why-questions in QA tasks
- The authors already suggest amendments
  - Notably feeding the final results to the start in a joint learning model

References

- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering 1 (2012):* 1-35.
- PDTB-Group. *The Penn Discourse Treebank 2.0 Annotation Manual.* The PDTB Research Group, 2007.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and BonnieWebber. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008),* 2008.



Extra slides

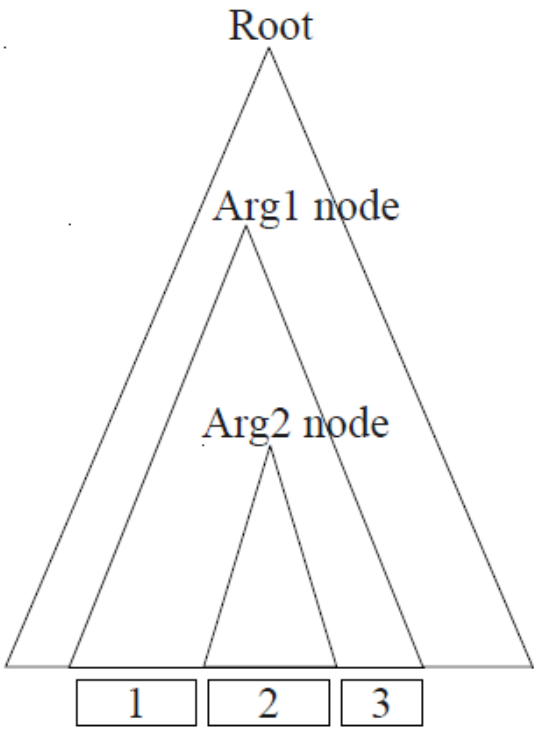
Some details on
the Argument Extractor
component

The SS case

- When Arg1 and Arg2 are in the same sentence, extraction is not trivial
  - Sentence is splitted into clauses
- Can be connected in three ways:
  - Subordination
  - Coordination
  - Adverbials

Subordination

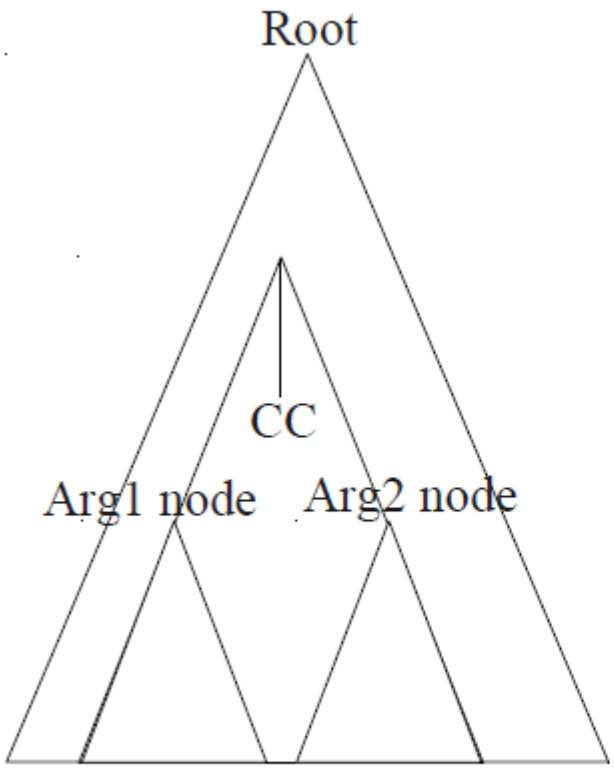
- This scheme is always the case (Dinesh et al., 2005):

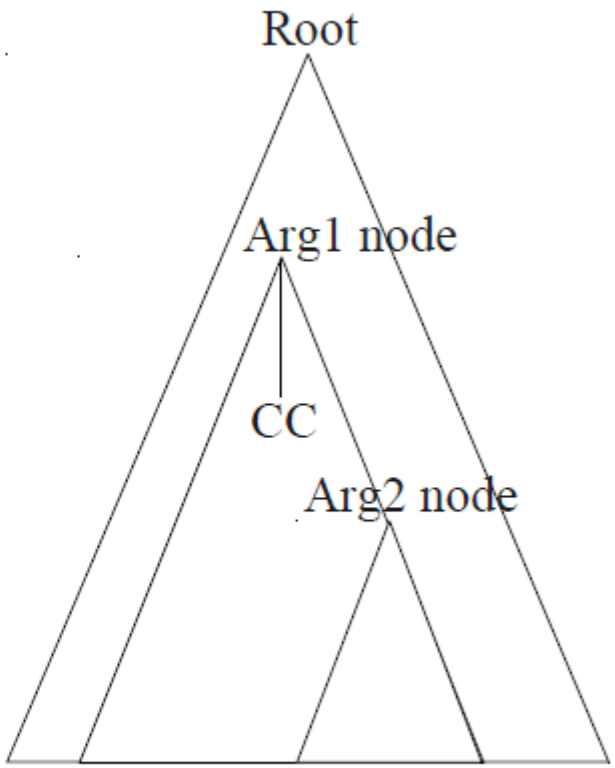


A rule-based algorithm is sufficient for identifying the respective spans

Coordination

- Arg1 and Arg2 mainly related in two ways:





Adverbials

- Adverbials do not demonstrate so strong syntactic constraints
- Still syntactically bound to some extent

The classifier

- Each internal node of the tree is labelled with three probablilities:
  - Arg1 node
  - Arg2 node
  - None
- Tree subtraction from Arg2 node is applied to get Arg1
- The connective is subtracted from the Arg2 node to get Arg2

The PS case

- When Arg1 is located in a previous sentence, the one preceding Arg2 is automatically labelled as Arg1
- This already has a decent performance
  - Anyway sentences further than the previous one would not be considered