Mining Discourse Treebanks with XQuery

Xuchen Yao[1] Gosse Bouma[2]

Johns Hopkins University[1]

University of Groningen[2]

TLT dec 2010

Penn Discourse Treebank (Pdtb)

Large scale discourse annotation

- Discourse annotation for part of the Penn Treebank
- Discourse segments are linked to (sequence of) corresponding syntactic constituents in the Penn Treebank

Prasad et al., *The Penn Discourse TreeBank 2.0*, LREC 2008

Discourse Annotation

Discourse

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 .

Syntax

```
(S
    (SBAR-ADV (IN Although)
      (S
        (NP-SBJ-2 (JJ preliminary)
                          (NNS findings) )
        (VP ....
              (IN ago) )))))
    (, ,)
    (NP-SBJ (DT the) (JJS latest)
                              (NNS results) )
    (VP (VBP appear)
      (PP-LOC (IN in)
        ......
         (NP (DT the) (NN problem) ))))))))))
    (. .) )
```

Pdtb

Statistics

	× 1000
Words	1.000
Sentences	47
Relations	33
– Explicit connective	17.7
– Implicit connective	15.6

Corpus Format

- Proprietary, text based, format
- up to 48 fields per discourse relation

```
Explicit|00|03|534..542|3,0,0|
Although|||although|||Comparison.Contrast||||Wr|Comm|Null|Null||||
600..722|3,1;3,2;3,3;3,4|the latest results appear in today’s New
England Journal of Medicine, a forum likely to bring new attention
to the problem|Inh|Null|Null|Null||||543..598|3,0,1|preliminary
findings were reported more than a year ago|Inh|Null|Null|Null|
||||||||
```

rijksuniversiteit
groningen

Pdtb distribution

Penn Treebank

- Text based labelled bracketing

Pdtb/Penn Treebank Integration

- Discourse segments linked to sequence of tokens in treebank
- Discourse segments linked to syntactic nodes using Gorn addresses (numbered tree nodes)

Drawbacks

- Token ids and node ids are absent in Penn Treebank
- Limited support for queries addressing discourse and syntax at the same time
- Corpus format not easily extendable/modifiable

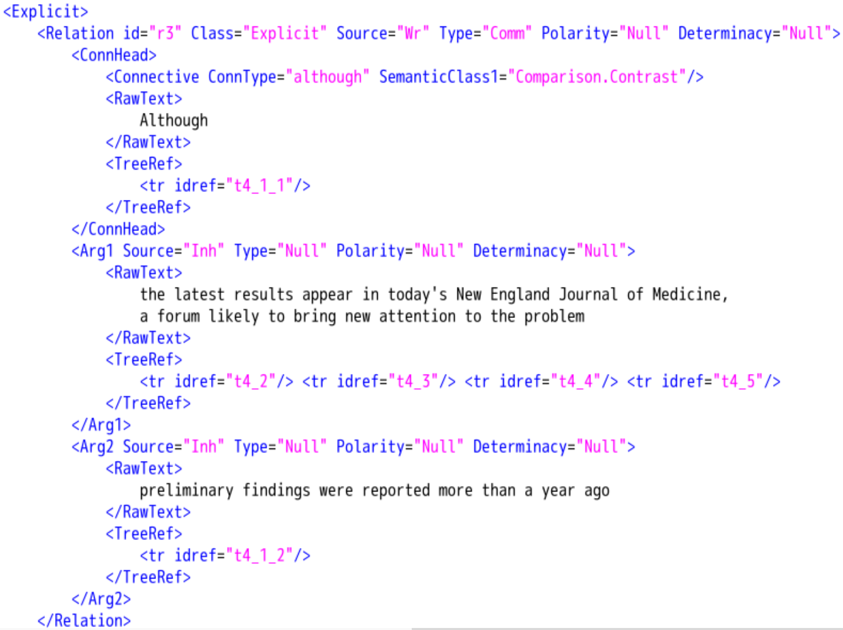
PDTB-XML

Unified corpus format

- Discourse annotation converted to XML
- Penn Treebank converted to XML tree format and Tiger XML
- Syntax and Discourse annotation present in a single document
- Gorn address added to all nodes in syntax as id (index) attributes
- idref attributes of discourse segments point to id of nodes in syntax

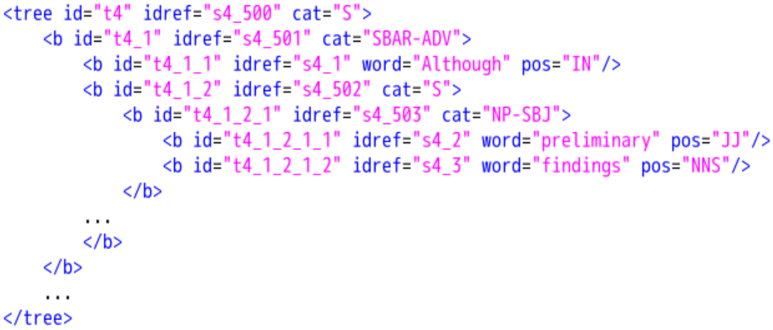
Yao et al, *PDTB XML: The XMLization of the Penn Discourse TreeBank 2.0,* LREC 2010

PDTB XML



```
<Explicit>
    <Relation id="r3" Class="Explicit" Source="Wr" Type="Comm" Polarity="Null" Determinacy="Null">
        <ConnHead>
            <Connective ConnType="although" SemanticClass1="Comparison.Contrast"/>
            <RawText>
                Although
            </RawText>
            <TreeRef>
                <tr idref="t4_1_1"/>
            </TreeRef>
        </ConnHead>
        <Arg1 Source="Inh" Type="Null" Polarity="Null" Determinacy="Null">
            <RawText>
                the latest results appear in today's New England Journal of Medicine,
                a forum likely to bring new attention to the problem
            </RawText>
            <TreeRef>
                <tr idref="t4_2"/> <tr idref="t4_3"/> <tr idref="t4_4"/> <tr idref="t4_5"/>
            </TreeRef>
        </Arg1>
        <Arg2 Source="Inh" Type="Null" Polarity="Null" Determinacy="Null">
            <RawText>
                preliminary findings were reported more than a year ago
            </RawText>
            <TreeRef>
                <tr idref="t4_1_2"/>
            </TreeRef>
        </Arg2>
    </Relation>
```

PDTB XML



```
<tree id="t4" idref="s4_500" cat="S">
    <b id="t4_1" idref="s4_501" cat="SBAR-ADV">
        <b id="t4_1_1" idref="s4_1" word="Although" pos="IN"/>
        <b id="t4_1_2" idref="s4_502" cat="S">
            <b id="t4_1_2_1" idref="s4_503" cat="NP-SBJ">
                <b id="t4_1_2_1_1" idref="s4_2" word="preliminary" pos="JJ"/>
                <b id="t4_1_2_1_2" idref="s4_3" word="findings" pos="NNS"/>
            </b>
        ...
        </b>
    </b>
    ...
</tree>
```

XQuery and XPath

XQuery

- Official and de facto standard for querying XML databases
- Functional (Declarative)
- Uses XPath for navigating in XML documents (tree structures)
- RegEx support, functions, modules, ...

FLWOR Expressions

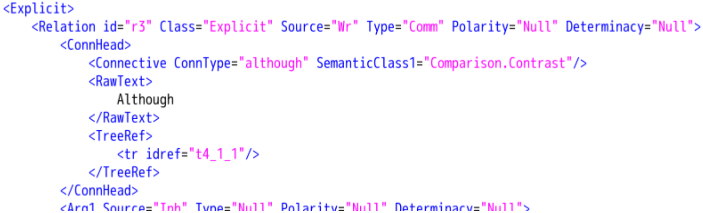
For	Identify elements to be searched
Let	Assign value to variables
Where	Constraints on results
Order	Order results
Return	Results (as XML or text)

XQuery and XPath

Finding all Relations with connective Although

```
for $rel in
  //Relation[@Class="Explicit" and
    ConnHead/Connective[@ConnType="although"] ]

return $rel
```



```
<Explicit>
    <Relation id="r3" Class="Explicit" Source="Wr" Type="Comm" Polarity="Null" Determinacy="Null">
        <ConnHead>
            <Connective ConnType="although" SemanticClass1="Comparison.Contrast"/>
            <RawText>
                Although
            </RawText>
            <TreeRef>
                <tr idref="t4_1_1"/>
            </TreeRef>
        </ConnHead>
        <Arg1 Source="Inh" Type="Null" Polarity="Null" Determinacy="Null">
```

Treebank Query Languages

Dedicated treebank query languages

- Tgrep2, TIGERsearch, Emu, CorpusSearch, NiteQL, LPath
- dedicated treebank query languages
- Syntax of various languages varies considerably
- Expressive power of languages varies considerably

Lai and Bird, *Querying Linguistic Trees*, J Log Lang Inf, 2010

Some more drawbacks

- Corpora tend to support only a single query language: need to learn multiple languages
- Query languages do not support complicated extraction tasks ('list verb-object pairs')

Navigation in XML Trees

XPath Functionality

- Child, Parent, (Last, First, Nth) Child
- Descendant, Ancestor,
- (Preceding, Following) Sibling

Q2: Find noun phrases whose rightmost child is a noun

```
for $np in collection("pdtb")//tree//
          b[ @cat="NP"    and
             b[last()][matches(@pos,"NN")]
           ]

return
$np
```

Navigation in XML Trees

XQuery: Write your own Functions

- Leftmost-descendant,
- Immediately Follows,
- Shortest-path between two nodes in a graph (*Dijkstra's Algorithm*) (Strömback & Schmidt, 2009)

Q3: VP containing V immediately followed by NP immediately followed by PP

```
for $v in collection("pdtb")//tree//b[@cat ="VP"]/
       b[matches(@pos, "VB")]
for $np in pdtb:imm-follow($v)[matches(@cat, "NP")]
for $pp in pdtb:imm-follow($np)[matches(@cat, "PP")]
where $pp

return
$v/..
```

Pdtb Xquery Module

Immediately Follows

```
declare function
    pdtb:imm-follow($node as el(b)) as el(b)*
{ let $followers :=
     if  ( $node/following-sibling::b )
     then pdtb:leftmost-desc(
              $node/following-sibling::b[1])
     else ()
  return $followers
};

declare function
   pdtb:leftmost-desc($node as el(b)) as el(b)*
{ let $descendants :=
     if   ($node/b)
     then local:leftmost-desc($node/b[1])
     else ()
  return ($node, $descendants)
} ;
```

Querying discourse and syntax

Case Study: Range Relations

To what extent can discourse segments introduced by a subordinating conjunction be arguments of a following discourse relation?

GM also had dismal results in the first 10 days of the month, while other auto makers reported mixed results. All of the Big Three suffered in the just-ended period, however. (wsj_1139)

Lee et al., *Departures from Tree Structures in Discourse*, Constraints in Discourse workshop, 2008

Querying discourse and syntax

```
for $c in collection($dir)/corpus

for $rel in $c/Relations/*/Relation[ConnHead/RawText[
     matches(.,"(although|however|after|as|....)","i")]]

let $shared := $c/Relations/*/Relation[
               pdtb:gorn2tree(Arg1/TreeRef) =
                  pdtb:gorn2tree($rel/Arg2/TreeRef)/.. ]

where $shared
return
  <shared>
      <first>$rel</first>
      <second>$shared</second>
  </shared>
```

Improved Query: no lexical selection

```
for $r in collection("pdtb")/corpus/Relations/*/Relation

let $tree := pdtb:gorn2tree($r/Arg2/TreeRef/tr[1])[
             ( @cat = "S" and starts-with(../@cat, "SBAR") ) or
             ( @CAT = "S-NOM" and ../@cat="PP-TMP") ]

 let $shared :=

$r/../../*/Relation[Arg1/TreeRef/tr[1]/@idref =
                       $tree/../@id ]

where $shared
return
  <shared>
      <first>$rel</first>
      <second>$shared</second>
  </shared>
```

Performance

Pdtb-XML

files	2159
size	376MB

Saxon vs XML Databases

- Saxon processes all files on the fly
  - Reading in data
  - Limited optimizations
  - Memory requirements: approx 5Gb for Pdtb-XML
- XML Databases
  - eXist, Berkeley Db, Sedna, ...
  - Corpus processed and indexed off-line
  - Various optimizations possible
  - Small memory requirements

rijksuniversiteit
groningen

Performance

Q1 sentences that include the word *saw*

Q2 NPs whose rightmost child is a noun

Q3 VPs that contain a verb immediately followed by an NP immediately followed by a PP

Q4 all *Explicit* relations whose connective type is *because*

Q5 connectives and corresponding POS tags of all *Explicit* relations

Q6 all words with POS='CC' that function as connective

Q7 *shared arguments* case study (cf. Lee et al 2008)

Performance

Experiments

CPU time in Minutes:Seconds

Paper Intel Xeon X5355, 2.66 Ghz, 16GB

Groningen Intel Xeon E5410, 2.33GHz, 64GB

	Paper			**Groningen**	
Q	saxon	exist	bdb	saxon	sedna
1	5:51	0:19	0:15	1:36	0:02
2	6:23	0:55	1:20	1:33	0:27
3	6:43	1:18	1:20	1:45	0:23
4	2:09	0:01	0:01	1:33	0:01
5	7:17	2:27	30:30	3:05	**0:15**
6	7:03	15:21	21:33	2:57	**0:08**
7	32:26	dnf	7:13	1:57	**0:21**

Conclusions

PDTB XML

- XML supports structuring and querying Discourse Annotation
- Merging Syntax and Discourse in single XML document supports tight integration

XQuery and XPath

- Widely supported standards
- XPath allows (XML) tree navigation
- XQuery modules can support corpus specific functionality
- XML Databases enable efficient querying

rijksuniversiteit groningen