Automatically Evaluating Text Coherence Using Discourse Relations

Ziheng Lin, Hwee Tou Ng and Min-Yen Kan

Department of Computer Science
National University of Singapore
Introduction

• Textual coherence ↔ discourse structure
• Canonical orderings of relations:
  – Satellite before nucleus
  – Nucleus before satellite

• Preferential ordering generalizes to other discourse frameworks
Two examples

1. [ Everyone agrees that most of the nation’s old bridges need to be repaired or replaced. ]\textsubscript{S1} [ But there’s disagreement over how to do it. ]\textsubscript{S2}
   - **Swapping S1 and S2 without rewording**
   - **Disturbs intra-relation ordering**

2. [ The Constitution does not expressly give the president such power. ]\textsubscript{S1} [ However, the president does have a duty not to violate the Constitution. ]\textsubscript{S2} [ The question is whether his only means of defense is the veto. ]\textsubscript{S3}
   - **Contrast-followed-by-Cause is common in text**
   - **Shuffling these sentences**
   - **Disturbs inter-relation ordering**

\[ S1 \xrightarrow{\text{Contrast}} S2 \]

\[ \text{Contrast}\rightarrow\text{Cause} \]}
Assess coherence with discourse relations

• Measurable preferences for intra- and inter-relation ordering
• Key idea: use statistical model of this phenomenon to assess text coherence
• Propose a model to capture text coherence
  • Based on statistical distribution of discourse relations
• Focus on relation transitions
Outline

- Introduction
- Related work
  - Using discourse relations
  - A refined approach
  - Experiments
  - Analysis and discussion
  - Conclusion
Coherence models

• **Barzilay & Lee (’04)**
  – Domain-dependent HMM model to capture topic shift
  – Global coherence = overall prob of topic shift across text

• **Barzilay & Lapata (’05, ’08)**
  – Entity-based model to assess local text coherence
  – Motivated by Centering Theory
  – Assumption: coherence = sentence-level local entity transitions
    • Captured by an entity grid model

• **Soricut & Marcu (’06), Elsner et al. (’07)**
  – Combined entity-based and HMM-based models: complementary

• **Karamanis (’07)**
  – Tried to integrate discourse relations into Centering-based metric
  – Not able to obtain improvement
Discourse parsing

- **Penn Discourse Treebank (PDTB)** (Prasad et al. ’08)
  - Provides discourse level annotation on top of PTB
  - Annotates arguments, relation types, connectives, attributions
- **Recent work in PDTB**
  - Focused on explicit/implicit relation identification
  - Wellner & Pustejovsky (’07)
  - Elwell & Baldridge (’08)
  - Lin et al. (’09)
  - Pitler et al. (’09)
  - Pitler & Nenkova (’09)
  - Lin et al. (’10)
  - Wang et al. (’10)
  - ...
Outline

• Introduction
• Related work
• Using discourse relations
  • A refined approach
  • Experiments
  • Analysis and discussion
• Conclusion
Parsing text

• First apply discourse parsing on the input text
  – Use our automatic PDTB parser (Lin et al., ’10)
    http://www.comp.nus.edu.sg/~linzihen
  – Identifies the relation types and arguments (Arg1 and Arg2)

• Utilize 4 PDTB level-1 types: Temporal, Contingency, Comparison, Expansion; as well as EntRel and NoRel
First attempt

• A simple approach: sequence of relation transitions
• Text (2) can be represented by:

\[ \text{S1} \xrightarrow{\text{Comp}} \text{S2} \xrightarrow{\text{Cont}} \text{S3} \]

• Compile a distribution of the n-gram sub-sequences
• E.g., a bigram for Text (2): Comp\(\rightarrow\)Cont
• A longer transition: Comp\(\rightarrow\)Exp\(\rightarrow\)Cont\(\rightarrow\)nil\(\rightarrow\)Temp
  • N-grams: Comp\(\rightarrow\)Exp, Exp\(\rightarrow\)Cont\(\rightarrow\)nil, ...
• Build a classifier to distinguish coherent text from incoherent one, based on transition n-grams
Shortcomings

• Results of our pilot work was poor
  – < 70% on text ordering ranking

• Shortcomings of this model:
  – Short text has short transition sequence
    • Text (1): Comp → Cont
      • Sparse features
  – Models inter-relation preference, but not intra-relation preference
    • Text (1): S1 < S2 vs. S2 < S1
Outline

• Introduction
• Related work
• Using discourse relations

→ • A refined approach
  • Experiments
  • Analysis and discussion
  • Conclusion
An example: an excerpt from wsj_0437

3. [ Japan normally depends heavily on the Highland Valley and Cananea mines as well as the Bougainville mine in Papua New Guinea. ]_{S1}
   [ Recently, Japan has been buying copper elsewhere. ]_{S2}
   [ But as Highland Valley and Cananea begin operating, ]_{C3.1}
   [ they are expected to resume their roles as Japan’s suppliers. ]_{C3.2}_{S3}
   [ According to Fred Demler, metals economist for Drexel Burnham Lambert, New York, ]_{C4.1}
   [ “Highland Valley has already started operating ]_{C4.2}
   [ and Cananea is expected to do so soon.” ]_{C4.3}_{S4}

- Definition: a term's discourse role is a 2-tuple of <relation type, argument tag> when it appears in a discourse relation.
  - Represent it as RelType.ArgTag
- E.g., discourse role of 'cananea' in the first relation:
  - Comp.Arg1
Discourse role matrix

• Discourse role matrix: represents different discourse roles of the terms across continuous text units
  – Text units: sentences
  – Terms: stemmed forms of open class words

• Expanded set of relation transition patterns

• Hypothesis: the sequence of discourse role transitions → clues for coherence

• Discourse role matrix: foundation for computing such role transitions
**Discourse role matrix**

- A fragment of the matrix representation of Text (3)
- A cell $C_{T_i,S_j}$: discourse roles of term $T_i$ in sentence $S_j$

<table>
<thead>
<tr>
<th>$S#$</th>
<th>Terms</th>
<th>Terms</th>
<th>Terms</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>copper</td>
<td>cananea</td>
<td>operat</td>
<td>depend</td>
</tr>
<tr>
<td>$S_1$</td>
<td>nil</td>
<td>Comp.Arg1</td>
<td>nil</td>
<td>Comp.Arg1</td>
</tr>
<tr>
<td>$S_2$</td>
<td>Comp.Arg2 Comp.Arg1</td>
<td>nil</td>
<td>nil</td>
<td>nil</td>
</tr>
<tr>
<td>$S_3$</td>
<td>nil</td>
<td>Comp.Arg2 Temp.Arg1 Exp.Arg1</td>
<td>Comp.Arg2 Temp.Arg1 Exp.Arg1</td>
<td>nil</td>
</tr>
<tr>
<td>$S_4$</td>
<td>nil</td>
<td>Exp.Arg2</td>
<td>Exp.Arg1</td>
<td>nil</td>
</tr>
</tbody>
</table>

- $C_{\text{cananea},S_3} = \{\text{Comp.Arg2, Temp.Arg1, Exp.Arg1}\}$
Sub-sequences as features

• Compile sub-sequences of discourse role transitions for every term
  – How the discourse role of a term varies through the text

• 6 relation types (Temp, Cont, Comp, Exp, EntRel, NoRel) and 2 argument tags (Arg1 and Arg2)
  – 6 x 2 = 12 discourse roles, plus a nil value
Sub-sequence probabilities

- Compute the probabilities for all sub-sequences
- E.g., \( P(\text{Comp.Arg2} \rightarrow \text{Exp.Arg2}) = \frac{2}{25} = 0.08 \)
- Transitions are captured locally per term, probabilities are aggregated globally
  - Capture distributional differences of sub-sequences in coherent and incoherent texts
- Barzilay & Lapata ('05): salient and non-salient matrices
  - Salience based on term frequency
Preference ranking

• The notion of coherence is relative
  – Better represented as a ranking problem rather than a classification problem
• Pairwise ranking: rank a pair of texts, e.g.,
  – Differentiating a text from its permutation
  – Identifying a more well-written essay from a pair
• Can be easily generalized to listwise
• Tool: SVM\textsuperscript{light}
  – Features: all sub-sequences with length $\leq n$
  – Values: sub-sequence prob
Outline

• Introduction
• Related work
• Using discourse relations
• A refined approach

• Experiments
• Analysis and discussion
• Conclusion
Task and data

- **Text ordering ranking (Barzilay & Lapata ’05, Elsner et al. ’07)**
  - Input: a pair of text and its permutation
  - Output: a decision on which one is more coherent
- **Assumption: the source text is always more coherent than its permutation**

\[
\text{Accuracy} = \frac{\text{# times the system correctly chooses the source text}}{\text{total # of test pairs}}
\]

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>Earthquakes</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Articles</td>
<td>1040</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td># Pairs</td>
<td>19120</td>
<td>1862</td>
<td>1996</td>
</tr>
<tr>
<td>Avg. # Sents</td>
<td>22.0</td>
<td>10.4</td>
<td>11.5</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Articles</td>
<td>1079</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td># Pairs</td>
<td>19896</td>
<td>1956</td>
<td>1986</td>
</tr>
</tbody>
</table>
Human evaluation

• 2 key questions about text ordering ranking:
  1. To what extent is the assumption that the source text is more coherent than its permutation correct?
     → Validate the correctness of this synthetic task
  2. How well do human perform on this task?
     → Obtain upper bound for evaluation

• Randomly select 50 pairs from each of the 3 data sets
• For each set, assign 2 human subjects to perform the ranking
  – The subjects are told to identify the source text
Results for human evaluation

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>Earthquakes</th>
<th>Accidents</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>90.0</td>
<td>90.0</td>
<td>94.0</td>
<td>91.3</td>
</tr>
</tbody>
</table>

1. Subjects’ annotation highly correlates with the gold standard
   → The assumption is supported

2. Human performance is not perfect
   → Fair upper bound limits
Evaluation and results

• Baseline: entity-based model (Barzilay & Lapata ’05)

• 4 questions to answer:

  Q1: Does our model outperform the baseline?
  Q2: How do the different features derived from using relation types, argument tags and salience information affect performance?
  Q3: Can the combination of the baseline and our model outperform the single models?
  Q4: How does system performance of these models compare with human performance on the task?
Q1: Does our model outperform the baseline?

- Type+Arg+Sal: makes use of relation types, argument tags and salience information
- Significantly outperform baseline on WSJ and Earthquakes (p < 0.01)
- On Accidents, not significantly different

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>Earthquakes</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>85.71</td>
<td>83.59</td>
<td>89.93</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>88.06**</td>
<td>86.50**</td>
<td>89.38</td>
</tr>
</tbody>
</table>
Q2: How do the different features derived from using relation types, argument tags and salience information affect performance?

Delete Type info, e.g., Comp.Arg2 becomes Arg2
- Performance drops on Earthquakes and Accidents

Delete Arg info, e.g., Comp.Arg2 becomes Comp
- A large performance drop across all 3 data sets

Remove Salience info
- Also markedly reduces performance

<table>
<thead>
<tr>
<th>Full model</th>
<th>WSJ</th>
<th>Earthquakes</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>85.71</td>
<td>83.59</td>
<td>89.93</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>88.06**</td>
<td>86.50**</td>
<td>89.38</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>88.28**</td>
<td>85.89*</td>
<td>87.06</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>87.06**</td>
<td>82.98</td>
<td>86.05</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>85.98</td>
<td>82.67</td>
<td>87.87</td>
</tr>
</tbody>
</table>

Support the use of all 3 feature classes
Q3: Can the combination of the baseline and our model outperform the single models?

- Different aspects: local entity transition vs. discourse relation transition
- Combined model gives highest performance
  - 2 models are synergistic and complementary
  - The combined model is linguistically richer

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>Earthquakes</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>85.71</td>
<td>83.59</td>
<td>89.93</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>88.06**</td>
<td>86.50**</td>
<td>89.38</td>
</tr>
<tr>
<td>Baseline &amp; Type+Arg+Sal</td>
<td>89.25**</td>
<td>89.72**</td>
<td>91.64**</td>
</tr>
</tbody>
</table>
Q4: How does system performance of these models compare with human performance on the task?

- Gap between baseline & human: relatively large
- Gap between full model & human: more acceptable on WSJ and Earthquakes
- Combined model: error rate significantly reduced

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>Earthquakes</th>
<th>Accidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>85.71 (-4.29)</td>
<td>83.59 (-6.41)</td>
<td>89.93 (-4.07)</td>
</tr>
<tr>
<td>Type+Arg+Sal</td>
<td>88.06 (-1.94)</td>
<td>86.50 (-3.50)</td>
<td>89.38 (-4.62)</td>
</tr>
<tr>
<td>Baseline &amp; Type+Arg+Sal</td>
<td>89.25 (-0.75)</td>
<td>89.72 (-0.28)</td>
<td>91.64 (-2.36)</td>
</tr>
<tr>
<td>Human</td>
<td>90.00</td>
<td>90.00</td>
<td>94.00</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• Related work
• Using discourse relations
• A refined approach
• Experiments

⇒ • Analysis and discussion
  • Conclusion
Performance on data sets

<table>
<thead>
<tr>
<th></th>
<th>Accidents</th>
<th>WSJ</th>
<th>Earthquakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type+Arg+Sal Acc.</td>
<td>89.38</td>
<td>&gt;</td>
<td>88.06</td>
</tr>
<tr>
<td>Ratio</td>
<td></td>
<td>&gt;</td>
<td>86.50</td>
</tr>
</tbody>
</table>

- Performance gaps between data sets
- Examine the relation/length ratio for source articles
  \[
  \text{Ratio} = \frac{\# \text{ relations in the article}}{\# \text{ sentences in the article}}
  \]
- The ratio gives an idea how often a sentence participates in discourse relations
- Ratios correlate with accuracies
Correctly vs. incorrectly ranked permutations

- Expect that: when a text contains more level-1 discourse types (Temp, Cont, Comp, Exp), less EntRel and NoRel
  - Easier to compute how coherent this text is
- These 4 relations can combine to produce meaningful transitions, e.g., Comp→Cont in Text (2)
- Compute the relation/length ratio for the 4 level-1 types for permuted texts

\[
\text{Ratio} = \frac{\# \text{ 4 discourse relations in the article}}{\# \text{ sentences in the article}}
\]

- Ratio: 0.58 for those that are correctly ranked, 0.48 for those that are incorrectly ranked
  - Hypothesis supported
Revisit Text (2)

2  [ The Constitution does not expressly give the president such power. ]_S1
   [ However, the president does have a duty not to violate the Constitution. ]_S2
   [ The question is whether his only means of defense is the veto. ]_S3

• 3 sentences $\rightarrow$ 5 (source, permutation) pairs
• Apply the full model on these 5 pairs
  – Correctly ranks 4
  – The failed permutation is $S3 < S1 < S2$
• A very good clue of coherence: explicit Comp relation between S1 and S2 (signaled by however)
  – Not retained in the other 4 permutations
  – Retained in $S3 < S1 < S2$ $\rightarrow$ hard to distinguish
Conclusion

• Coherent texts preferentially follow certain discourse structures
  – Captured in patterns of relation transitions
• First demonstrated that simply using the transition sequence does not work well
• Transition sequence $\rightarrow$ discourse role matrix
• Outperforms the entity-based model on the task of text ordering ranking
• The combined model outperforms single models
  – Complementary to each other
Backup
Discourse role matrix

• In fact, each column corresponds to a lexical chain
• Difference:
  – Lexical chain: nodes connected by WordNet rel
  – Matrix: nodes connected by same stemmed form
    • Further typed with discourse relations

<table>
<thead>
<tr>
<th>cananea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.Arg1</td>
</tr>
<tr>
<td>nil</td>
</tr>
<tr>
<td>Comp.Arg2</td>
</tr>
<tr>
<td>Temp.Arg1</td>
</tr>
<tr>
<td>Exp.Arg1</td>
</tr>
<tr>
<td>Exp.Arg2</td>
</tr>
</tbody>
</table>
Learning curves

• On WSJ:
  – Acc. Increases rapidly from 0—2000
  – Slowly increases from 2000—8000
  – Full model consistently outperforms baseline with a significant gap
  – Combined model consistently and significantly outperformance the other two

• On Earthquakes:
  – Always increase as more data are utilized
  – Baseline better at the start
  – Full & combined models catch up at 1000 and 400, and remain consistently better

• On Accidents:
  – Full model and baseline do not show difference
  – Combined model shows significant gap after 400
• **Combined model vs human:**
  – Avg error rate reduction against 100%:
    • 9.57% for full model and 26.37% for combined model
  – Avg error rate reduction against human upper bound:
    • 29% for full model and 73% for combined model