Automatically Evaluating Text Coherence Using Discourse Relations

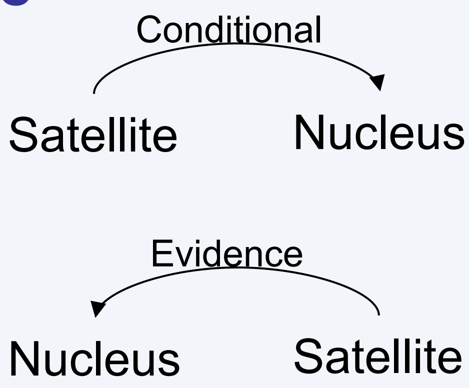
***Ziheng Lin*, Hwee Tou Ng and Min-Yen Kan**

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.]$_{S1}$ [*But* there's disagreement over how to do it.]$_{S2}$

- **Swapping S1 and S2 without rewording**
- **Disturbs intra-relation ordering**

$S1 \xrightarrow{\text{Contrast}} S2$

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.]$_{S3}$

- **Contrast-followed-by-Cause is common in text**
- **Shuffling these sentences**
- **Disturbs inter-relation ordering**

Contrast→Cause

Incoherent text

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

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.]$_{S3}$

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

$$S1 \xrightarrow{\text{Comp}} S2 \xrightarrow{\text{Cont}} S3$$

- **Compile a distribution of the n-gram sub-sequences**
- **E.g., a bigram for Text (2): Comp→Cont**
- **A longer transition: Comp→Exp→Cont→nil→Temp**
  - N-grams: Comp→Exp, Exp→Cont→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   Text (2): 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." $]_{C4.3}$ $]_{S4}$

Implicit Comp
Explicit Comp
Explicit Temp
Implicit Exp
Explicit Exp

- **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

S#	Terms				
	copper	cananea	operat	depend	...
S_1	nil	Comp.Arg1	nil	Comp.Arg1	
S_2	Comp.Arg2 Comp.Arg1	nil	nil	nil	
S_3	nil	Comp.Arg2 Temp.Arg1 Exp.Arg1	Comp.Arg2 Temp.Arg1 Exp.Arg1	nil	
S_4	nil	Exp.Arg2	Exp.Arg1 Exp.Arg2	nil	

  - $C_{cananea,S3}$ = {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

cananea
Comp.Arg1
nil
Comp.Arg2 Temp.Arg1 Exp.Arg1
Exp.Arg2

Sub-sequence probabilities

- **Compute the probabilities for all sub-sequences**
- **E.g., P(Comp.Arg2→Exp.Arg2) = 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^{light}**
  - Features: all sub-sequences with length <= 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**

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

new

		WSJ	Earthquakes	Accidents
Train	# Articles	1040	97	100
	# Pairs	19120	1862	1996
	Avg. # Sents	22.0	10.4	11.5
Test	# Articles	1079	99	100
	# Pairs	19896	1956	1986

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

WSJ	Earthquakes	Accidents	Overall
90.0	90.0	94.0	91.3

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?

		WSJ	Earthquakes	Accidents
	Baseline	85.71	83.59	89.93
Full model	Type+Arg+Sal	88.06**	86.50**	89.38

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**

		WSJ	Earthquakes	Accidents
	Baseline	85.71	83.59	89.93
Full model	Type+Arg+Sal	88.06**	86.50**	89.38
	Type+Arg+Sal	88.28**	85.89*	87.06
	Type+Arg+Sal	87.06**	82.98	86.05
	Type+Arg+Sal	85.98	82.67	87.87

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

→ Support the use of all 3 feature classes

		WSJ	Earthquakes	Accidents
	Baseline	85.71	83.59	89.93
Full model	Type+Arg+Sal	88.06**	86.50**	89.38
	Baseline & Type+Arg+Sal	89.25**	89.72**	91.64**

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

		WSJ	Earthquakes	Accidents
	Baseline	85.71 (-4.29)	83.59 (-6.41)	89.93 (-4.07)
Full model	Type+Arg+Sal	88.06 (-1.94)	86.50 (-3.50)	89.38 (-4.62)
	Baseline & Type+Arg+Sal	89.25 (-0.75)	89.72 (-0.28)	91.64 (-2.36)
	Human	**90.00**	**90.00**	**94.00**

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

Outline

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

Performance on data sets

	Accidents		WSJ		Earthquakes
Type+Arg+Sal Acc.	89.38	>	88.06	>	86.50
Ratio					

- **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 → 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 → hard to distinguish

$$S1 \xrightarrow[\textit{however}]{\text{Comp}} S2$$

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 → 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

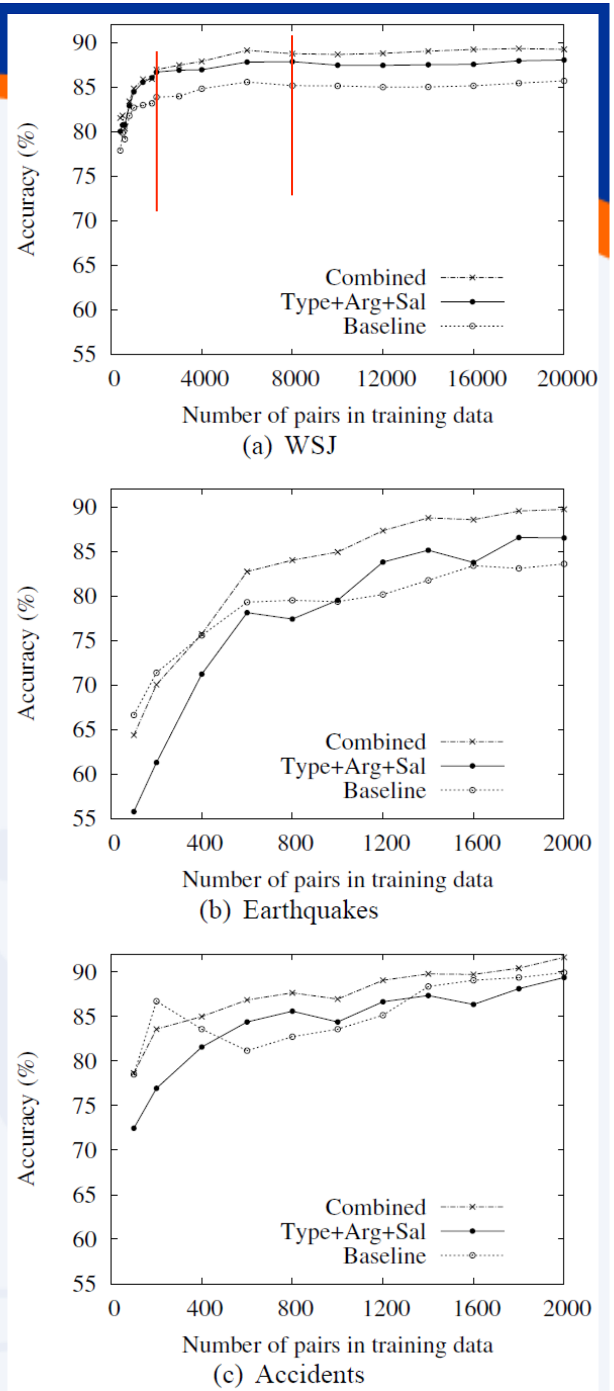
Discourse role matrix

- **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

cananea
Comp.Arg1
nil
Comp.Arg2 Temp.Arg1 Exp.Arg1
Exp.Arg2

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