Coherence in Text

Coherence:

- is a property of well-written texts;
- makes them easier to read and understand;
- ensures that sentences are meaningfully related;
- and that the reader can work out what expressions mean;
- the text is thematically organized;
- temporally organized;
- rather than a random concatenation of sentences.

In this lecture, we will discuss Barzilay and Lapata’s (2008) entity-based model of coherence.

Summary A

Britain said he did not have diplomatic immunity. The Spanish authorities contend that Pinochet may have committed crimes against Spanish citizens in Chile. Baltasar Garzón filed a request on Wednesday. Chile said, President Fidel Castro said Sunday he disagreed with the arrest in London.

Summary B

Former Chilean dictator Augusto Pinochet, was arrested in London on 14 October 1998. Pinochet, 82, was recovering from surgery. The arrest was in response to an extradition warrant served by a Spanish judge. Pinochet was charged with murdering thousands, including many Spaniards. Pinochet is awaiting a hearing, his fate in the balance. American scholars applauded the arrest.
The way entities are introduced and discussed influences coherence (Grosz et al., 1995).

- Entities in an utterance are ranked according to salience.
  - Is an entity pronominalized or not?
  - Is an entity in a prominent syntactic position?
- Each utterance has one center (≈ topic or focus).
- Coherent discourses have utterances with common centers.
- Entity transitions capture degrees of coherence (e.g., in Centering theory \textit{continue} > \textit{shift}).

\textit{Notions of salience, utterance, ranking are left unspecified.}

---

**Entity-based Local Coherence**

John went to his favorite music store to buy a piano.
He had frequented the store for many years.
He was excited that he could finally buy a piano.
He arrived just as the store was closing for the day.

Can we compute a discourse representation automatically?

- Does it capture coherence characteristics?
- What linguistic information matters for coherence?
- Is it robust across domains and genres?

What is an appropriate coherence model?

- View coherence rating as a machine learning problem.
- Learn a ranking function without manual involvement.
- Apply to text-to-text generation tasks.

\textit{Inspired from entity-based theories, not a direct implementation of any theory in particular.}
1. Former Chilean dictator Augusto Pinochet, was arrested in London on 14 October 1998.
2. Pinochet, 82, was recovering from surgery.
3. The arrest was in response to an extradition warrant served by a Spanish judge.
4. Pinochet was charged with murdering thousands, including many Spaniards.
5. He is awaiting a hearing, his fate in the balance.
6. American scholars applauded the arrest.
The Entity Grid

<table>
<thead>
<tr>
<th>Entity Grid</th>
<th>Discourse Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>The Entity Grid Evaluation</td>
</tr>
<tr>
<td>Rank Model</td>
<td>Entity Transitions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Entity</th>
<th>Discourse Representation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinochet</td>
<td>London</td>
<td>October</td>
<td>Surgery</td>
</tr>
<tr>
<td>1</td>
<td>S</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>S</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>O</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>S</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>S</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The Entity Grid

The Entity Grid

The Entity Grid

The Entity Grid

Introduction
The Entity Grid
Evaluation
Discourse Representation
Entity Transitions
Ranking Model

Entity Transitions

Definition
A local entity transition is a sequence \( \{S, O, X, -\}^n \) that represents entity occurrences and their syntactic roles in \( n \) adjacent sentences.

Feature Vector Notation
Each grid \( x_{ij} \) for document \( d_i \) is represented by a feature vector:

\[
\Phi(x_{ij}) = (p_1(x_{ij}), p_2(x_{ij}), \ldots, p_m(x_{ij}))
\]

\( m \) is the number of predefined entity transitions

\( p_t(x_{ij}) \) the probability of transition \( t \) in grid \( x_{ij} \)

Example (transitions of length 2)
Entity Transitions

Example (transitions of length 2)

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>O</th>
<th>X</th>
<th></th>
<th>S</th>
<th>O</th>
<th>X</th>
<th></th>
<th>S</th>
<th>O</th>
<th>X</th>
<th></th>
<th>S</th>
<th>O</th>
<th>X</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>d_1</td>
<td>0</td>
<td>0</td>
<td>.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.02</td>
<td>.07</td>
<td>0</td>
<td>0</td>
<td>.12</td>
<td>.02</td>
<td>.02</td>
<td>.05</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>d_2</td>
<td>0</td>
<td>0</td>
<td>.02</td>
<td>.07</td>
<td>0</td>
<td>0</td>
<td>.02</td>
<td>0</td>
<td>.06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d_3</td>
<td>.02</td>
<td>0</td>
<td>.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.05</td>
<td>.03</td>
<td>.07</td>
<td>.07</td>
<td>.29</td>
<td></td>
</tr>
</tbody>
</table>

Learning a Ranking Function

Training Set
Ordered pairs \((x_{ij}, x_{ik})\), where \(x_{ij}\) and \(x_{ik}\) represent the same document \(d_i\), and \(x_{ij}\) is more coherent than \(x_{ik}\) (assume \(j > k\)).

Goal
Find a parameter vector \(\vec{w}\) such that:

\[
\vec{w} \cdot (\Phi(x_{ij}) - \Phi(x_{ik})) > 0 \quad \forall j, i, k \text{ such that } j > k
\]

Support Vector Machines
Constraint optimization problem can be solved using the search technique described in Joachims (2002).

Linguistic Dimensions

Salience: Are some entities more important than others?
- Discriminate between salient (frequent) entities and the rest.
- Collect statistics separately for each group.

Coreference: What is its contribution?
- Entities are coreferent if they have the same surface form.
- Apply a coreference resolution system.

Syntax: Does syntactic knowledge matter?
- Use four categories \(\{S, O, X, –\}\).
- Reduce categories to \(\{X, –\}\).
Introduction
The Entity Grid
Evaluation
Text Ordering
Summarization

Introduction
The Entity Grid
Evaluation
Text Ordering
Summarization

Comparison with State of the Art

Vector-based Model (LSA, Foltz et al., 1998):
- Meaning of individual words is represented in vector space.
- Sentence meaning is the mean of the vectors of its words.
- Average distance of adjacent sentences.
- Unsupervised, local, lexicalized, domain independent.
HMM-based Content Models (Barzilay and Lee, 2004):
- Model topics and their order in texts.
- Model is an HMM: states correspond to topics (≈ sentences).
- Model selects sentence order with highest probability.
- Supervised, global, lexicalized, domain dependent.
Omission of coreference causes performance drop.
- Syntax and Salience have more effect on Accidents corpus.
- Linguistically poor model generally worse.
- Entity model is better than LSA.
- HMM-based content models exhibit high variability.
- Models seem to be complementary.

Motivation
- Summaries naturally exhibit coherence violations.
- Compare model against rankings elicited by human judges.
- Useful for automatic evaluation of machine generated text.

Data
- Participants assign readability score on a seven point scale.
- 144 summaries, 177 participants (23 per summary).

Results: Summarization

- Coreference decreases accuracy (machine generated texts).
- Salience seems to have more of an impact here.
- Linguistically poor model is generally worse.
- Entity model performs better than LSA.
- LSA is unsupervised and exposed only to human texts.
- Training corpus is unsuitable for HMM-based content models.
Summary

Strengths:

- Novel framework for representing and measuring coherence.
- Entity grid and cross-sentential transitions.
- Suited for learning appropriate ranking function.
- Fully automatic and robust, useful for system development.

Weaknesses:

- Entity grid doesn’t contain lexical information.
- Doesn’t contain a notion of global coherence.
- Can’t model multi-paragraph text.