What’s in a style?

What does it mean to write journalistically?
...for students?
...for academics?
How do these styles differ?

Can we learn to detect compliance with a style?
Translate one style into another?
Studying style

**Summarization** is a stylistic task (sort of):
- Translate from one style (news articles)...
- ...to another (really short news articles)
- Remove news-specific structures (explanations, quotes, etc)

**Readability measurement** is another:
- Does a text conform to “simple English” style? *(Napoles+Dredze ‘10)*
- “Grade level” style? *(lots of work!)*
- Intelligible for general readers? *(Chae+Nenkova ‘09)*
Why editing?

**Summarization**: paraphrase a text to make it **shorter**

**Editing**: paraphrase a text to make it **better journalism**

**Editors**
- Trained professionals
- Stay close to original texts
- Produce a specific style for a specific audience
- Exist for many styles and domains

Can we learn to do what they do?
The data

500 article pairs processed by professional editors:

Novel dataset courtesy of Thomson Reuters

Each article in two versions: original and edited

We align originals with edited versions to find:

▸ Five thousand sentences unchanged
▸ Three thousand altered inline
▸ Six hundred inserted or deleted
▸ Three hundred split or merged
Editing is hard!

Tasks we tried:

- Predicting which sentences the editor will edit:
  - Mostly syntactic readability features from (Chae+Nenkova ‘08)
  - Significantly better than random, but not by much
Editing is hard!

Tasks we tried:

▶ Predicting which sentences the editor will edit:
  ▶ Mostly syntactic readability features from (Chae+Nenkova ‘08)
  ▶ Significantly better than random, but not by much

▶ Distinguishing “before” from “after” editing
  ▶ Major trend: News editing makes stories shorter...
  ▶ ...and individual sentences too!
  ▶ Hard to do better than this, though
Editing is hard!

Tasks we tried:

- Predicting which sentences the editor will edit:
  - Mostly syntactic readability features from (Chae+Nenkova ‘08)
  - Significantly better than random, but not by much
- Distinguishing “before” from “after” editing
  - Major trend: News editing makes stories shorter...
  - ...and individual sentences too!
  - Hard to do better than this, though
- Our most successful study: **sentence fusion**
Overview

Editing

Sentence fusion: motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
The bodies showed signs of torture. They were left on the side of a highway in Chilpancingo, in the southern state of Guerrero, state police said.

The bodies of the men, which showed signs of torture, were left on the side of a highway in Chilpancingo, state police told Reuters.
Motivation

Humans fuse sentences:

- Multidocument summaries (Banko+Vanderwende ‘04)
- Single document summaries (Jing+McKeown ‘99)
- Editing (this study)
Motivation

Humans fuse sentences:

- Multidocument summaries (Banko+Vanderwende ‘04)
- Single document summaries (Jing+McKeown ‘99)
- Editing (this study)

Previous work: multidocument case:

- Similar sentences (themes)
- Goal: summarize common information
  (Barzilay+McKeown ‘05), (Krahmer+Marsi ‘05), (Filippova+Strube ‘08)
Which sentences?

Our fusion examples

Sentences from our dataset that were fused or merged.

- Probably similar to cases from single-document summary
- Not as similar to multidocument case
  - Sentences are *not* mostly paraphrases of each other
- ...Poses problems for standard approaches
Generic framework for sentence fusion

NLP is often fun
NLP is useful

root NLP is often fun

root NLP is useful

root NLP is
often fun
useful

selection

fun

root NLP is
useful

linearization/read-out

NLP is fun and useful
## Issues with the generic framework

### Selection
What content do we keep?
- Convey the editor’s desired information
- Remain grammatical

### Merging
Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

### Learning
Can we learn to imitate human performance?
Issues with the generic framework

Selection
What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
  - Remain grammatical

Merging
Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

Learning
Can we learn to imitate human performance?
## Issues with the generic framework

### Selection

**What content do we keep?**
- Convey the editor’s desired information
  - Requires discourse; not going to address
- Remain grammatical
  - **Constraint satisfaction** (Filippova+Strube ‘08)

### Merging

**Which nodes in the graph match?**
Dissimilar sentences: correspondences are noisy!

### Learning

**Can we learn to imitate human performance?**
## Issues with the generic framework

### Selection

What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
- Remain grammatical
  - Constraint satisfaction (Filippova+Strube ‘08)

### Merging

Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!
**Contribution: Solve jointly with selection**

### Learning

Can we learn to imitate human performance?
Issues with the generic framework

Selection
What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
  - Remain grammatical
    - Constraint satisfaction (Filippova+Strube ‘08)

Merging
Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!
Contribution: Solve jointly with selection

Learning
Can we learn to imitate human performance?
Contribution: Use structured learning
Overview

Editing

Sentence fusion: motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
The content selection problem

Which content to select:

Many valid choices (Daume+Marcu ‘04), (Krahmer+al ‘08)

Input

Uribe appeared unstoppable after the rescue of Betancourt.
His popularity shot to over 90 percent, but since then news has been bad.
The content selection problem

Which content to select:

Many valid choices (Daume+Marcu ‘04), (Krahmer+al ‘08)

Input

Uribe appeared unstoppable after the rescue of Betancourt.
His popularity shot to over 90 percent, but since then news has been bad.

Output

Uribe’s popularity shot to over 90 percent after the rescue of Betancourt.
The content selection problem

Which content to select:

Many valid choices (Daume+Marcu ‘04), (Krahmer+al ‘08)

Input

Uribe appeared unstoppable after the rescue of Betancourt.
His popularity shot to over 90 percent, but since then news has been bad.

Output

Uribe used to appear unstoppable, but since then news has been bad.
Use simple dynamic programming to align input with truth...

Provide true alignments to both **system** and **human judges**.

**Input**

*Uribe appeared unstoppable* after the rescue of Betancourt.

*His popularity shot to over 90 percent*, but since then news has been bad.

**True output**

Uribe appeared unstoppable and his popularity shot to over 90 percent.
Faking content selection: finding alignments

Use simple dynamic programming to align input with truth...

Provide true alignments to both system and human judges.

**Input**

Uribe appeared unstoppable after the rescue of Betancourt.

His popularity shot to over 90 percent, but since then news has been bad.

**True output**

Uribe appeared unstoppable and his popularity shot to over 90 percent.

Still not easy– grammaticality!

Aligned regions often just fragments:

**Input**

...the Berlin speech will be a centerpiece of the tour...
Overview

Editing

Sentence fusion: motivation

Setting up the problem

**Fusion as optimization**
- Jointly finding correspondences
- Staying grammatical

Learning to fuse
- Defining an objective
- Structured learning

Evaluation
Merging dependency graphs

Previous:

Merge nodes deterministically:
- Lexical similarity
- Local syntax tree similarity

For disparate sentences, these features are noisy!
Previous:

Merge nodes deterministically:
- Lexical similarity
- Local syntax tree similarity

For disparate sentences, these features are noisy!

Our work:

Soft merging: add \textit{merge arcs} to graph
System decides whether to use or not!
Simple paraphrasing

Add relative clause arcs between subjects and verbs
(Alternates “police said” / “police, who said”)

bodies ← showed ← root

merge?  they ← left ← root

signs ← torture

side ← highway ← chilpancingo

north ← hour ← resort ← acapulco

state ← police ← root

were ← said ← root

police said
“The bodies, which showed signs of torture, were left by the side of a highway”
Finding a good fusion

Put weights on all words and arcs, then maximize the sum for selected items

Weights determine the solution—we will learn them!
Not every set of selected arcs is valid...

Constraints

Unconnected fragment

Merged node with two heads

Cycle

Missing argument (subject)
Solving with ILP

Integer Linear Programming (ILP)

Maximize a linear function
subject to:
  linear constraints
  integrality constraints

NP-hard, but well-studied practical solutions (Ilog CPLEX)

Our ILP based on (Filippova+Strube ‘08), generalized for soft merging...
Similar setup for sentence compression (Clarke+Lapata ‘08)
Very efficient for this size problem
Overview

Editing

Sentence fusion: motivation

Setting up the problem

Fusion as optimization
   Jointly finding correspondences
   Staying grammatical

Learning to fuse
   Defining an objective
   Structured learning

Evaluation
How to fuse?

ILP tells us what fusions are allowed...
The weights tell us which ones are good.

Recipe for structured learning, (Collins ‘02), others:

- Define a feature representation
- Define a loss function
- For each datapoint:
  - Compute current solution
  - Compute best possible solution
  - Update weights to push away from current, proportionally to loss
Same thing, with picture

- **space of accessible solutions**
- **oracle (best accessible)**
- **current solution**
- **loss relative to oracle**
- **editor's solution**
- **structural loss**
- **direction of update**
Features

Features for dependencies
Keep this arc?
- Parent/child POS tags
- Dependency label
- Parent/child word retained by editor?
- Dependency is inserted relative clause

Features for words
Keep this word?
- POS tag
- Word retained by editor?
Features 2

Features for merge arcs
Do these two words correspond?

- Same POS tag
- Same word
- Same arc type to parent
- WordNet similarity (Resnik ‘95), (Pedersen+al ‘04)
- Thesaurus similarity (Jarmasz+Szpakowicz ‘03)
- Hand-annotated pronoun coreference
Measure similarity to the editor’s sentence...

- Not just lexically (the editor can paraphrase, we can’t!)

Look at **connections between** the retained content

```
 bodies of the men, which showed signs of torture

 were left on the side of a highway...

 state police told Reuters root
```
Finding the oracle

Match this structure:

- **bodies** of the men, which **showed** signs of torture
- were left on the side of a highway...
- **state police** told Reuters

On this graph:

- **bodies**
- **showed**
- **root**
- **signs**
- **torture**
- **side**
- **highway**
- **chilpancingo**
- **north**
- **hour**
- **resort**
- **acapulco**
- **state**
- **police**
- **root**
- **said**
Our loss function

Penalty for:

- Bad/missing connections
- Leaving out words the editor used
- Words the editor didn’t use

Can actually find the oracle (minimize loss) with ILP...
Using polynomial number of auxiliary variables.
We have **features**, the **loss** and the **oracle**...
So we can learn...
Just need to choose an update rule:

Use the **perceptron** update with averaging \(\text{(Freund+Schapire '99)}\) and committee \(\text{(Elsas+al '08)}\)
Overview

Editing

Sentence fusion: motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

**Human**

The editor’s fused sentence
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

<table>
<thead>
<tr>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>The editor’s fused sentence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Readability upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our parsing and linearization on the editor’s sentence</td>
</tr>
</tbody>
</table>
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

<table>
<thead>
<tr>
<th>Human</th>
<th>The editor’s fused sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability upper bound</td>
<td>Our parsing and linearization on the editor’s sentence</td>
</tr>
<tr>
<td>“and”-splice</td>
<td>All input sentences, spliced with the word “and”</td>
</tr>
</tbody>
</table>
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

<table>
<thead>
<tr>
<th>Human</th>
<th>The editor’s fused sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability upper bound</td>
<td>Our parsing and linearization on the editor’s sentence</td>
</tr>
<tr>
<td>“and”-splice</td>
<td>All input sentences, spliced with the word “and”</td>
</tr>
<tr>
<td>System</td>
<td>Our system output</td>
</tr>
<tr>
<td></td>
<td>Only abstractive system we tested</td>
</tr>
<tr>
<td>System</td>
<td>Avg</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----</td>
</tr>
<tr>
<td>Editor</td>
<td>4.6</td>
</tr>
<tr>
<td>Readability UB</td>
<td>4.0</td>
</tr>
<tr>
<td>“And”-splice</td>
<td>3.7</td>
</tr>
<tr>
<td>System</td>
<td>3.1</td>
</tr>
</tbody>
</table>

- Poor linearization: gap of .6
- System: additional loss of .9
- Average system score still 3, “fair”
<table>
<thead>
<tr>
<th>System</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor</td>
<td>4.6</td>
</tr>
<tr>
<td>Readability UB</td>
<td>4.3</td>
</tr>
<tr>
<td>“And”-splice</td>
<td>3.8</td>
</tr>
<tr>
<td>System</td>
<td>3.8</td>
</tr>
</tbody>
</table>

- Score close to 4, “good”
Comparison with “and”-splice

“and”-splice content scores comparable to ours, but...

- Spliced sentences too long
  - 49 words vs human 34, system 33
- Our system has more extreme scores

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>“And”-splice</td>
<td>3</td>
<td>43</td>
<td>60</td>
<td>57</td>
<td>103</td>
<td>266</td>
</tr>
<tr>
<td>System</td>
<td>24</td>
<td>24</td>
<td>39</td>
<td>58</td>
<td>115</td>
<td>260</td>
</tr>
</tbody>
</table>
The bodies who showed signs of torture were left on the side of a highway in Chilpancingo state police said.
The suit claims the company helped fly terrorism suspects abroad to secret prisons.

Holder’s review was disclosed the same day as Justice Department lawyers repeated a Bush administration state-secret claim in a lawsuit against a Boeing Co unit.

Review was disclosed the same day as Justice Department lawyers repeated a Bush administration claim in a lawsuit against a Boeing Co unit that helped fly terrorism suspects abroad to secret prisons.
Biden a veteran Democratic senator from Delaware that Vice president-elect and Joe had contacted to lobby was quoted by the Huffington Post as saying Obama had made a mistake by not consulting Feinstein on the Panetta choice.

Vice President-elect Joe Biden, a veteran Democratic senator from Delaware who had contacted...
Our system

The White House that took when Israel invaded Lebanon in 2006 showed no signs of preparing to call for restraint by Israel and the stance echoed of the position.

Missing arguments

took, position
A sentence-fusion technique:

- Trained on naturally occurring data
- Finds correspondences jointly with selection
- Supervised structured learning
Future work

New data:
- Data elicited from humans (McKeown ‘10 corpus)
- Single-document summary

Better techniques:
- Automatic coreference
- Paraphrasing rules
Editing data provides:

▶ Information about style
▶ Natural examples of how to improve text

In principle, it should be easy to obtain—though news corporations may not agree!

Learning to edit is hard (but possible):

▶ We can’t always predict what will be edited and how.
▶ Automatic editing and style translation are still far from solved.
Acknowledgements

Thompson-Reuters: Alan Elsner, Howard Goller, Thomas Kim
BLLIP labmates: Eugene Charniak, Stu Black, Rebecca
Mason, Ben Swanson
Funds: Google Fellowship for NLP
All of you!