

Models for Sentence Compression

A Comparison across Domains, Training Requirements and
Evaluation Measures

James Clarke and Mirella Lapata

School of Informatics
University of Edinburgh

July 2006
ACL 2006

What is Sentence Compression?

Sentence Compression

Can be viewed as producing a summary of a single sentence.

What is Sentence Compression?

Sentence Compression

Can be viewed as producing a summary of a single sentence.

More formally

A compressed sentence should:

- Use **less** words than the original sentence.
- Preserve the most **important information**.
- Remain **grammatical**.

Simplification

Sentence compression can involve...

- word reordering
- word deletion
- word substitution
- word insertion

Simplification

Sentence compression can involve...

- word reordering
- word deletion
- word substitution
- word insertion

Ideally we want to exploit all of these operations but let's start simple:

Knight and Marcu (2002)

Given an input sentence of words $W = w_1, w_2, \dots, w_n$, a compression is formed by dropping any subset of these words.

Example Compression

Original

Prime Minister Tony Blair today insisted the case for holding terrorism suspects without trial was “absolutely compelling” as the government published new legislation allowing detention for 90 days without charge.

Example Compression

Original

Prime Minister Tony Blair today insisted the case for holding terrorism suspects without trial was “absolutely compelling” as the government published new legislation allowing detention for 90 days without charge.

Compression

Tony Blair insisted the case for holding terrorism suspects without trial was “compelling”.

Example Compression

Original

Prime Minister Tony Blair today insisted the case for holding terrorism suspects without trial was “absolutely compelling” as the government published new legislation allowing detention for 90 days without charge.

Compression

Tony Blair insisted the case for holding terrorism suspects without trial was “compelling”.

Outline

1 Sentence Compression

- Motivation
- Previous Work

2 Our Work

- How do humans compress sentences?
- Do existing methods port well across domains?
- What about automatic evaluation measures?

3 Discussion

Outline

1 Sentence Compression

- Motivation
- Previous Work

2 Our Work

- How do humans compress sentences?
- Do existing methods port well across domains?
- What about automatic evaluation measures?

3 Discussion

Applications

Within summarisation:

- Current systems contain manually written rules for sentence compression.

Other Applications include:

- Subtitle generation.
- Text compression for display on small screens.
- Audio scanning devices for the blind.

Outline

1 Sentence Compression

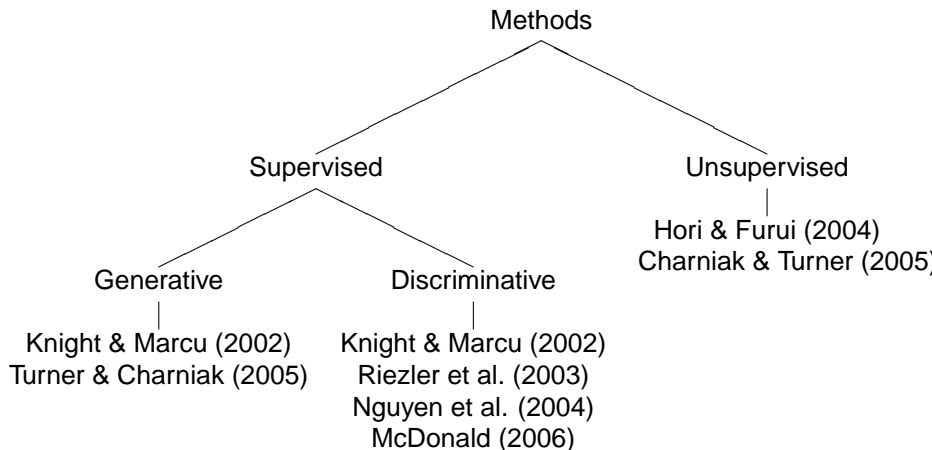
- Motivation
- Previous Work

2 Our Work

- How do humans compress sentences?
- Do existing methods port well across domains?
- What about automatic evaluation measures?

3 Discussion

Previous Work



Data Requirements

Parallel Corpora

- Most approaches rely on a **parallel corpus**.
- Automatically produced Ziff-Davis (Knight and Marcu, 2002).
- Domain: newspaper articles.
- There is no 'natural' resource of original-compressed sentences.

Data Requirements

Parallel Corpora

- Most approaches rely on a **parallel corpus**.
- Automatically produced Ziff-Davis (Knight and Marcu, 2002).
- Domain: newspaper articles.
- There is no ‘natural’ resource of original-compressed sentences.

Abstract

Blah blah blah. **The documentation is excellent.**
Blah blah blah ...

Document

... blah blah blah. The documentation is excellent – it is clearly written with numerous drawings, cautions and tips, and includes an entire section on troubleshooting. Blah ...

Data Requirements

Parallel Corpora

- Most approaches rely on a **parallel corpus**.
- Automatically produced Ziff-Davis (Knight and Marcu, 2002).
- Domain: newspaper articles.
- There is no ‘natural’ resource of original-compressed sentences.

Abstract

Blah blah blah. **The documentation is excellent.**
Blah blah blah ...

Document

... blah blah blah. **The documentation is excellent – it is clearly written with numerous drawings, cautions and tips, and includes an entire section on troubleshooting.** Blah ...

Evaluation

Methodology

- Algorithms are evaluated on small sample (32 sentences).
- Humans are asked to assess grammaticality and information content.
- Typically four participants are used.
- Unlike machine translation, no established automatic measure.
- Comparisons across systems and system-configurations?

Outline

1 Sentence Compression

- Motivation
- Previous Work

2 Our Work

- **How do humans compress sentences?**
- Do existing methods port well across domains?
- What about automatic evaluation measures?

3 Discussion

Human-authored Compression Corpus

Spoken Text

- Natural domain for compression applications.
- Speech is challenging (ungrammatical, incomplete).
- No naturally occurring compression corpora.

Human-authored Compression Corpus

Spoken Text

- Natural domain for compression applications.
- Speech is challenging (ungrammatical, incomplete).
- No naturally occurring compression corpora.

Methodology

- 50 Broadcast news documents.
- 3 annotators remove tokens from original transcript:
 - preserve most important information in original sentence.
 - preserve grammaticality of the compressed sentence.
- Could also leave a sentence uncompressed.

Example Human Compressions

Original

President Boris Yeltsin has won the most votes in Russia 's hotly contested presidential election , one watched around the world .

Compressions

- 1 Boris Yeltsin has the most votes in Russia 's presidential election .
- 2 Boris Yeltsin has won the most votes in Russia 's presidential election , watched around the world .
- 3 Boris Yeltsin has won the most votes in Russia 's presidential election .

Analysis: Compression Rate

	A1	A2	A3	Av	Ziff-Davis
% compressed	88	79	87	84.4	97
CompRate	73.1	79.0	70.0	73.03	47

Analysis: Compression Rate

	A1	A2	A3	Av	Ziff-Davis
% compressed	88	79	87	84.4	97
CompRate	73.1	79.0	70.0	73.03	47

- Similar compression rates for annotators.

Analysis: Compression Rate

	A1	A2	A3	Av	Ziff-Davis
% compressed	88	79	87	84.4	97
CompRate	73.1	79.0	70.0	73.03	47

- Similar compression rates for annotators.
- Ziff-Davis corpus is compressed much more aggressively.

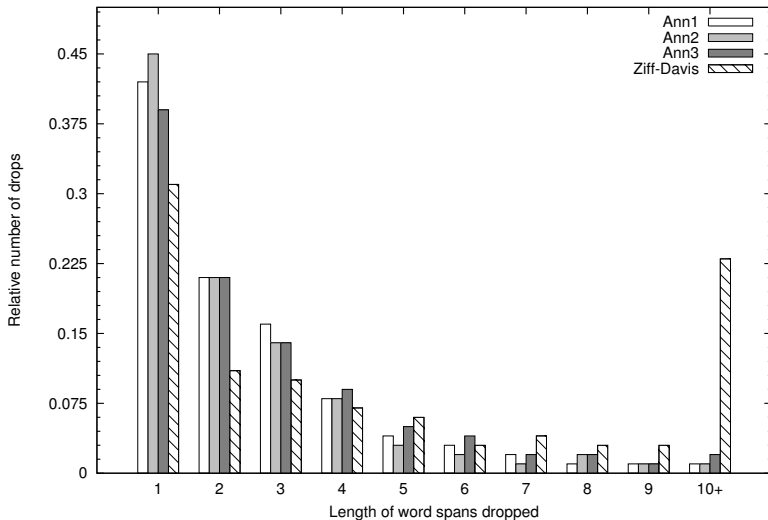
Analysis: Compression Rate

	A1	A2	A3	Av	Ziff-Davis
% compressed	88	79	87	84.4	97
CompRate	73.1	79.0	70.0	73.03	47

- Similar compression rates for annotators.
- Ziff-Davis corpus is compressed much more aggressively.
- Ziff-Davis corpus may not be comparable with human performance.

Analysis: Spans

Distribution of lengths of words spans dropped



Outline

1 Sentence Compression

- Motivation
- Previous Work

2 Our Work

- How do humans compress sentences?
- **Do existing methods port well across domains?**
- What about automatic evaluation measures?

3 Discussion

Decision-based Sentence Compression

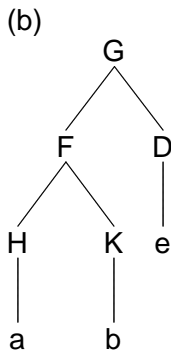
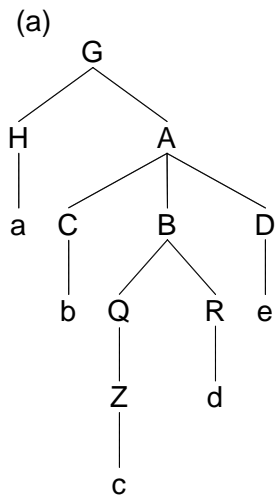
Compression as a rewriting problem

Decompose rewriting process into sequence of shift-reduce-drop actions (Knight and Marcu, 2002) following an extended shift-reduce parsing paradigm.

Operations

- SHIFT** transfers the first word from the input list to the stack.
- ASSIGNTYPE** changes the label of trees at the top of the stack.
- REDUCE** combines syntactic trees from the stack to form a new tree.
- DROP** deletes from the input list subsequences of words that correspond to a syntactic constituent.

Decision-based Example



Decision-based Example

Stack	Input List	Operation
	G H a G A C b G A B Q Z c G A B R d G A D e	SHIFT ASSIGNTYPE H

Decision-based Example

Stack	Input List	Operation
H a	G A C b G A B Q Z c G A B R d G A D e	SHIFT ASSIGNTYPE K

Decision-based Example

Stack	Input List	Operation
<pre> H a K b </pre>	<pre> G A B Q Z c G A B R d G A D e </pre>	<p>REDUCE 2 F</p>

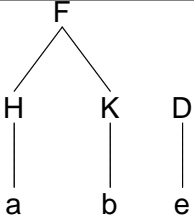
Decision-based Example

Stack	Input List	Operation
<pre> graph TD F --> H F --> K H --> a K --> b </pre>	<p>G A B Q Z c</p> <p>G A B R d</p> <p>G A D e</p>	<p>DROP B</p>

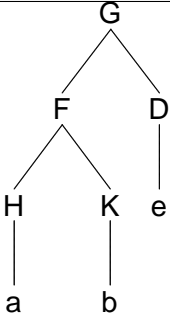
Decision-based Example

Stack	Input List	Operation
<pre>graph TD; F --> H; F --> K; H --> a; K --> b;</pre>	G A D e	SHIFT ASSIGNTYPE D

Decision-based Example

Stack	Input List	Operation
 <pre>graph TD; F --> H; F --> K; H --> a; K --> b; D --> e;</pre>		REDUCE 2 G

Decision-based Example

Stack	Input List	Operation
 <pre>graph TD; G --> F; G --> D; F --> H; F --> K; H --> a; K --> b; D --> e;</pre>		

Decision-based Compression

- Learning cases are automatically generated from a **parallel corpus**.
- 99 features are extracted from each learning case.
- Decision tree model learnt from the data.
- Model determines which operation to perform given a set of features.

Word-based Model

Original Model (Hori, 2002)

- Word-based score maximisation model.
- Score based on corpus knowledge.
- Maximised for fixed compression length using dynamic programming.
- Does **not** require a parallel corpus.

Word-based Model

Original Model (Hori, 2002)

- Word-based score maximisation model.
- Score based on corpus knowledge.
- Maximised for fixed compression length using dynamic programming.
- Does **not** require a parallel corpus.

Modifications

- Removed the length parameter.
- Added more **linguistic knowledge** into the scoring function.

Score

$$\arg \max_{\mathbf{V}} S(\mathbf{V}) = \sum_{m=1}^M \lambda_I I(V_m) + \lambda_L L(V_m | V_{m-1} V_{m-2}) + \lambda_{SOV} SOV(V_m)$$

Score

$$\arg \max_V S(V) = \sum_{m=1}^M \lambda_I I(V_m) + \lambda_L L(V_m | V_{m-1} V_{m-2}) + \lambda_{SOV} SOV(V_m)$$

Score

$$\arg \max_V S(V) = \sum_{m=1}^M \lambda_I I(V_m) + \lambda_L L(V_m | V_{m-1} V_{m-2}) + \lambda_{SOV} SOV(V_m)$$

- Significance score is designed to include important nouns and verbs.

Score

$$\arg \max_V S(V) = \sum_{m=1}^M \lambda_I I(V_m) + \lambda_L L(V_m | V_{m-1} V_{m-2}) + \lambda_{SOV} SOV(V_m)$$

- Significance score is designed to include important nouns and verbs.
- Language model's task is to preserve grammaticality.

Score

$$\arg \max_V S(V) = \sum_{m=1}^M \lambda_I I(V_m) + \lambda_L L(V_m | V_{m-1} V_{m-2}) + \lambda_{sov} \text{SOV}(V_m)$$

$$\text{SOV}(w_i) = \begin{cases} \text{freq} & \text{if } w_i \text{ in subject, object or verb role} \\ \lambda_{default} & \text{otherwise} \end{cases}$$

- Significance score is designed to include important nouns and verbs.
- Language model's task is to preserve grammaticality.
- Subjects, objects and verbs should not be dropped.
- Words in other syntactic roles can be considered for removal.

Score

$$\arg \max_V S(V) = \sum_{m=1}^M \lambda_I I(V_m) + \lambda_L L(V_m | V_{m-1} V_{m-2}) + \lambda_{sov} SOV(V_m)$$

$$SOV(w_i) = \begin{cases} \text{freq} & \text{if } w_i \text{ in subject, object or verb role} \\ \lambda_{default} & \text{otherwise} \end{cases}$$

- Significance score is designed to include important nouns and verbs.
- Language model's task is to preserve grammaticality.
- Subjects, objects and verbs should not be dropped.
- Words in other syntactic roles can be considered for removal.

Comparison

Experimental Setup

- Compare decision-tree and word-based model on Ziff-Davis and Broadcast news corpus.
- Evaluate against human judgements:
 - Sixty unpaid volunteers.
 - Instructions and examples define the compression task.
 - Rate each sentence on a five point scale.
 - Take into account information retained and grammaticality.

Results

Broadcast News	CompR	Ratings
Decision-tree	0.55	2.04
Word-based	0.72	2.78
gold standard	0.71	3.87

Ziff-Davis	CompR	Ratings
Decision-tree	0.58	2.34
Word-based	0.60	2.43
gold standard	0.54	3.53

Results

Broadcast News	CompR	Ratings
Decision-tree	0.55	2.04
Word-based	0.72	2.78
gold standard	0.71	3.87

Ziff-Davis	CompR	Ratings
Decision-tree	0.58	2.34
Word-based	0.60	2.43
gold standard	0.54	3.53

Decision-tree sensitive to training data.

Results

Broadcast News	CompR	Ratings
Decision-tree	0.55	2.04
Word-based	0.72	2.78
gold standard	0.71	3.87

Ziff-Davis	CompR	Ratings
Decision-tree	0.58	2.34
Word-based	0.60	2.43
gold standard	0.54	3.53

Rebuilds original sentence 75% of the time.

Results

Broadcast News	CompR	Ratings
Decision-tree	0.55	2.04
Word-based	0.72	2.78
gold standard	0.71	3.87

Ziff-Davis	CompR	Ratings
Decision-tree	0.58	2.34
Word-based	0.60	2.43
gold standard	0.54	3.53

Word-based model produces compression rate similar to gold-standard.

Results

Broadcast News	CompR	Ratings
Decision-tree	0.55	2.04
Word-based	0.72	2.78
gold standard	0.71	3.87

Ziff-Davis	CompR	Ratings
Decision-tree	0.58	2.34
Word-based	0.60	2.43
gold standard	0.54	3.53

Word-based model sig. better than decision-tree; both sig. worse than humans

Results

Broadcast News	CompR	Ratings
Decision-tree	0.55	2.04
Word-based	0.72	2.78
gold standard	0.71	3.87

Ziff-Davis	CompR	Ratings
Decision-tree	0.58	2.34
Word-based	0.60	2.43
gold standard	0.54	3.53

No sig. difference between models; both sig. worse than humans.

Outline

1 Sentence Compression

- Motivation
- Previous Work

2 Our Work

- How do humans compress sentences?
- Do existing methods port well across domains?
- What about automatic evaluation measures?

3 Discussion

Simple String Accuracy

Based on the edit distance between two strings (Bangalore, Rambow, and Whittaker, 2000).

$$\text{Simple String Accuracy (SSA)} = \left(1 - \frac{I + D + S}{R}\right)$$

I = Insertions

D = Deletions

S = Substitutions

R = Length of gold-standard

Relation-based Evaluation

- Proposed by Riezler et al. (2003).
- Compares the grammatical relations between compression and gold-standard.
- This allows us “to measure the semantic aspects of summarisation quality in terms of grammatical-functional information”
- Use standard IR measure of F-score.

Correlation Analysis

Measure	Ziff-Davis	Broadcast News
SSA	0.171	0.348*
F-score	0.575**	0.532**
IntSubj	0.679	0.746
<i>*p < 0.05 **p < 0.01</i>		

- SSA does not correlate with human judgements on both corpora.
- Relation F-score correlates significantly with human ratings on both corpora.

Example System Compressions

Example System Compressions

- o: Apparently Fergie very much wants to have a career in television.
- d: A career in television.
- w: Fergie wants to have a career in television.
- g: Fergie wants a career in television.

Example System Compressions

- o: Many debugging features, including user-defined break points and variable-watching and message-watching windows, have been added.
- d: Many debugging features.
- w: Debugging features, and windows, have been added.
- g: Many debugging features have been added.

Example System Compressions

- o: As you said, the president has just left for a busy three days of speeches and fundraising in Nevada, California and New Mexico.
- d: As you said, the president has just left for a busy three days.
- w: You said, the president has left for three days of speeches and fundraising in Nevada, California and New Mexico.
- g: The president left for three days of speeches and fundraising in Nevada, California and New Mexico.

Discussion

Findings

- Decision-tree model is **sensitive** to the style of training data and does not generalise to our new corpus.
- Word-based model performs significantly better than decision-tree on broadcast news.
- Both systems are comparable on written text.
- F-Score correlates with human judgements.

Future Work

Sentence Compression as Optimisation

- Underlying model: Trigram language model.
- Decoding: Integer Programming.
- Advantage: Include linguistically motivated constraints.
 - Compressions are **structurally** and **semantically** valid.
- **See my poster today!**