Generating a Table of Contents

S.R.K.Branavan, Pawan Deshpande, Regina Barzilay

MIT

Introduction

Current Summarization Methods :

Focus on relatively short articles

e.g. Newspaper articles ≈ 1,200 words (DUC 2006)

May not be applicable to longer texts

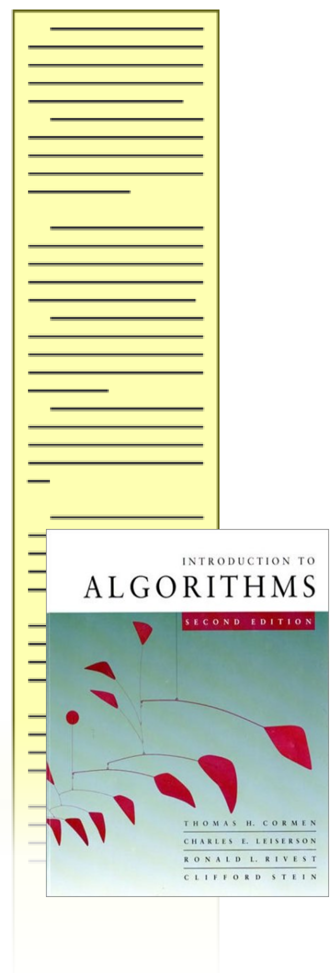
e.g. Books, lecture transcripts, ...

Our Goal :

Summarize large texts into Tables of Contents.

Automatic Generation of Tables of Contents

Unstructured Document



Structured Document

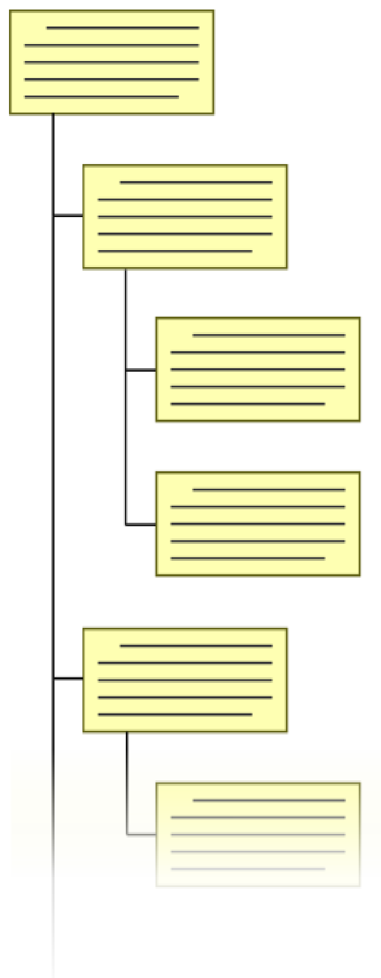


Table of Contents

Dictionary operations

Direct address table

Computer memory

Worst case running time

Hash table with load factor

Address table

Hash function

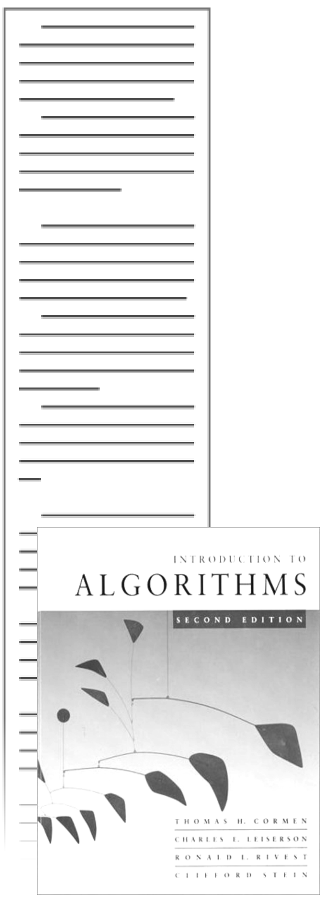
Quadratic probing

Double hashing

An actual generated Table of Contents

Our Focus - *The Generation Step*

Unstructured Document



Structured Document

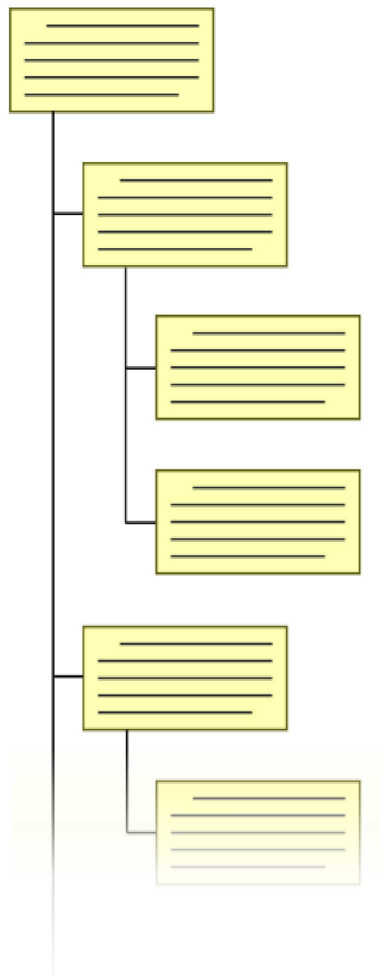


Table of Contents

Dictionary operations

Direct address table

Computer memory

Worst case running time

Hash table with load factor

Address table

Hash function

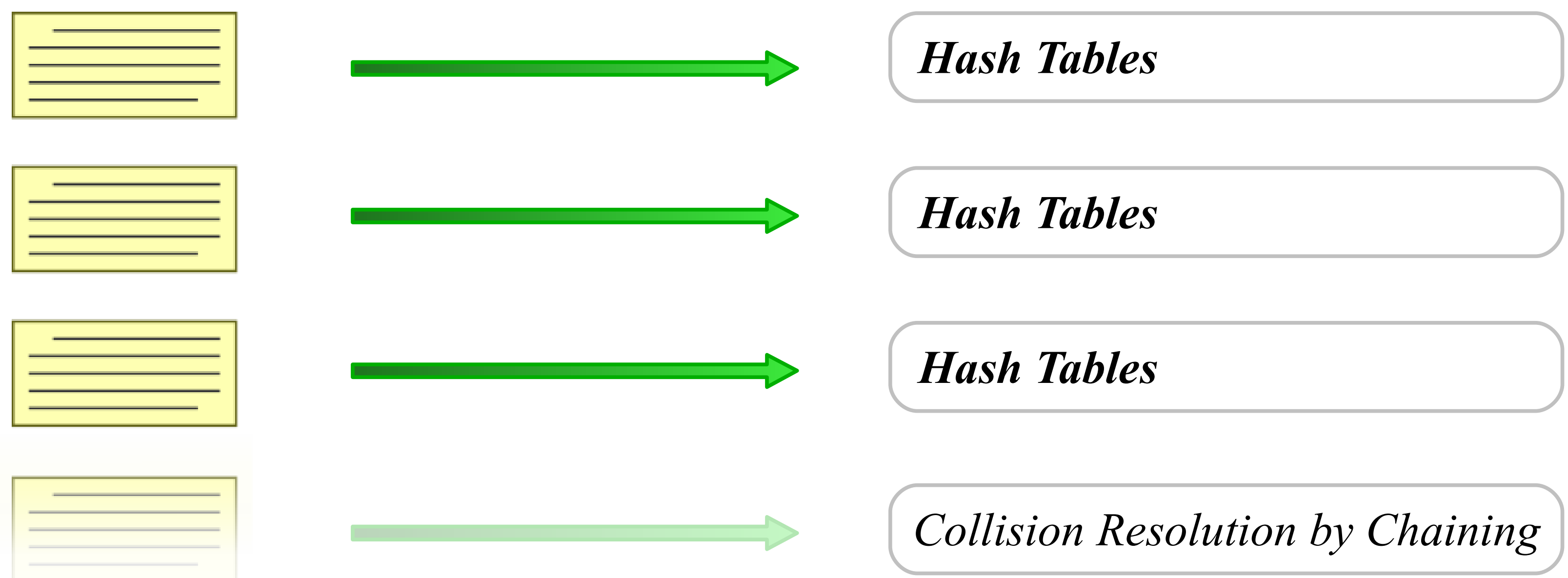
Quadratic probing

Double hashing

An actual generated Table of Contents

A Simple Approach
One Title at a Time

Document → Titles



Problems :

- Duplication of titles
- Lack of cohesion between titles

A Better Alternative

Our Approach : A Hierarchical Model

- Model relationships between titles
- Integrate a wide variety of local & global constraints

Challenges

Title Generation:

Generated titles should be :

- Representative
- Grammatical

Search space is exponential in title length

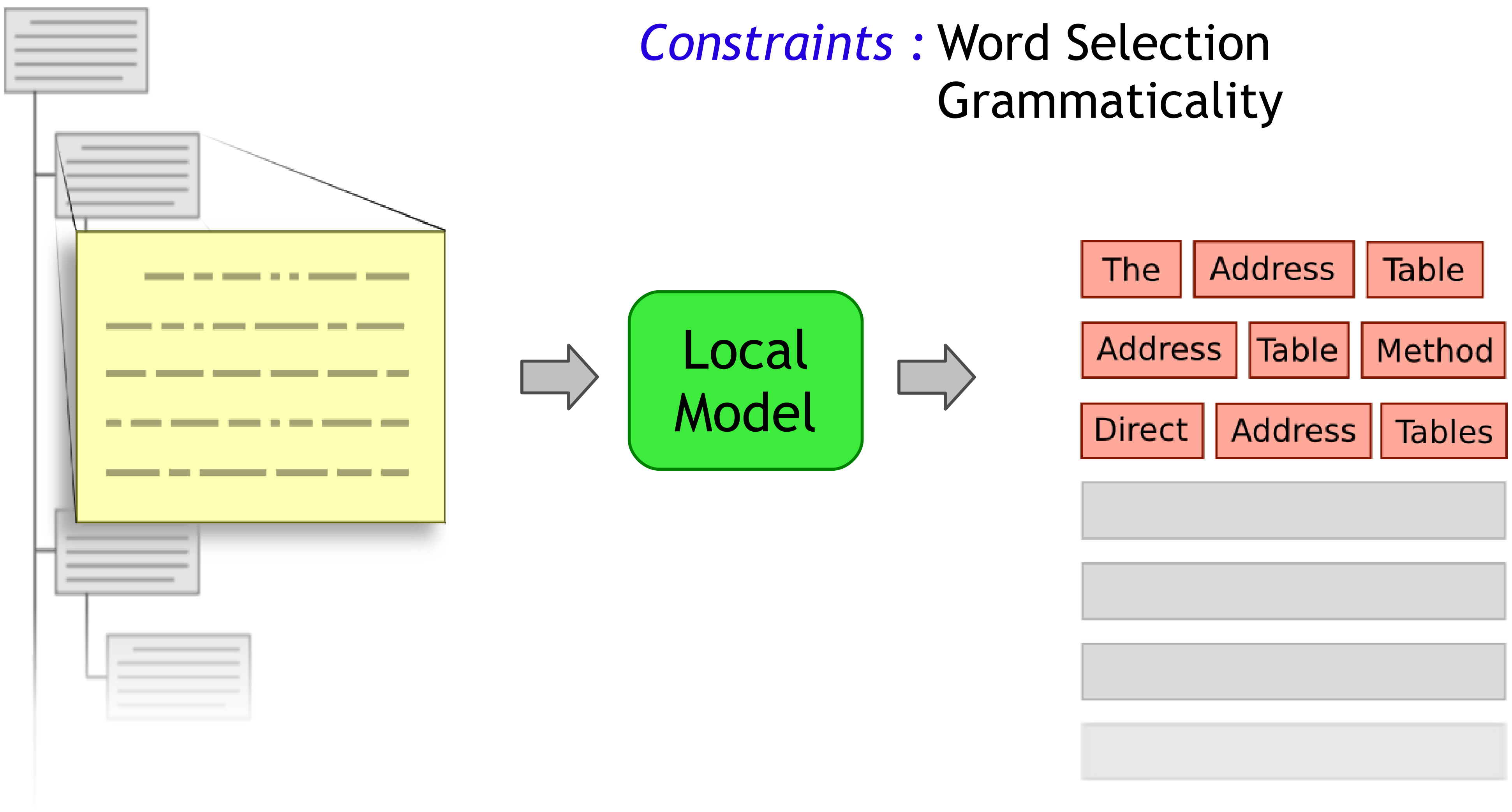
Table of Contents Generation:

Tables of Contents have rich internal structure

Search space exponential in number of titles

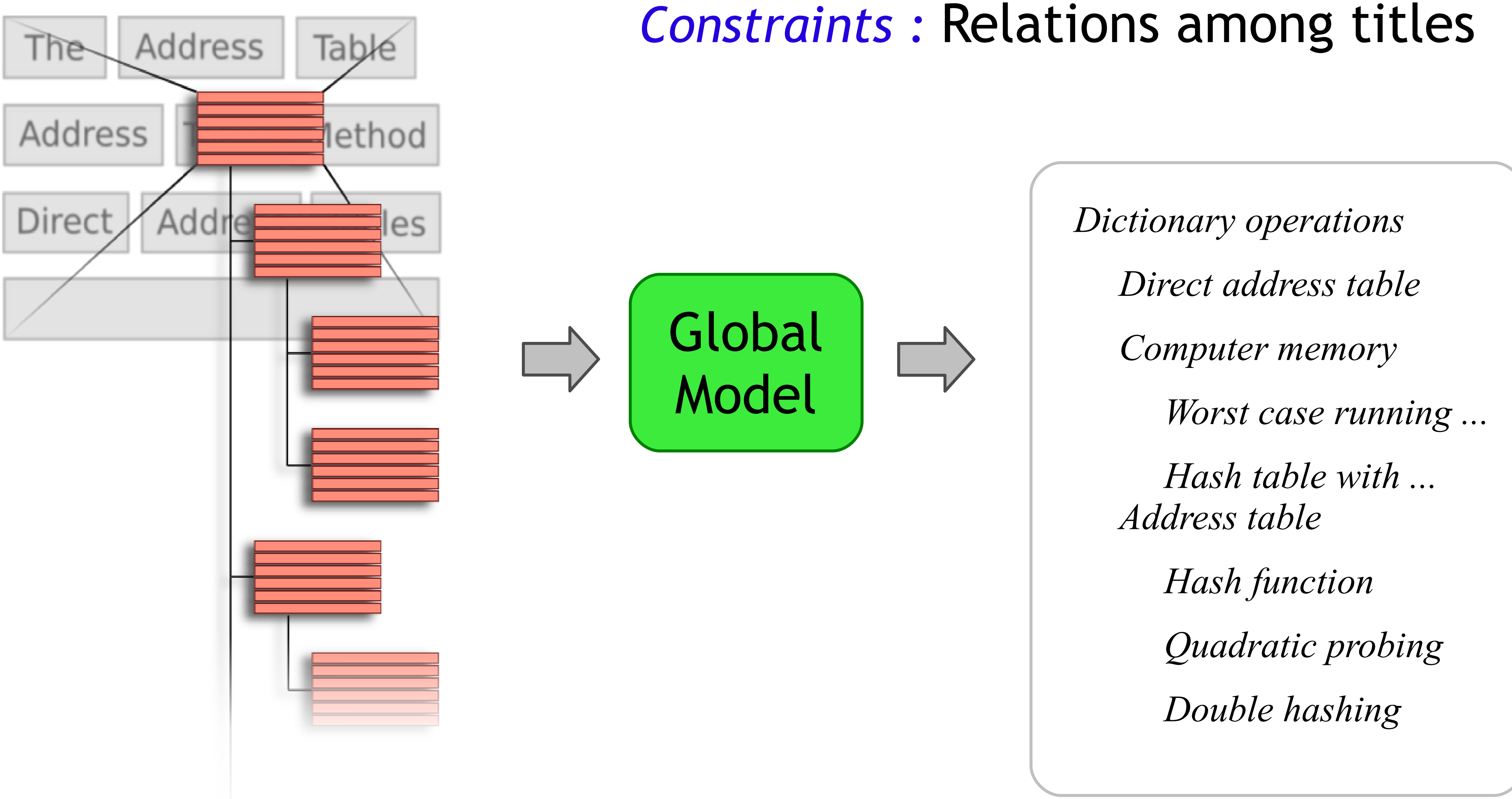
Model Architecture

Constraints : Word Selection
Grammaticality



Model Architecture

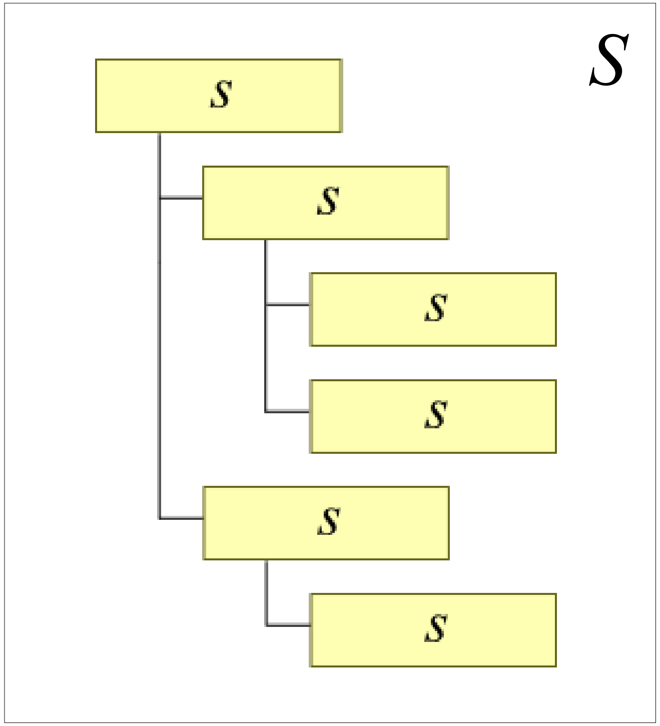
Constraints : Relations among titles

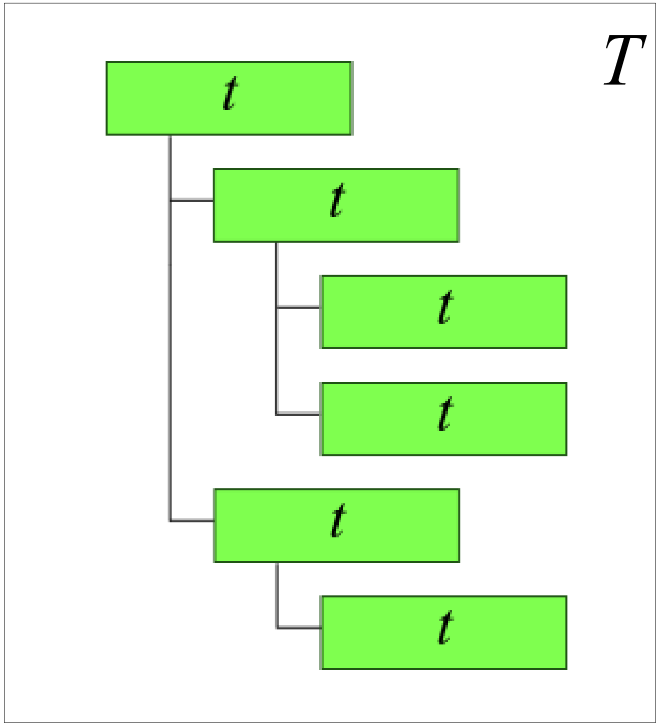


Two Linear Models

Local Model : $(s, t)_i \longrightarrow \Phi_{loc}(s, t)$

Global Model : $(S, T)_j \longrightarrow \Phi_{glob}(S, T)$





Features – *Local Model*

Selection Features :

- TF*IDF
- Part of speech
- Position of word in section
- Occurrence of word in parent / sibling sections

Contextual Features :

- Bigram + trigram language model scores

  – lexical & part of speech
- Noun phrase relative frequency

Features – *Global Model*

- Local model rank of title
- Redundancy
  - Title duplication
  - Title similarity
- Parallel construction
  - First word match with parent / sibling titles
  - Last word match with parent / sibling titles
- Local model feature averages

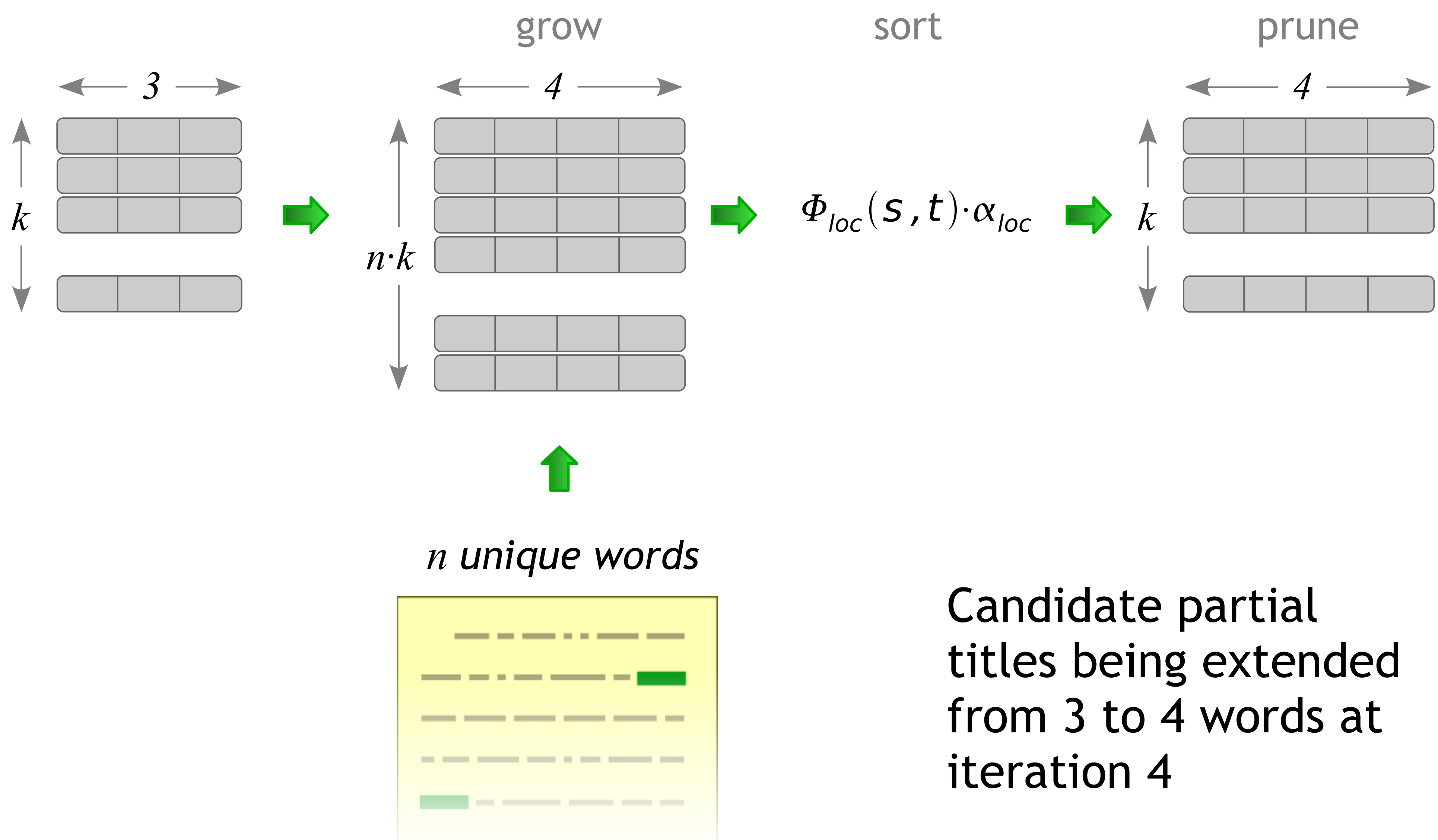
Generating Titles

$\Phi_{loc}(s,t)\cdot\alpha_{loc}$	- Local linear model
$\Phi_{loc}(s,t)$	- Local model feature vector
α_{loc}	- Local model parameter vector

All possible titles	Score
Candidate Title 1	0.154
Candidate Title 2	0.102
Candidate Title 3	0.099
Candidate Title 4	0.099
Candidate Title 5	0.098
Candidate Title 6	0.081
Candidate Title 7	0.080
Candidate Title 8	0.077

$$\Phi_{loc}(s,t)\cdot\alpha_{loc}$$

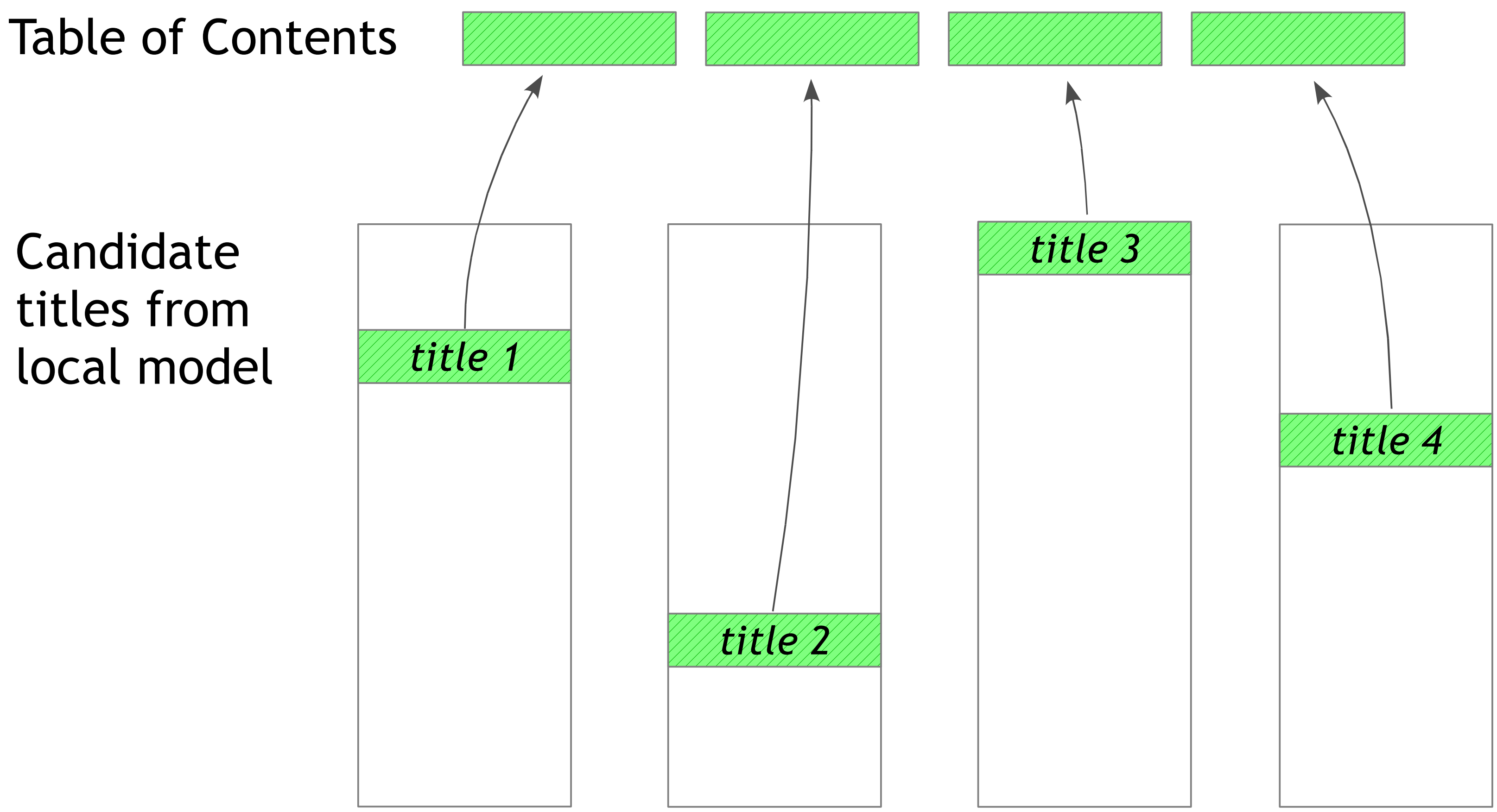
Generating Titles



Candidate partial titles being extended from 3 to 4 words at iteration 4

Generating Tables of Contents

$\Phi_{glob}(s,t)\cdot\alpha_{glob}$ - Global linear model



Training

Local & Global Models:

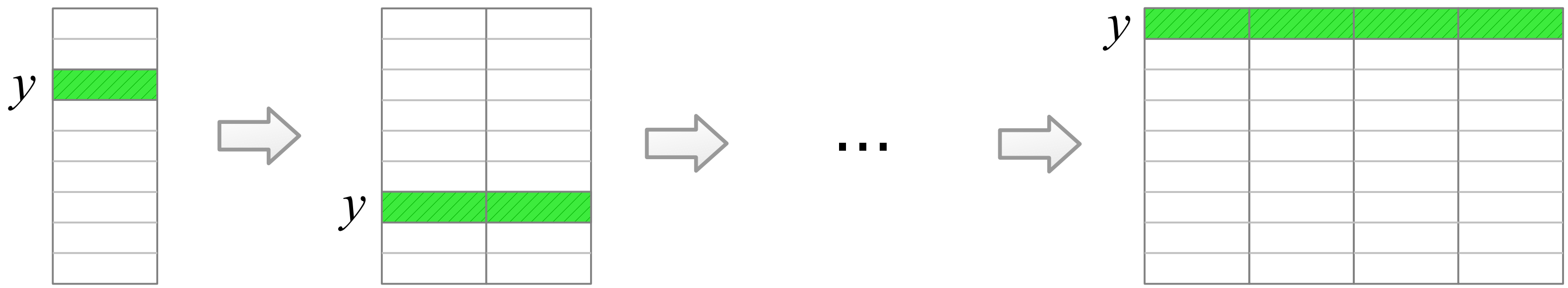
The Incremental Perceptron Algorithm
(Collins & Roark, 2004)

$$\alpha = \alpha + \Phi(s, y) - \Phi(s, z)$$

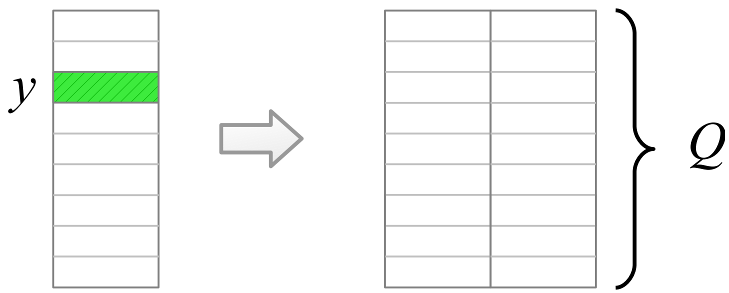
α - Parameter Vector
Φ - Feature Function
s - Input (Text Segment)
y - Target Output (Reference Title)
z - Actual Output (Generated Title)

Training - *Local Model*

y - reference title

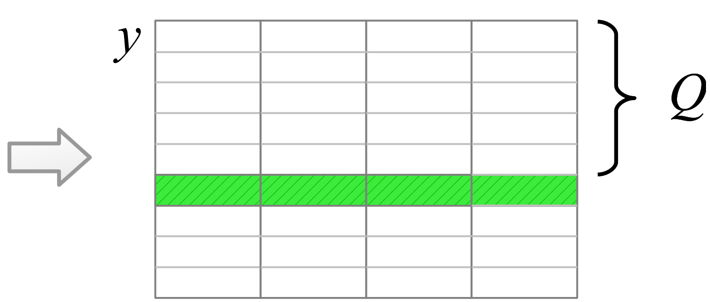


1



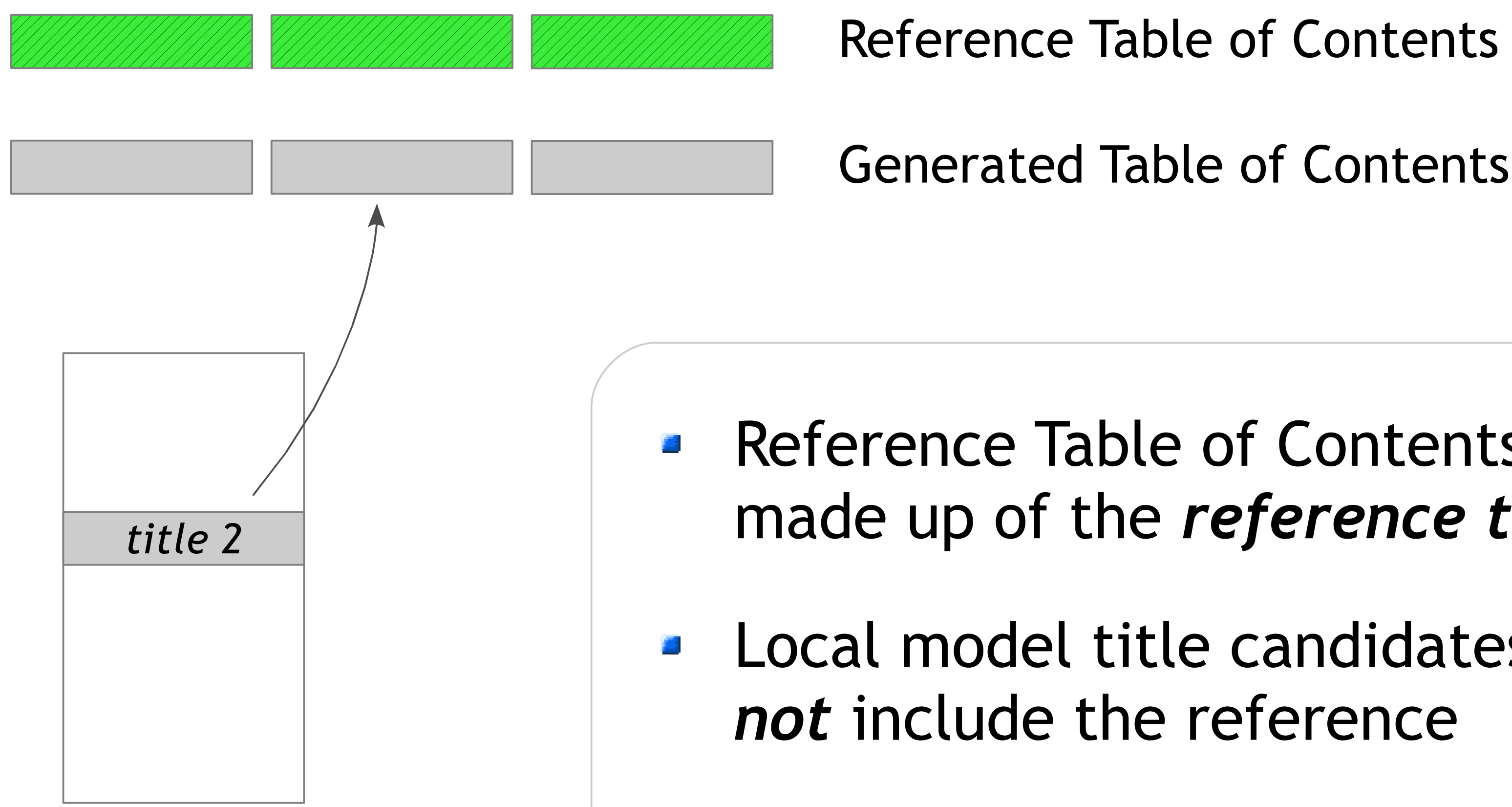
$$\alpha = \alpha + \Phi(s, y_j) - \frac{1}{|Q|} \sum_{z \in Q} \Phi(s, z)$$

2



$$\alpha = \alpha + \Phi(s, y) - \frac{1}{|Q|} \sum_{z \in Q} \Phi(s, z)$$

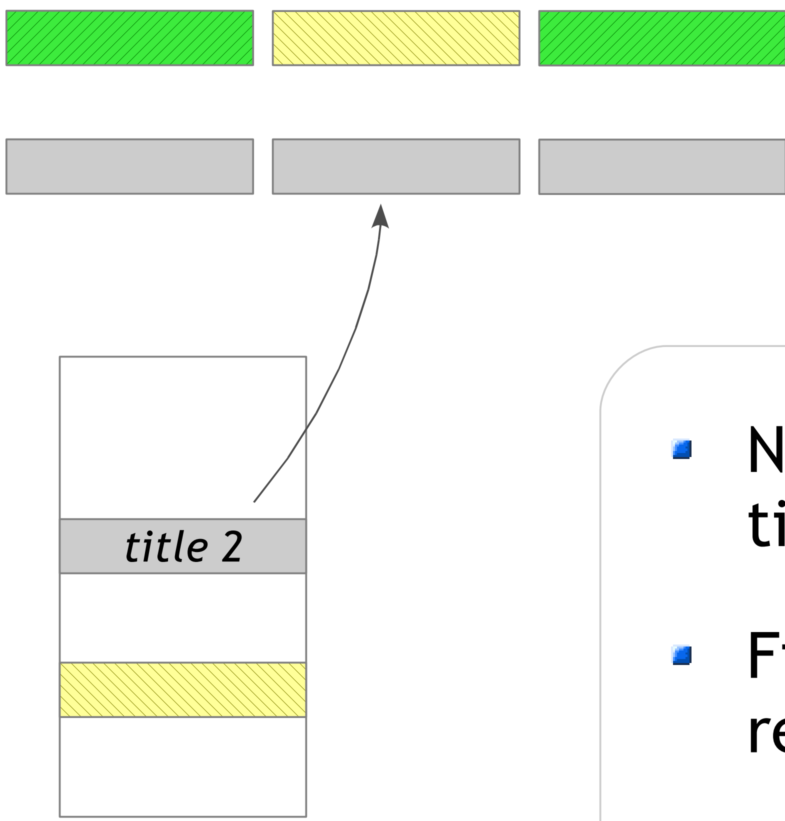
Training – *Global Model*



Local model title candidates don't include ***reference title*** !

- Reference Table of Contents is made up of the ***reference titles***
- Local model title candidates may ***not*** include the reference
- Then the global model can never produce the reference

Training – *Global Model*



Local model title candidates don't include ***reference title*** !

- Need an ***alternate*** if reference title is not present
- Find the title ***most similar*** to the reference among the candidates

$$\hat{y} = \underset{z \in Q}{argmin}\, L_1(z, y)$$

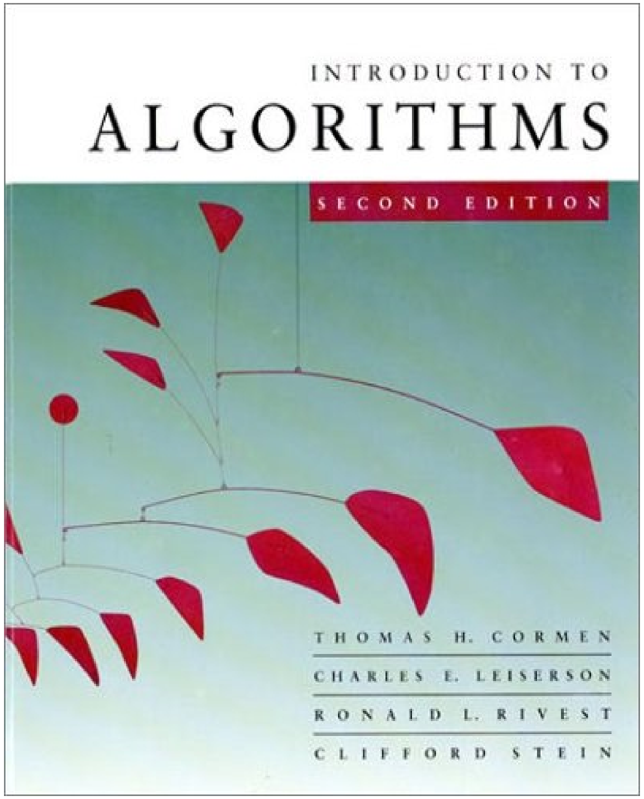
Corpus

Number of Sections	540
Number of Trees	39
Tree Depth	4

Number of Words	269,650
Avg. Section Length (words)	609.2
Avg. Title Length (words)	3.64
Avg. Title Duplicates	21
Avg. Branching	3.29

Training : 80% of trees

Testing : 20% of trees



Experimental Set-up

Data :

10 randomization of the 540 segments.

80% training, 20% testing

Automatic Evaluation :

Rouge

Human Evaluation :

Comparative assessment of title quality

- 3 alternative titles – reference, HD, best baseline
- 6 judges
- Total of 498 titles for 166 unique segments ranked

Comparisons

NP - Noun Phrase with Highest TF*IDF

FG - Noisy-channel Generative (Banko et al. 2000)

HG - Hierarchical Generative (***FG*** + hierarchical model)

FD - Flat Discriminative

HD - Hierarchical Discriminative (***FD*** + hierarchical model)

Results - *Rouge Evaluation*

	Rouge 1	Rouge L	Rouge W	Exact Match
Discriminative				
Hierarchical (HD)	***0.256***	***0.249***	***0.216***	***13.5***
Flat (FD)	0.241	0.234	0.203	13.1
Generative				
Hierarchical (HG)	***0.139***	***0.133***	***0.117***	***5.8***
Flat (FG)	0.094	0.090	0.079	4.1
Noun Phrase	0.168	0.168	0.157	6.3

Improvement given by HD is significant on the Sign test at $p \leq 0.03$

Results - *Human Evaluation*

Overall pairwise comparisons of the judges' rankings.

	Better	Worse	Equal
HD *vs* **FD**	***68***	***32***	***49***
Reference *vs* **HD**	115	13	22
Reference *vs* **FD**	123	7	20

Improvement given by **HD** is significant on the Sign test at $p \leq 0.0002$

Related Work

Title Generation :

- Dorr et al. (2003)
- Jin & Hauptmann (2001)
- Banko et al. (2000)

Domain Specific Summarization :

- Teufel & Moens (2002)
- Elhadad & McKeown (2001)

Domain Independent Summarization :

- Angheluta et al. (2002)
- Boguraev & Neff (2000)

Conclusions

- Tables of contents can be automatically generated
- Global constraints significantly improve title quality
- Human & automatic evaluations confirm the benefits of joint tree learning

Future Work

Table of contents generation on :

- Automatically hierarchically segmented text
- Lecture transcripts

Code & feature vectors at:

`http://people.csail.mit.edu/branavan/code/toc`