Image and Natural Language Processing for Multimedia Information Retrieval

Mirella Lapata

School of Informatics
University of Edinburgh

ECIR 2010, Milton Keynes

Joint work with Yansong Feng

Image Retrieval

- Number of image collections is rapidly growing!
- Flickr hosts more than 3 billion images (2.5 million every day).
- CNN, Yahoo!, and BBC publish images with their stories, also photo feeds related to current events.
- Need to browse and find images in large-scale collections.
- Build a computer system that does this automatically.
- Two flavors: content-based retrieval vs. text-based.

Content-based Image Retrieval

- **User** enters an image
- **System** returns image most similar to query

Content-based Image Retrieval

- **User** enters an image
- **System** returns image most similar to query
- **Can users create good images as queries?**

Text-based Image Retrieval

- **User** types a query (`rose`)
- **System** returns images with keywords most similar to query

`rose, flower, leaf`

`rose, beetle, leaf`

`rose, church, room`

`rose, flower, leaf`

Text-based Image Retrieval

- **User** types a query (`rose`)
- **System** returns images with keywords most similar to query
- **Who will annotate the images?**

`rose, flower, leaf`

`rose, beetle, leaf`

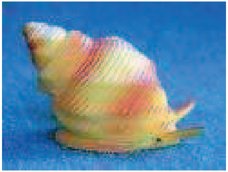
`rose, church, room`

`rose, flower, leaf`

Google Images

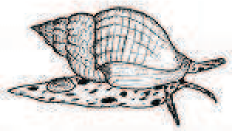
- **User** types a query (`whelk`)
- **System** matches query against text found near image (meta-data, file name, captions, user tags)

`waved whelk`



`Plastic Whelk (Sea Snail)`

`kellets-whelk.jpg`

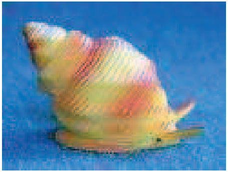


`Common Whelk`

Google Images

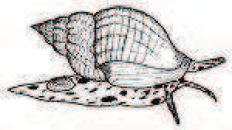
- **User** types a query (`whelk`)
- **System** matches query against text found near image (meta-data, file name, captions, user tags)
- **No annotation, no image processing**

`waved whelk`



`Plastic Whelk (Sea Snail)`

`kellets-whelk.jpg`



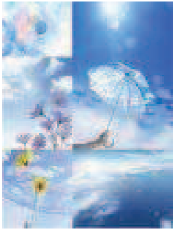
`Common Whelk`

Google Images

- Problematic for specific queries (`car, blue, sky`)
- Problematic for images without collateral text

The Solor-Blue Sky Hands Free Car Kit

The Golfer 2014 Blue Sky Hands Free Car Kit



Blue Sky HQ Stock

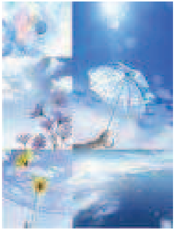
cable car terminal building corner

Google Images

- Problematic for specific queries (`car, blue, sky`)
- Problematic for images without collateral text
- **Popular:** Cyclo.ps, Pixsy, Spffy, Incogna, PicSearch

The Solor-Blue Sky Hands Free Car Kit

The Golfer 2014 Blue Sky Hands Free Car Kit



Blue Sky HQ Stock

cable car terminal building corner

Image Annotation

Solution

Obtain set of images with human annotations (clean, reliable) and train a model to do labeling task **automatically**.

Definition

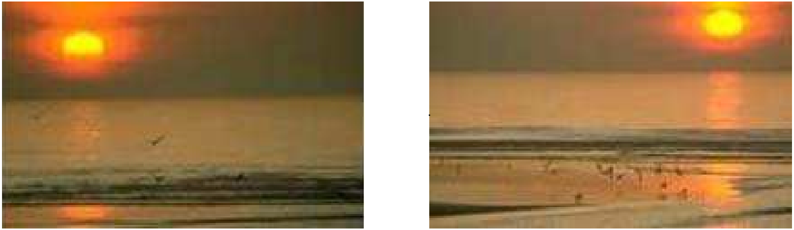
Given image I with visual features $V_i = \{v_1, v_2, \ldots, v_N\}$ and set of keywords $W = \{w_1, w_2, \ldots, w_M\}$ find subset $W_I \subset W$ which appropriately describes image I.

Model

Learn correspondence of keywords and image segments under assumption that words correspond to concepts in image.

The Corel Database

- 600 CD-ROMs, each has 100 images on same topic
- Topic is associated with keywords
- Keywords (370 in total) apply to all images in topic
- Contains many related images which share keywords



`birds, sea, sun, waves`

Modeling

Key idea: model joint probability of images and keywords based on underlying semantic concepts.

- Introduce set of latent variables $\approx$ semantic concepts
- Joint probability model describes image-word relationship based on each latent variable
- Are image features and words compatible based on concept set?

$$P(V_I, W_I) = \sum_{s \in D} P(V_I, W_I | s_i) P(s)$$

D is the number of latent variables, $P(s)$ prior probability of s

Related Work

- The co-occurrence model (Mori et al., 1999)
- Alignment model (Duygulu et al., 2002)
- LSA and PLSA models (Monay et al., 2003)
- Hierarchical latent model (Banard et al., 2002)
- Gaussian mixture model and CorrLDA (Blei and Jordan, 2003)
- Information retrieval model (Lavrenko et al., 2003)
- Many other models (Wang et al., 2002)

Issues: **scalability**, **portability**, easy to do well on Corel (Tang and Lewis, 2007), database is neither diverse nor noisy.

This Talk

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

This Talk

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

A_1: Perhaps Google is not so wrong! Exploit resources where images and their annotations co-occur naturally, but **with image and text processing**.

This Talk

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

A_1: Perhaps Google is not so wrong! Exploit resources where images and their annotations co-occur naturally, but **with image and text processing**.

Q_2: Wouldn't it be better if we generate a description for an image rather than keywords?

This Talk

Q$_1$: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

A$_1$: Perhaps Google is not so wrong! Exploit resources where images and their annotations co-occur naturally, but **with image and text processing**.

Q$_2$: Wouldn't it be better if we generate a description for an image rather than keywords?

A$_2$: Yes, it would reduce ambiguity and help with more specific queries, but we need **natural language generation** for that.

Outline

1. Introduction
   - Motivation
   - Image Annotation
2. Topic Modeling for Image Annotation
   - BBC News Database
   - Annotation Model
   - Evaluation
3. Automatic Caption Generation
   - Caption Generation Model
   - Evaluation
4. Conclusions

BBC News Database

Michelle Obama fever hits the UK
By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

GETTY IMAGES

It is reported that the Queen asked to stay in touch with Mrs Obama

BBC News Database

Michelle Obama fever hits the UK
By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

It is reported that the Queen asked to stay in touch with Mrs Obama

BBC News Database

Michelle Obama fever hits the UK

By Rajini Vaidyanathan

BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

It is reported that the Queen asked to stay in touch with Mrs Obama

BBC News Database

Michelle Obama fever hits the UK

By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

It is reported that the Queen asked to stay in touch with Mrs Obama

BBC News Database

- 3,361 news articles from the BBC News website (`http://news.bbc.co.uk/`)
- Each article has an image and caption
- Wide range of topics (e.g., politics, technology, education)
- Images: 203 pixels wide and 152 pixels high
- Avg caption length: 5.35 tokens
- Avg document length: 133.85 tokens
- Caption vocabulary: 2,167 tokens
- Document vocabulary: 6,253 tokens
- Shared vocabulary: 2,056 tokens

BBC News Database

- 3,361 news articles from the BBC News website (`http://news.bbc.co.uk/`)
- Each article has an image and caption
- Wide range of topics (e.g., politics, technology, education)
- Images: 203 pixels wide and 152 pixels high
- Avg caption length: 5.35 tokens
- Avg document length: 133.85 tokens
- Caption vocabulary: 2,167 tokens
- Document vocabulary: 6,253 tokens
- Shared vocabulary: 2,056 tokens

Other resources: **Yahoo! news, CNN news, Wikipedia.**

Problem Formulation

Image Annotation

1. **Training:** document-image-caption tuples
2. **Testing:** document-image pairs
3. **Task:** infer description keywords for image

Modeling Assumptions

1. Caption describes image content directly or indirectly.
2. We cannot annotate **all objects** present in the image.
3. Document describes the content of the image.

Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$\begin{aligned} W_I^* &= \arg\max_W P(W|I,D) \\ &= \arg\max_W \prod_{w \in W} P(w|I,D) \end{aligned}$$

Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$
\begin{aligned}
W_I^* &= \arg\max_W P(W|I, D) \\
&= \arg\max_W \prod_{w \in W} P(w|I, D)
\end{aligned}
$$

1. I and D are distinct modalities! (continuous vs. discrete).

Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$\begin{aligned} W_I^* &= \arg\max_W P(W|I, D) \\ &= \arg\max_W \prod_{w \in W} P(w|I, D) \end{aligned}$$

1. I and D are distinct modalities! (continuous vs. discrete).
2. We will represent them **jointly** as bag-of-terms.

Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

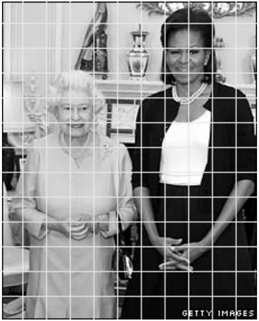
$$W_I^* = \arg\max_W P(W|I, D)$$

$$= \arg\max_W \prod_{w \in W} P(w|I, D)$$

1. I and D are distinct modalities! (continuous vs. discrete).
2. We will represent them **jointly** as bag-of-terms.
3. I and D describe common underlying concepts and are generated by mixture of latent topics.

Image Processing

Normalized cuts
(20 regions)



Uniform Grid
(11×13 regions)

SIFT point detector
(240 points)

- Obtain non-sparse feature representation.
- Use SIFT algorithm (Lowe, 1999) to compute local descriptors.
- Quantize SIFT descriptors (K-means).
- Obtain discrete set of visiterms $\approx$ visual vocabulary.

Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$\begin{aligned} W_I^* &= \arg\max_W P(W|I,D) \\ &= \arg\max_W \prod_{w \in W} P(w|I,D) \end{aligned}$$

Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$\begin{aligned} W_I^* &= \arg\max_W P(W|I, D) \\ &= \arg\max_W \prod_{w \in W} P(w|I, D) \\ &= \arg\max_W \prod_{w \in W} P(w|d_{Mix}) \end{aligned}$$

1. I and D are concatenation of textual and visual terms (d_{Mix})

Annotation Model

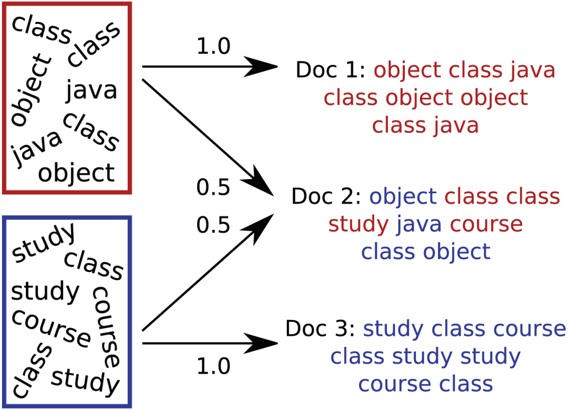
Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$\begin{aligned} W_I^* &= \arg\max_W P(W|I,D) \\ &= \arg\max_W \prod_{w \in W} P(w|I,D) \\ &= \arg\max_W \prod_{w \in W} P(w|d_{Mix}) \end{aligned}$$

1. I and D are concatenation of textual and visual terms (d_{Mix})
2. $P(w|d_{Mix})$ is multimodal word distribution over topics.

Latent Dirichlet Allocation

- Blei et al. (2003), Griffiths and Steyvers, (2002, 2003, 2004).
- Topics are mixtures of words and words are mixtures of topics.
- Infer topic information from word-document co-occurrences.



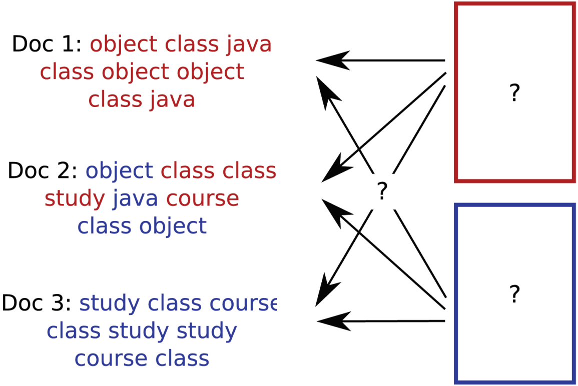
Doc 1: object class java class object object class java

Doc 2: object class class study java course class object

Doc 3: study class course class study study course class

Latent Dirichlet Allocation

- Blei et al. (2003), Griffiths and Steyvers, (2002, 2003, 2004).
- Topics are mixtures of words and words are mixtures of topics.
- Infer topic information from word-document co-occurrences.

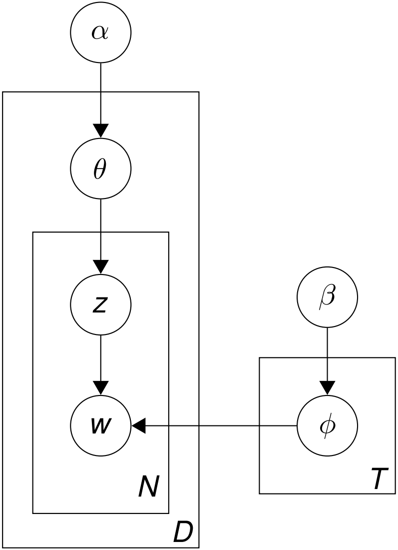


Doc 1: object class java class object object class java

Doc 2: object class class study java course class object

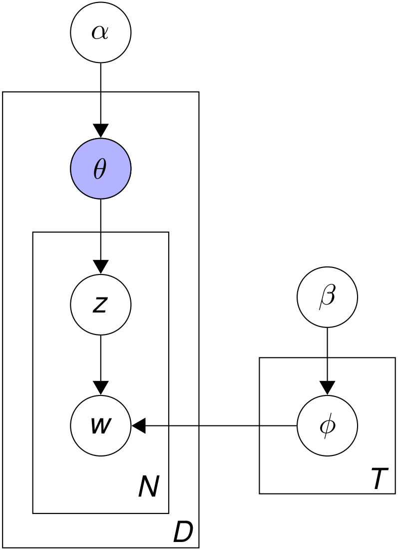
Doc 3: study class course class study study course class

Latent Dirichlet Allocation



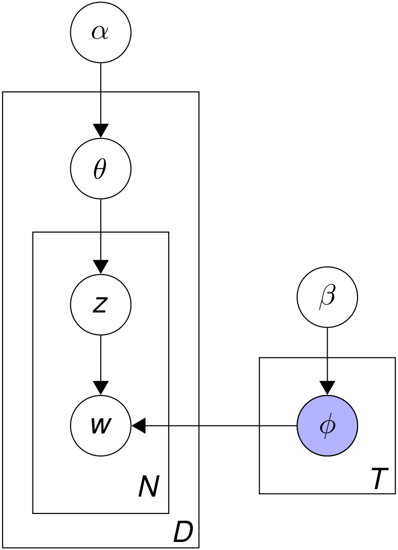
Latent Dirichlet Allocation

- For each document *d*, draw a topic mixture θ_d from $Dir(\theta_d; \alpha)$



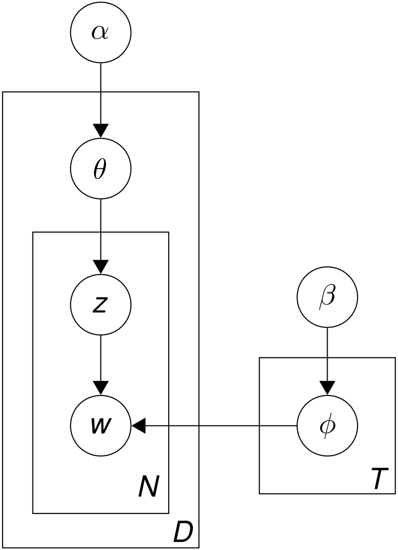
Latent Dirichlet Allocation

- For each document *d*, draw a topic mixture θ_d from $Dir(\theta_d; \alpha)$
- For each topic *t*, draw a distribution over words ϕ_t from $Dir(\phi_t; \beta)$



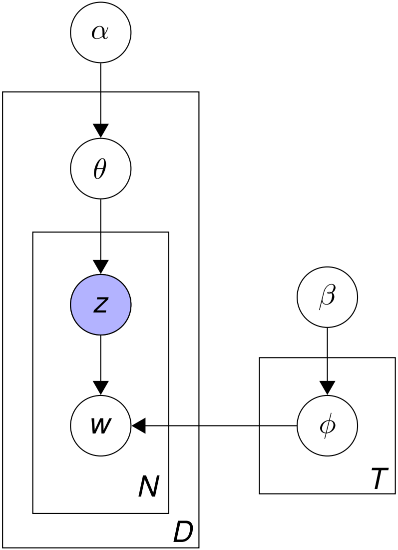
Latent Dirichlet Allocation

- For each document *d*, draw a topic mixture θ_d from $Dir(\theta_d; \alpha)$
- For each topic *t*, draw a distribution over words ϕ_t from $Dir(\phi_t; \beta)$
- For each position *i* in document *d*:



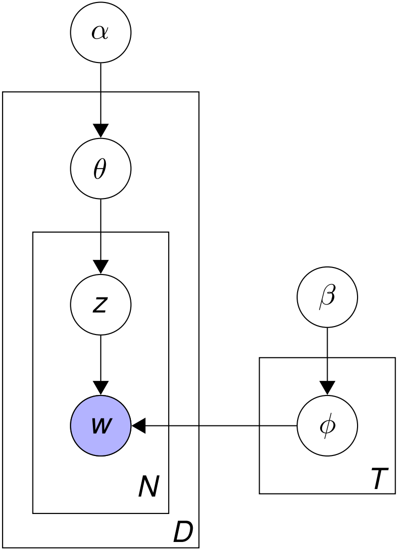
Latent Dirichlet Allocation

- For each document *d*, draw a topic mixture θ_d from $Dir(\theta_d; \alpha)$
- For each topic *t*, draw a distribution over words ϕ_t from $Dir(\phi_t; \beta)$
- For each position *i* in document *d*:
    - Draw a topic z_i from θ_d



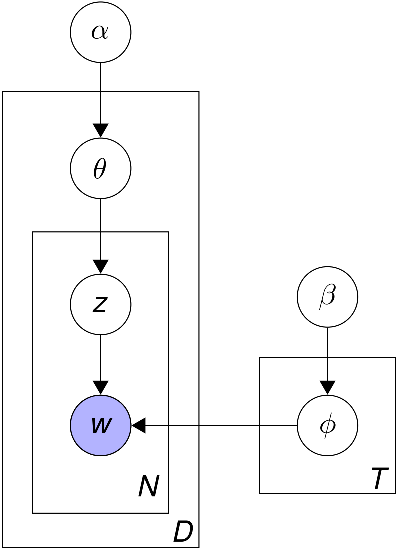
Latent Dirichlet Allocation

- For each document *d*, draw a topic mixture θ_d from $Dir(\theta_d; \alpha)$
- For each topic *t*, draw a distribution over words ϕ_t from $Dir(\phi_t; \beta)$
- For each position *i* in document *d*:
  - Draw a topic z_i from θ_d
  - Draw a word w_i from ϕ_{z_i}



Latent Dirichlet Allocation

- For each document *d*, draw a topic mixture θ_d from $Dir(\theta_d; \alpha)$
- For each topic *t*, draw a distribution over words ϕ_t from $Dir(\phi_t; \beta)$
- For each position *i* in document *d*:
  - Draw a topic z_i from θ_d
  - Draw a word w_i from ϕ_{z_i}
- *w* is either a visual or textual word



Annotation Model

Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

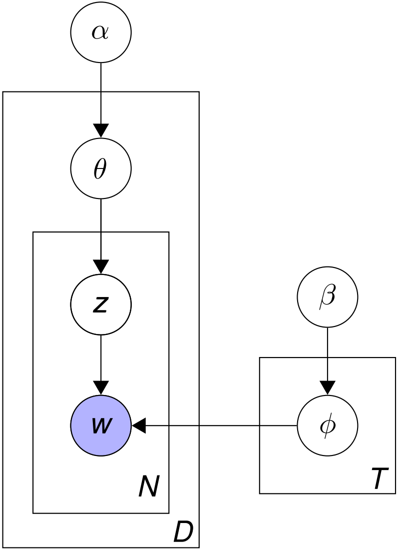
$$\begin{aligned} W_I^* &= \arg\max_W P(W|I, D) \\ &= \arg\max_W \prod_{w \in W} P(w|I, D) \\ &\approx \arg\max_W \prod_{w \in W} P(w|d_{Mix}) \end{aligned}$$

Annotation Model

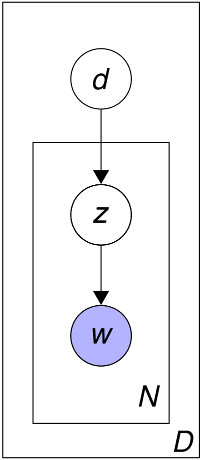
Given image I, keywords W, and document D, find subset W_I ($W_I \subseteq W$) which describes I **and** D.

$$\begin{aligned}
W_I^* &= \arg\max_W P(W|I,D) \\
&= \arg\max_W \prod_{w \in W} P(w|I,D) \\
&\approx \arg\max_W \prod_{w \in W} P(w|d_{Mix}) \\
&\approx \arg\max_W \prod_{w \in W} \sum_{k=1}^{K} P(w|z_k) P(z_k|d_{Mix})
\end{aligned}$$

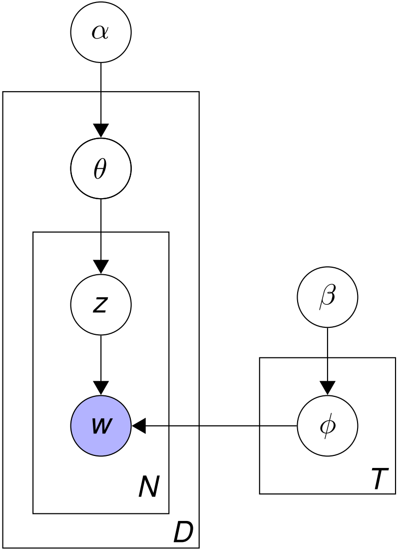
Topic Models



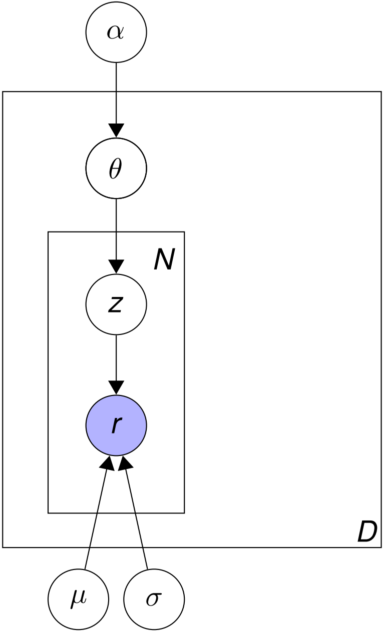
Topic Models: PLSA



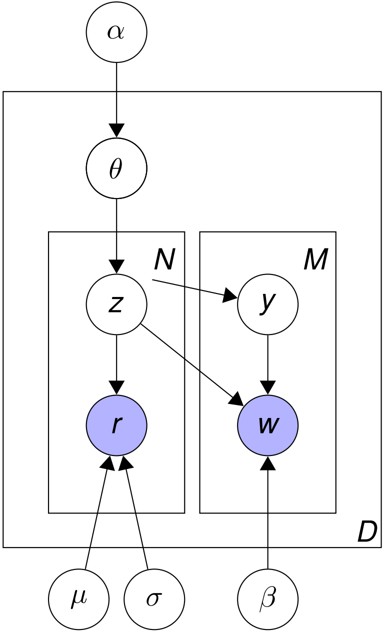
Topic Models



Topic Models: Correspondence LDA



Topic Models: Correspondence LDA



Experimental Setup

Preprocessing

- POS-tag and lemmatize database, nouns, adjectives, verbs.
- extract on average 150 SIFT features per image
- 1,000 topics and 750 visual terms

Model Comparisons

- Vanilla LDA model without images
- PLSA-based model (Monay and Gatica-Perez, 2007)
- Correspondence LDA (Blei and Jordan, 2003)
- Extension of Relevance model (Lavrenko et al. 2003, Feng and Lapata, 2008)

Evaluation

- consider *m*-best words as annotations for image *I*
- precision, recall, F1 against caption words

Results

Model	Top 10		
	Precision	Recall	F1
CorrLDA	5.33	11.80	7.36
TxtLDA	7.30	16.90	10.20
PLSA	10.26	22.60	14.12
ExtRel	14.70	27.90	19.80
MixLDA	16.30	33.10	21.60

- All differences between models statistically significant.
- CorrLDA worst performing model, MixLDA best performing.
- Visual information generally improves performance.
- Results in the same ballpark with Corel-based models.

Results

Model	Top 10		
	Precision	Recall	F1
CorrLDA	5.33	11.80	7.36
TxtLDA	7.30	16.90	10.20
PLSA	10.26	22.60	14.12
ExtRel	14.70	27.90	19.80
MixLDA	16.30	33.10	21.60

- All differences between models statistically significant.
- CorrLDA worst performing model, MixLDA best performing.
- Visual information generally improves performance.
- Results in the same ballpark with Corel-based models.

Results

Model	Top 10		
	Precision	Recall	F1
CorrLDA	5.33	11.80	7.36
TxtLDA	7.30	16.90	10.20
PLSA	10.26	22.60	14.12
ExtRel	14.70	27.90	19.80
MixLDA	16.30	33.10	21.60

- All differences between models statistically significant.
- CorrLDA worst performing model, MixLDA best performing.
- Visual information generally improves performance.
- Results in the same ballpark with Corel-based models.

Example Output

TxtLDA	Afghanistan, Taleban, soldier, British, zone, kill, force, Microsoft, **troop**, NATO
MixLDA	Afghanistan, **troop**, Blair, British, NATO, **helicopter**, soldier, support, **operation**, commander
Caption	Troops need more Chinook helicopters to carry out operations

Example Output

TxtLDA	police, Burgess, time, letter, **crash**, case, death, operation, investigation, jail
MixLDA	**Diana**, police, case, **crash**, **Princess**, report, **death**, inquest, **Paris**, Burgess
Caption	Princess Diana died in a car crash in Paris in 1997

Automatic Caption Generation

- Keywords are ambiguous (`car, blue, sky`)
- Caption makes relations between objects explicit.
- Increase accessibility of web for visually impaired.

Automatic Caption Generation

- Keywords are ambiguous (`car, blue, sky`)
- Caption makes relations between objects explicit.
- Increase accessibility of web for visually impaired.

Automatic Caption Generation

- Keywords are ambiguous (`car, blue, sky`)
- Caption makes relations between objects explicit.
- Increase accessibility of web for visually impaired.

www.shutterstock.com · 37917220

Automatic Caption Generation

- Keywords are ambiguous (`car, blue, sky`)
- Caption makes relations between objects explicit.
- Increase accessibility of web for visually impaired.

Automatic Caption Generation

- Keywords are ambiguous (`car, blue, sky`)
- Caption makes relations between objects explicit.
- Increase accessibility of web for visually impaired.
- **Assist journalists in caption creation.**

Automatic Caption Generation

- Task is challenging, even for humans!
- Captions most commonly read in article together with title, lead and section headings.
- A good caption must be succinct and informative.
- Identify the subject of the picture.
- Establish the picture's relevance to the article.
- Provide context for the picture.
- Draw the reader into the article.
- Journalists rely on general world knowledge beyond document.

Extractive Summarization

Michelle Obama fever hits the UK
By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas’ arrival, Mrs Obama’s star appeal was apparent.

Extractive Summarization

Michelle Obama fever hits the UK
By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.


GETTY IMAGES

Extractive Summarization

Michelle Obama fever hits the UK
By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

Extractive Summarization

Michelle Obama fever hits the UK
By Rajini Vaidyanathan
BBC NEWS

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

It is reported that the Queen asked to stay in touch with Mrs Obama

Abstractive Summarization

Caveats with Extracts

- Extracted sentences are grammatical but long
- Neither concise, nor catchy as human captions
- The caption should describe the image's content
- But often no single document sentence can do that.

Create Abstracts

- Generate a new sentence as a caption
- Use visual information (output of image annotation model)
- Content selection and surface realization

Caption Generation Model

$P(w_1, w_2, ..., w_n) =$

$$\prod_{i=1}^{n} P(w_i \in C | I, D) \quad \text{image annotation probability}$$

$$\cdot P(len(C) = n) \quad \text{caption length distribution}$$

$$\cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \quad \text{trigram language model}$$

Caption Generation Model

$P(w_1, w_2, ..., w_n) =$

$$\prod_{i=1}^{n} P(w_i \in C | I, D) \qquad \text{image annotation probability}$$

$$\cdot P(len(C) = n) \qquad \text{caption length distribution}$$

$$\cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \qquad \text{trigram language model}$$

Caption Generation Model

$P(w_1, w_2, ..., w_n) =$

$$\prod_{i=1}^{n} P(w_i \in C | I, D) \qquad \text{image annotation probability}$$

$$\cdot P(len(C) = n) \qquad \text{caption length distribution}$$

$$\cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \qquad \text{trigram language model}$$

Caption Generation Model

$P(w_1, w_2, ..., w_n) =$

$$\prod_{i=1}^{n} P(w_i \in C|I, D)$$ image annotation probability

$$\cdot P(len(C) = n)$$ caption length distribution

$$\cdot \prod_{i=3}^{n} P(w_i|w_{i-1}, w_{i-2})$$ trigram language model

Caption Generation Model

$P(w_1, w_2, ..., w_n) =$

$$\prod_{i=1}^{n} P(w_i \in C | I, D) \qquad \text{image annotation probability}$$
$$\cdot P(len(C) = n) \qquad \text{caption length distribution}$$
$$\cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \qquad \text{trigram language model}$$

- Adapted from Banko et al. (2000).
- This model will not output any function words.
- Generated caption will be incoherent.

Caption Generation Model

$$P(w_1, w_2, ..., w_n) =$$

$$\prod_{i=1}^{n} P(w_i \in C | I, D) \qquad \text{image annotation probability}$$

$$\cdot P(len(C) = n) \qquad \text{caption length distribution}$$

$$\cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \qquad \text{trigram language model}$$

- Adapted from Banko et al. (2000).
- This model will not output any function words.
- Generated caption will be incoherent.
- **Consider image in surface realization.**

Caption Generation Model

$P(w_1, w_2, ..., w_n) =$

$$\prod_{i=1}^{n} P(w_i \in C | w_i \in D)$$ caption generation probability

$$\cdot P(len(C) = n)$$ caption length distribution

$$\cdot \prod_{i=3}^{n} P_{adap}(w_i | w_{i-1}, w_{i-2})$$ adapted trigram model

- This model will output function words.
- Generated caption will be less incoherent.
- Considers image in surface realization.

Caption Generation Model

$$P(w_1, w_2, ..., w_n) =$$

$$\begin{aligned}
&\prod_{i=1}^{n} P(w_i \in C | w_i \in D) && \text{caption generation probability} \\
&\cdot P(len(C) = n) && \text{caption length distribution} \\
&\cdot \prod_{i=3}^{n} P_{adap}(w_i | w_{i-1}, w_{i-2}) && \text{adapted trigram model}
\end{aligned}$$

- This model will output function words.
- Generated caption will be less incoherent.
- Considers image in surface realization.

Caption Generation Model

$$P(w_1, w_2, ..., w_n) =$$

$$\prod_{i=1}^{n} P(w_i \in C | w_i \in D) \qquad \text{caption generation probability}$$

$$\cdot P(len(C) = n) \qquad \text{caption length distribution}$$

$$\cdot \prod_{i=3}^{n} P_{adap}(w_i | w_{i-1}, w_{i-2}) \qquad \text{adapted trigram model}$$

- This model will output function words.
- Generated caption will be less incoherent.
- Considers image in surface realization.

Caption Generation Model

$$P(w_1, w_2, ..., w_n) =$$

$$\begin{aligned}
&\prod_{i=1}^{n} P(w_i \in C | w_i \in D) && \text{caption generation probability} \\
&\cdot P(len(C) = n) && \text{caption length distribution} \\
&\cdot \prod_{i=3}^{n} P_{adap}(w_i | w_{i-1}, w_{i-2}) && \text{adapted trigram model}
\end{aligned}$$

- This model will output function words.
- Generated caption will be less incoherent.
- Considers image in surface realization.
- **But phrases could capture long-range dependencies.**

Caption Generation Model

$$P(\rho_1, \rho_2, ..., \rho_m) =$$

$$\prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in D)$$

caption generation probability

$$\cdot P(len(C) = \sum_{j=1}^{m} len(\rho_j))$$

caption length distribution

$$\cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1})$$

attachment constraints

$$\cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2})$$

adapted trigram model

Caption Generation Model

$$P(\rho_1, \rho_2, ..., \rho_m) =$$

$$\begin{aligned}
&\prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in D) && \text{caption generation probability} \\
&\cdot P(len(C) = \textstyle\sum_{j=1}^{m} len(\rho_j)) && \text{caption length distribution} \\
&\cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1}) && \text{attachment constraints} \\
&\cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2}) && \text{adapted trigram model}
\end{aligned}$$

Caption Generation Model

$$P(\rho_1, \rho_2, ..., \rho_m) =$$

$$\prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in D)$$ caption generation probability

$$\cdot P(len(C) = \textstyle\sum_{j=1}^{m} len(\rho_j))$$ caption length distribution

$$\cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1})$$ attachment constraints

$$\cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2})$$ adapted trigram model

Caption Generation Model

$$P(\rho_1, \rho_2, ..., \rho_m) =$$

$$\begin{aligned}
&\prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in D) && \text{caption generation probability} \\
&\cdot P(len(C) = \textstyle\sum_{j=1}^{m} len(\rho_j)) && \text{caption length distribution} \\
&\cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1}) && \text{attachment constraints} \\
&\cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2}) && \text{adapted trigram model}
\end{aligned}$$

Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).

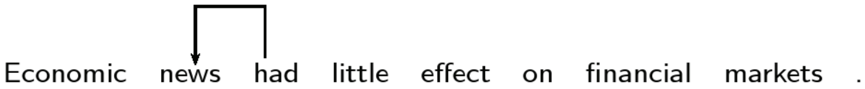
Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).

Economic news had little effect on financial markets .

Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).



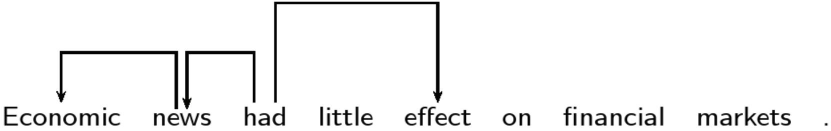
Economic news had little effect on financial markets .

Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).

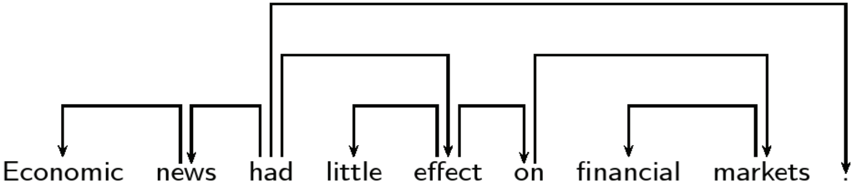
Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).



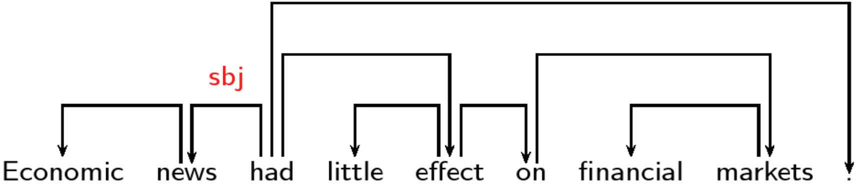
Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).



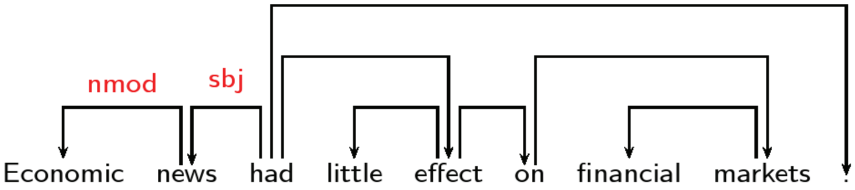
Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).



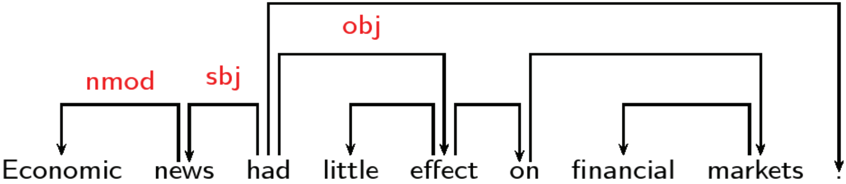
Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).



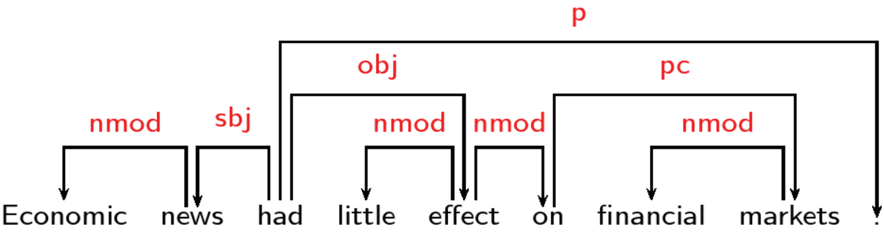
Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).



Dependency Structure

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies.
- A phrase is a head and its dependent(s).
- *economic news*, *on financial markets*, *had effect*.



Experimental Setup

Preprocessing

- Annotation keywords are nouns, adjectives, verbs.
- Extract on average 150 SIFT features per image
- 1,000 topics and 750 visual terms, 15 keywords

Model Comparisons

- Extract lead sentence, using KL Divergence
- Word-based, phrase-based abstractive models

Evaluation

- $\text{TER}(E, E_r) = \frac{\text{Ins}+\text{Del}+\text{Sub}+\text{Shft}}{N_r}$ (Snover et al., 2006)
- judgment elicitation study (grammaticality, relevance)

Results

Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Results

Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Results

Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Results

Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Results

Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Results

Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Results

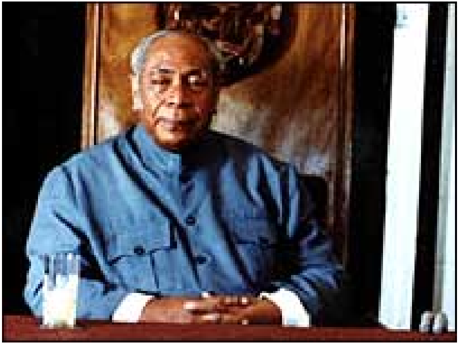
Model	TER	AvgL
LeadS	2.12	21.0
KLDiv	1.77	18.4
Words	1.11	10.0
Phrases	1.06	10.1

- LeadS sig worse than KLDiv
- KLDiv takes visual info into account
- Abstractive models seem better

Model	Gram	Relv
KLDiv	6.42	4.10
Words	2.08	3.20
Phrases	4.80	4.96
Gold	6.39	5.55

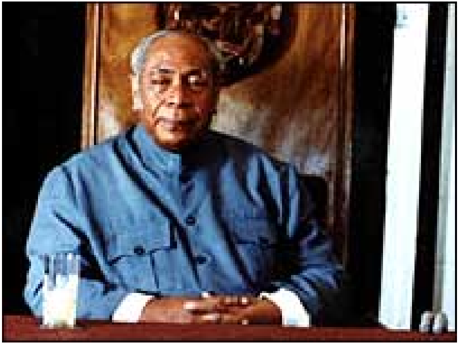
- KLDiv most grammatical model
- Words model least grammatical
- Phrases best model wrt relevance
- And as good as gold standard

Example Output



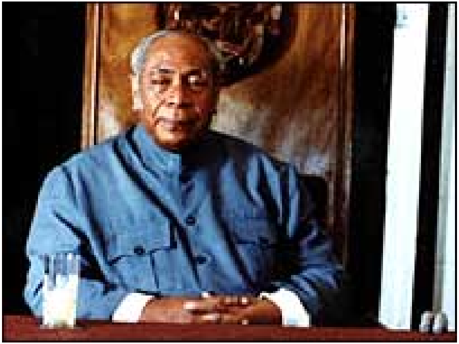
G **King Tupou, who was 88, died a week ago.**

Example Output



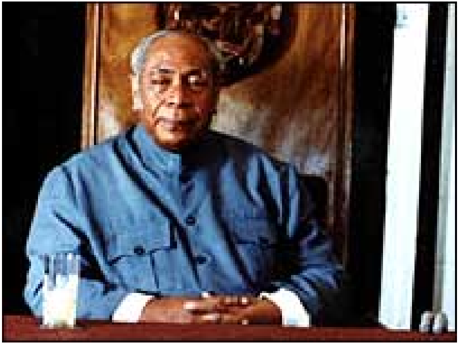
G	**King Tupou, who was 88, died a week ago.**
KL	Last year, thousands of Tongans took part in unprecedented demonstrations to demand greater democracy and public ownership of key national assets.

Example Output



G	**King Tupou, who was 88, died a week ago.**
KL	Last year, thousands of Tongans took part in unprecedented demonstrations to demand greater democracy and public ownership of key national assets.
A_W	King Toupou IV died at the age of Tongans last week.

Example Output



G	**King Tupou, who was 88, died a week ago.**
KL	Last year, thousands of Tongans took part in unprecedented demonstrations to demand greater democracy and public ownership of key national assets.
A_W	King Toupou IV died at the age of Tongans last week.
A_P	King Toupou IV died at the age of 88 last week.

Example Output

G **Children were found to be far more internet-wise than parents.**

Example Output

G **Children were found to be far more internet-wise than parents.**
KL That’s where parents come in.

Example Output

G	**Children were found to be far more internet-wise than parents.**
KL	That's where parents come in.
A_W	The survey found a third of children are about mobile phones.

Example Output

G **Children were found to be far more internet-wise than parents.**

KL That’s where parents come in.

A_W The survey found a third of children are about mobile phones.

A_P The survey found a third of children in the driving seat.

Conclusions

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

Conclusions

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

A_1: Yes, need better image processing and perhaps some form of supervision!

Conclusions

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

A_1: Yes, need better image processing and perhaps some form of supervision!

Q_2: Wouldn't it be better if we generate a description for an image rather than keywords?

Conclusions

Q_1: Can we relieve the data acquisition bottleneck associated with image annotation and scale the task onto real-world images and noisy data?

A_1: Yes, need better image processing and perhaps some form of supervision!

Q_2: Wouldn't it be better if we generate a description for an image rather than keywords?

A_2: Yes, task is feasible, need more NLP, and a joint image annotation and caption generation model!