Language Modeling for Information Retrieval

Ananthakrishnan R
(anand@cse)

Outline of the Seminar

- What is Language Modeling?
- Language Modeling in NLP
    - Speech Recognition, Statistical MT, Classification, IR …
- The Language Modeling approach to IR
  – the conceptual model
  – smoothing
  – semantic smoothing
    - IR as Statistical Machine Translation

Language Modeling

- A ***language model*** is a probabilistic mechanism for generating text

- Language models estimate the probability distribution of various natural language phenomena
  - sentences, utterances, queries …

Language Modeling Techniques

- N-grams
- Class-based N-grams
- Probabilistic CFGs
- Decision Trees

Claude Shannon *was probably the first language modeler*

> *"How well can we predict or compress natural language text through simple n-gram models?"*

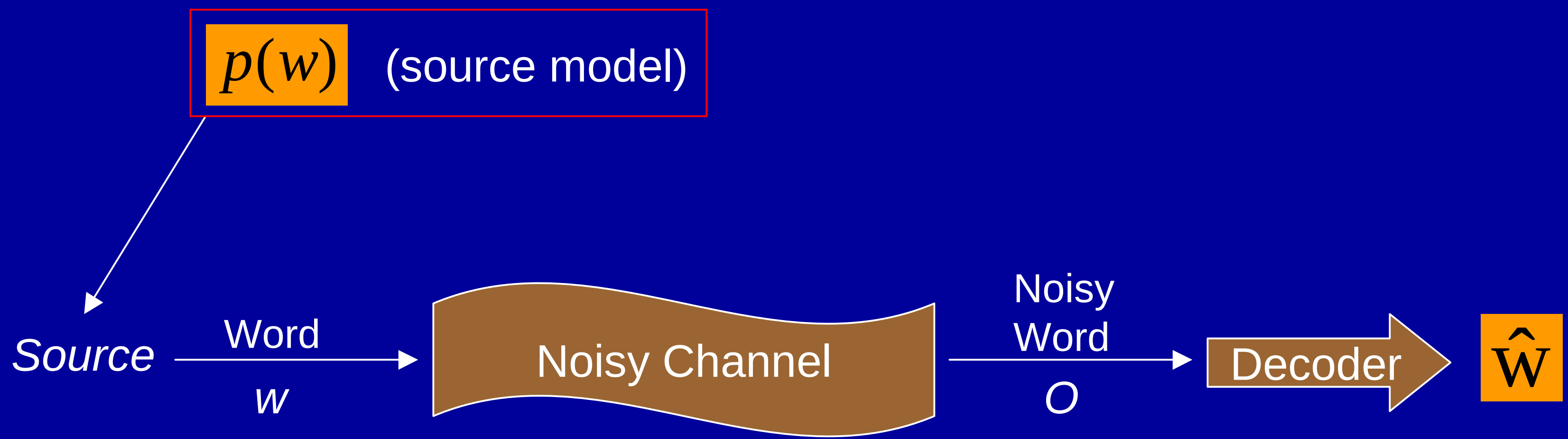
Language Modeling for Other NLP Tasks

Applications

- Speech Recognition
- Statistical Machine Translation
- Classification
- Spell Checking

*Play the role of either the **prior** or the **likelihood***

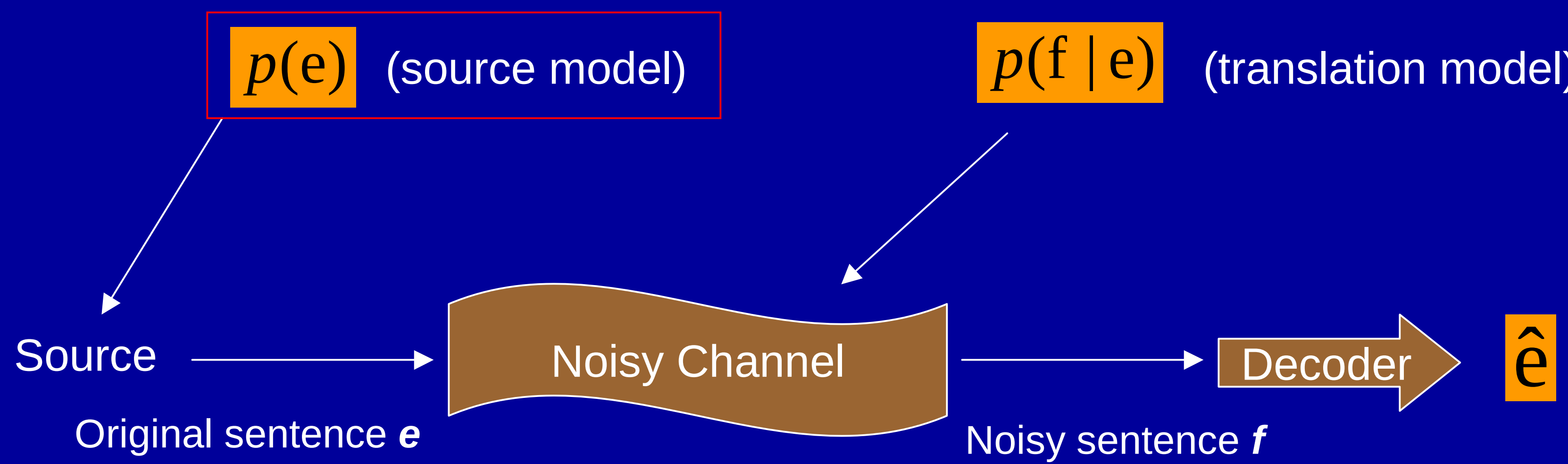
Language Models in *Speech Recognition*



$$\hat{w} = \arg\max_{w \in Vocabulary} p(w \mid O)$$

$$= \arg\max_{w \in Vocabulary} p(O \mid w)\, p(w)$$

Language Models in *Statistical Machine Translation*



$$
\begin{aligned}
\hat{\mathrm{e}} &= \arg\max_{\mathrm{e}\in E} p(\mathrm{e}\,|\,\mathrm{f}) \\
&= \arg\max_{\mathrm{e}\in E} p(\mathrm{f}\,|\,\mathrm{e})\,p(\mathrm{e})
\end{aligned}
$$

Language Models in *Text Classification*

$$\hat{c} = \arg\max_{c \in C} p(c \mid d)$$
$$= \arg\max_{c \in C} p(d \mid c)\, p(c)$$

Bayesian Classifier

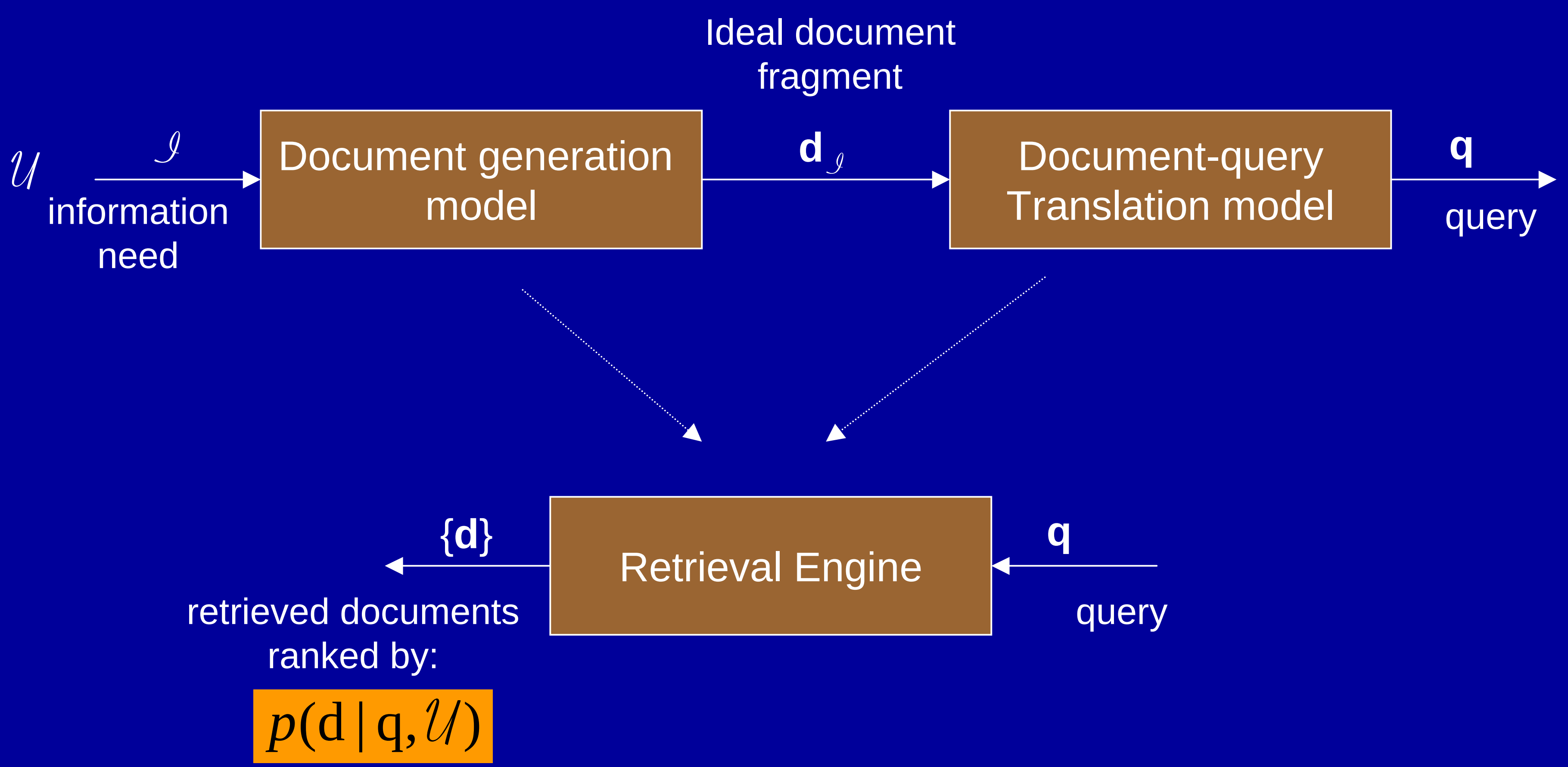
a language model for each class

The Language Modeling Approach to Information Retrieval

The Conceptual Model of IR

- The user has an information need $\mathcal{I}$
- From this need he generates an ideal document fragment $\mathbf{d}_{\mathcal{I}}$
- He selects a set of key terms from $\mathbf{d}_{\mathcal{I}}$, and generates a query **q** from this set


Information need $\mathcal{I}$
Analysis of Australia's defeat to England in the Ashes. Reason for defeat: overconfidence?
Ideal document $\mathbf{d}_{\mathcal{I}}$
Eng bt. Aus: **Analysis**
Australia's **defeat** to England in the **Ashes** is largely due to their overconfidence. They don't seem to have figured out that they are facing a far superior team …
query
australia ashes defeat analysis
q
The **Ashes** Review
Brash batting by **Australia** has lost them a test match. Their opposition is not Zimbabwe, and not many of the Aussies seem to remember that! …
d
A candidate document


$\mathcal{U}$
$\mathcal{I}$
information need
Document generation model
Ideal document fragment
$\mathbf{d}_{\mathcal{I}}$
Document-query Translation model
q
query
{**d**}
Retrieval Engine
q
query
retrieved documents ranked by:
$p(\mathrm{d} \mid \mathrm{q}, \mathcal{U})$

The Language Modeling Approach

Using Bayes' law,

$$p(\mathrm{d} \mid \mathrm{q}, \mathcal{U}) = \frac{p(\mathrm{q} \mid \mathrm{d}, \mathcal{U})\, p(\mathrm{d} \mid \mathcal{U})}{p(\mathrm{q} \mid \mathcal{U})}$$

For the purpose of ranking,

$$p_{\mathrm{q}}(\mathrm{d}) = p(\mathrm{q} \mid \mathrm{d}, \mathcal{U})\, p(\mathrm{d} \mid \mathcal{U})$$

$$p_{\mathrm{q}}(\mathrm{d}) = p(\mathrm{q} \mid \mathrm{d})\, p(\mathrm{d})$$ (user independent)

Language Modeling for IR:
Using document language models to assign likelihood scores to queries
(Ponte and Croft, 1998)

query-likelihood: comes from the **language model**

$$p_q(d) = \overbrace{p(q \mid d)}^{\text{query-likelihood}}\,\underbrace{p(d)}_{\text{document prior}}$$

document prior
(query-independent)

E.g., PageRank

Information need $\mathcal{I}$

Analysis of Australia's defeat to England in the Ashes. Reason for defeat: overconfidence?

*Ideal document **d** $_{\mathcal{I}}$*

<u>Eng bt. Aus: **Analysis**</u>
Australia's **defeat** to England in the **Ashes** is largely due to their overconfidence. They don't seem to have figured out that they are facing a far superior team …

query

australia ashes defeat analysis

q

<u>The **Ashes** Review</u>
Brash batting by **Australia** has lost them a test match. Their opposition is not Zimbabwe, and not many of the Aussies seem to remember that! …

d

$$p(\mathrm{q} \mid \mathrm{d}, \mathcal{U}) = ?$$

A candidate document

The *Language Model*

- For each document **d** in the collection C, for each term *w,* estimate:

$$p(\mathrm{w}\mid\mathrm{d})$$

- To find the query likelihood, assume query terms to be independent:

$$p(\mathrm{q}\mid\mathrm{d})=\prod_{i=1}^{m}p(q_i\mid\mathrm{d})$$

The Need for Smoothing

- To eliminate “zero” probabilities
  - Correct the error in the MLE due to the problem of *data sparsity*

q

australia ashes defeat analysis

d

<u>The **Ashes Review**</u>
Brash batting by **Australia** has **lost** them a test match. Their opposition is not Zimbabwe, and not many of the Aussies seem to remember that! …

$p(\text{defeat}|\mathbf{d}) = 0$

$p(\mathbf{q}|\mathbf{d}) = 0$

Smoothing

- Jelinek-Mercer Method: Linear interpolation of the **maximum likelihood model** with the **collection model**

$$p_{\lambda}(\mathrm{w} \mid \mathrm{d}) = (1-\lambda)\, p(\mathrm{w} \mid \mathrm{d}) + \lambda p(\mathrm{w} \mid \mathcal{C})$$

- (Zhai & Lafferty, 2003) compares **Jelinek-Mercer**, **Dirichlet** and the **absolute discounting** methods for Language Models applied to IR.

The Need for Semantic Smoothing

Analysis of Australia's defeat to England in the Ashes. Reason for defeat: overconfidence?

Information need $\mathcal{I}$

Ideal document $\mathbf{d}_{\mathcal{I}}$

Eng bt. Aus: **Analysis**

Australia's **defeat** to **England** in the **Ashes** is largely due to their **overconfidence**. They don't seem to have figured out that they are **facing** a far **superior team** …

query

australia ashes defeat overconfidence

q

The **Ashes Review**

Brash batting by **Australia** has **lost** them a **test match**. Their **opposition** is not Zimbabwe, and not many of the **Aussies** seem to remember that! …

d

A candidate document

Document-Query Translation Model

<u>Eng bt. Aus: **Analysis**</u>
Australia’s **defeat** to **England** in the **Ashes** is largely due to their **overconfidence**. They don’t seem to have figured out that they are **facing** a far **superior team** …

Ideal document $\mathbf{d}_{\mathcal{I}}$

$p(\mathrm{q} \mid \mathrm{d}) = ?$

Must be sophisticated enough to handle polysemy and synonymy

Document-query Translation model

query

australia ashes defeat overconfidence

Semantic Smoothing

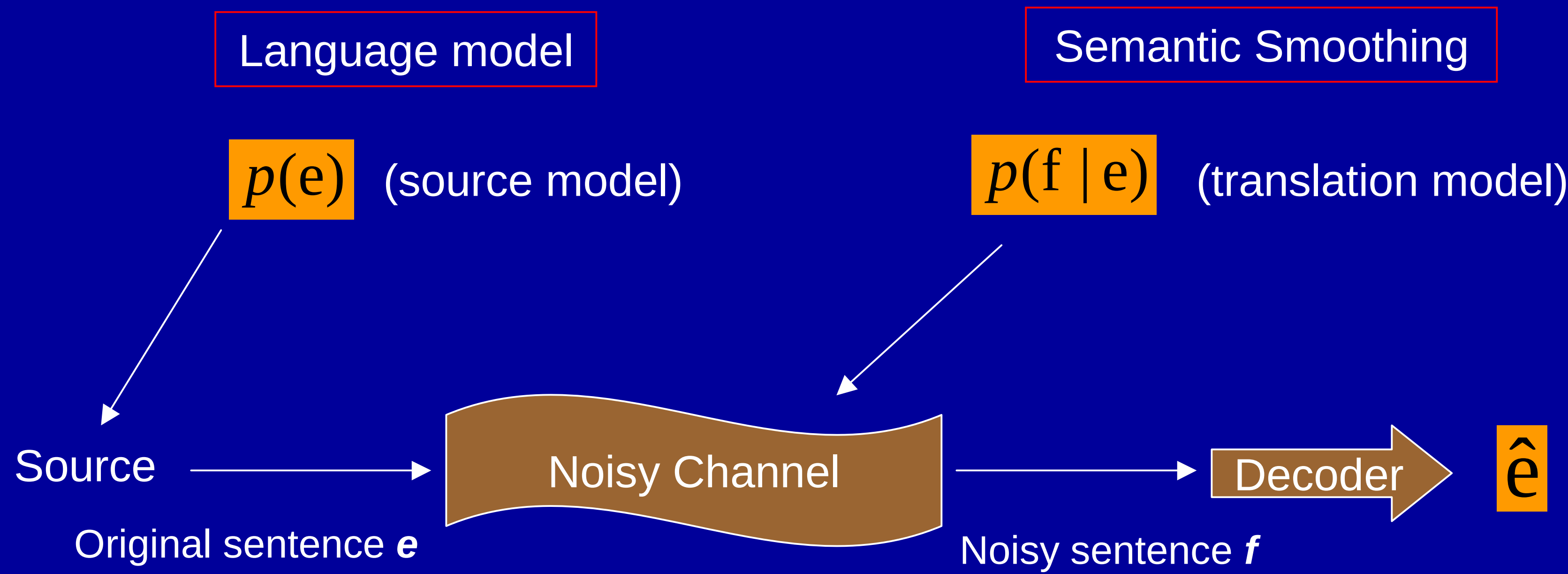
Estimate translation probabilities $\boldsymbol{t}(w_q|w_d)$ for mapping a document term w_d to a query term w_q

$$p(q \mid d) = \prod_{i=1}^{m} \sum_{w} t(q_i \mid w)\, p(w \mid d)$$

(Berger and Lafferty, 1999)
(Lafferty and Zhai, 2001)

$p(\mathit{analysis} \mid d)$ $\Rightarrow$ $p(\mathit{analysis} \mid \text{review})\, p(\mathit{review} \mid d)$

Compare with *The Statistical MT model*



$$\hat{\mathrm{e}} = \arg\max_{\mathrm{e}\in E} p(\mathrm{e}\,|\,\mathrm{f}) = \arg\max_{\mathrm{e}\in E} p(\mathrm{f}\,|\,\mathrm{e})\,p(\mathrm{e})$$

Language Modeling *vis-à-vis* *the* 'traditional probabilistic model'

$$p(r \mid \mathrm{q}, \mathrm{d}) = ?$$

The Probability Ranking Principle
(Robertson, 1977)

$$log\ \frac{p(\mathrm{d} \mid \mathrm{q}, r)}{p(\mathrm{d} \mid \mathrm{q}, \bar{r})}$$

The Robertson-Sparck Jones Model
(Sparck Jones et al., 2000)

Language Modeling approach:
- *Where's the relevance?*
- *Is the ranking optimal?*

(Lafferty & Zhai, 2002)
(Laverenko and Croft, 2001)

Discussion: Advantages of the Language Modeling Approach

- Conditioning on **d** provides a larger 'foothold' for estimation
- $p(\mathbf{d})$: an explicit notion of the importance of a document
- Document normalization is not an issue

"*When designing a statistical model … the most natural route is to apply a generative model which builds up the output step-by-step. The source channel perspective suggests a different approach: turn the search problem around to predict the input. In speech recognition, NLP, and MT, researchers have time and again found that predicting what is known from competing hypothesis can be easier than directly predicting all of the hypothesis*"

Discussion: Disadvantage of the Language Modeling Approach

- The Robertson-Sparck Jones model can directly use relevance judgements to improve the estimation of $p(A_i\,|\mathbf{q},r)$

  This is not possible in the Language Modeling approach

References

- (Robertson, 1977) The probability ranking principle in IR
- (Sparck Jones et al., 2000) A probabilistic model of information retrieval: development and comparative experiments.
- (Ponte and Croft, 1998) A language modeling approach to information retrieval
- (Zhai and Lafferty, 2001) A study of smoothing methods for language models applied to ad hoc information retrieval

References …

- (Berger and Lafferty, 1999) Information retrieval as statistical translation
- (Lafferty and Zhai, 2001) Risk minimization and language modeling in information retrieval
- (Lavrenko and Croft, 2001) Relevance-based language models
- (Lafferty and Zhai, 2001) Probabilistic IR models based on document and query generation