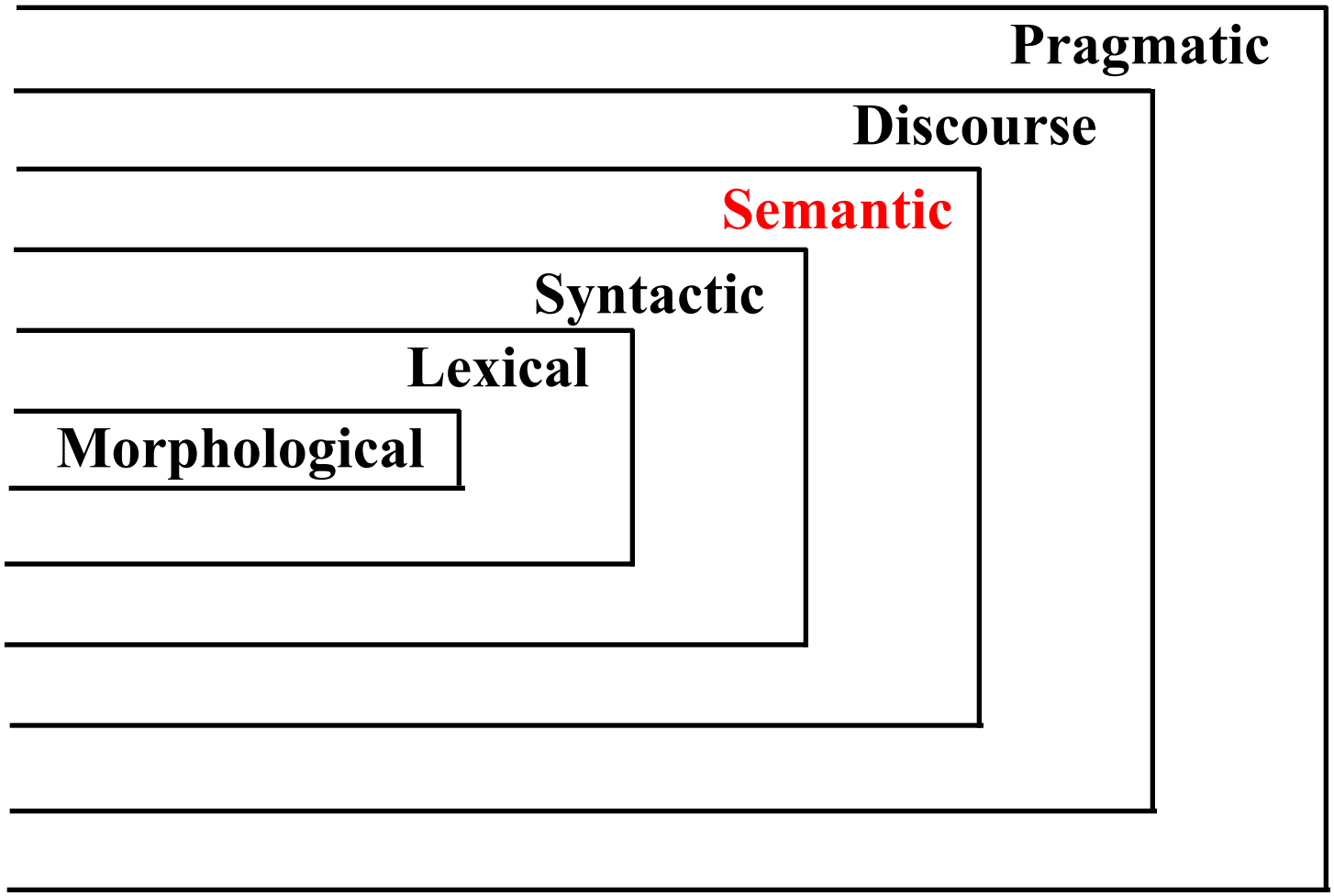
Semantics:
Topic Models (LDA)

This topic is treated in more detail in
IST 736 Text Mining

Synchronic Model of Language



Topic Models

- Topic Modeling takes a corpus and looks for sets of similar words across the collection (the topics) in such a way that each document can be represented by one or more topics
- Widely used algorithms are based on LDA (Latent Dirichlet Allocation)
  - Goes beyond unigram models, where word presence or frequency is considered independently of other words
  - Looks at intra-document statistical structure that can reveal a model of word co-occurrence within the document (similar to the kind of co-occurrence used for mutual information)
    - Blei, Ng, Jordan "Latent Dirichlet Allocation", Journal of Machine Learning Research, 2003.

Topic Models

- Topic model
  - Each topic is a collection of words together with the probabilities that they are used in that particular topic
- Generative model for documents
  - A document is generated as a mixture of topics, with words used according to the probability distribution of the topic and to the percentage that the topic contributes to the mixture
- Problem of statistical inference
  - Given the words of a document, find the topic model that is most likely to have generated those words
    - Find the probability distribution of words in each topic
    - Find the distribution of topics over the document
- Requires significant amounts of data to work well
  - Not successful on short documents such as tweets

Documents as mixture of topics

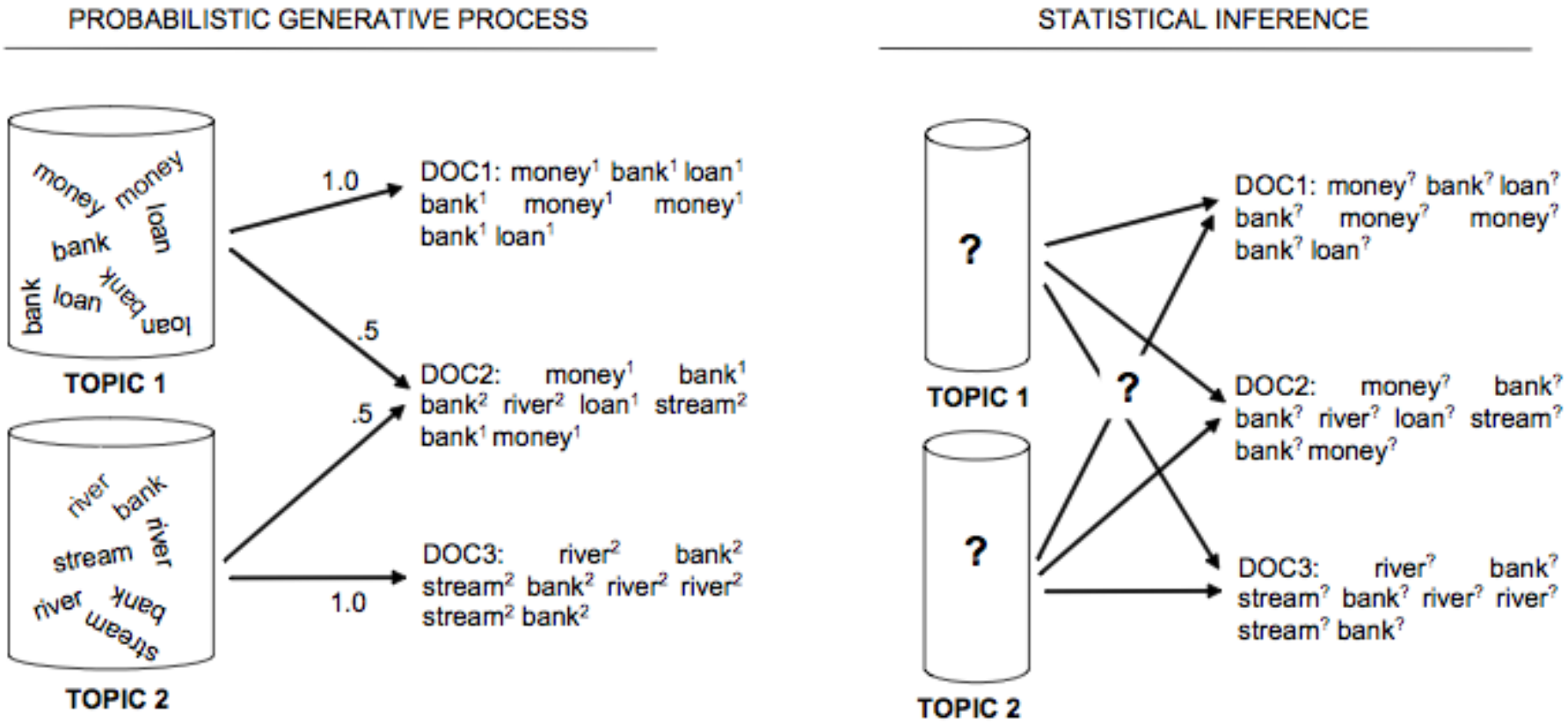


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

From Steyvers and Griffiths "Probabilistic Topic Models"

LDA applications

- LDA software infers topic models from a document collection
- Typically, you must specify how many topics you want (and possibly experiment to get satisfactory ones)
  - Also may be necessary to tune parameters alpha and beta
- Topics are represented as lists of words, sometimes with the probabilities of the topic model
  - Note that in the example on the next slide, humans added the overall "topic" heading at the top

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

From Blei et al

Software: Mallet

- Topic modeling software from CMU
- Tutorial on using Mallet:
  - The Programming Historian: Getting Started with Topic Modeling and Mallet
  - https://programminghistorian.org/lessons/topic-modeling-and-mallet
- Software also available from:
    - Blei
    - Stanford