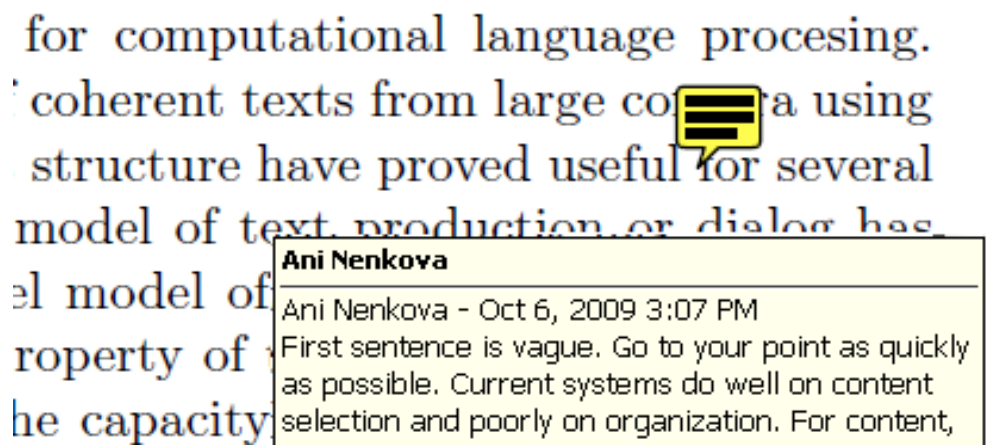
Document-level Text Quality: Models for Organization and Reader Interest

Annie Louis
October 16, 2015
Université Catholique de Louvain

Joint work with Ani Nenkova

People spontaneously respond to differences in writing



http://www.publishersweekly.com

The Top 10 Most Difficult Books

"*Finnegans Wake* is long, dense, and linguistically knotty, yet hugely rewarding, if you're willing to learn how to read it..."

http://www.cnn.com

“My Faith: Why I don't sing the 'Star Spangled Banner’ ”

“What a poorly written article. Strays off topic and hardly even addresses the point of the article.

The only brief mention of why they don’t play the national anthem is that they believe in church and state. This just was one long rant about his religion.”

http://www.vocabula.com

The Best Words

hubris (HYOO-bris) — excessive pride or self-confidence; arrogance.

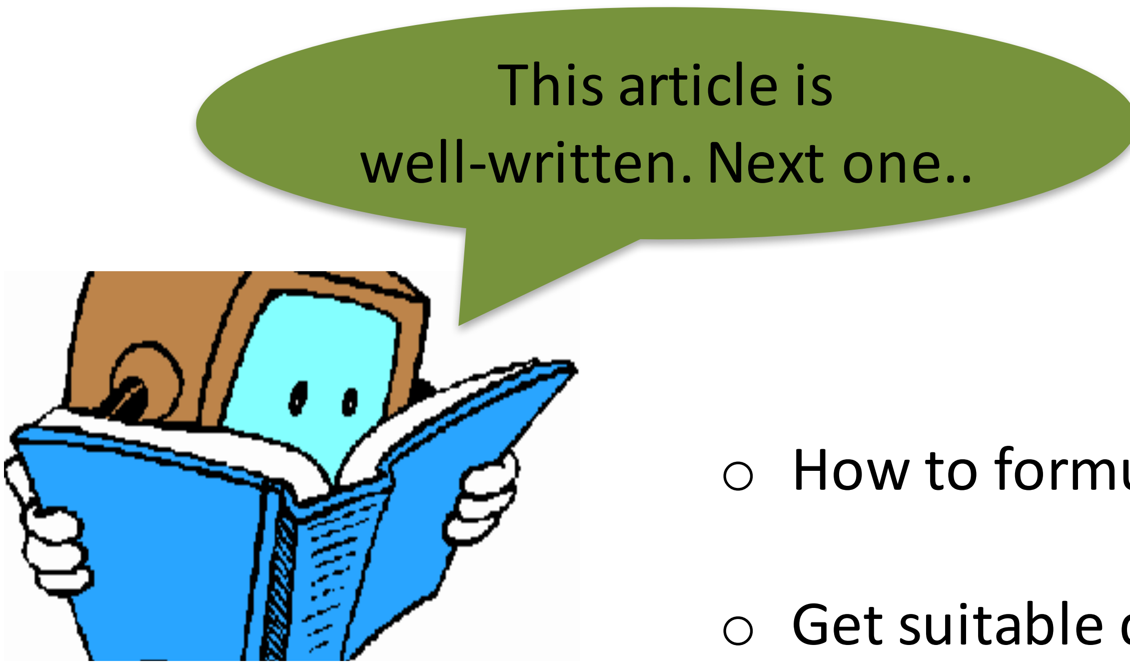
This word is striking, bold and its meaning is completely unexpected.

dodecahedron (doh-dek-ah-HEE-dren) — any polyhedron having twelve plane faces.

It's almost musical.

Text Quality Prediction

Can we teach computers to make similar judgements?



- How to formulate the task?
- Get suitable data with distinctions
- Find correlates in text

Why do we care?

- Information retrieval, article recommendation
  - All articles are not of the same quality
  - Can filter by quality in addition to relevance
- Authoring support, educational assessment
  - Automatic assessment is cheap, consistent and quick
  - Spelling and grammar correction are commercially successful
- Text generation systems
  - Systems can understand how to generate coherent text
  - Can evaluate system output

This talk

- Defining text quality and creating a corpus of overall article ratings
  - Large scale realistic sample of writing differences
- Two models
  - A model for organization using syntax patterns
  - A model for reader interest
- Document-level quality prediction
  - In contrast to spelling and grammar
  - Often not a binary, correct/in-correct distinction

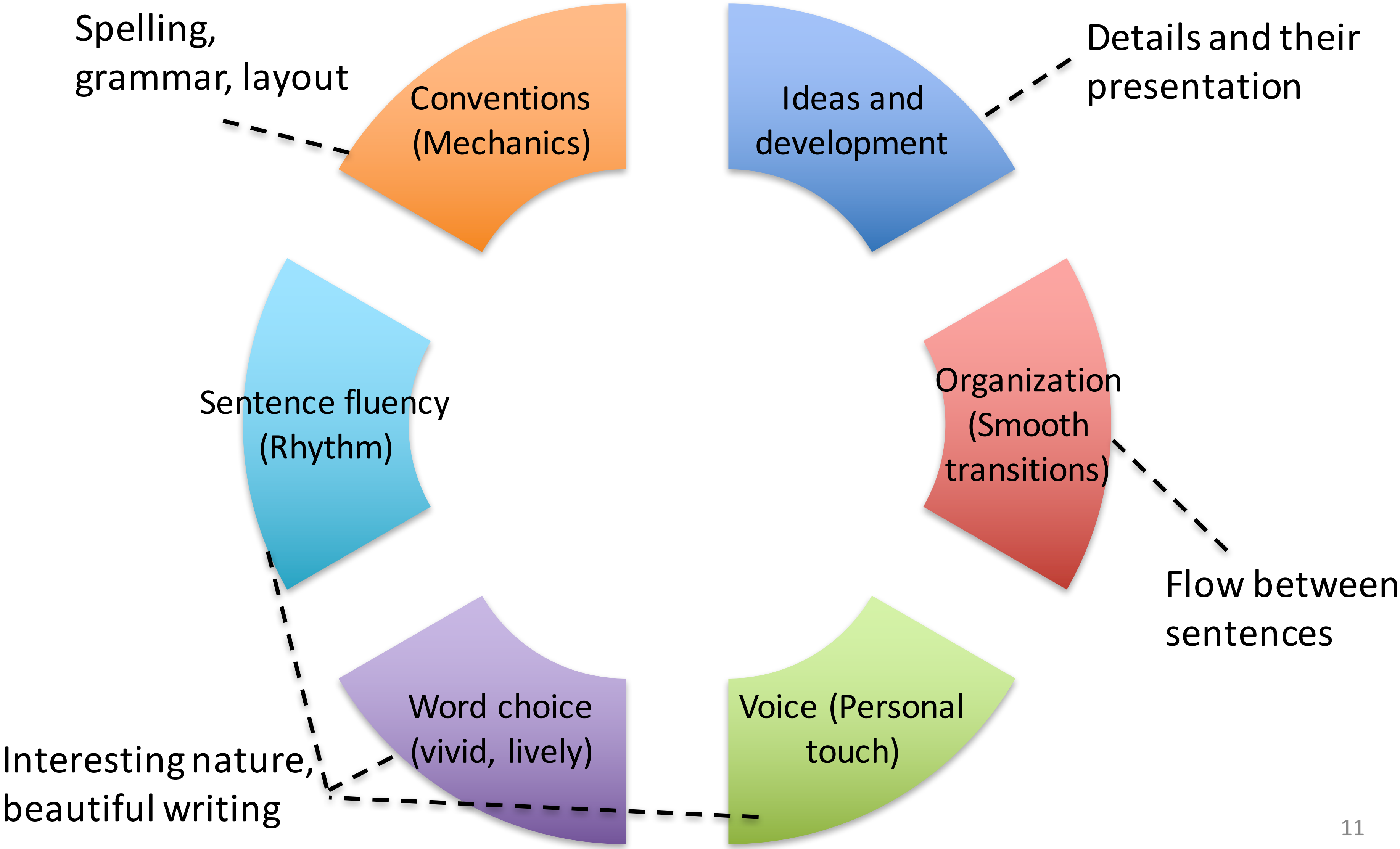
>> Defining Text Quality

- Aspects of quality
- Who is the audience?

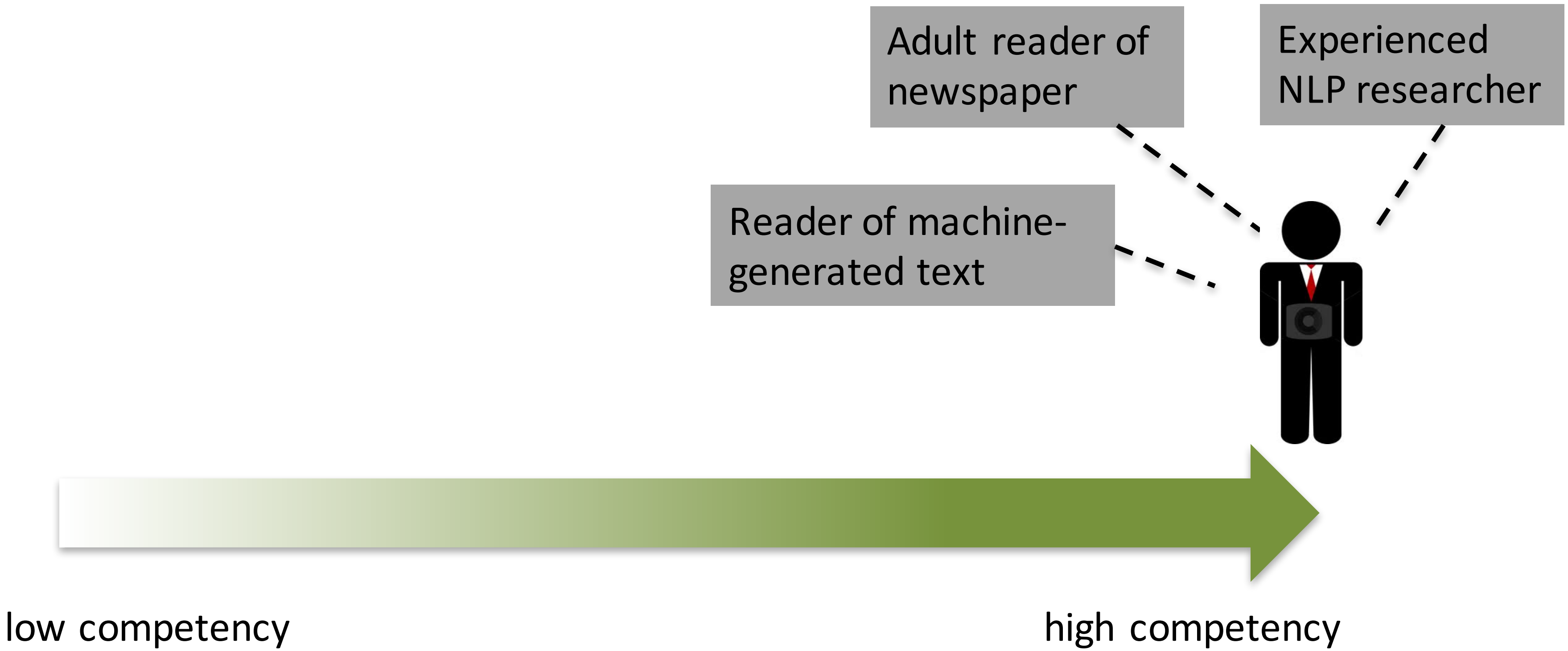
Aspects of quality

- We adopt a definition from the education field

Six Traits [Spandel 2004]



Audience for text quality – An expert

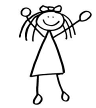


- Increased focus on linguistic properties of the text

Relationship to readability

- Readability has a strong focus on comprehension

Grade level 1



Grade level 2

..

Grade level 12

- Audience distinctions
  - child vs. adult, novice vs. expert, cognitive disability or not

>> A Corpus for Document-level Quality

Louis & Nenkova, Discourse and Dialogue, 2013


The New York Times

Science

Thursday, April 7, 2011

As Dinosaurs Waned and Mammals Rose, the Lowly Louse Kept Pace

By NICHOLAS WADE

Lice are expert evolvers, and a new family tree of lice stretches so far back that the host of the first louse would have been a dinosaur.

Science journalism: example snippet

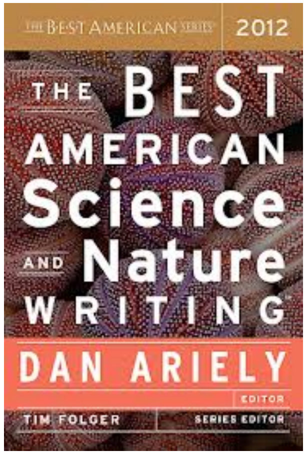
Sarah Lewis is fluent in firefly.

On this night she walks through a farm field in eastern Massachusetts, watching the first fireflies of the evening rise into the air and begin to blink on and off.

Dr. Lewis, an evolutionary ecologist at Tufts University…

Category 1 : VERY GOOD articles

- Seed set = 63 New York Times articles that appeared in the Best American Science Writing series



- We choose only the NYT articles
  - We use the NYT Corpus to expand our category
  - Normalize for differences in writing due to source

Topics in the seed set

Tag	Appearance
Medicine and Health	22
Space	14
Physics	10
Biology and Biochemistry	8
Genetics and Heredity	8
Archaeology and Anthropology	7
…	
Computers and the Internet	4

Expanding the VERY GOOD set

- Assume: ~40 authors of the seed set are excellent writers

- Other articles from the NYT written by the same authors
  - which are research related
  - during the same 10 year period
  - on similar topics
  - similar lengths

Category 2: TYPICAL writing in the NYT

- Other science articles around the same time, but not written by the popular authors

The general corpus:

Category	Total Articles
VERY GOOD	3,530
TYPICAL	20,242

A topic-paired corpus

- The general categories mix different topics
  - geography, biology, astronomy, linguistics…
- But an IR system compares articles on the same topic

- For each VERY GOOD article, get 10 most similar TYPICAL articles (based on the content)
- Enumerate all pairs of (VERY GOOD, TYPICAL)

- 35,300 pairs

Two quality prediction tasks

2 categories
GOOD (~3500)
TYPICAL (~3500)

→ `Any-topic’

– is this article VERY GOOD or TYPICAL?

Topically similar pairs
<VERY GOOD, TYPICAL>
~35,000 pairs

→ `Same-topic’

– which article in the pair is the VERY GOOD one?

Properties of the dataset

- Distinguishes average writing from very good
- Allow to focus on aspects such as beautiful writing
  - Less likely to have spelling and grammar errors
- Large scale and realistic sample of writing differences
  - Previous work often used machine generated text or artificially manipulated text

>> Predicting organization quality

Louis & Nenkova, EMNLP 2012

Some sequences of sentence types convey the overall purpose better

Motivation

Solving X is useful for many applications.

Introduce approach

We present a new approach to address X.

Results

Results show that our method works well.

Intentional structure of an article

- Every text has a purpose that the author wishes to convey

- Influential early theories discuss it at length

  [Grosz & Sidner 1986]

- Particularly for academic writing, it is popular to see articles as a sequence of intentions

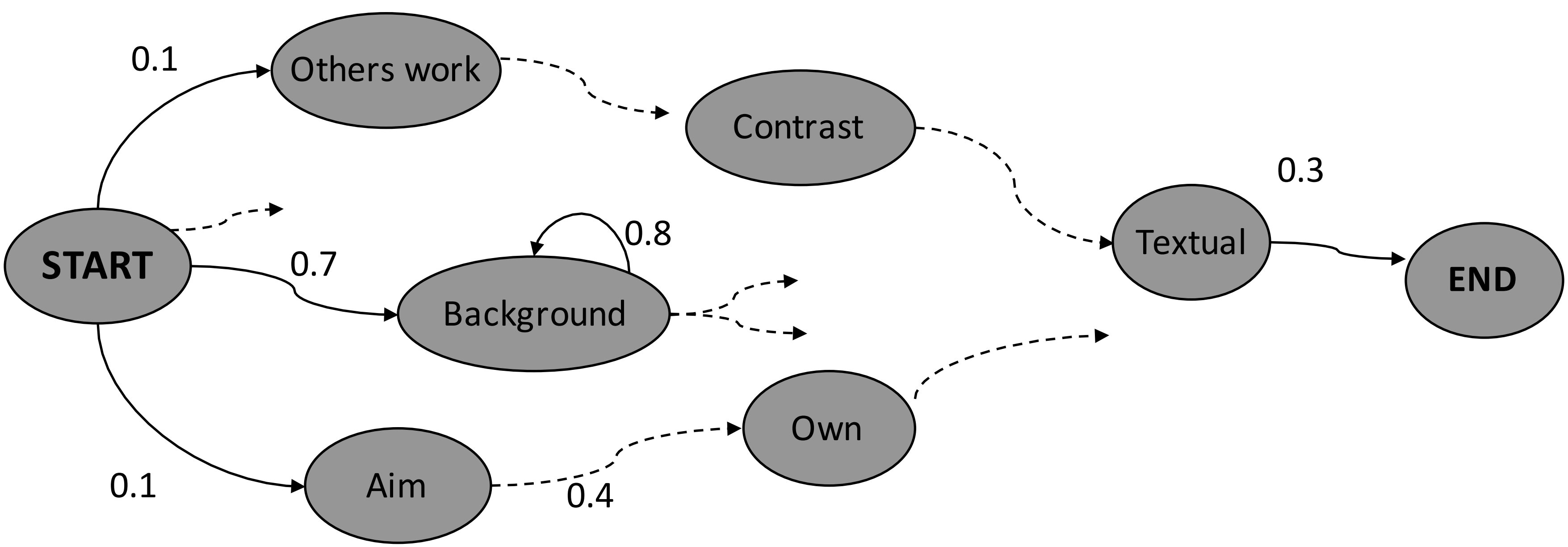
  [Swales 1990, Teufel 2000]

Oracle model of intentional structure

- Using manual annotations of intentions on ACL articles

[corpus by Teufel, 2000]

Markov Chain on Introduction sections



Main idea of the work

- Annotating sentence types is hard. Pre-defining the set of sentence types is harder

- Assume

Syntax ~ rough proxy for sentence type

Syntactic patterns in explanations

indefinite article

term to define

is/are

NP

More specific:
topicalized PP

Definitions look like this

- An **aqueduct** is a water supply or navigable channel constructed to convey water.

  Relative clause

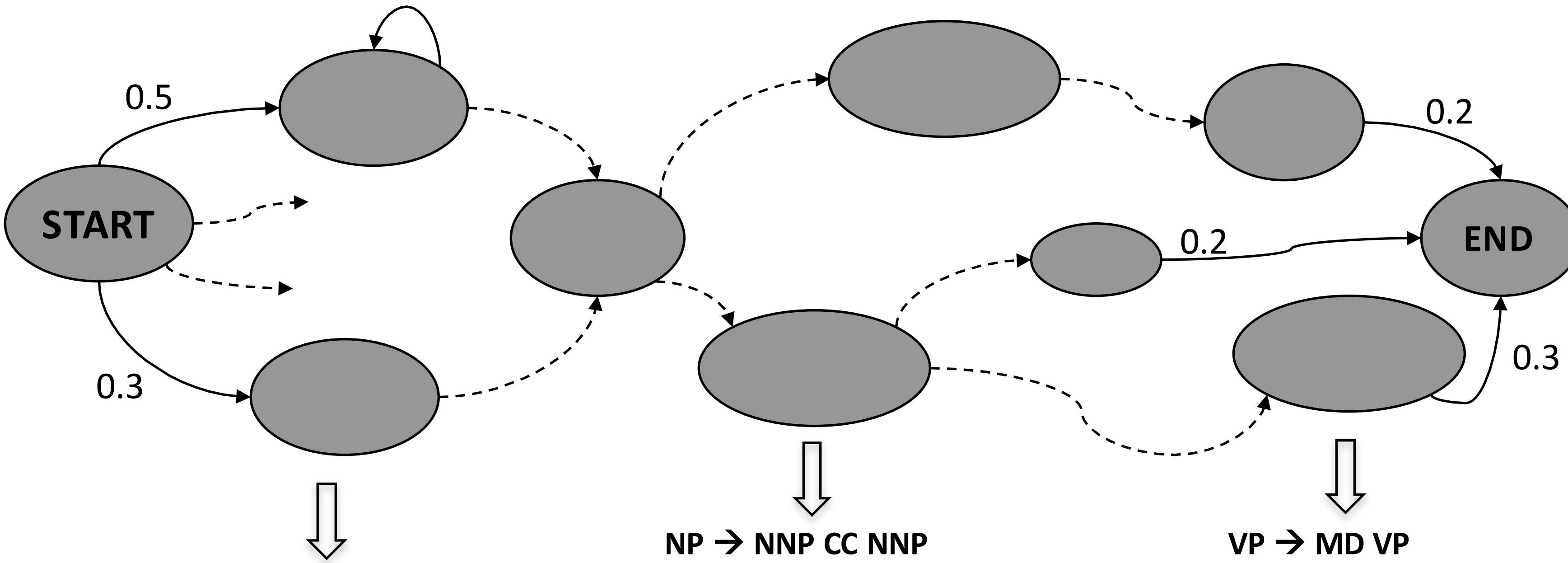
  In modern engineering, the term is used for any system of pipes, canals, tunnels, and other structures used for this purpose.

Descriptive articles look like this

- A **cytokine receptor** is a receptor that binds cytokines.

  In recent years, the cytokine receptors have come to demand more attention because their deficiency has now been directly linked to certain debilitating immunodeficiency states.

Syntax-based HMM model



VP → VBZ NP

NP → DT ADJP

NP → NP PP

….

“Definitions”

NP → NNP CC NNP

NP → CD

NP → NP , NP

…

“Citations”

VP → MD VP

VP → VB VP

VP → VB PP

…

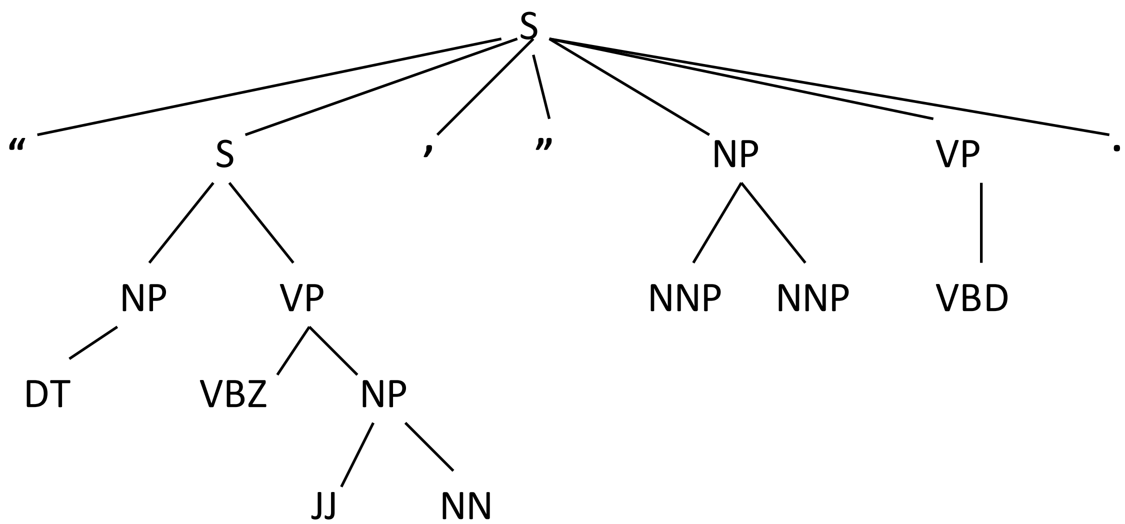
“Speculations”

* Uses grammatical productions

A second model: based on *d*-sequences

- More information about adjacent constituents
- A POS tag sequence loses all abstraction

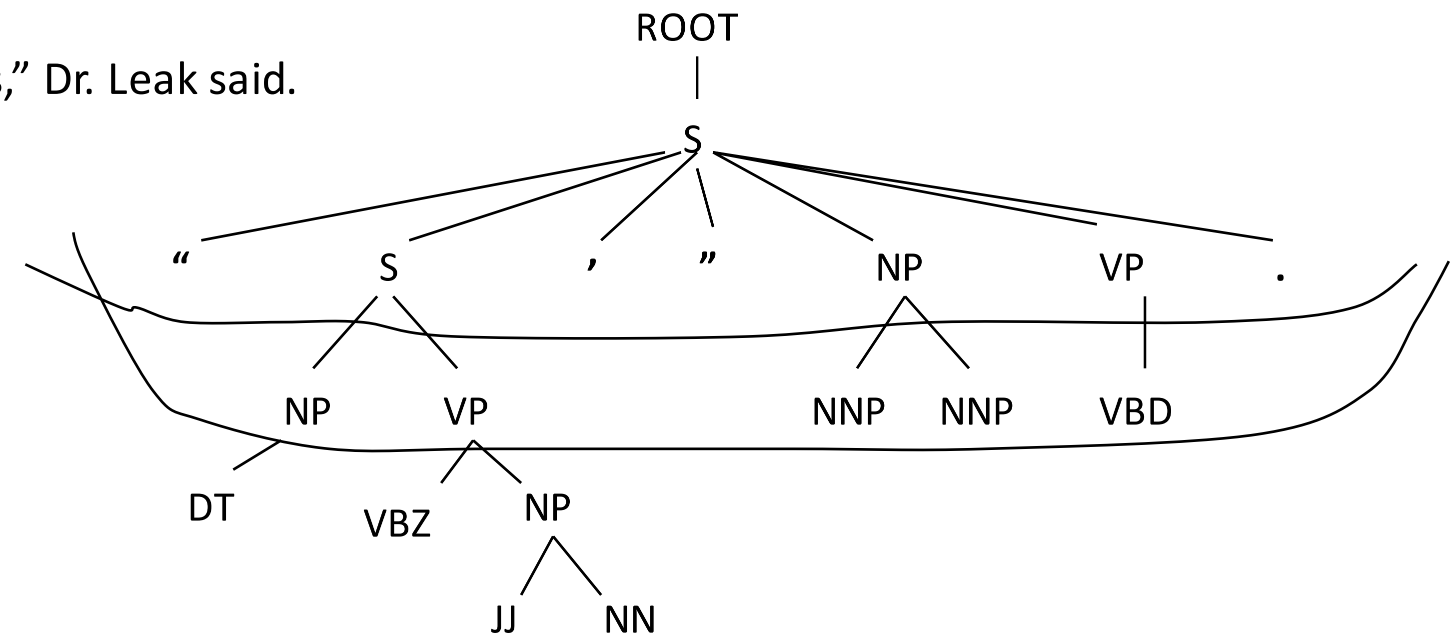
[“ DT VBZ JJ NN , ” NNP NNP VBD .]



- D-sequence
  - control abstraction using a parameter “depth” (d)

Step 1 – depth cutoff

“That’s good news,” Dr. Leak said.



d = 2

“ S , ” NP VP .

d = 3

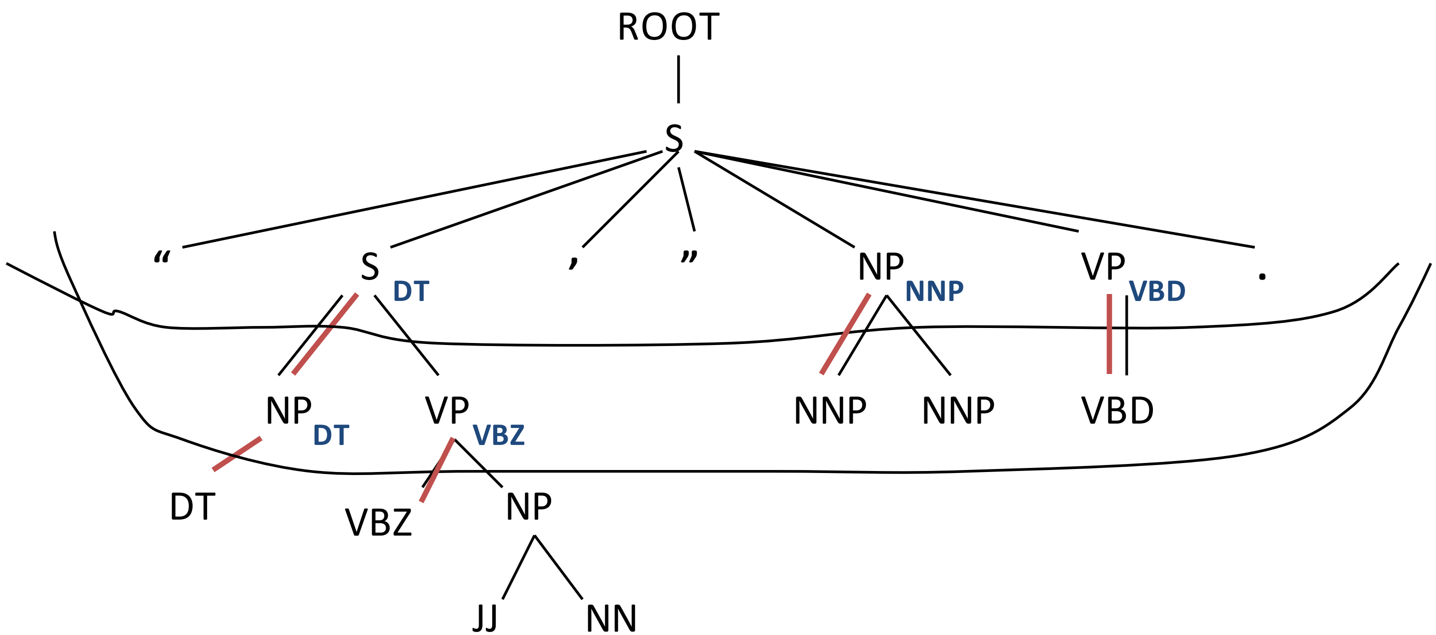
“ NP VP , ” NNP NNP VBD .

Choose a *depth d*

Terminate tree at d

Read off *new* leaves from left to right

Step 2: Node augmentation



For phrasal nodes in d-sequence,

- Annotate with left most leaf in full tree

d = 2

“ S_{DT} , ” NP_{NNP} VP_{VBD} .

d = 3

“ NP_{DT} VP_{VBZ} , ” NNP NNP VBD .

Evaluation task on academic writing

- ACL anthology corpus
  - abstract, introduction, related work
- Approximate distinction for organization quality
  - Original article → well-organized
  - Random permutation of original → poorly-organized
  - Create pairs <original, permutation?
- Task: identify the original version in the pair
  - Baseline 50% accuracy

Summary of results on academic writing

- Correct = higher likelihood for original article
  - versus permuted article

- D-seq model

ACL conference	Accuracy
Abstract	62.9
Introduction	68.8
Related work	72.7

Baseline = 50%

Do sentence types distinguish VERY GOOD and TYPICAL science news?

- Create the HMM on VERY GOOD training articles
- Get likelihood and most likely state sequence for a new article
  - Compute features based on these
- A classifier is trained to predict the VERY GOOD article

Results on our corpus

Any Topic: Given an article, is it "VERY GOOD" or "TYPICAL" ?

System	Accuracy
Baseline (random)	50%
HMM-productions	61%

- 10 fold cross validation results
- SVM classifier

Same Topic: Given a pair of articles on the same topic, which one is "VERY GOOD"?

System	Accuracy
Baseline (random)	50%
HMM-productions	63%

>> Predicting reader interest

Louis & Nenkova, TACL 2013

Predicting interest: A new task

- A lot of work on identifying what is wrong with a text
  - Spelling mistakes, grammar errors, incoherent writing

- It is not known how to characterize writing that is engaging, interesting and nice

Approach to feature development

- Focus on interpretable features
  - Only 41 features
  - Each feature is a composite one: indicates an aspect directly
  - Linguistically interesting
- Confirm that features represent the intended aspect
  - Tune by checking feature values on random snippets of text

1. Unusual words and phrases

Is the phrasing and language use unique?

- Word-based
  - high perplexity under a phoneme n-gram model
  - Eg: ‘undersheriff’, ‘powwow’, ‘chihuahua’, ‘qipao’

- Word pairs--based
  - adjective-noun, noun-noun, adverb-verb, subject-verb pairs
  - perplexity under a language model
  - Eg: ‘plasticky woman’, ‘so-called superkids’

2. Visual nature

Is there scene setting?

- Creating a large lexicon of visual terms
  - Source: an image-tagged corpus
  - Large source of potentially visual words, but noisy

- Create LDA-based topics on the tag set
  - Use the manual MRC terms to filter out non-visual topics

grass, mountain, green, hill, blue, field, sand...
round, ball, circles, logo, dots, square, sphere...
silver, white, diamond, gold, necklace, chain...

Human interest and text structure

3. Use of people in the story

   Does the story revolve around a person?

   – animacy information from NEs, pronouns, ngram patterns

4. Sub-genre

   Is the article is a narrative, interview or dialog

   – Eg: narrative score ~ past tense verbs, pronouns, proper names

Sentiment and Research

5. Affect

   Is there an emotional angle to the story?

   – using sentiment word dictionaries

6. Research content

   How much explicit research description is present?

   – using a hand-built dictionary of research words

How the features vary in a random sample of very good and typical articles (t-test)

Higher values in VERY GOOD set

- ✓ Visual words in beginning and end of articles
- ✓ Unusual words and phrases
- ✓ Sentiment words, negative polarity
- ✓ Research words

- ✗ Total visual words
- ✗ Animacy counts
- ✗ Narrative, interview or dialog format

Accuracies on the two tasks

Any Topic: Given an article, is it "VERY GOOD" or "TYPICAL" ?

System	Accuracy
Baseline (random)	50%
Interesting-science features	75%

- 10 fold cross validation results
- SVM classifier

Same Topic: Given a pair of articles on the same topic, which one is "VERY GOOD"?

System	Accuracy
Baseline (random)	50%
Interesting-science features	68%

Combining interest with other aspects

Feature set	any topic	same topic
Interesting science	**75.3**	**68.0**

Genre-specific measures are stronger than generic ones

Different aspects of writing have complementary strengths

Conclusions

- Text quality is an interesting and challenging task
- More success on the topic recently
  - application to novels, tweets, essays
- Future work
  - A lot to be done in terms of formalizing the tasks, collecting data, models and evaluation
  - Transferring the knowledge to generating texts

Thank you!