A Tree-Position Kernel for Document Compression

Hal Daumé III and Daniel Marcu
Information Sciences Institute

Special thanks to Eric Horvitz

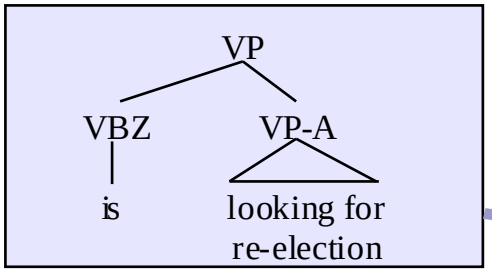
Talk Outline

- Our previous model
- What was wrong with it
- A slide or two on reproducing kernels
- A (novel) tree-position kernel
- The standard tree kernel
- Putting it all together
- Results

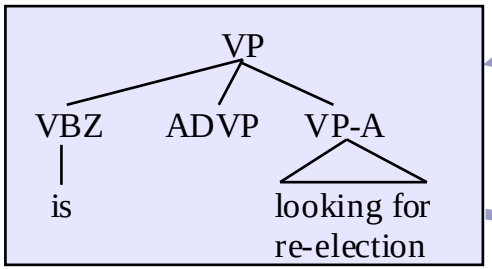
Our Previous Model

- Noisy-channel model for document compression; generative story:
  - Begin with a summary:
    - The mayor is looking for re-election.
  - Add syntactic units:
    - The mayor is now looking for re-election.
    - The mayor is not looking for re-election.
  - Add discourse constituents:
    - The mayor is now looking for re-election. But without the support of the governer, he is still on shaky grounds.
    - The mayor is now looking for re-election. Sharks have sharp teeth.

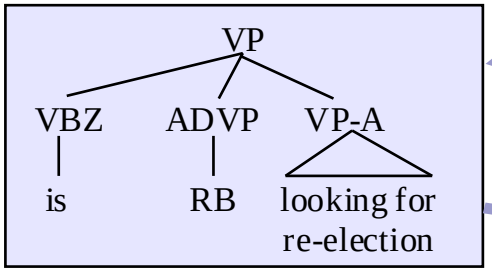
Syntax Probabilities



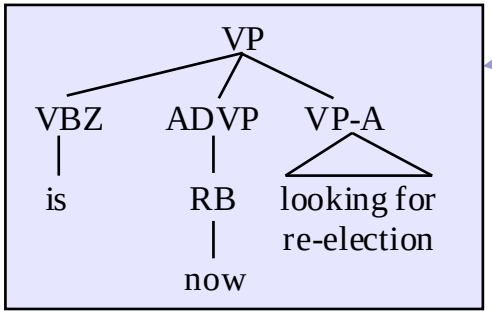
$$P(VP \rightarrow VBZ\ \ ADVP\ \ VP-A \mid VP \rightarrow VBZ\ \ VP-A)$$



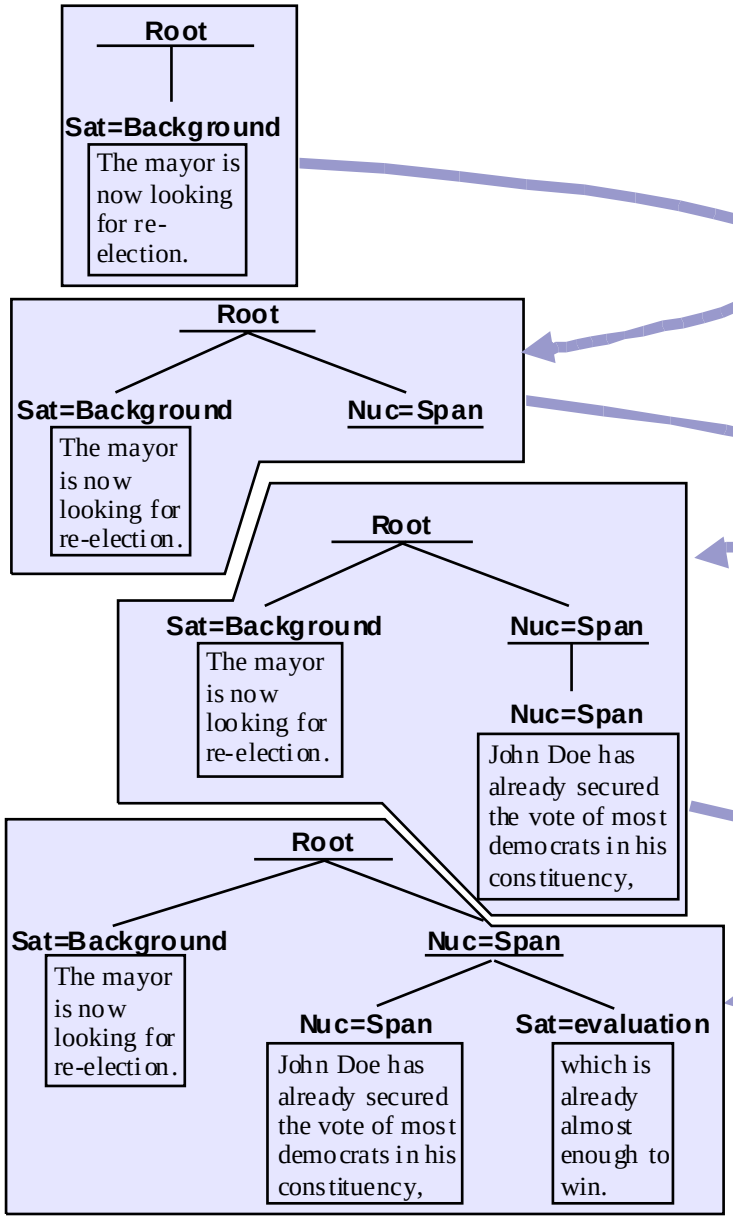
$$P(ADVP \rightarrow RB \mid ADVP)$$



$$P(RB \rightarrow \text{now} \mid RB)$$



Discourse Probabilities



$$P(\text{Root} \rightarrow \text{Sat=Background Nuc=Span} \mid \text{Root} \rightarrow \text{Sat=Background})$$

$$P(\text{Nuc=Span} \rightarrow \text{Nuc=Span} \mid \text{Nuc=Span})$$

$$P(\text{Nuc=Span} \rightarrow \text{Nuc=Span Sat=Background} \mid \text{Nuc=Span} \rightarrow \text{Nuc=Span})$$

Sources of Data

- Syntax:
  - Mined Ziff-Davis document/abstract corpus for pairs of compressions (~2k sentences)
  - Mined MSN for several weeks (~2 sents)
- Discourse:
  - RST Corpus has EDU-level annotations for relevance (~132 documents)

What's Wrong with It?

- CFG-style rules are too coarse:
  - Does this branch dominate 10 words or 500?
  - Is my grandparent the root?
  - Are the words I dominate important?
- CFG-style rules are too fine:
  - 708/1061 syntax rules are singletons
  - 629/1146 discourse rules are singletons
- How can we address *both* of these problems?

KERNEL METHODS!!!

A Slide on Kernels

- Inputs $x \in X$, $y \in \{-1,+1\}$
  - Many learners only use $(x_i \cdot x_j)$
  - Replace with $(\phi(x_i) \cdot \phi(x_j))$
  - Can often compute using a *kernel:*

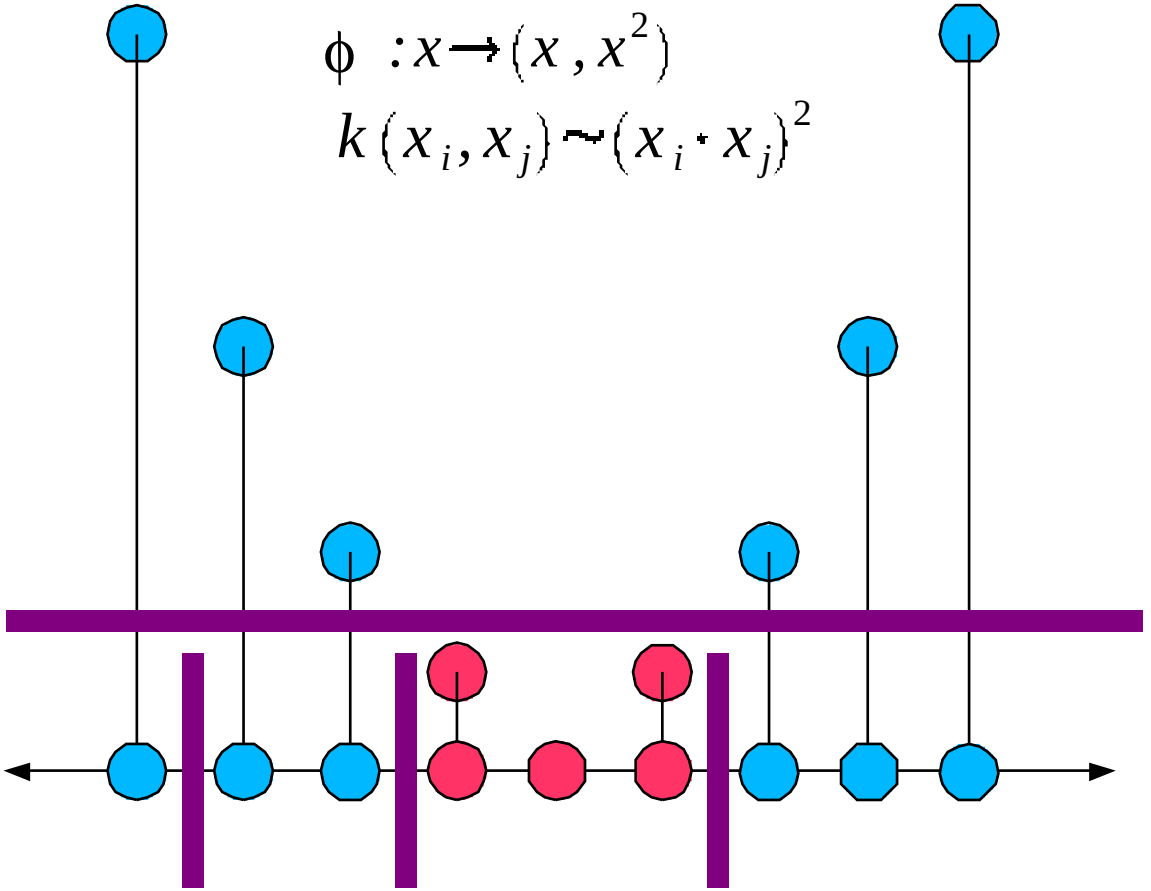
$$(\phi(x_i) \cdot \phi(x_j)) = k(x_i, x_j)$$

- Sufficient and necessary condition:
  - k is positive semi-definite:

$$\iint f(x_i)\, k(x_i, x_j)\, f(x_j)\, dx_i\, dx_j \geq 0$$

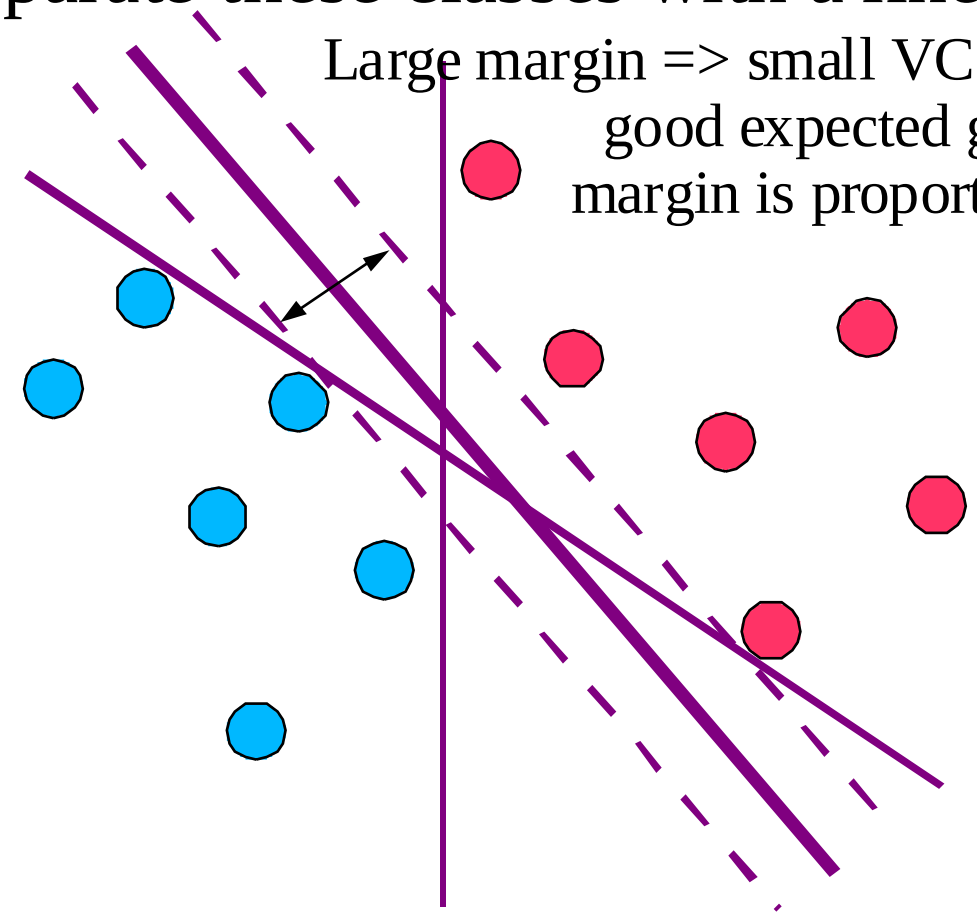
...or two: a simple example

- Find a line to seperate these classes:



...or two.5: margins

- Separate these classes with a line



Large margin => small VC dimension => good expected generalization; margin is proportional to 1/||w||

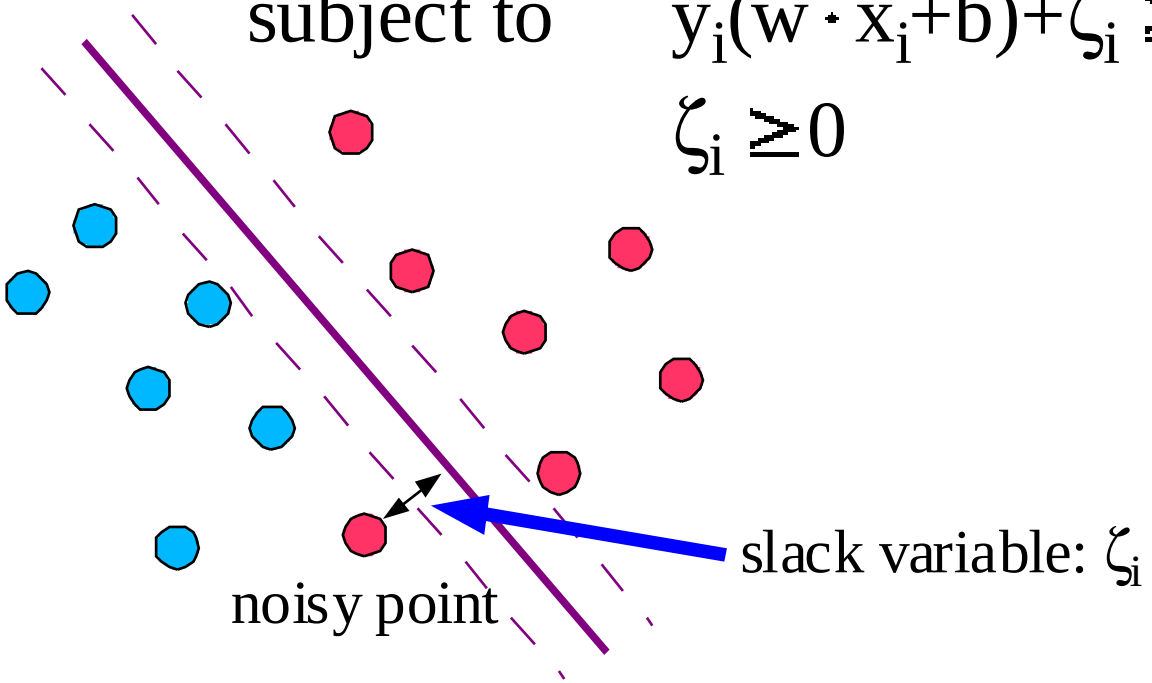
...or two.75: SVMs

- Support vector machines maximize margins in kernel space:

$$\text{minimize} \quad \|w\|^2 + C\sum_i \zeta_i$$

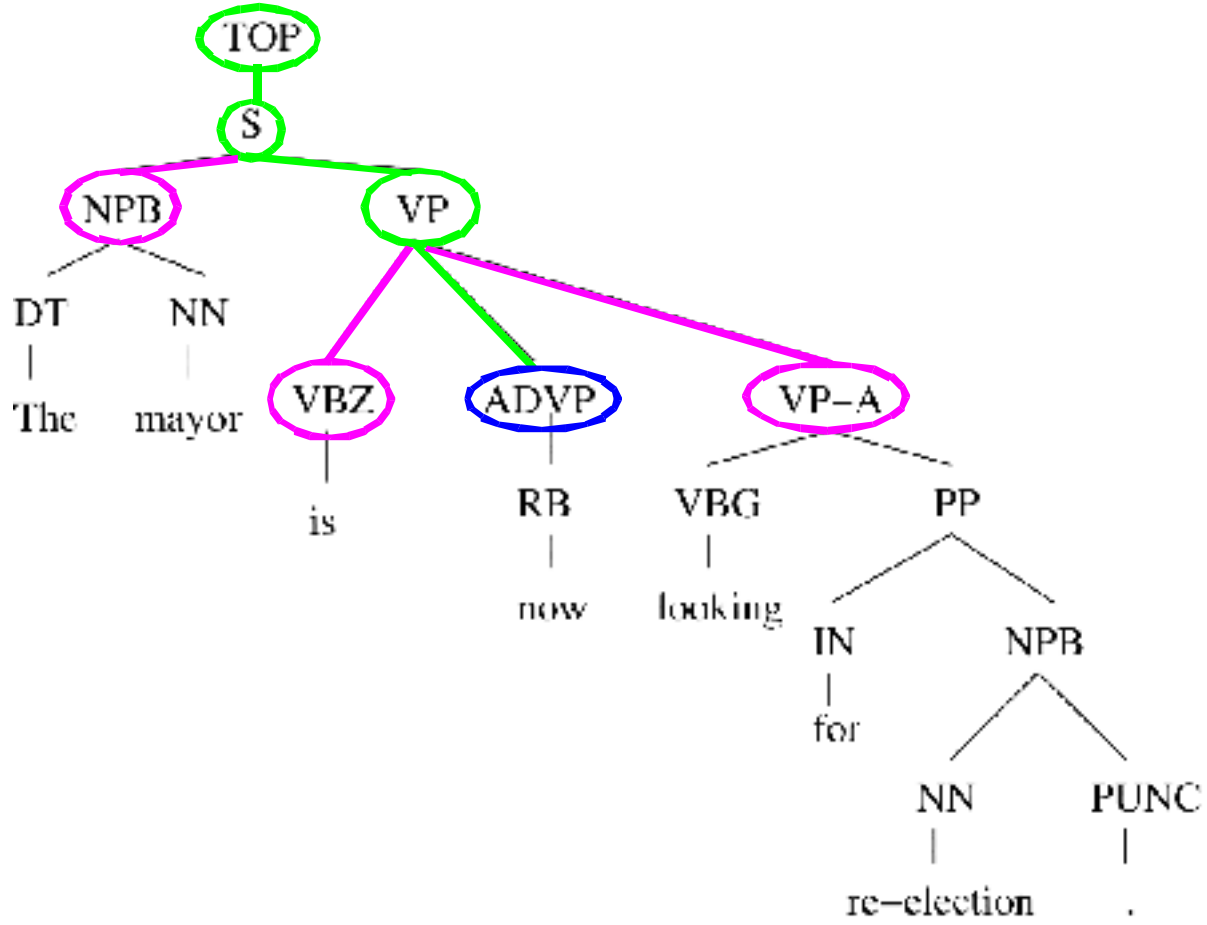
$$\text{subject to} \quad y_i(w \cdot x_i + b) + \zeta_i \geq 1$$

$$\zeta_i \geq 0$$



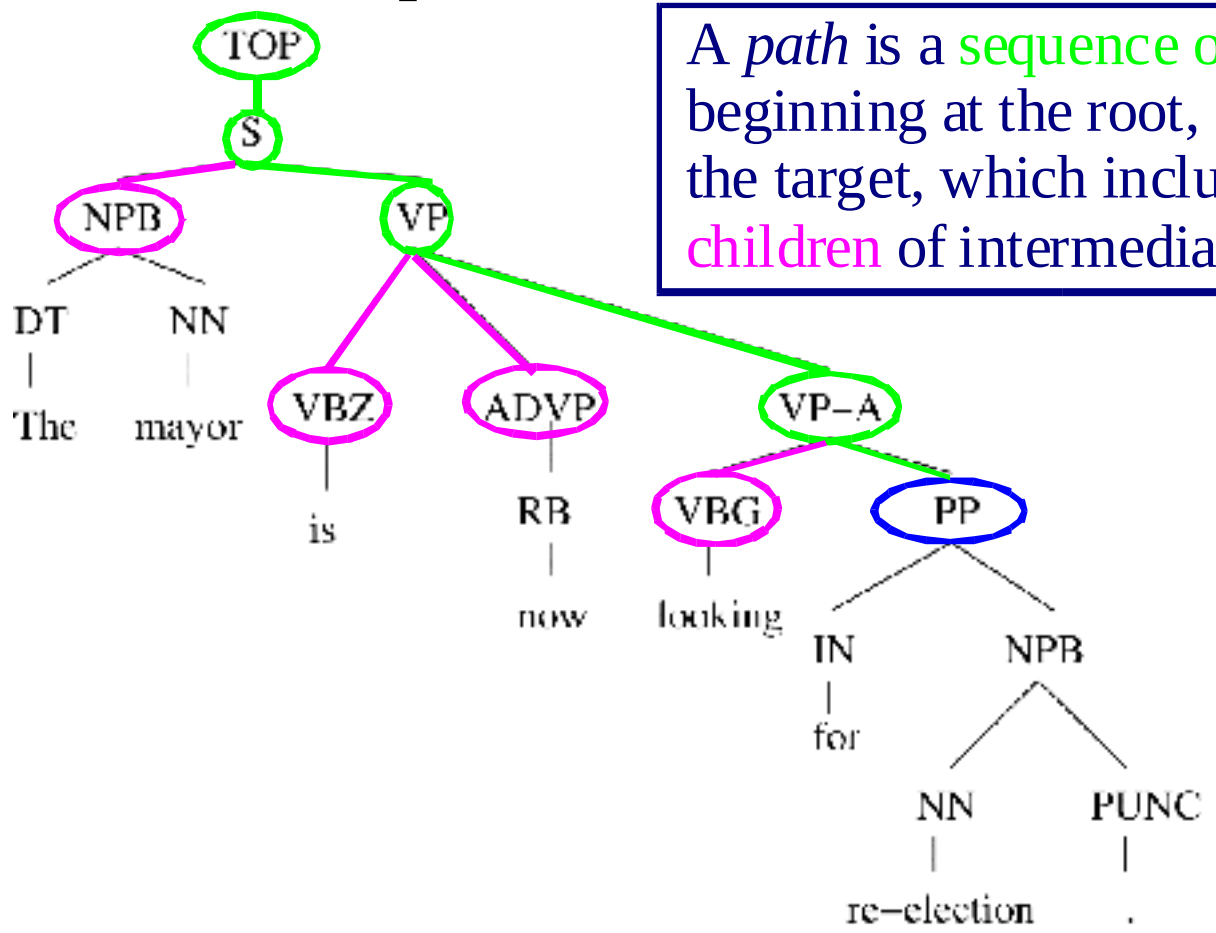
Tree-Position Kernel

- What ancestral features should help decide whether to keep a node or not?



Tree-Position Kernel

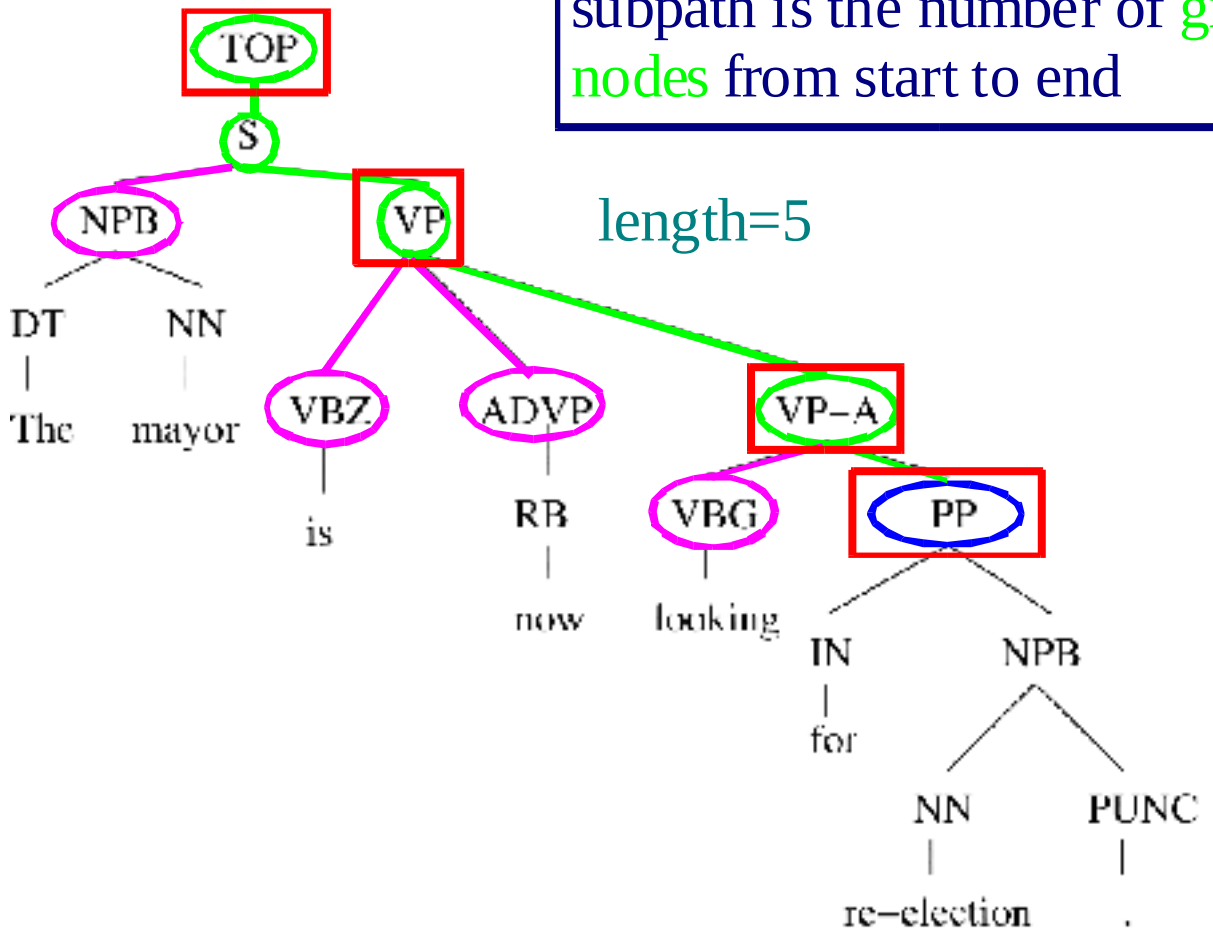
- What ancestral features should help decide whether to keep a node or not?



A *path* is a sequence of nodes, beginning at the root, ending at the target, which includes all children of intermediate nodes

Tree-Position Kernel

A *subpath* is any subsequence of a path; the *length* of a subpath is the number of green nodes from start to end



length=5

Tree-Position Kernel

- For *any* possible subpath p, define ϕ_p by:

  $\phi_p(t) = \lambda^{\text{length of } p \text{ in } t}$ if p occurs in t

  $\phi_p(t) = 0$ otherwise

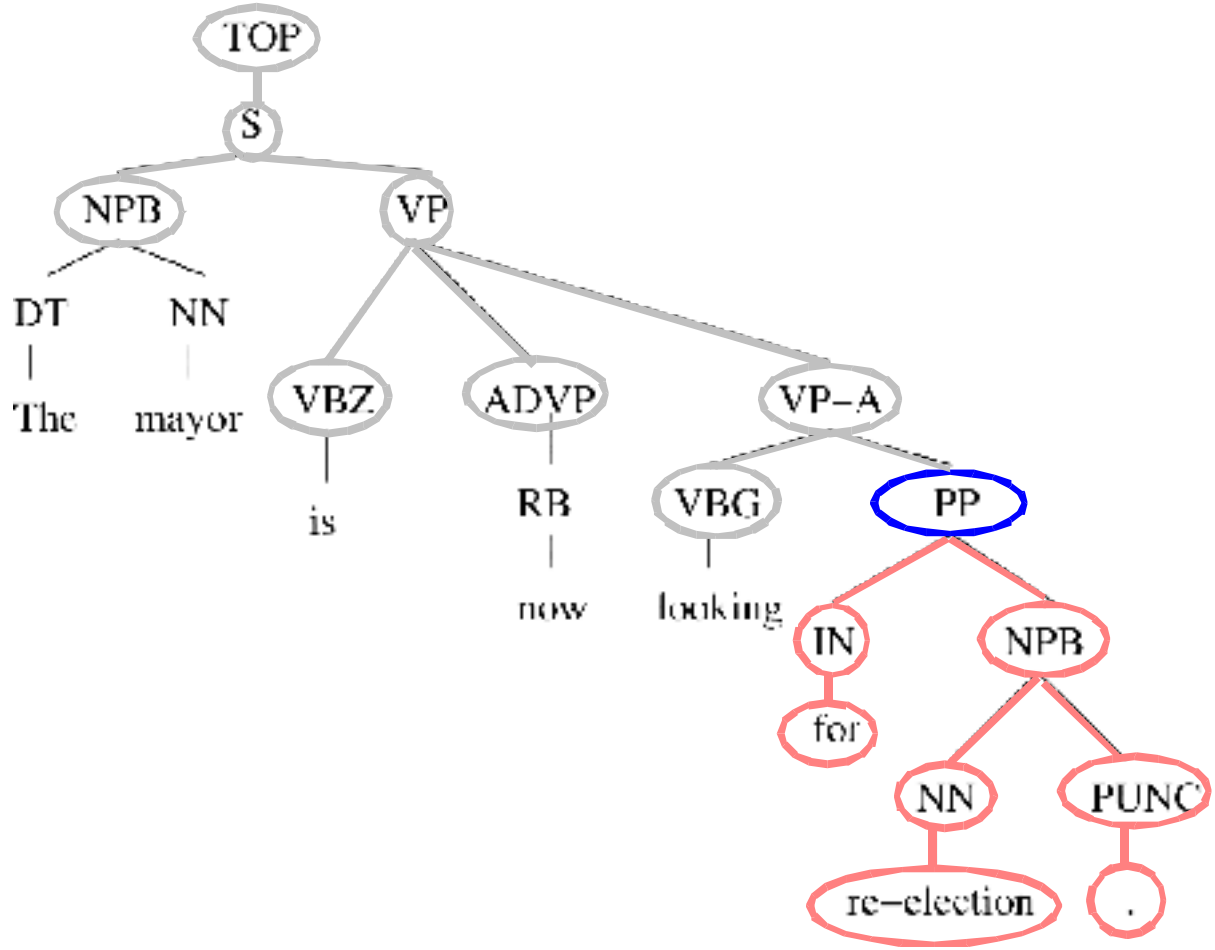
  for some λ in $(0,1]$

- Clearly cannot enumerate all such ϕ_p

- *But*, we can compute $k_\phi(t,t')$ in time linear in the length of the longest path in t or t'!

Sub-Tree Kernel

- What descendent features should help decide whether to keep a node or not?



Sub-Tree Kernel

- Defined similarly to tree-position kernel: each possible subtree gets it's own feature, weigted by size
- Can compute kernel in time linear in the product of the size of the two trees [Collins and Duffy] or linear in the sum of the sizes [Vishwarathan] -- we use the fast method

Additional Features

- Also include global features:
  - Length of dominated string
  - Absolute sentence/word position
  - {min,max,avg} tf-idf scores of dominated words
  - Total document length
  - Tag of current node
  - Nuclearity of current node (discourse only)

Putting It All Together

- Recall that k is a kernel iff for all f:

$$\iint f(x_i)\,k(x_i,x_j)\,f(x_j)\,dx_i\,dx_j \geq 0$$

- Easy to see that if k and l are kernels, and α is a positive real number, then:

  $k(\bullet,\bullet) + l(\bullet,\bullet)$ and

  $\alpha k(\bullet,\bullet)$

  are both kernels

- We combine the three linearly

Training the System

- Use our own iterative optimization; each iteration is **linear** in number of examples
- Since discourse data and syntax data are disjoint, each is trained separately
- Results of classifier:

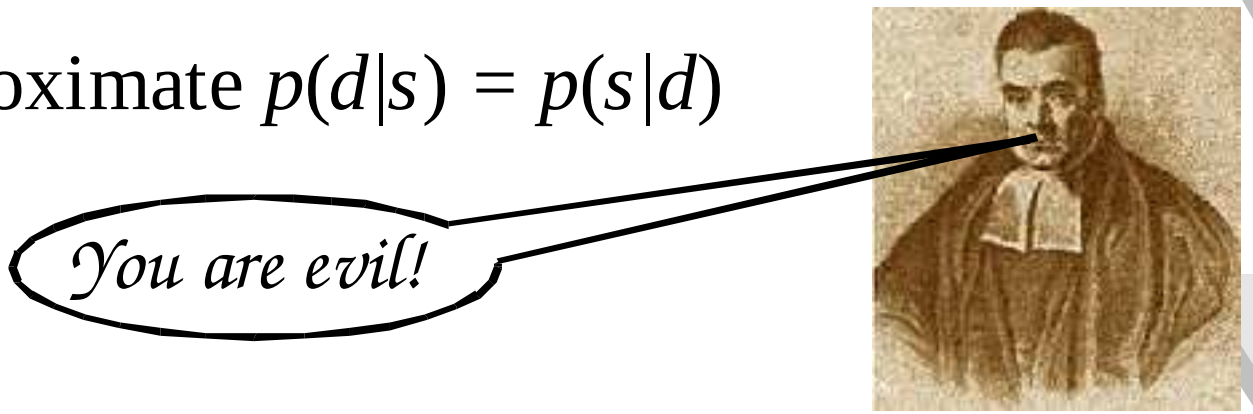
	Syntax	Discourse
# of training points	19,702	8,735
baseline accuracy	53.2%	51.9%
accuracy on training	82.7%	71.2%
x-val accuracy	76.7%	65.8%

Language Model

- Trigrams
- Linear interpolation of 300mw plain English model and 200kw headline model (EM on held-out headlines)
- Kneser-Ney smoothing

Channel Model

- Approximate $p(d|s) = p(s|d)$



- Convert classifier outputs to probabilities using sigmoids [Lin et al.]
- Classifier lets us calculate $p(s|d)$ by, eg,

$$p(\mathrm{p}\rightarrow\mathrm{y} \mid \mathrm{p}\rightarrow\mathrm{xyz}) =$$
$$p(\mathrm{keep}|\mathrm{y})(1-p(\mathrm{keep}|\mathrm{x}))(1-p(\mathrm{keep}|\mathrm{z}))$$

Decoder

- Irene's forest-ranker
  - Nodes with $\log(p)$<-6 are pruned
  - ~1400w document gives a forest with:
    - ~6k nodes
    - ~3.1 options per node
- Produce best output at any length
- Take best one under 75 bytes
  - modulo some ugly heuristics

Example Outputs (1)

- References:
  - Cambodian party defends leader Hun Sen against criticism of U.S. House
  - CPP defends Hun Sen to US Senate. Asks rejection of non-binding resolution.
  - Cambodia's ruling party seeks to counter human right's criticism
  - US House seeks probe of Cambodian rights violations; Ruling party responds
- Our output:
  - Cambodia's ruling party responded criticized a resolution passed.

Example Outputs (2)

- References:
  - Uganda faces rebel forces on west (Congo) and north (Sudan)
  - Rebels, likely ADF, attack Chiondo, killing 6. Soldiers kill 2 rebels.
  - Reels kill 6 civilians in Congolese villiage; Ugandan troops aid revels
  - Congolese rebels kill 6 people before Ugandans drive them across border
- Our output:
  - Rebels attacked and killed six civilians said occurred overnight Wednesday

Example Outputs (3)

- References:
  - Asia-Pacific economic summit in Kuala Lumpur faces sever problems
  - Financial officials advise reform; topic likely to dominate at APEC talks
  - Asian countries advised to restructure economies and corporations
  - Gloom faces up-coming 18-nation Asia-Pac meet; turmoil in host nation.
- Our output:
  - To adopt further reforms in their effor Among the necessary steps is

Example Outputs (4)

- References:
  - President's next premier-designate likely to be widely trusted lawmaker.
  - President Demirel seeking new candidate to form new government
  - Ecevit fails 3-wk try to form majority. Long left-right split in parliament
  - Speaker of Turkish parliament likely to form government; Islamic party out
- Our output:
  - President Suleyman Demirel appeared to persuade bickering political leaders

DUC Evaluation Results

- System 75 on Task 1

	Human	Baseline	Best	Us
Rouge1	0.29	0.21	0.22	0.12
Rouge2	0.08	0.06	0.06	0.03
Rouge3	0.033	0.018	0.018	0.006
Rouge4	0.013	0.007	0.005	0.0013
RougeL	0.25	0.19	0.19	0.11
RougeW	0.14	0.12	0.12	0.07

- Bottom ~8 out of ~50 in all measures :)

Conclusions

- Tried something new & interesting
- Worked well as a standalone component
- Did not work well in a system...why?
  - Poorly combined kernels
  - Use of Bayes' un-rule
  - Components trained separately
  - Data mismatch:
    - 10% extracts on discourse (news), 53% on syntax (Ziff-Davis)
    - Expected production: 1% extracts, headlines, sometimes keywords
  - Poor language model
- ...or perhaps it did...we'll never know

Future Work (if any)

- Use automatic alignments to get data
- Include syntax-based decoder
  - ...or don't use a language model at all!
- Use more global features
- Evaluate by *hand*
  - ...on something other than news
  - ...with more than 75 bytes
- Shameless plug:

  `www.isi.edu/~hdaume/SVMsequel`