Machine Learning for NLP: New Developments and Challenges

Dan Klein
Computer Science Division
University of California at Berkeley

NOTE

- These slides are still incomplete
- A more complete version will be posted at a later date at:

http://www.cs.berkeley.edu/~klein/nips-tutorial

What is NLP?

- Fundamental goal: *deep* understand of *broad* language
- End systems that we want to build:
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Modest: spelling correction, text categorization…
- Sometimes we're also doing computational linguistics

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% for digit strings, 5% dictation, 50%+ TV

"Speech Lab"

- Text to Speech (TTS)
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)

Machine Translation

Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

- Translation systems encode:
  - Something about fluent language
  - Something about how two languages correspond
- SOTA: for easy language pairs, better than nothing, but more an understanding aid than a replacement for human translators

Information Extraction

- Information Extraction (IE)
  - Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 70% accuracy for multi-sentence temples, 90%+ for single easy fields

Question Answering

- Question Answering:
  - More than search
  - Ask general comprehension questions of a document collection
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the global warming debate?"
- SOTA: Can do factoids, even when text isn't a perfect match

Goals of this Tutorial

- Introduce some of the core NLP tasks
- Present the basic statistical models
- Highlight recent advances
- Highlight recurring constraints on use of ML techniques
- Highlight ways this audience could really help out

Recurring Issues in NLP Models

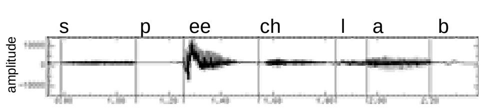
- Inference on the training set is slow enough that discriminative methods can be prohibitive
- Need to scale to millions of features
  - Indeed, we tend to have more features than data points, and it all works out ok
- Kernelization is almost always too expensive, so everything's done with primal methods
- Need to gracefully handle unseen configurations and words at test time
- Severe non-stationarity when systems are deployed in practice
- Pipelined systems, so we need relatively calibrated probabilities, also errors often cascade

Outline

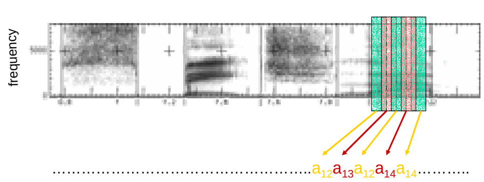
- Language Modeling
- Syntactic / Semantic Parsing
- Machine Translation
- Information Extraction
- Unsupervised Learning

Speech in a Slide

- Frequency gives pitch; amplitude gives volume



- Frequencies at each time slice processed into observation vectors



The Noisy-Channel Model

- We want to predict a sentence given acoustics:

$$w^* = \arg\max_w P(w|a)$$

- The noisy channel approach:

$$w^* = \arg\max_w P(w|a)$$

$$= \arg\max_w P(a|w)P(w)/P(a)$$

$$\propto \arg\max_w P(a|w)P(w)$$

Acoustic model: HMMs over word positions with mixtures of Gaussians as emissions

Language model: Distributions over sequences of words (sentences)

Language Models

- In general, we want o place a distribution over sentences
- Classic solution: n-gram models

$$P(w) = \prod_i P(w_i | w_{i-1} \ldots w_{i-k})$$

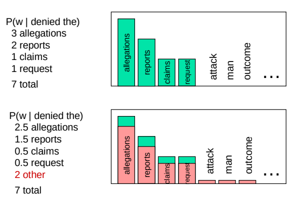
- N-gram models are (weighted) regular languages
- Natural language is not regular
  - Many linguistic arguments
  - Long-distance effects:
    "The computer which I had just put into the machine room on the fifth floor crashed."
- N-gram models often work well anyway (esp. with large n)

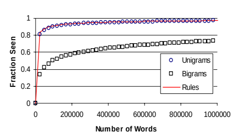
Language Model Samples

- Unigram:
  - [fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter]
  - [that, or, limited, the]
  - []
  - [after, any, on, consistently, hospital, lake, of, of, other, and, factors, raised, analyst, too, allowed, mexico, never, consider, fall, bungled, davison, that, obtain, price, lines, the, to, sass, the, the, further, board, a, details, machinists, …… , nasdaq]
- Bigram:
  - [outside, new, car, parking, lot, of, the, agreement, reached]
  - [although, common, shares, rose, forty, six, point, four, hundred, dollars, from, thirty, seconds, at, the, greatest, play, disingenuous, to, be, reset, annually, the, buy, out, of, american, brands, vying, for, mr., womack, currently, share, data, incorporated, believe, chemical, prices, undoubtedly, will, be, as, much, is, scheduled, to, conscientious, teaching]
  - [this, would, be, a, record, november]
- PCFG (later):
  - [This, quarter, 's, surprisingly, independent, attack, paid, off, the, risk, involving, IRS, leaders, and, transportation, prices, .]
  - [It, could, be, announced, sometime, .]
  - [Mr., Toseland, believes, the, average, defense, economy, is, drafted, from, slightly, more, than, 12, stocks, .]

Smoothing

- Dealing with sparsity well: smoothing / shrinkage
  - For most histories P(w | h), relatively few observations
  - Very intricately explored for the speech n-gram case
  - Easy to do badly





Interpolation / Dirichlet Priors

- Problem: $\hat{P}(w|w_{-1}, w_{-2})$ is supported by few counts
- Solution: share counts with related histories, e.g.:

$$\lambda \hat{P}(w|w_{-1}, w_{-2}) + \lambda' \hat{P}(w|w_{-1}) + \lambda' \hat{P}(w)$$

- Despite classic mixture formulation, can be viewed as a hierarchical Dirichlet prior [MacKay and Peto, 94]
  - Each level's distribution drawn from prior centered on back-off
  - Strength of prior related to mixing weights
- Problem: this kind of smoothing doesn't work well empirically
- All the details you could ever want: [Chen and Goodman, 98]

Kneser-Ney: Discounting

- N-grams occur more in training than they will later:

Count in 22M Words	Avg in Next 22M	Good-Turing c*
1	0.448	0.446
2	1.25	1.26
3	2.24	2.24
4	3.23	3.24

- Absolute Discounting
  - Save ourselves some time and just subtract 0.75 (or some d)
  - Maybe have a separate value of d for very low counts

$$P(w|w') = \frac{c(w,w') - d}{c(w')} + \alpha P'(w)$$

Kneser-Ney: Details

- Kneser-Ney smoothing combines several ideas
  - Absolute discounting

$$P(w|w') = \frac{c(w,w') - d}{c(w')} + \alpha P'(w)$$

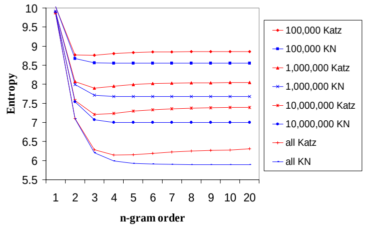
  - Lower order models take a special form

$$P'(w) \propto |w' : c(w,w') > 0|$$

- KN smoothing repeatedly proven effective
  - But we've never been quite sure why
  - And therefore never known how to make it better
- [Teh, 2006] shows KN smoothing is a kind of approximate inference in a hierarchical Pitman-Yor process (and better approximations are superior to basic KN)

Data >> Method?

- Having more data is always better…



- … but so is using a better model
- Another issue: N > 3 has huge costs in speech recognizers

Beyond N-Gram LMs

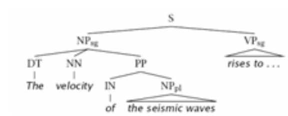
- Lots of ideas we won't have time to discuss:
  - Caching models: recent words more likely to appear again
  - Trigger models: recent words trigger other words
  - Topic models
- A few recent ideas I'd like to highlight
  - Syntactic models: use tree models to capture long-distance syntactic effects [Chelba and Jelinek, 98]
  - Discriminative models: set n-gram weights to improve final task accuracy rather than fit training set density [Roark, 05, for ASR; Liang et. al., 06, for MT]
  - Structural zeros: some n-grams are syntactically forbidden, keep estimates at zero [Mohri and Roark, 06]

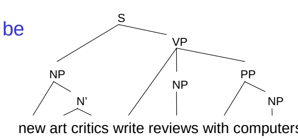
Outline

- Language Modeling
- Syntactic / Semantic Parsing
- Machine Translation
- Information Extraction
- Unsupervised Learning

Phrase Structure Parsing

- Phrase structure parsing organizes syntax into *constituents* or *brackets*
- In general, this involves nested trees
- Linguists can, and do, argue about what the gold structures should be
- Lots of ambiguity
- Not the only kind of syntax…





Syntactic Ambiguities

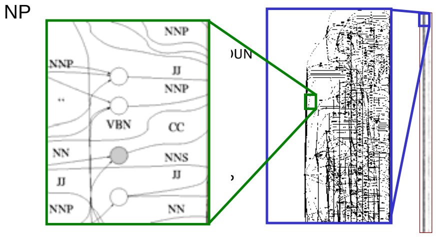
- Prepositional phrases:
  They cooked the beans in the pot on the stove with handles.
- Particle vs. preposition:
  The puppy tore up the staircase.
- Complement structures
  The tourists objected to the guide that they couldn't hear.
- Gerund vs. participial adjective
  Visiting relatives can be boring.
- Many more ambiguities
- Note that most incorrect parses are structures which are permitted by the grammar but not salient to a human listener like the examples above

Probabilistic Context-Free Grammars

- A context-free grammar is a tuple *<N, T, S, R>*
  - *N* : the set of non-terminals
    - Phrasal categories: S, NP, VP, ADJP, etc.
    - Parts-of-speech (pre-terminals): NN, JJ, DT, VB
  - *T* : the set of terminals (the words)
  - *S* : the start symbol
    - Often written as ROOT or TOP
    - *Not* usually the sentence non-terminal S
  - *R* : the set of rules
    - Of the form $X \rightarrow Y_1 Y_2 \ldots Y_k$, with $X, Y_i \in N$
    - Examples: $S \rightarrow NP\ VP$, $VP \rightarrow VP\ CC\ VP$
    - Also called rewrites, productions, or local trees
- A PCFG adds:
  - A top-down production probability per rule $P(Y_1 Y_2 \ldots Y_k \mid X)$

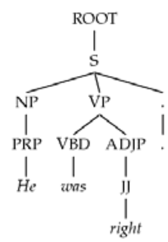
Treebank Grammar Scale

- Treebank grammars can be enormous
  - As FSAs, the raw grammar has ~10K states, excluding the lexicon
  - Better parsers usually make the grammars larger, not smaller.



Treebank Parsing

- Typically get grammars (and parameters) from a treebank of parsed sentences

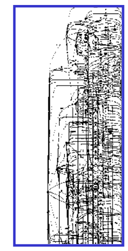


ROOT → S

S → NP VP .

NP → PRP

VP → VBD ADJ

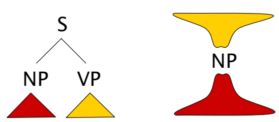


PCFGs and Independence

- Symbols in a PCFG imply conditional independence:

S → NP VP

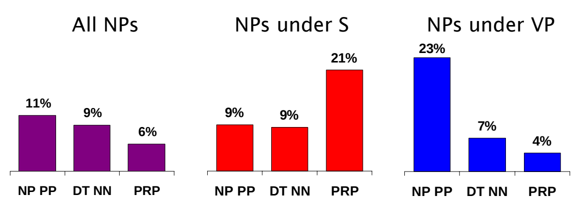
NP → DT NN



- At any node, the productions inside that node are independent of the material outside that node, given the label of that node.
- Any information that statistically connects behavior inside and outside a node must be encoded into that node's label.

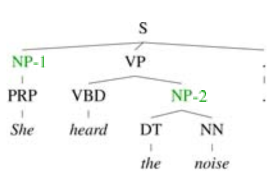
Non-Independence

- Independence assumptions are often too strong.



- Example: the expansion of an NP is highly dependent on the parent of the NP (i.e., subjects vs. objects).
- Also: the subject and object expansions are correlated

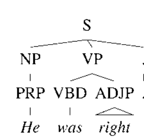
The Game of Designing a Grammar



- Symbol refinement can improve fit of the grammar
  - Parent annotation [Johnson '98]
  - Head lexicalization [Collins '99, Charniak '00]
  - Automatic clustering [Matsuzaki 05, Petrov et. al. 06]

Manual Annotation

- Manually split categories
  - Examples:
    - NP: subject vs object
    - DT: determiners vs demonstratives
    - IN: sentential vs prepositional

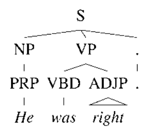


- Fairly compact grammar
- Linguistic motivations

Model	F1
Naïve Treebank Grammar	72.6
Klein & Manning '03	86.3

Automatic Annotation Induction

- Advantages:
  - Automatically learned:
    Label *all* nodes with latent variables.
    Same number *k* of subcategories for all categories.
- Disadvantages:
  - Grammar gets too large
  - Most categories are oversplit while others are undersplit.



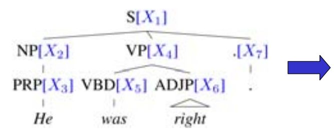
[Matsuzaki et. al '05, Prescher '05]

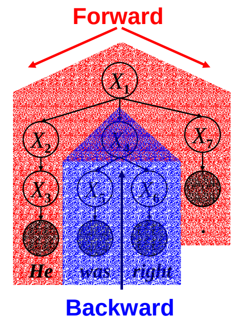
Model	*F1*
Klein & Manning '03	86.3
Matsuzaki et al. '05	86.7

Learning Latent Annotations

EM algorithm:

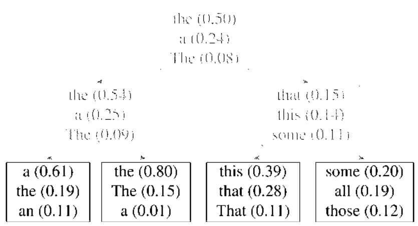
- Brackets are known
- Base categories are known
- Only induce subcategories





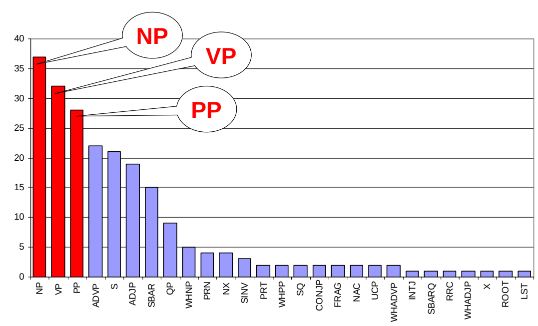
Just like Forward-Backward for HMMs.

Hierarchical Split / Merge



Model	*F1*
Matsuzaki et al. '05	86.7
Petrov et. al. 06	90.2

Number of Phrasal Subcategories



Linguistic Candy

- Proper Nouns (NNP):

NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street

- Personal pronouns (PRP):

PRP-0	It	He	I
PRP-1	it	he	they
PRP-2	it	them	him

Linguistic Candy

- Relative adverbs (RBR):

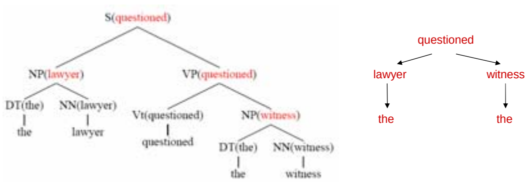
RBR-0	further	lower	higher
RBR-1	more	less	More
RBR-2	earlier	Earlier	later

- Cardinal Numbers (CD):

CD-7	one	two	Three
CD-4	1989	1990	1988
CD-11	million	billion	trillion
CD-0	1	50	100
CD-3	1	30	31
CD-9	78	58	34

Dependency Parsing

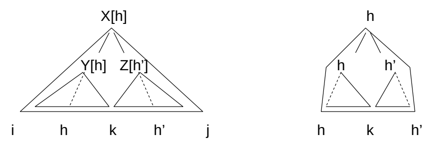
- Lexicalized parsers can be seen as producing *dependency trees*



- Each local binary tree corresponds to an attachment in the dependency graph

Dependency Parsing

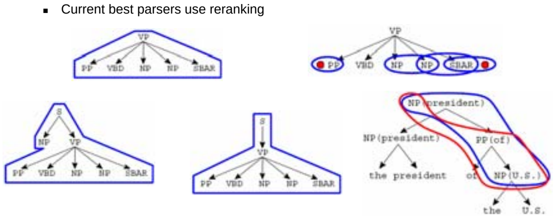
- Pure dependency parsing is only cubic [Eisner 99]



- Some work on *non-projective* dependencies
  - Common in, e.g. Czech parsing
  - Can do with MST algorithms [McDonald and Pereira, 05]

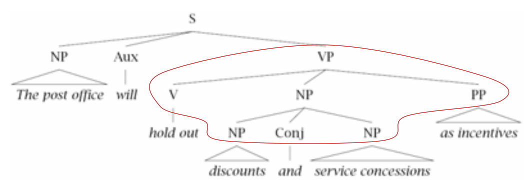
Parse Reranking

- Assume the number of parses is very small
- We can represent each parse T as an arbitrary feature vector φ(T)
  - Typically, all local rules are features
  - Also non-local features, like how right-branching the overall tree is
  - [Charniak and Johnson 05] gives a rich set of features
  - Can use most any ML techniques
  - Current best parsers use reranking



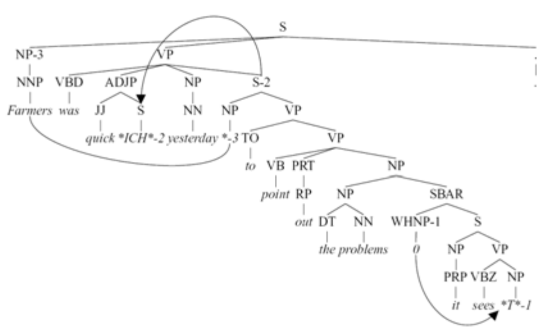
Tree Insertion Grammars

- Rewrite large (possibly lexicalized) subtrees in a single step [Bod 95]



- Derivational ambiguity whether subtrees were generated atomically or compositionally
- Most probable *parse* is NP-complete
- Common problem: ML estimates put all mass on large rules, and simple priors don't adequately fix the problem

Non-CF Phenomena



Semantic Role Labeling (SRL)

- Want to know more than which NP is a verb's subject:

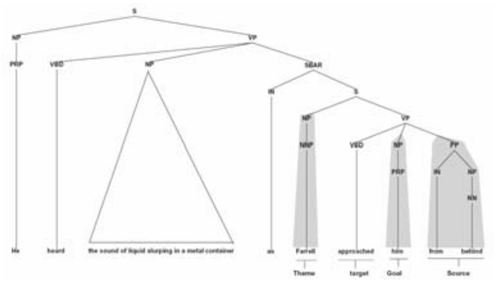
[$_{Judge}$ She] **blames** [$_{Evaluee}$ the Government] [$_{Reason}$ for failing to do enough to help] .

Holman would characterise this as **blaming** [$_{Evaluee}$ the poor] .

The letter quotes Black as saying that [$_{Judge}$ white and Navajo ranchers] misrepresent their livestock losses and **blame** [$_{Reason}$ everything] [$_{Evaluee}$ on coyotes] .

- Typical pipeline:
  - Parse then label roles
  - Almost all errors in parsing
  - Really, SRL is quite a lot easier than parsing

SRL Example



Propbank / FrameNet

- FrameNet: roles shared between verbs
- PropBank: each verb has its own roles
- PropBank more used, because it's layered over the treebank (and so has greater coverage, plus parses)
- Note: some linguistic theories postulate even fewer roles than FrameNet (e.g. 5-20 total: agent, patient, instrument, etc.)

Outline

- Language Modeling
- Syntactic / Semantic Parsing
- Machine Translation
- Information Extraction
- Unsupervised Learning

Machine Translation: Examples

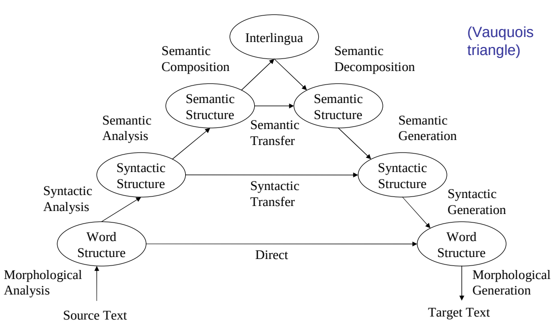
Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

Levels of Transfer



(Vauquois triangle)

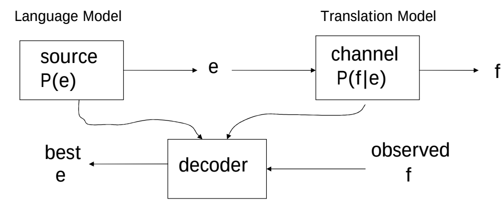
General Approaches

- Rule-based approaches
  - Expert system-like rewrite systems
  - Deep transfer methods (analyze and generate)
  - Lexicons come from humans
  - Can be very fast, and can accumulate a lot of knowledge over time (e.g. Systran)
- Statistical approaches
  - Word-to-word translation
  - Phrase-based translation
  - Syntax-based translation (tree-to-tree, tree-to-string)
  - Trained on parallel corpora
  - Usually noisy-channel (at least in spirit)

The Coding View

- "One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' "
  - Warren Weaver (1955:18, quoting a letter he wrote in 1947)

MT System Components



$$\underset{e}{\operatorname{argmax}} P(e|f) = \underbrace{\underset{e}{\operatorname{argmax}} P(f|e)P(e)}$$

Finds an English translation which is both fluent and semantically faithful to the foreign source

Pipeline of an MT System

- Data processing
- Sentence alignment
- Word alignment
- Transfer rule extraction
- Decoding

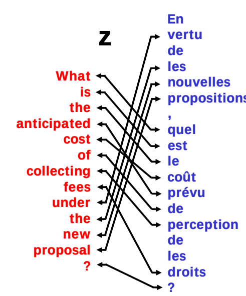
Word Alignment

x

What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

z

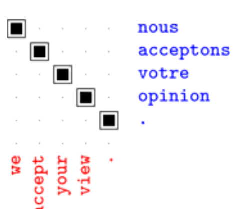


Unsupervised Word Alignment

- Input: a ***bitext***: pairs of translated sentences

```
nous acceptons votre opinion .
we accept your view .
```

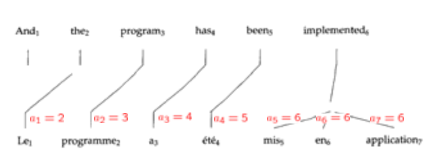
- Output: ***alignments***: pairs of translated words
  - When words have unique sources, can represent as a (forward) alignment function a from French to English positions



IBM Model 1 [Brown et al, 93]

- Alignments: a hidden vector called an *alignment* specifies which English source is responsible for each French target word.

$$a = a_1 \ldots a_J$$



$$P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)$$

$$= \prod_j \frac{1}{I+1} P(f_j|e_i)$$

Examples: Translation and Fertility

the

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
le	0.497	1	0.746
la	0.207	0	0.254
les	0.155		
l'	0.086		
ce	0.018		
cette	0.011		

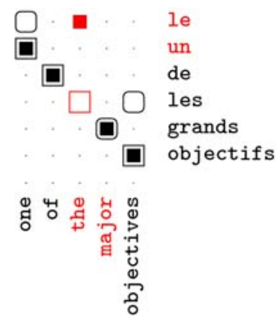
not

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
ne	0.497	2	0.735
pas	0.442	0	0.154
non	0.029	1	0.107
rien	0.011		

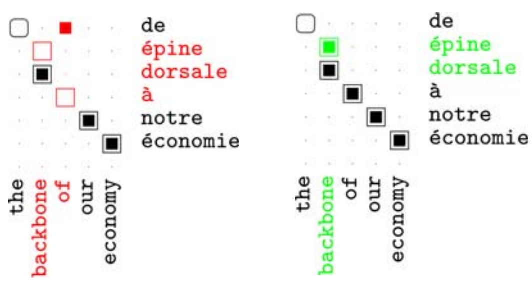
farmers

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
agriculteurs	0.442	2	0.731
les	0.418	1	0.228
cultivateurs	0.046	0	0.039
producteurs	0.021		

Example Errors



Fertility example



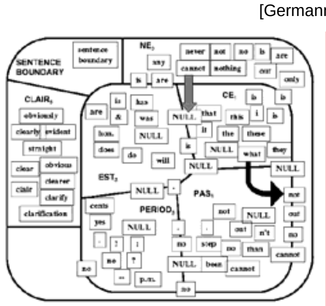
Decoding

- In these word-to-word models
  - Finding best alignments is easy
  - Finding translations (decoding) is hard

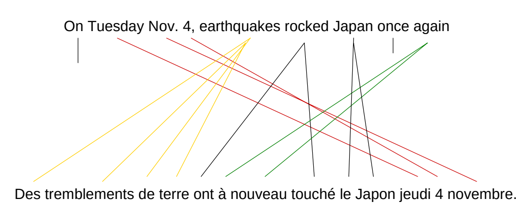
```
it is not clear .
|   \  |     \   \
|    \ +      \   \
|     \/ \     \   \
|     /\  \     \   \
CE NE EST PAS CLAIR  .
```

IBM Decoding as a TSP

[Germann et al, 01]



Phrase Movement

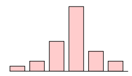


The HMM Alignment Model

- The HMM model (Vogel 96)

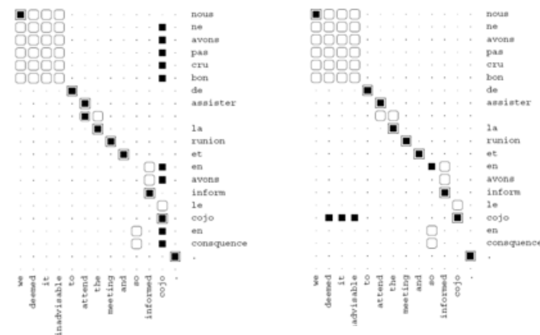
f	t(f | e)
nationale	0.469
national	0.418
nationaux	0.054
nationales	0.029

$$P(f, a|e) = \prod_j P(a_j|a_{j-1})P(f_j|e_i)$$



  - Re-estimate using the forward-backward algorithm
  - Handling nulls requires some care
- Note: alignments are not provided, but induced

HMM Examples

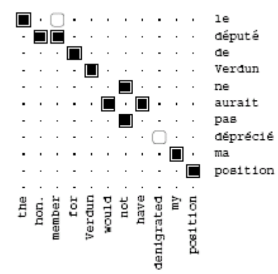


Intersection of HMMs

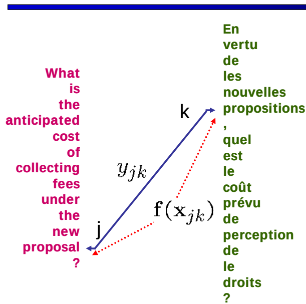
- Better alignments from intersecting directional results
- Still better if you train the two directional models to agree [Liang et. al., 06]

Model	AER
Model 1 INT	19.5
HMM E→F	11.4
HMM F→E	10.8
HMM AND	7.1
HMM INT	4.7
GIZA M4 AND	6.9

Complex Configurations



Feature-Based Alignment

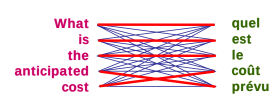


$score(x_{jk}) = w^\top f(x_{jk})$

$f(x_{jk})$

Features:
- Association
  - MI = 3.2
  - Dice = 4.1
- Lexical pair
  - ID(proposal, proposition) = 1
- Position
  - AbsDist = 5
  - RelDist = 0.3
- Orthography
  - ExactMatch = 0
  - Similarity = 0.8
- Neighborhood
  - Next pair Dice
- Resources
  - PairInDictionary
  - POS Tags Match
- IBM Models

Finding Viterbi Alignments



$$score(x, y) = \sum_{jk \in y} score(x_{jk})$$

- Complete bipartite graph
- Maximum score matching with node degree ≤ 1

$$y = \arg\max_{y' \in \mathcal{Y}} score(x, y') = \arg\max_{y' \in \mathcal{Y}} w^\top f(x, y')$$

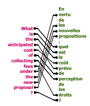
⇒ Weighted bipartite matching problem

[Lacoste-Julien, Taskar, Jordan, and Klein, 05]

Learning **w**

- Supervised training data

$(\mathbf{x}^i, \mathbf{y}^i)$



- Training methods
  - Maximum likelihood/entropy
  - Perceptron
  - Maximum margin

[Lacoste-Julien, Taskar, Jordan, and Klein, 05]

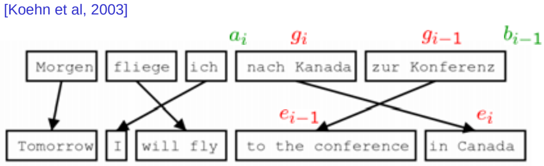
Problem: Idioms

nodding

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
signe	0.164	4	0.342
la	0.123	3	0.293
tête	0.097	2	0.167
oui	0.086	1	0.163
fait	0.073	0	0.023
que	0.073		
hoche	0.054		
hocher	0.048		
faire	0.030		
me	0.024		
approuve	0.019		
qui	0.019		
un	0.012		
faites	0.011		

A Phrase-Based Model

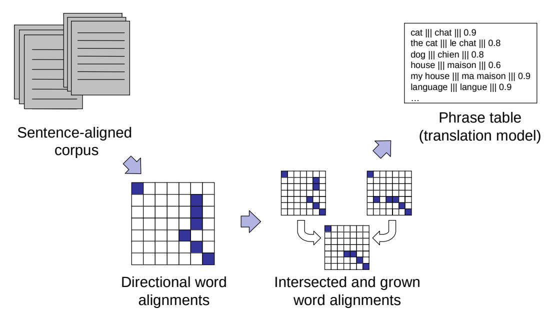
[Koehn et al, 2003]



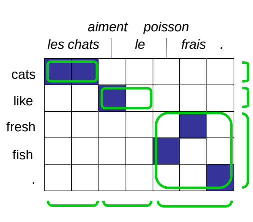
$$P(e|g) = P(\{\bar{g}_i\}|g) \prod_i \phi(\bar{e}_i|\bar{g}_i) d(a_i - b_{i-1})$$

Segmentation Translation Distortion

Overview: Extracting Phrases



Phrase Scoring



- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - … and others
- Seems not to work, for a variety of only partially understood reasons

Phrase-Based Decoding

Some Output

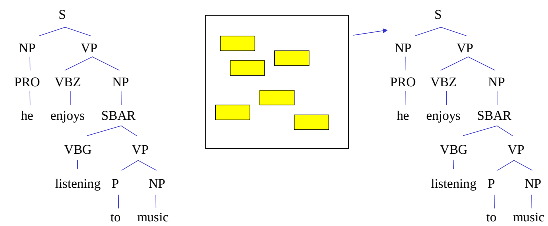
Madame la présidente, votre présidence de cette institution a été marquante.
Mrs Fontaine, your presidency of this institution has been outstanding.
Madam President, president of this house has been discoveries.
Madam President, your presidency of this institution has been impressive.

Je vais maintenant m'exprimer brièvement en irlandais.
I shall now speak briefly in Irish .
I will now speak briefly in Ireland .
I will now speak briefly in Irish .

Nous trouvons en vous un président tel que nous le souhaitons.
We think that you are the type of president that we want.
We are in you a president as the wanted.
We are in you a president as we the wanted.

Top-Down Tree Transducers

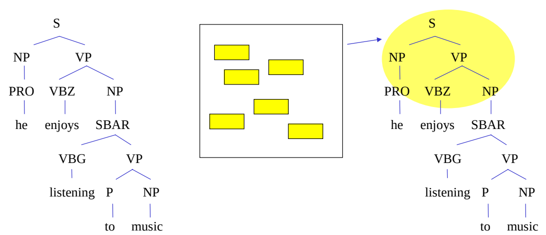
Original input: Transformation:



[Next slides from Kevin Knight]

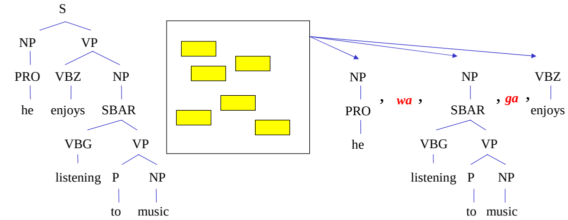
Top-Down Tree Transducers

Original input: Transformation:



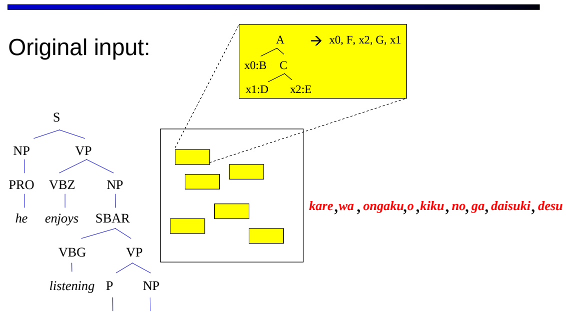
Top-Down Tree Transducers

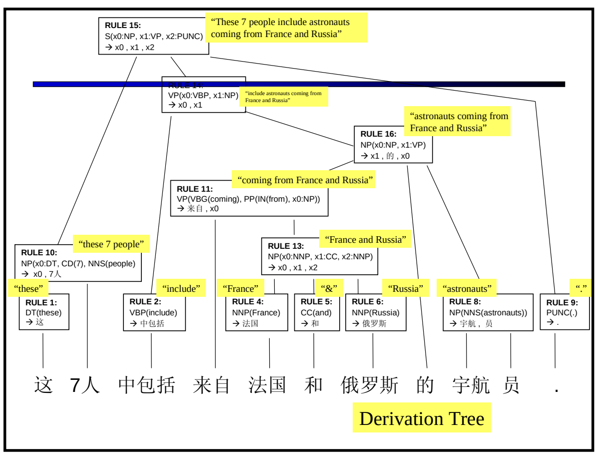
Original input: Transformation:



Top-Down Tree Transducers

Original input:





Outline

- Language Modeling
- Syntactic / Semantic Parsing
- Machine Translation
- Information Extraction
- Unsupervised Learning

Reference Resolution

- Noun phrases refer to entities in the world, many pairs of noun phrases co-refer:

John Smith, CFO of Prime Corp. since 1986,

saw his pay jump 20% to $1.3 million

as the 57-year-old also became

the financial services co.'s president.

Kinds of Reference

- Referring expressions
  - *John Smith*
  - *President Smith*
  - *the president*
  - *the company's new executive*
- Free variables
  - Smith saw *his pay* increase
- Bound variables
  - Every company trademarks its name.

More common in newswire, generally harder in practice

More interesting grammatical constraints, more linguistic theory, easier in practice

Grammatical Constraints

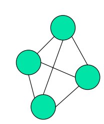
- Gender / number
  - Jack gave Mary a gift. She was excited.
  - Mary gave her mother a gift. She was excited.
- Position (cf. binding theory)
  - The company's board polices itself / it.
  - Bob thinks Jack sends email to himself / him.
- Direction (anaphora vs. cataphora)
  - She bought a coat for Amy.
  - In her closet, Amy found her lost coat.

Other Constraints

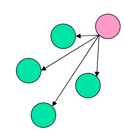
- Recency
- Salience
- Focus
- Centering Theory [Grosz et al. 86]
- Style / Usage Patterns
  - *Peter Watters* was named CEO. *Watters'* promotion came six weeks after his brother, *Eric Watters*, stepped down.
- Semantic Compatibility
  - Smith had bought *a used car* that morning. *The used car dealership* assured him it was in good condition.

Two Kinds of Models

- Mention Pair models
  - Treat coreference chains as a collection of pairwise links
  - Make independent pairwise decisions and reconcile them in some way (e.g. clustering or greedy partitioning)



- Entity-Mention models
  - A cleaner, but less studied, approach
  - Posit single underlying entities
  - Each mention links to a discourse entity [Pasula et al. 03], [Luo et al. 04]



Two Paradigms for NLP

Supervised Learning

Unsupervised Learning

Parts-of-Speech

- Syntactic classes of words
  - Useful distinctions vary from language to language
  - Tagsets vary from corpus to corpus [See M+S p. 142]
- Some tags from the Penn tagset

CD	numeral, cardinal	mid-1890 nine-thirty 0.5 one
DT	determiner	a all an every no that the
IN	preposition or conjunction, subordinating	among whether out on by if
JJ	adjective or numeral, ordinal	third ill-mannered regrettable
MD	modal auxiliary	can may might will would
NN	noun, common, singular or mass	cabbage thermostat investment subhumanity
NNP	noun, proper, singular	Motown Cougar Yvette Liverpool
PRP	pronoun, personal	hers himself it we them
RB	adverb	occasionally maddeningly adventurously
RP	particle	aboard away back by on open through
VB	verb, base form	ask bring fire see take
VBD	verb, past tense	pleaded swiped registered saw
VBN	verb, past participle	dilapidated imitated reunified unsettled
VBP	verb, present tense, not 3rd person singular	twist appear comprise mold postpone

Part-of-Speech Ambiguity

- Example

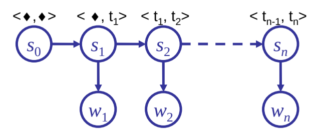
VBD		VB			
VBN	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN
Fed	raises	interest	rates	0.5	percent

- Two basic sources of constraint:
  - Grammatical environment
  - Identity of the current word

HMMs for Tagging

$$P(T,W) = \prod_i P(t_i \mid t_{i-1}, t_{i-2}) P(w_i \mid t_i)$$

$$P(T,W) = \prod_i P(s_i \mid s_{i-1}) P(w_i \mid s_i)$$



Domain Effects

- Accuracies degrade outside of domain
  - Up to triple error rate
  - Usually make the most errors on the things you care about in the domain (e.g. protein names)
- Open questions
  - How to effectively exploit unlabeled data from a new domain (what could we gain?)
  - How to best incorporate domain lexica in a principled way (e.g. UMLS specialist lexicon, ontologies)

Merialdo: Setup

- Some (discouraging) experiments [Merialdo 94]
- Setup:
  - You know the set of allowable tags for each word
  - Fix k training examples to their true labels
    - Learn initial P(w|t) on these examples
    - Learn initial P(t|t_{-1},t_{-2}) on these examples
  - On n examples, re-estimate with EM
- Note: we know allowed tags but not frequencies

Merialdo: Results

	Number of tagged sentences used for the initial model						
	0	100	2000	5000	10000	20000	all
Iter	Correct tags (% words) after ML on 1M words						
0	77.0	90.0	95.4	96.2	96.6	96.9	97.0
1	80.5	92.6	95.8	96.3	96.6	96.7	96.8
2	81.8	93.0	95.7	96.1	96.3	96.4	96.4
3	83.0	93.1	95.4	95.8	96.1	96.2	96.2
4	84.0	93.0	95.2	95.5	95.8	96.0	96.0
5	84.8	92.9	95.1	95.4	95.6	95.8	95.8
6	85.3	92.8	94.9	95.2	95.5	95.6	95.7
7	85.8	92.8	94.7	95.1	95.3	95.5	95.5
8	86.1	92.7	94.6	95.0	95.2	95.4	95.4
9	86.3	92.6	94.5	94.9	95.1	95.3	95.3
10	86.6	92.6	94.4	94.8	95.0	95.2	95.2

Distributional Clustering

♦ *the president said that the downturn was over* ♦

president	the __ of
president	the __ said
governor	the __ of
governor	the __ appointed
said	sources __ ♦
said	president __ that
reported	sources __ ♦

president governor

the a

said reported

[Finch and Chater 92, Shuetze 93, many others]

Distributional Clustering

- Three main variants on the same idea:
  - Pairwise similarities and heuristic clustering
    - E.g. [Finch and Chater 92]
    - Produces dendrograms
  - Vector space methods
    - E.g. [Shuetze 93]
    - Models of ambiguity
  - Probabilistic methods
    - Various formulations, e.g. [Lee and Pereira 99]

Nearest Neighbors

word	nearest neighbors
accompanied	submitted banned financed developed authorized headed canceled awarded barred
almost	virtually merely formally fully quite officially just nearly only less
causing	reflecting forcing providing creating producing becoming carrying particularly
classes	elections courses payments losses computers performances violations levels pictures
directors	professionals investigations materials competitors agreements papers transactions
goal	mood roof eye image tool song pool scene gap voice
japanese	chinese iraqi american western arab foreign european federal soviet indian
represent	reveal attend deliver reflect choose contain impose manage establish retain
think	believe wish know realize wonder assume feel say mean bet
york	angeles francisco sox rouge kong diego zone vegas inning layer
on	through in at over into with from for by across
must	might would could cannot will should can may does helps
they	we you i he she nobody who it everybody there

What Else?

- Various newer ideas:
  - Context distributional clustering [Clark 00]
  - Morphology-driven models [Clark 03]
  - Contrastive estimation [Smith and Eisner 05]
- Also:
  - What about ambiguous words?
  - Using wider context signatures has been used for learning synonyms (what's wrong with this approach?)

Early Approaches: Structure Search

- Incremental grammar learning, chunking [Wolff 88, Langley 82, many others]
  - Can recover synthetic grammars
- An (extremely good) result of incremental structure search:

N-bar or zero determiner NP
zNN → NN | NNS
zNN → JJ zNN
zNN → zNN zNN

NP with determiner
zNP → DT zNN
zNP → PRP$ zNN

Proper NP
zNNP → NNP | NNPS
zNNP → zNNP zNNP

Transitive VPs
(complementation)
zVP → zV JJ
zVP → zV zNP
zVP → zV zNN
zVP → zV zPP

Transitive VPs
(adjunction)
zVP → zRB zVP
ZVP → zVP zPP

PP
zPP → zIN zNN
zPP → zIN zNP
zPP → zIN zNNP

verb groups / intransitive VPs
zV → VBZ | VBD | VBP
zV → MD VB
zV → MD RB VB
zV → zV zRB
zV → zV zVBG

Intransitive S
zS → PRP zV
zS → zNP zV
zS → zNNP zV

Transitive S
zSt → zNNP zVP
zSt → zNN zVP
zSt → PRP zVP

- Looks good, … but can't parse in the wild.

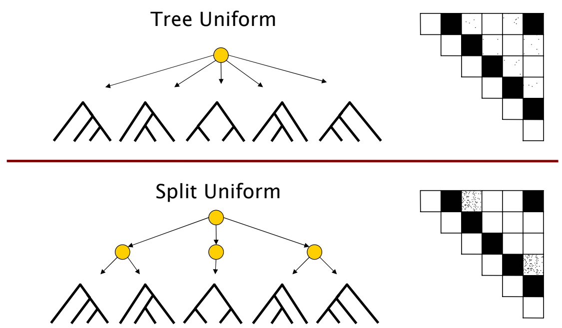
Idea: Learn PCFGs with EM

- Classic experiments on learning PCFGs with Expectation-Maximization [Lari and Young, 1990]

$\{ X_1 , X_2 \ldots X_n \}$ $\qquad$ $X_i \rightarrow X_j \; X_k$

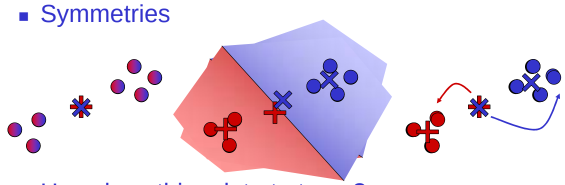
  - Full binary grammar over n symbols
  - Parse uniformly/randomly at first
  - Re-estimate rule expectations off of parses
  - Repeat
- Their conclusion: it doesn't really work.

Problem: "Uniform" Priors

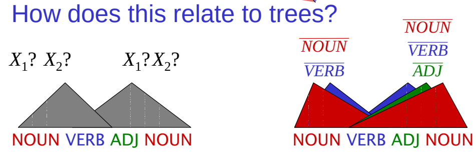


Problem: Model Symmetries

- Symmetries

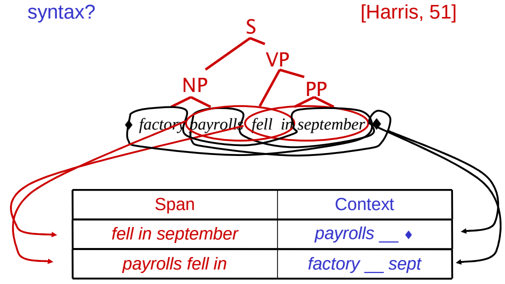


- How does this relate to trees?



Idea: Distributional Syntax?

- Can we use distributional clustering for learning syntax? [Harris, 51]

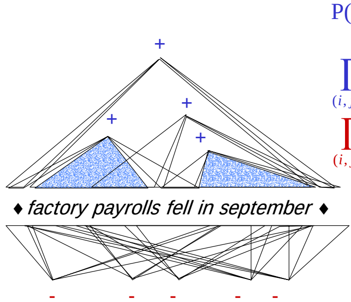


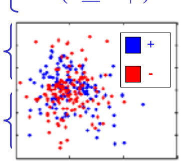
Span	Context
fell in september	*payrolls* __ ♦
payrolls fell in	*factory* __ *sept*

Constituent-Context Model (CCM)

$$P(S|T) = \prod_{(i,j)\in T} P(\phi_{ij}\,|+)P(\chi_{ij}\,|+) \prod_{(i,j)\notin T} P(\phi_{ij}\,|-)P(\chi_{ij}\,|-)$$

P(fpfis|+) P(♦__♦|+) P(♦__ *fell*|+)





Conclusions

- NLP includes many large-scale learning problems
  - Places constraints on what methods are possible
- Active interaction between the NLP and ML communities
  - Many cases where NLP could benefit from latest ML techniques (and does)
  - Also many cases where new ML ideas could come from empirical NLP observations and models
- Many NLP topics we haven't even mentioned
  - Check out the ACL and related proceedings, all online

References

- REFERENCE SECTION STILL TO COME