Syntax-based Language Modeling

April 12, 2012

many of today's examples were taken from
Syntactic Theory: A formal introduction, 2nd Ed *(Sag, Wasow, & Bender)*

Today's goals

- Review some issues with MT output
- Examine past approaches to incorporating syntax
  - ...in speech recognition
  - ...in machine translation
- Understand how linguists approach grammars and the critical ways standard CFGs differ from them
- Look into current language modeling work

Evaluating translation

- Adequacy (faithfulness): *was the meaning preserved?*
- **Fluency** (grammaticality): *is the sentence well-formed?*
- 我们 有 一个 共同 的 认识

	adequate	not adequate
fluent	*we have a common understanding*	*we do not agree*
disfluent	*have an agreement*	*them owning compatibility*

Poor grammar is common

MT output

- still to define who is the winner
- not to mention of the parades .
- certainly will not regret , because the clothes that feels perfectly is invaluable .
- begins a new era of crisis
- the study shows that in the families of obese children are consumed much more often the drink chips .
- survey to 900 children

human reference

- *it is time to define the winners .*
- *not to mention fashion shows .*
- *you will definitely not regret the investment , as perfectly fitting clothes are priceless .*
- *new era of crisis commences*
- *a survey has shown that fries are consumed more often in the families of obese children .*
- *the research was performed among 900 children .*

Poor grammar can obscure meaning

of games of this kind can not be expected that recreated with deformities and collisions complicated , but in fact before a coup against any object , you can not predict how will your car , so not everything is in order .

reference:
from a game of this type , one does not expect complicated deformations and collisions , but when you have no idea , before crashing into any object , how your car will act , something is not right .

Another example

not to stand in the passive listening and put something in place , we have learned of the suela shoes .

reference:

to have some change from listening , and gain some practical experience , we learned how to properly underlay shoe soles .

Why is the output so disfluent?

- **One reason:** we're not even modeling the grammar
- N-grams condition the probability of a word based on the previous n-1 words, but it is easy to show this is problematic:

  The dog bit the goat. P(bit | dog)

  The dog with the missing eye bit the goat P(bit | eye)

- With no concept of sentence structure (an intervening PP), the n-gram model fails here

Why is the output so disfluent?

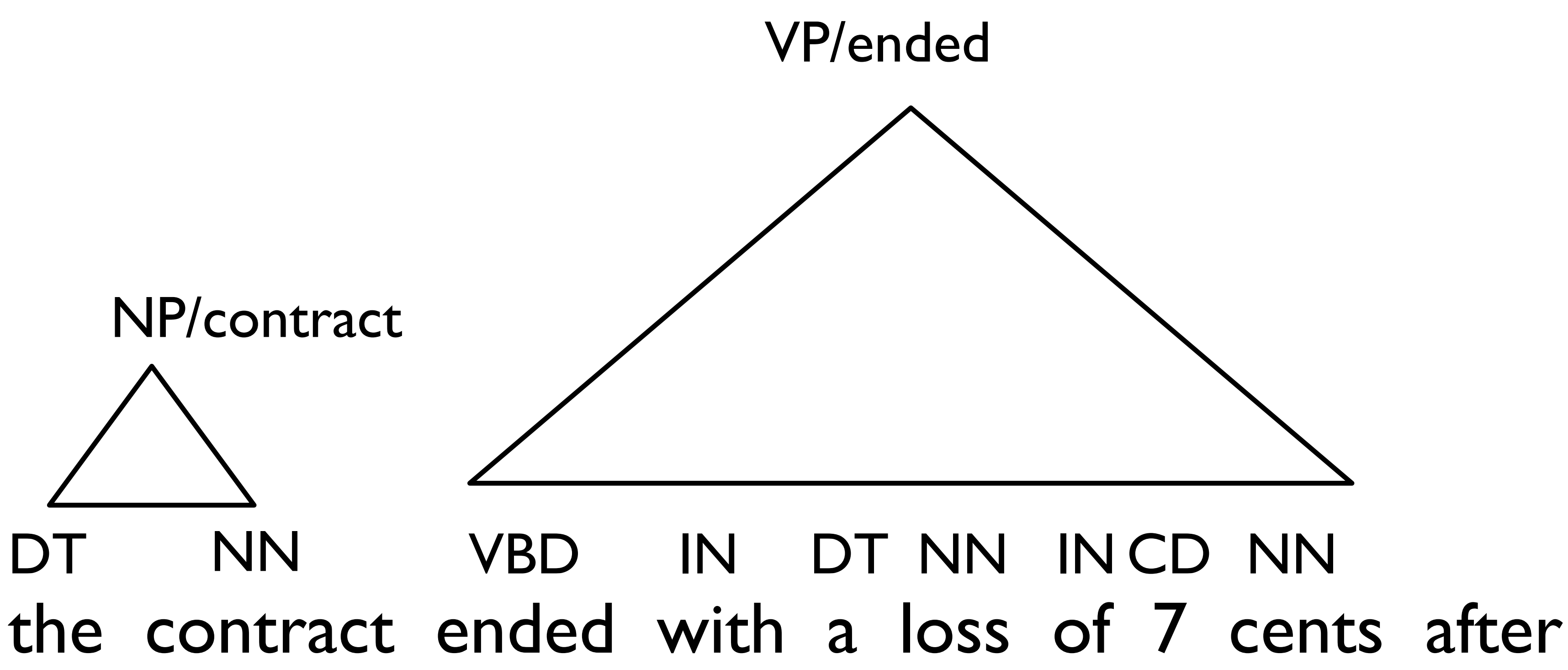
- Review: options for encoding languages
  - Lists
  - Regular expressions
  - Context-free grammars
  - *Context sensitive grammars*
  - *Unrestricted grammars*
- N-grams are essentially lists!
- So let's model structure!

Syntax-based LMs for ASR

- Speech recognition is like MT but without reordering
  - the translation model describes how acoustic signals get translated into phoneme and then words
  - the language model selects among the alternatives
- Since hypotheses are generated left-to-right, this integrates fairly naturally with ngrams.

Syntax-based LMs for ASR

- Chelba & Jelinek (1998) proposed a model that maintains constituents as part of the hypothesis representation
- When predicting words, we can now condition them on the labeled heads instead of just the previous few words



Syntax-based LMs for MT

- Charniak, Yamada, & Knight (2003): string-to-tree decoding
  - Words are translated and parsed at the same time
  - The dynamic programming forest is the *rescored* with the Charniak parser
- Charniak parser
  - state-of-the-art bilexical context-free parser

Bilexical parsing models

- So far, our CFG rules have looked like this:

  S → NP VP

- But this isn't nearly detailed enough. Why not?
- *Example on the board.*

Bilexical parsing models

- Annotates CFG productions with head words

  S → NP VP

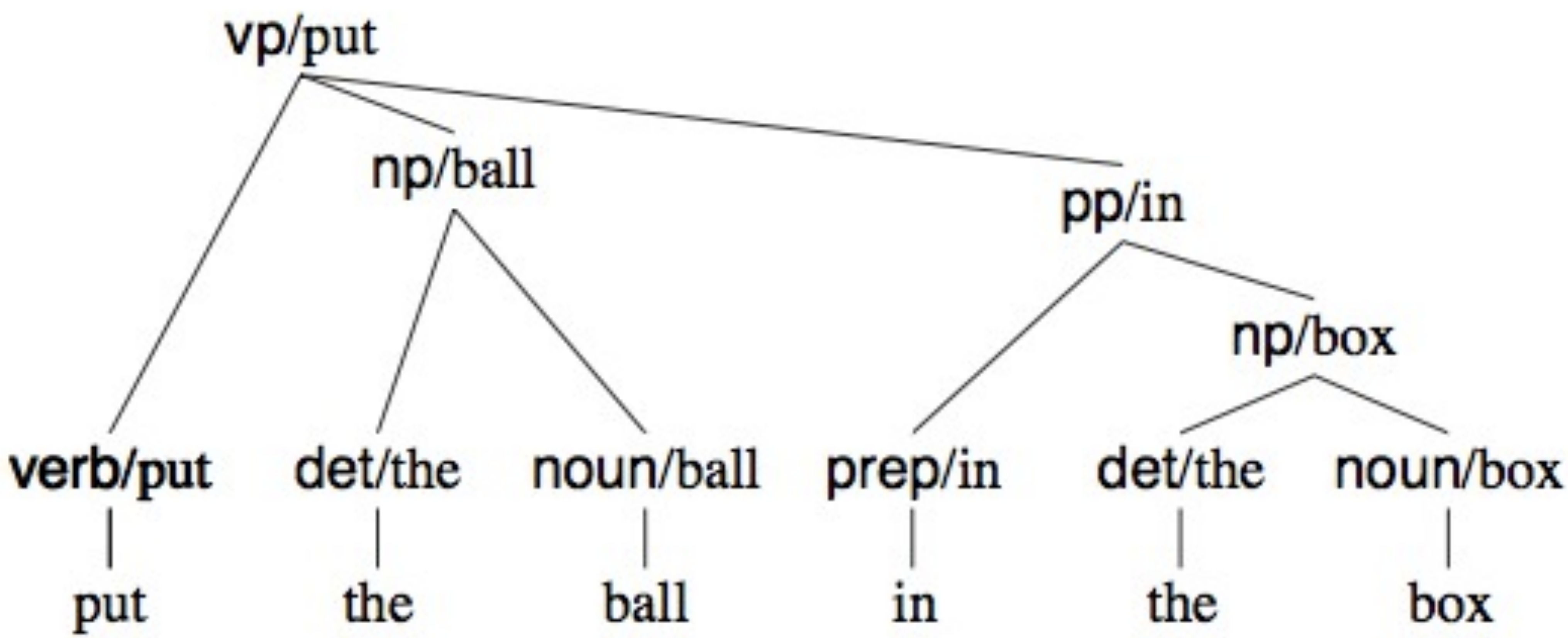
  becomes

  S/walked → NP/boy VP/walked

- Nonterminals are annotated with words that correspond to the constituent's head
- You can think of such models as supplementing normal CFG productions with long-distance bigrams
  - These bigrams capture head-argument relationships

An example

- Also called "immediate-head" parsing models
- Here's an example from Charniak (2001)



Charniak, Yamada, & Knight (2003)

- Part of the difficulty is a metric mismatch

System	Perfect Translation	Syntactically Correct but Semantically Wrong	Semantically Correct Syntactically Wrong	Wrong	BLEU
syntax TM + LM	45	67	70	164	0.0717
syntax TM only	31	19	87	209	0.1031
word-based	26	11	87	223	0.0722

- But that's not the whole story

General observations

- It is hugely expensive to incorporate syntax in this way
- The gains are marginal and come at huge expense
  - (papers rarely report running time or resource consumption)
- Part of the reason is search, but a big part of the reason is also the model

Samples

5-gram LM

- Grammars are supposed to define languages
- Which of these is a sample from an ngram model, and which from a CFG?
  - *the commissioner for labour , water transport the great hall of the people in beijing .*
  - *Wilson Protestantism Herald Of the fire settled $ 7.52 million " at financial reviews .*

latent variable PCFG (Petrov et al., 2006)

Syntax in language

- Studying the structure of a language is an interesting empirical task!
    - It treats *inherent, inscrutable linguistic judgments of native speakers* as the gold standard!

    It is April 12.

* It are April 12.

- Syntacticians form hypotheses about a language generalization and then test it by looking for examples and counterexamples

Syntax as science: An example

- * We like us.
  We like ourselves.
  She likes her.
  She likes herself.
  Nobody likes us.
  * Leslie likes ourselves.
- Hypothesis 1: *A reflexive pronoun can appear in a clause if that clause also contains a preceding coreferent expression.*

Example adapted from Sag, Wasow, & Bender, itself borrowed from David Perlmutter.

Syntax as science: An example

- Hypothesis 1: *A reflexive pronoun can appear in a clause if that clause also contains a preceding coreferent expression.*
- But what about:

  Our friends like us.
  * Our friends like ourselves.
  Those pictures of us offended us.
  * Those pictures of us offended ourselves.

- Hypothesis 2: *A reflexive pronoun must be an argument of a verb that has another preceding argument with the same referent.*

Example adapted from Sag, Wasow, & Bender, itself borrowed from David Perlmutter.

English linguistic phenomena

- What are some other facts about language that we would like to encode?

  Come up with a small list with your neighbor.

English linguistic phenomena

- Unbounded productivity
- Categories of words (noun, verb, preposition)
- Constraints on word order (* *taught Matt class)*
- High-level patterns (*subject-verb-object*)
- Agreement (*I eat,* * *I eats)*
- Predicate argument structure (*"give" is ditransitive)*
- Patterns of inflection (past: *verb* + *ed*; gerund: *verb* + *ing*)
- Noncompositional interpretations (*threw under the bus)*
- Exceptions (**The dog sleeped in the hallway*)

English linguistic phenomena

Phenomenon	ngrams	context-free grammars	immediate-head models
infinite	✓	✓	✓
word categories		✓	✓
word order		✓	✓
high-level patterns		✓	✓
agreement			✓
predicate-argument structure			
morphology			

Problems with the models

- There are still many phenomena not captured by these models
- The generative process assumes vastly more independence than is warranted
- Independence assumptions of parsers are too permissive

model	task	difficulties
parsers	discriminate *structures* (grammaticality assumed)	PP attachment, coordination
language models	discriminate *strings*	ensuring global coherence

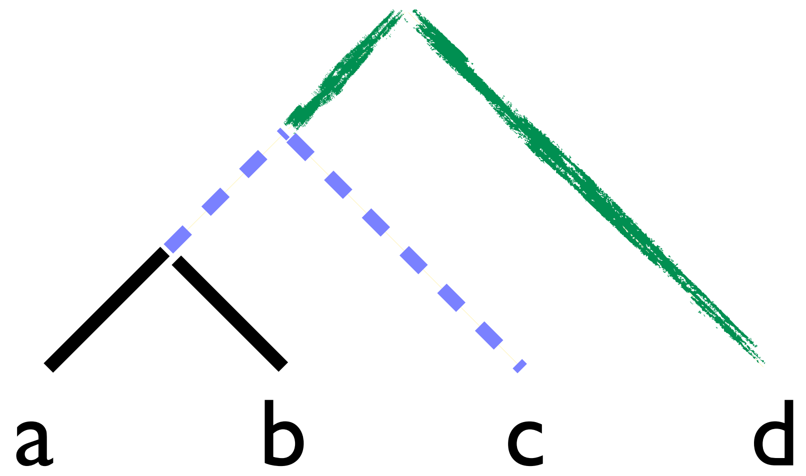
Current work

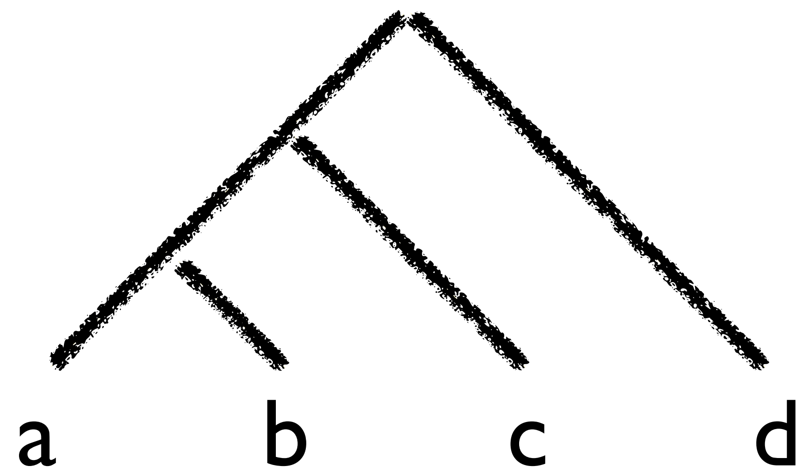
- Current work: extending the domain of locality
- Basic idea
  - Longer ngrams work by memorizing longer pieces of the text
  - The longer the ngram you use, the more likely it is that the text you are producing will be grammatical
- Apply the same idea to parse trees

In the meantime

- Desiderata
  - Inference no worse than it already is
  - Weak independence assumptions
- Search informed by grammar (so that grammatical candidates are not pruned)
- Syntax working as a language (and not a reordering) model

Is this sentence grammatical?

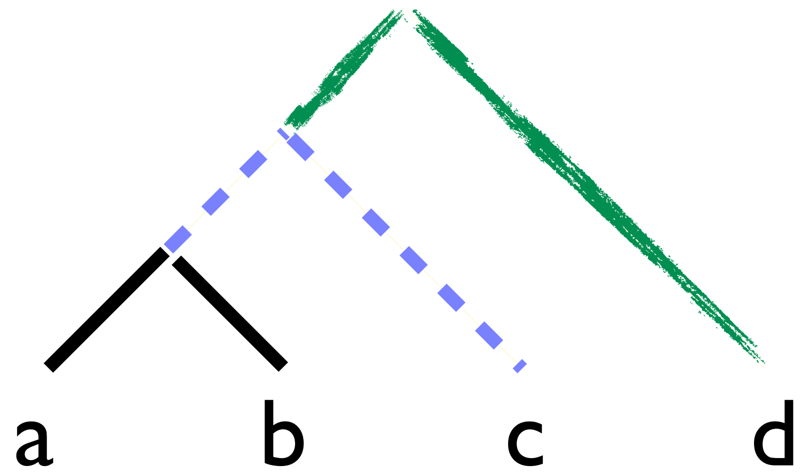


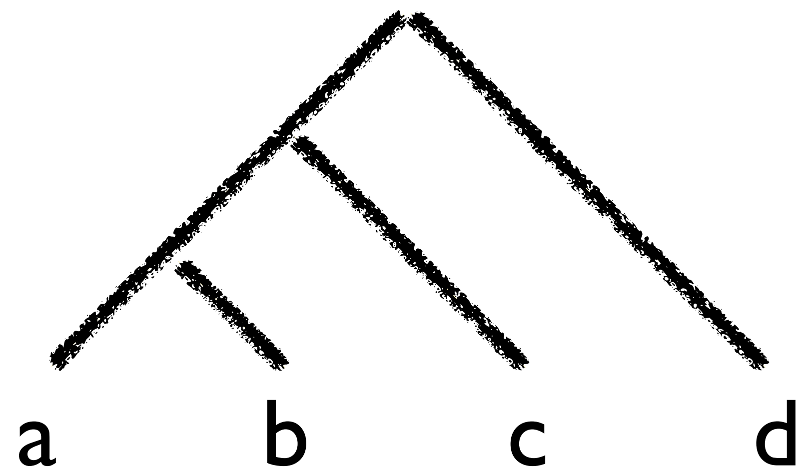


Is this sentence grammatical?

many little fragments *single large fragment*

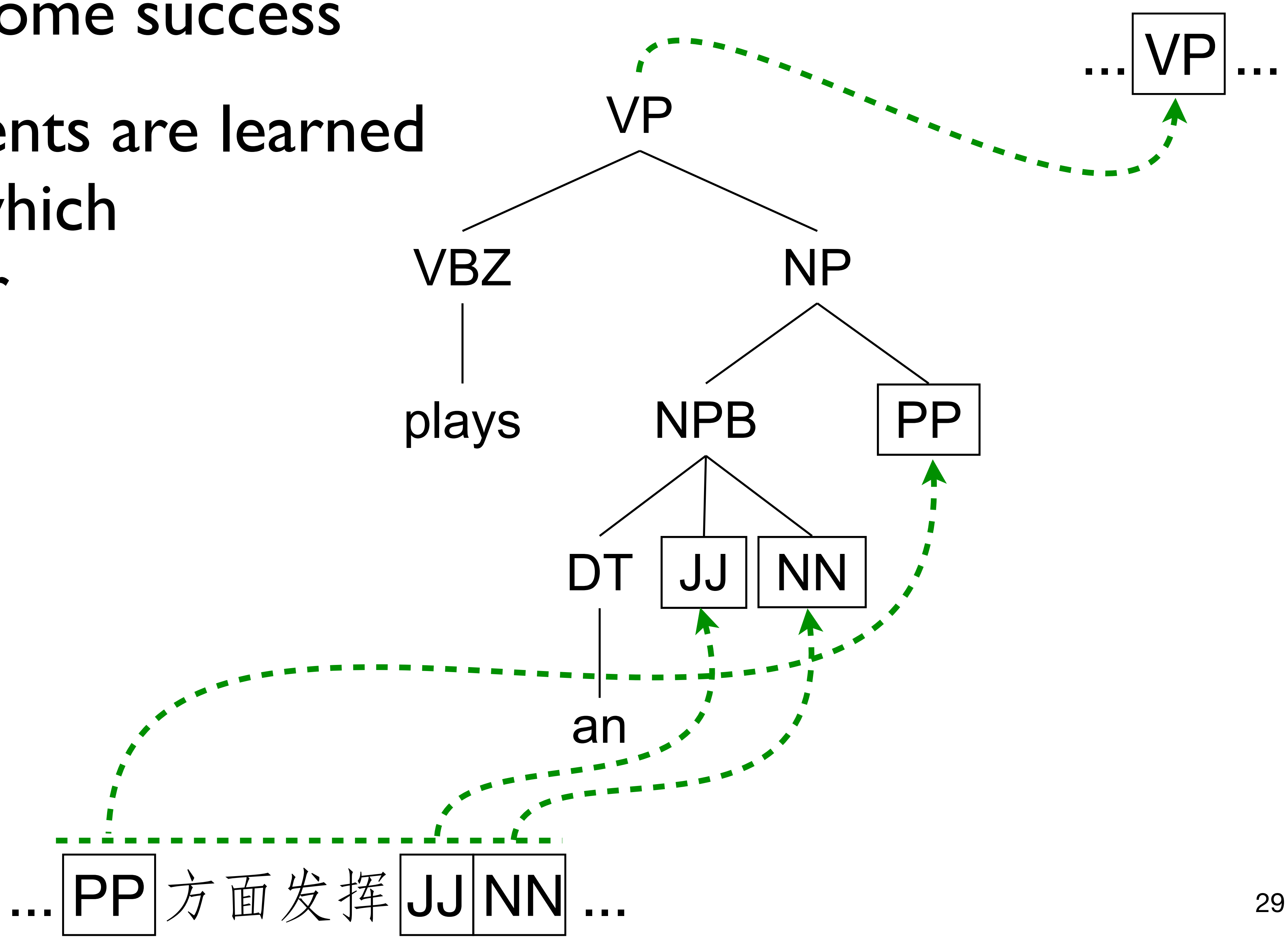
increased likelihood of grammaticality →





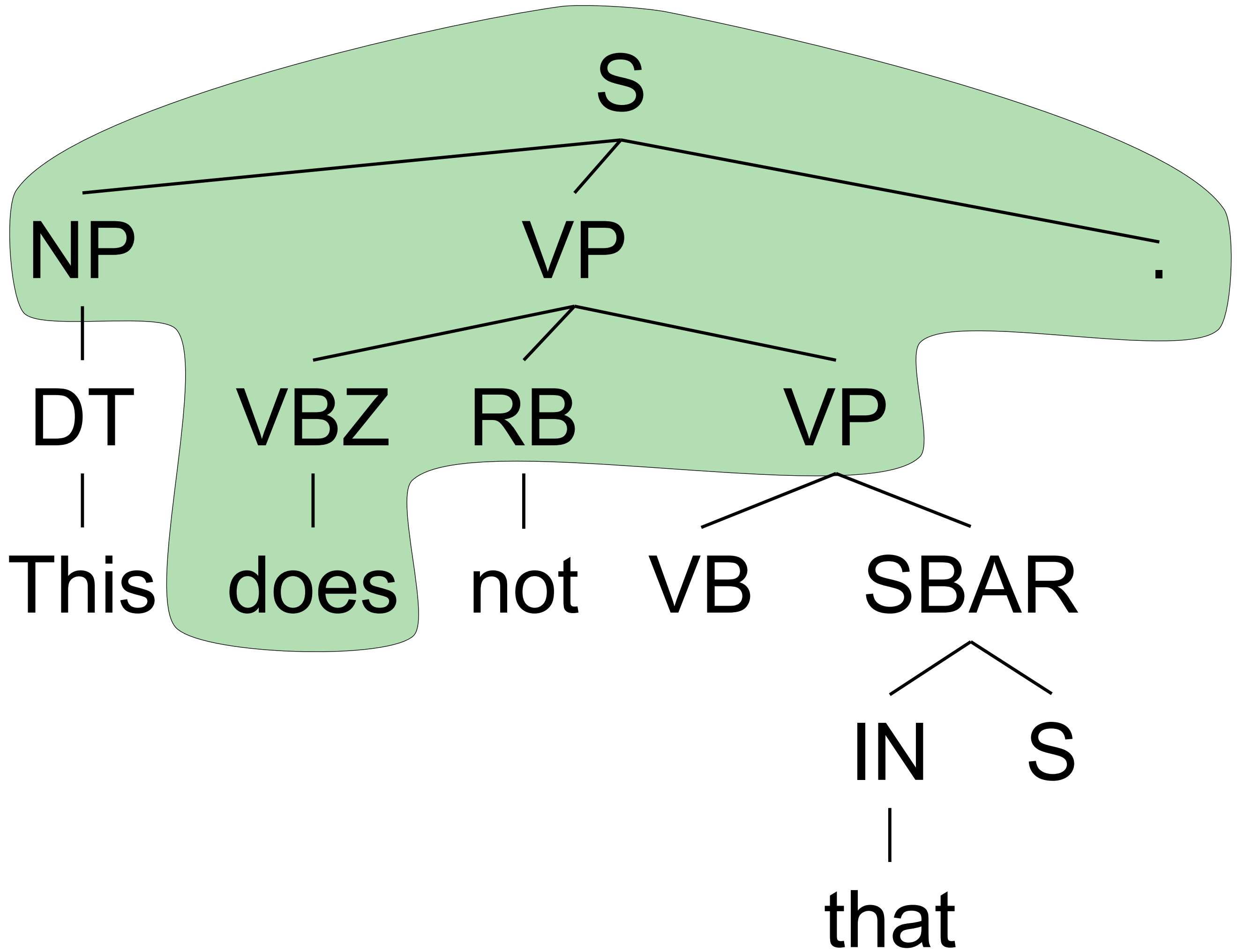
Tree substitution grammars

- This idea underlies translation approaches such as Galley et al. (2004, 2006), who use synchronous tree substitution grammars with some success
- But those fragments are learned for reordering, which complicates their utility as LMs



TSG example

- With TSGs, there is always a question of *what* fragments to use
    - With ngrams, we can just use all seen ones
- There are many techniques proposed for learning good fragments



A large hairy fragment and a more reasonable smaller one

Coarse language modeling

- It's difficult to incorporate syntax into search procedures
- We can evaluate the effectiveness of syntax on a much coarser level with a discriminative classification setup
  - Come up with positive and negative examples (grammatical and ungrammatical text)
  - Train models, see which ones do the best
- This should be an easier way to evaluate models

Two tasks

	positive	negative
coarse	WSJ text	samples from an n-gram model
MT	reference translations	machine translation output

Experimental setup

- Classification
  - L2-regularized support vector classifier (`liblinear`)
  - tune regularization tradeoff on development data
  - L1-regularization for feature reporting
- Tree kernels: SVM-TK toolkit, again tuned regularization parameter

Feature sets

feature set	example
length	*17*
Gigaword 5-gram LM score	*-12.045*
bigrams and trigrams	*"he further praised"*
CFG productions	S → NP VP .
Charniak & Johnson (2005) reranking features	*number of nodes in the parse tree* *head projections*
TSG (parse score, fragments, aggregate features)	(TOP (S NP (VP VBD said) NP SBAR) .)

Task 1: ngram samples from real text

The most troublesome report may be the August merchandise trade deficit due out tomorrow .

Good

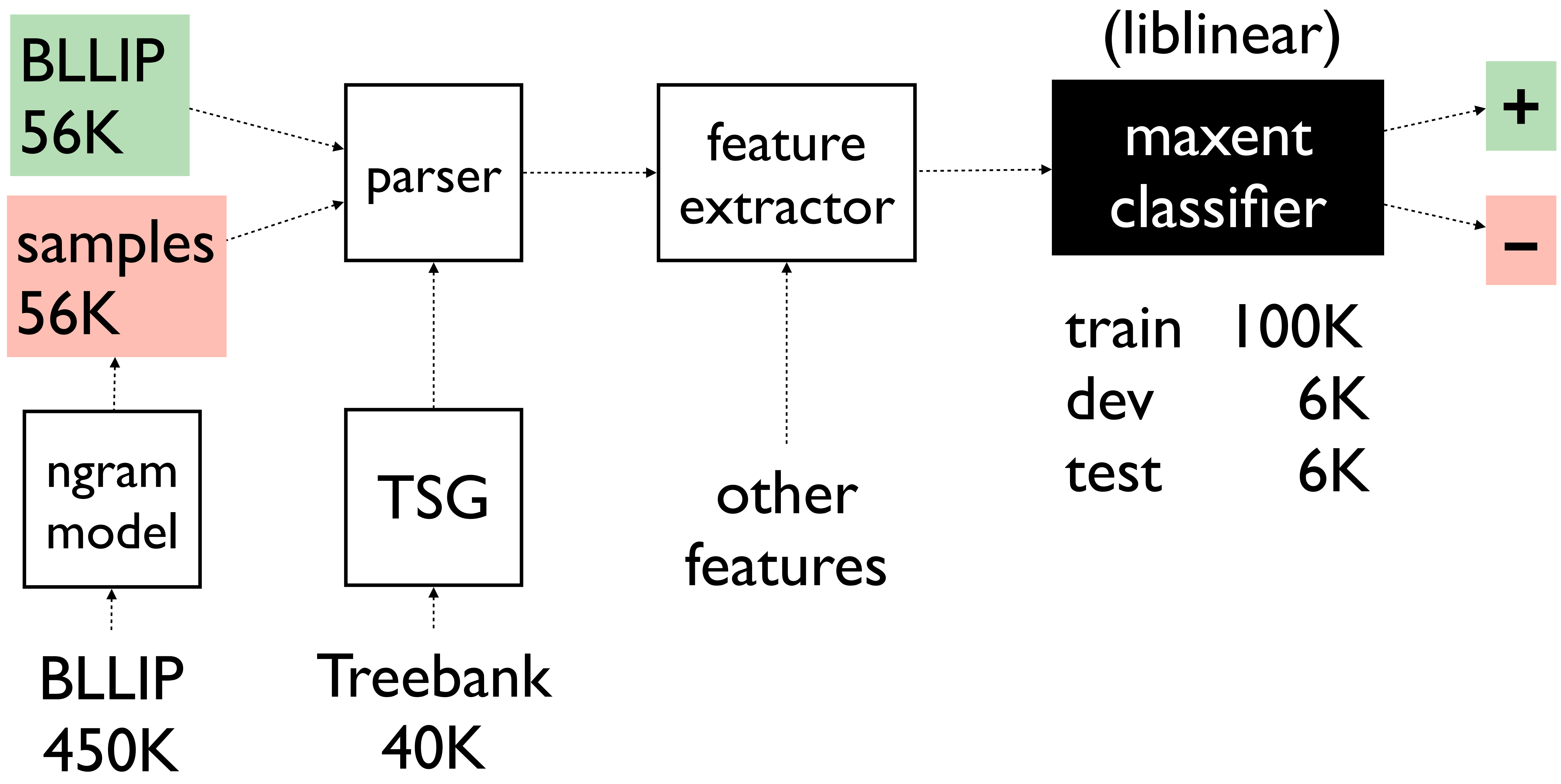
§24 #2

To and , would come Hughey Co. may be crash victims , three billion .

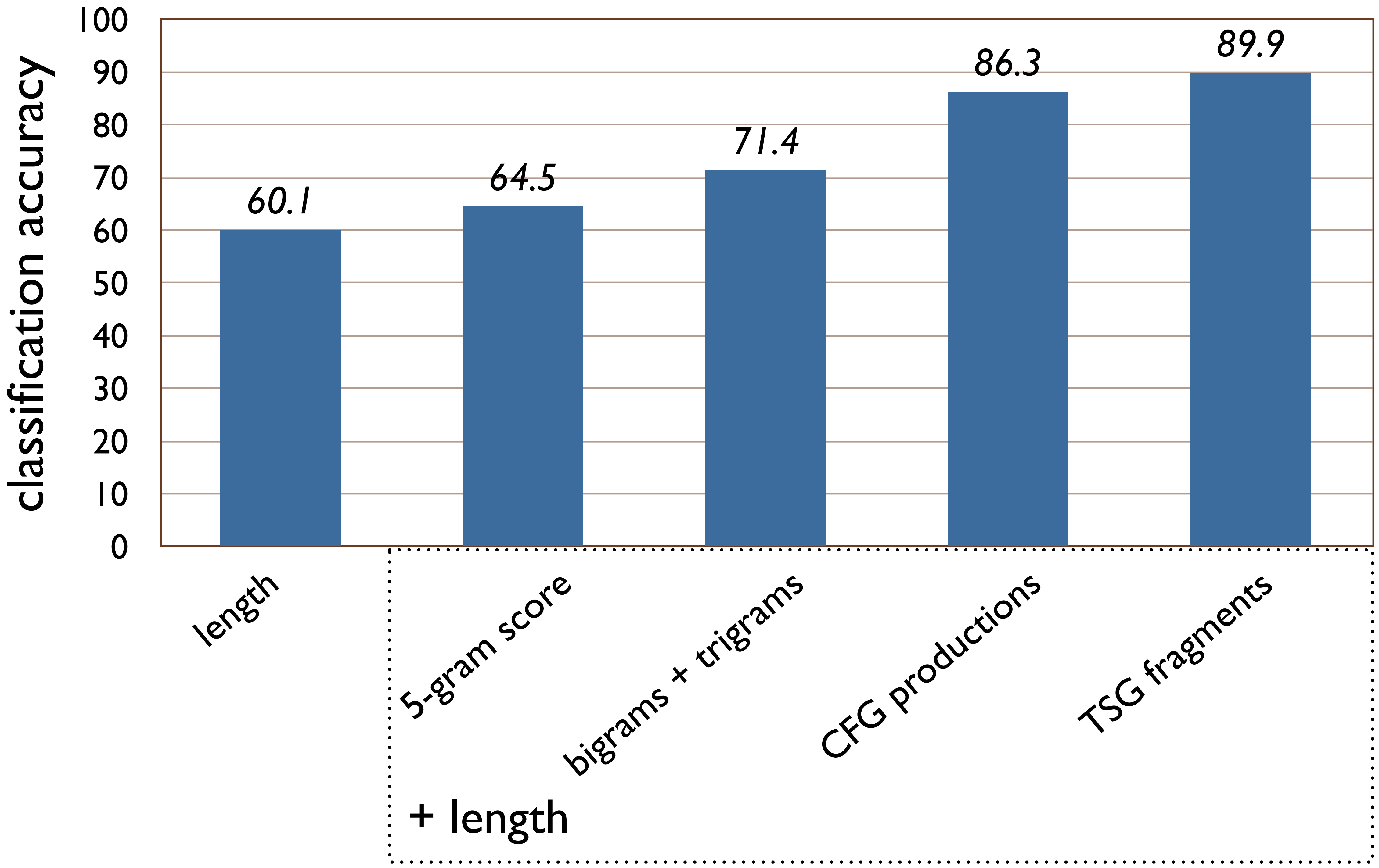
Bad

Experimental setup

- Following Cherry & Quirk (2008):



Classification results



What features are helpful?

Good	**Bad**
(TOP (S `` S , '' NP (VP (VBZ says) ADVP) .))	(NP (NP DT CD (NN %)) PP)
(FRAG (X SYM) VP .)	(NP DT)
(PRN (-LRB- -LRB-) S (-RRB- -RRB-))	(PP (IN of))
(PRN (-LRB- -LRB-) NP (-RRB- -RRB-))	[failed parse]
(S NP VP .)	(TOP (NP NP PP PP .))
(SBARQ WHADVP SQ (. ?))	(NP DT JJ NNS)
(NNP Mr)	(TOP (NP NP PP . ''))
(PRN (COLON --) PP (COLON --))	(TOP (S NP , NP VP . ('' '')))
(NNP Sons)	(VP PP)
(WHNP WP$ NN NN)	(PP (IN with))

Analysis

- What kinds of features are useful?
- Looking at the 100 top- and bottom-weighted features

	bad	good	example
unary productions	47	36	NP → DT
lexicalized fragments	37	60	(SBARQ WHADVP SQ (. ?))
bilexicalized fragments	1	10	(PRN (-LRB- -LRB-) S (-RRB- -RRB-))
fragment size >= 3	21	33	(TOP (S PP , NP (VP MD VP) .))

Observations

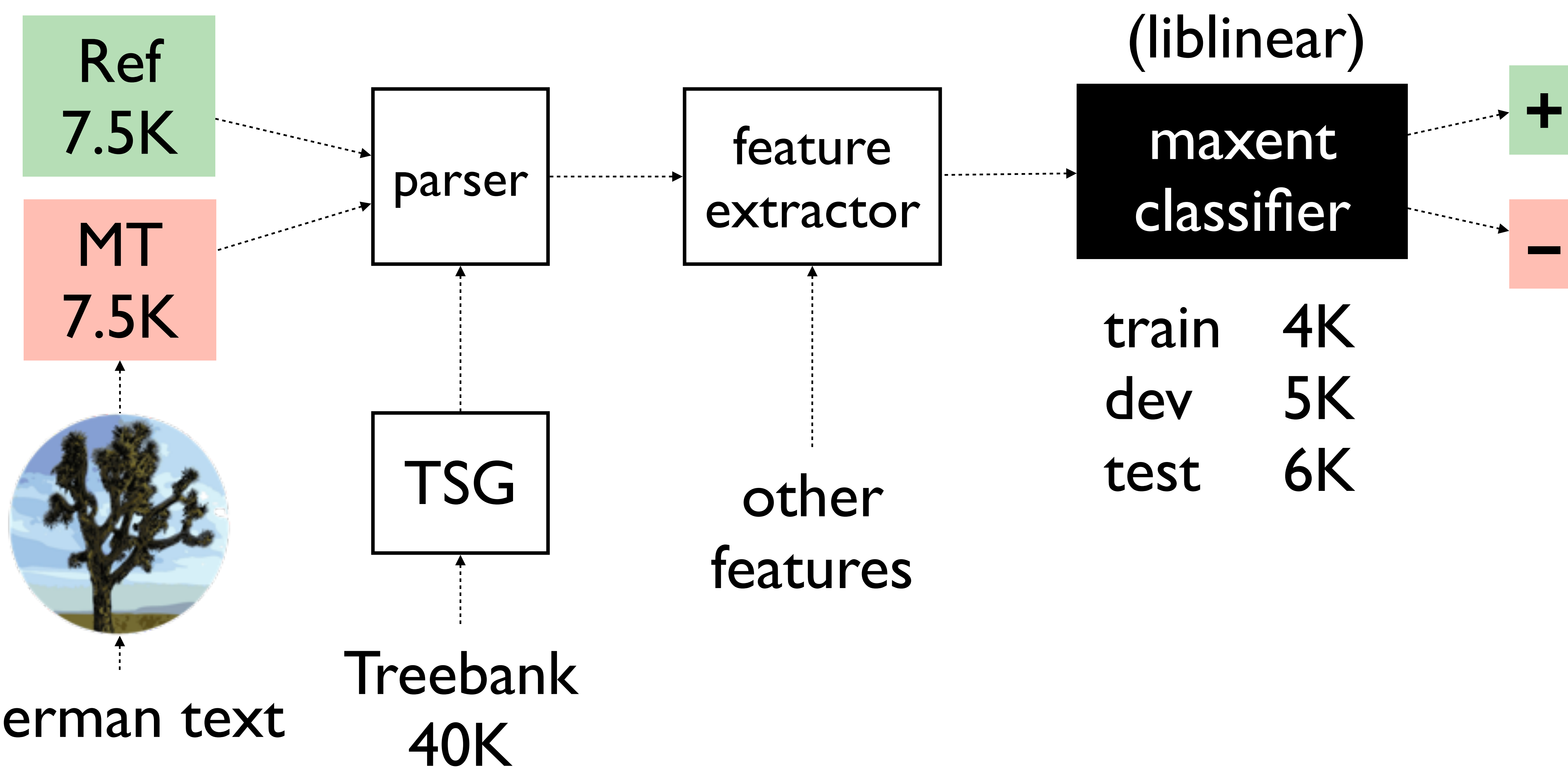
- TSGs performed well, weights are intuitive
- Shallow, unlexicalized rules correlate with ungrammaticality
- The C&J feature set performs the best, but at some cost in terms of model size

Task 2: MT output vs. human reference

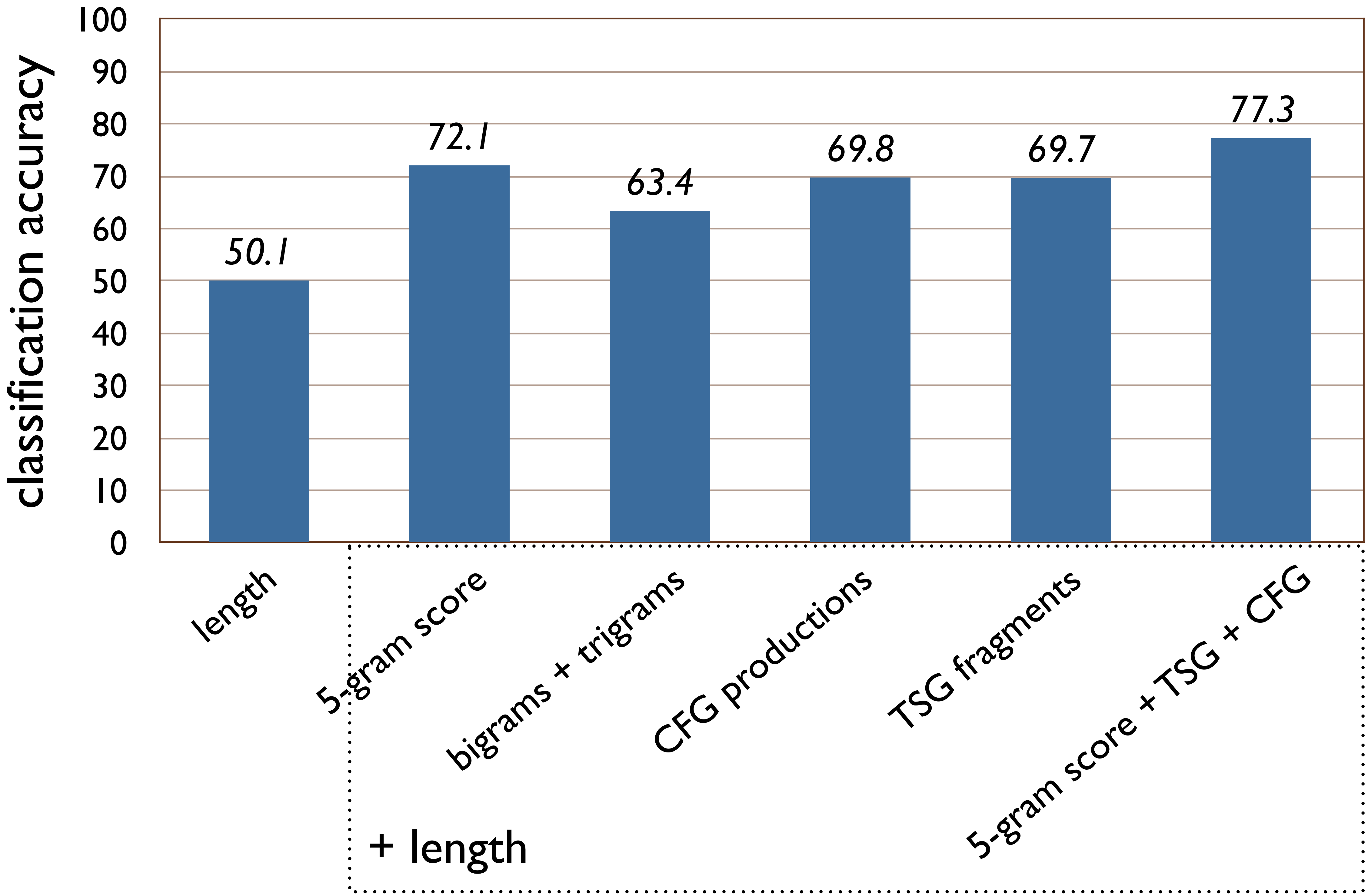
- Discriminate between MT output and a human reference translation (no access to the input)
- Some examples (MT — **reference**):
    - *a serious memory* — ***the weight of the past***
    - *at that time was warhol been dead for three years .* — ***at that point in time , warhol had already been dead for three years .***
    - *if the rally actually happened , the immobiliengesellschaften benefit from it .* — ***the constructors also will be able to benefit from this rally , in case it happens .***

Experiments

- Following Cherry & Quirk (2008):



Classification results



Observations

- TSG features alone didn't beat the baseline (as before), but were very complementary with the n-grams
  - But note that the n-gram model was used to produce the output in the first place

Closing observations

- Language is very complex, and we don’t know the rules (although we use them every day)
- Modeling always involves compromises
  - N-grams are wrong! But quite useful in accounting for local fluency
  - Similarly, CFGs are also wrong! But minor variations informed by linguistics can produce useful models that help account for global structure
- The use of syntax (for language modeling) in production systems is likely a ways off