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

<table>
<thead>
<tr>
<th></th>
<th>adequate</th>
<th>not adequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluent</td>
<td><em>we have a common understanding</em></td>
<td><em>we do not agree</em></td>
</tr>
<tr>
<td>disfluent</td>
<td><em>have an agreement</em></td>
<td><em>them owning compatibility</em></td>
</tr>
</tbody>
</table>
• 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

• 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:

  \[
  \text{The dog bit the goat.} \quad P(\text{bit} | \text{dog})
  \]
  
  \[
  \text{The dog with the missing eye bit the goat} \quad P(\text{bit} | \text{eye})
  \]

- 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.
• 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:

\[ S \rightarrow \text{NP VP} \]

• But this isn’t nearly detailed enough. Why not?

• *Example on the board.*
Bilexical parsing models

- Annotates CFG productions with head words

  \[ S \rightarrow NP \ VP \]

  becomes

  \[ S/\text{walked} \rightarrow NP/\text{boy} \ VP/\text{walked} \]

- 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)
Part of the difficulty is a metric mismatch.

<table>
<thead>
<tr>
<th>System</th>
<th>Perfect Translation</th>
<th>Syntactically Correct but Semantically Wrong</th>
<th>Semantically Correct Syntactically Wrong</th>
<th>Wrong</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>syntax TM + LM</td>
<td>45</td>
<td>67</td>
<td>70</td>
<td>164</td>
<td>0.0717</td>
</tr>
<tr>
<td>syntax TM only word-based</td>
<td>31</td>
<td>19</td>
<td>87</td>
<td>209</td>
<td>0.1031</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>11</td>
<td>87</td>
<td>223</td>
<td>0.0722</td>
</tr>
</tbody>
</table>

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
• 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.*
• 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 slepted in the hallway)
## English linguistic phenomena

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>ngrams</th>
<th>context-free grammars</th>
<th>immediate-head models</th>
</tr>
</thead>
<tbody>
<tr>
<td>infinite</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>word categories</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>word order</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>high-level patterns</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>agreement</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>predicate-argument structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>morphology</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>model</th>
<th>task</th>
<th>difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td>parsers</td>
<td>discriminate structures (grammaticality assumed)</td>
<td>PP attachment, coordination</td>
</tr>
<tr>
<td>language models</td>
<td>discriminate strings</td>
<td>ensuring global coherence</td>
</tr>
</tbody>
</table>
• 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
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.
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

<table>
<thead>
<tr>
<th></th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>coarse</strong></td>
<td>WSJ text</td>
<td>samples from an n-gram model</td>
</tr>
<tr>
<td><strong>MT</strong></td>
<td>reference translations</td>
<td>machine translation output</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>feature set</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>17</td>
</tr>
<tr>
<td>Gigaword 5-gram LM score</td>
<td>-12.045</td>
</tr>
<tr>
<td>bigrams and trigrams</td>
<td>“he further praised”</td>
</tr>
<tr>
<td>CFG productions</td>
<td>$S \rightarrow \text{NP VP}$</td>
</tr>
<tr>
<td>Charniak &amp; Johnson (2005) reranking features</td>
<td>number of nodes in the parse tree</td>
</tr>
<tr>
<td></td>
<td>head projections</td>
</tr>
<tr>
<td>TSG (parse score, fragments, aggregate features)</td>
<td>(TOP (S NP (VP VBD said) NP SBAR)) .)</td>
</tr>
</tbody>
</table>
Task 1: ngram samples from real text

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

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

Good

Bad
• Following Cherry & Quirk (2008):
Classification results

<table>
<thead>
<tr>
<th>Feature</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>60.1</td>
</tr>
<tr>
<td>5-gram score</td>
<td>64.5</td>
</tr>
<tr>
<td>Bigrams + Trigrams</td>
<td>71.4</td>
</tr>
<tr>
<td>CFG Productions</td>
<td>86.3</td>
</tr>
<tr>
<td>TSG Fragments</td>
<td>89.9</td>
</tr>
</tbody>
</table>

+ length
What features are helpful?

**Good**

(TOP (S `` S , " NP (VP (VBZ says) ADVP) .))
(FRAG (X SYM) VP .)
(PRN (-LRB- -LRB-) S (-RRB- -RRB-))
(PRN (-LRB- -LRB-) NP (-RRB- -RRB-))
(S NP VP .)
(SBARQ WHADVP SQ (. ?))
(NNP Mr)
(PRN (COLON --) PP (COLON --))
(NNP Sons)
(WHNP WP$ NN NN)

**Bad**

(NP (NP DT CD (NN %)) PP)
(NP DT)
(PP (IN of))
[failed parse]
(TOP (NP NP PP PP .))
(NP DT JJ NNS)
(TOP (NP NP PP ."))
(TOP (S NP , NP VP . (" ")))
(VP PP)
(PP (IN with))
• What kinds of features are useful?

• Looking at the 100 top- and bottom-weighted features

<table>
<thead>
<tr>
<th></th>
<th>bad</th>
<th>good</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>unary productions</td>
<td>47</td>
<td>36</td>
<td>NP → DT</td>
</tr>
<tr>
<td>lexicalized fragments</td>
<td>37</td>
<td>60</td>
<td>(SBARQ WHADVP SQ (. ?))</td>
</tr>
<tr>
<td>bilexicalized fragments</td>
<td>1</td>
<td>10</td>
<td>(PRN (-LRB- -LRB-) S (-RRB- -RRB-))</td>
</tr>
<tr>
<td>fragment size &gt;= 3</td>
<td>21</td>
<td>33</td>
<td>(TOP (S PP , NP (VP MD VP) .))</td>
</tr>
</tbody>
</table>
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):

Ref 7.5K

MT 7.5K

German text

Treebank 40K

parser

TSG

feature extractor

other features

(liblinear)

maxent classifier

train 4K
dev 5K
test 6K
Classification results

<table>
<thead>
<tr>
<th>Feature</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>50.1</td>
</tr>
<tr>
<td>5-gram score</td>
<td>72.1</td>
</tr>
<tr>
<td>bigrams + trigrams</td>
<td>63.4</td>
</tr>
<tr>
<td>CFG productions</td>
<td>69.8</td>
</tr>
<tr>
<td>TSG fragments</td>
<td>69.7</td>
</tr>
<tr>
<td>5-gram score + TSG + CFG</td>
<td>77.3</td>
</tr>
<tr>
<td>+ length</td>
<td></td>
</tr>
</tbody>
</table>
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