

Statistical Perspectives on Text-to-Text Generation

Noah Smith

Language Technologies Institute

Machine Learning Department

School of Computer Science

Carnegie Mellon University

`nasmith@cs.cmu.edu`

I'm A Learning Guy

- I use statistics for **prediction**
 - *Linguistic Structure Prediction* – my new book
 - Computational social science research: discovery via prediction
 - Predicting the future from text
- Ideal: inputs and outputs



Prediction-Friendly Problems

Predicting the *whole* output from the *whole* input:

- Linguistic Analysis
(morphology, syntax, semantics, discourse)
 - linguists can reliably annotate data (we think)
- Machine Translation
 - parallel data is abundant (in some cases)
- Generation?

But Generation is Unnatural!

- Relevant data do not occur in “nature.”
 - Consider the effort required to build datasets for paraphrase, textual entailment, factual question answering, summarization ...
 - Do people perform these tasks “naturally”?
- Datasets are small and highly task-specific.
- Do statistical techniques even make sense?

Three Kinds of Predictions


Assume a text-text relation of interest.

- Given a pair, does the relationship hold?
(**Yes or no.**) easier
- Given an input, **rank** a set of candidates.
- Given an input, **generate** an output. harder



Three Kinds of Predictions

Assume a text-text relation of interest.

- Given a pair, does the relationship hold?
(**Yes or no.**) boys/girls
 - Given an input, **rank** a set of candidates.
 - Given an input, **generate** an output. men/women
- 

Outline

1. Quasi-synchronous grammars
2. Tree edit models
3. A foray into text-to-text generation



© DISNEY

Synchronous Grammar

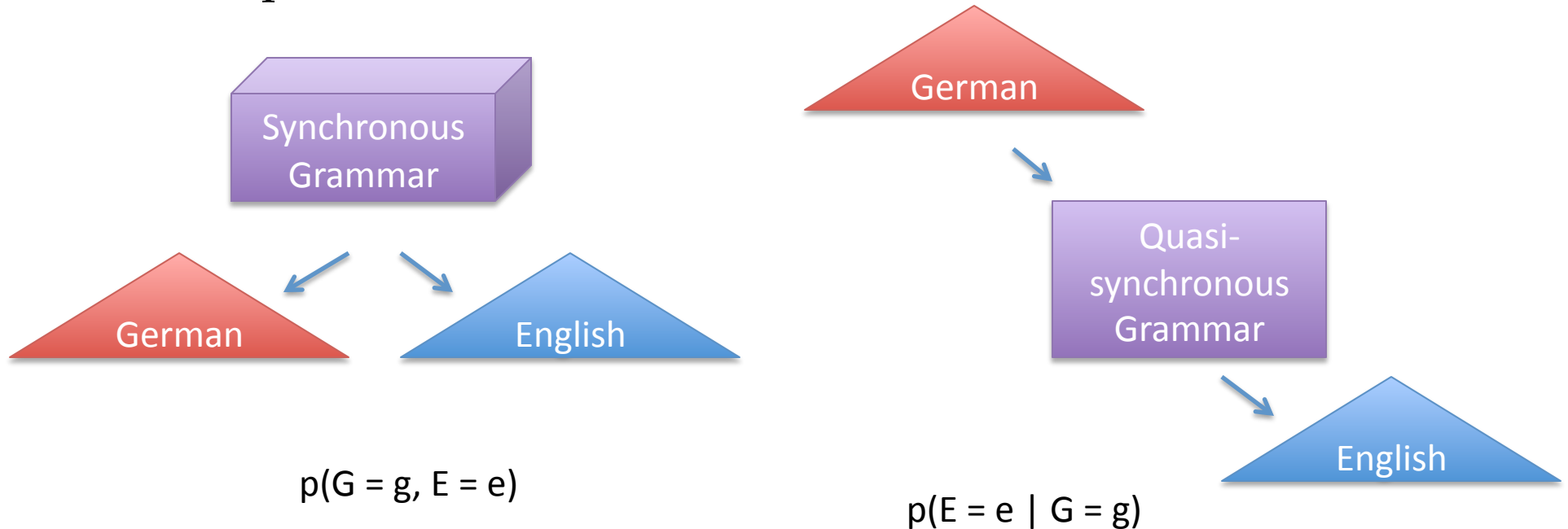
- Basic idea: one grammar, two languages.

$$\begin{aligned} \text{VP} &\rightarrow \text{ne } V_1 \text{ pas } \text{VP}_2 / \text{not } V_1 \text{ VP}_2 \\ \text{NP} &\rightarrow N_1 A_2 / A_2 N_1 \end{aligned}$$

- Many variations:
 - formal richness (rational relations, context-free, ...)
 - rules from experts, treebanks, heuristic extraction, rich statistical models, ...
 - linguistic nonterminals or not

Quasi-Synchronous Grammar

- Compare:



- Developed by David Smith and Jason Eisner (SMT workshop 2006).

Quasi-Synchronous Grammar

- Basic idea: one grammar per source sentence.

(S₁ Je (VP₄ ne₅ (V₆ veux) pas₇
(VP₈ aller à l' (NP₁₂ (N₁₃ usine) (A₁₄ rouge))) .)

$$\begin{aligned} \text{VP}_{\{4\}} &\rightarrow \text{not}_{\{5, 7\}} \text{V}_{\{6\}} \text{VP}_{\{8\}} \\ \text{NP}_{\{12\}} &\rightarrow \text{A}_{\{14\}} \text{N}_{\{13\}} \end{aligned}$$

- Doesn't have to be CFG! We use dependency grammar.

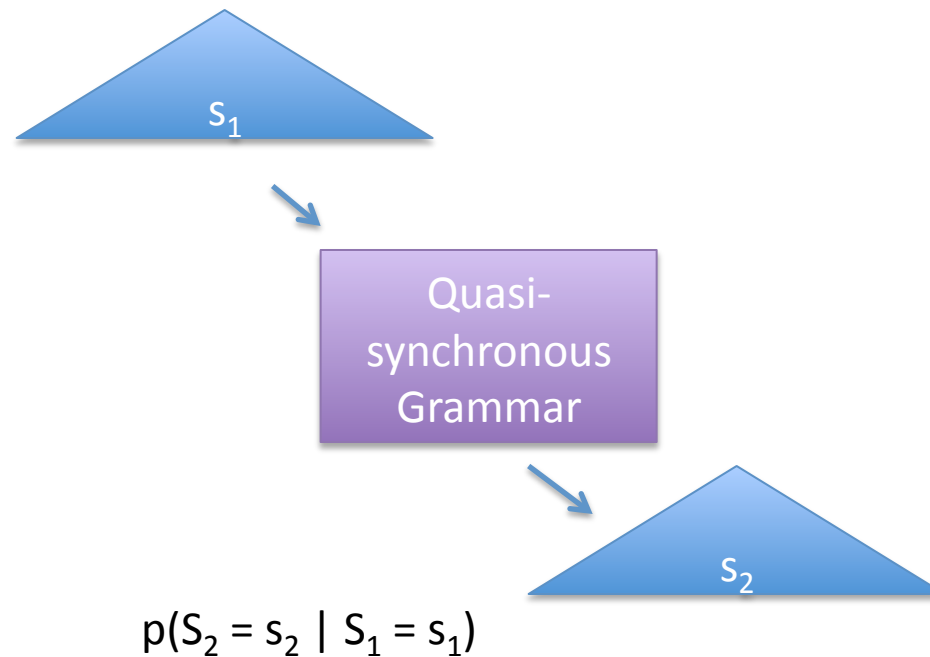
Quasi-Synchronous Grammar

- The grammar is determined by the input sentence and only models output language.
 - Generalizes IBM models.
- Allows loose relationship between input and output.
 - “Divergences,” which we think of as non-standard configurations.
 - By disallowing some relationships, we can simulate stricter models; we explored this a good bit in MT ...

Aside: Machine Translation

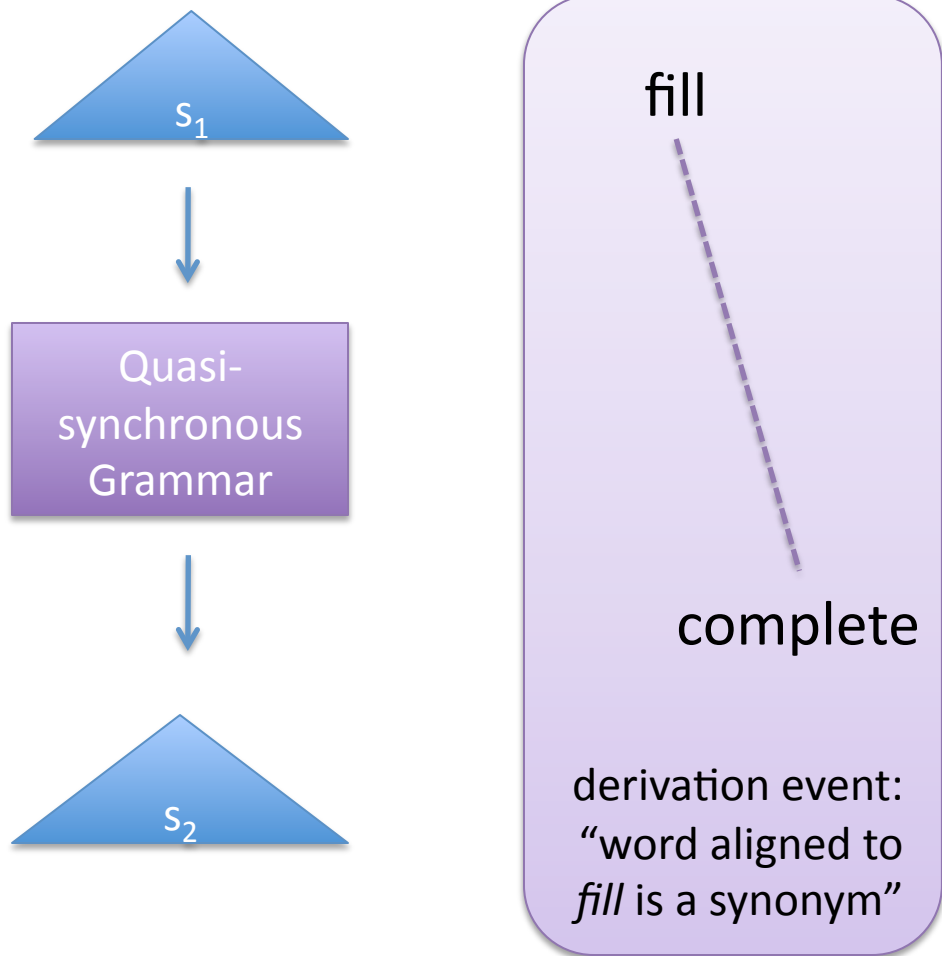
- The QG formalism originated in translation research (D. Smith and Eisner, 2006).
- Gimpel and Smith (EMNLP 2009): QG as a framework for translation with a blend of dependency syntax features and phrase features. *Generation by lattice parsing.*
- Gimpel and Smith (EMNLP 2011): QG on *phrases* instead of words shown competitive for Chinese-English and Urdu-English.

Paraphrase (Basic Model)

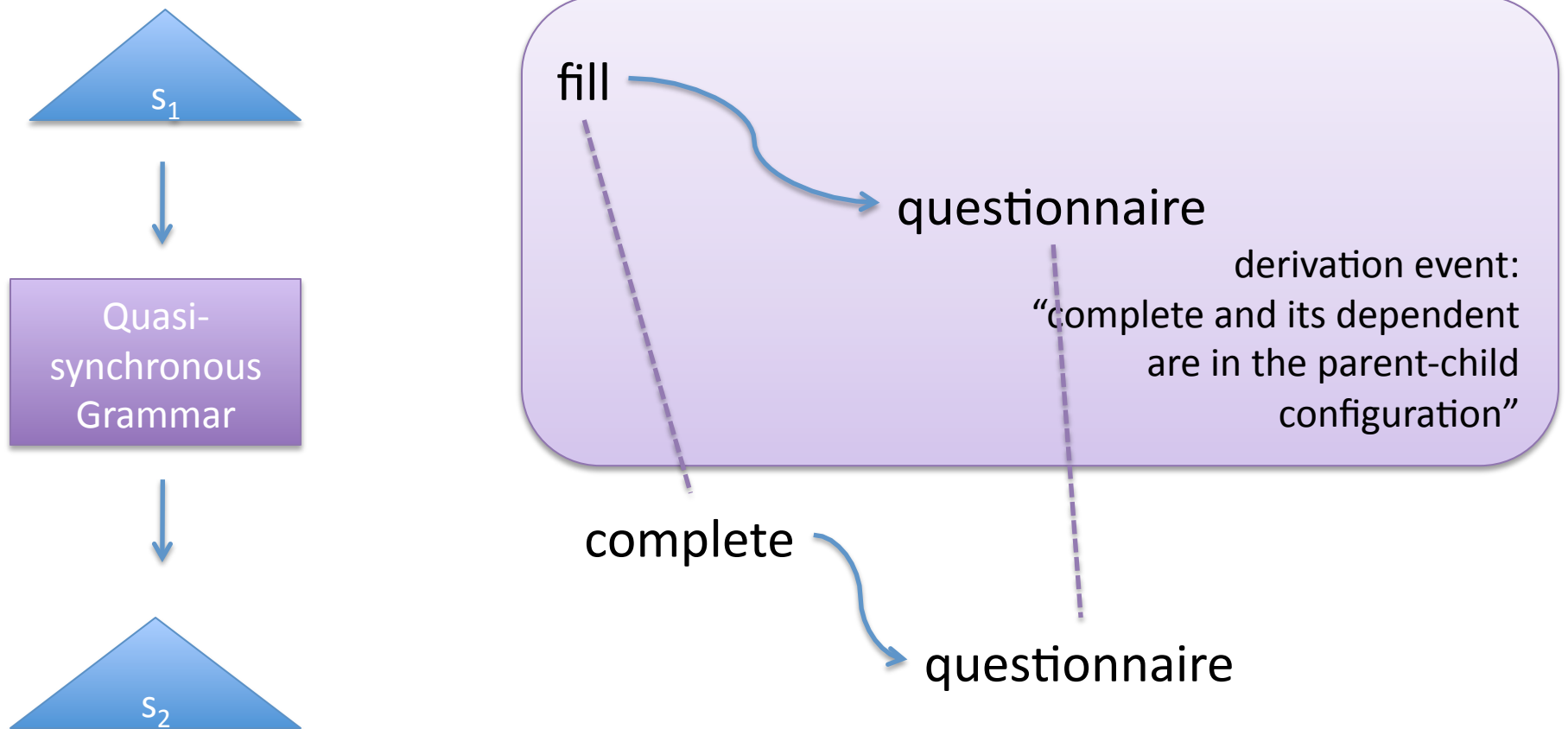


Note: Wu (2005) explored a *synchronous* grammar for this problem.

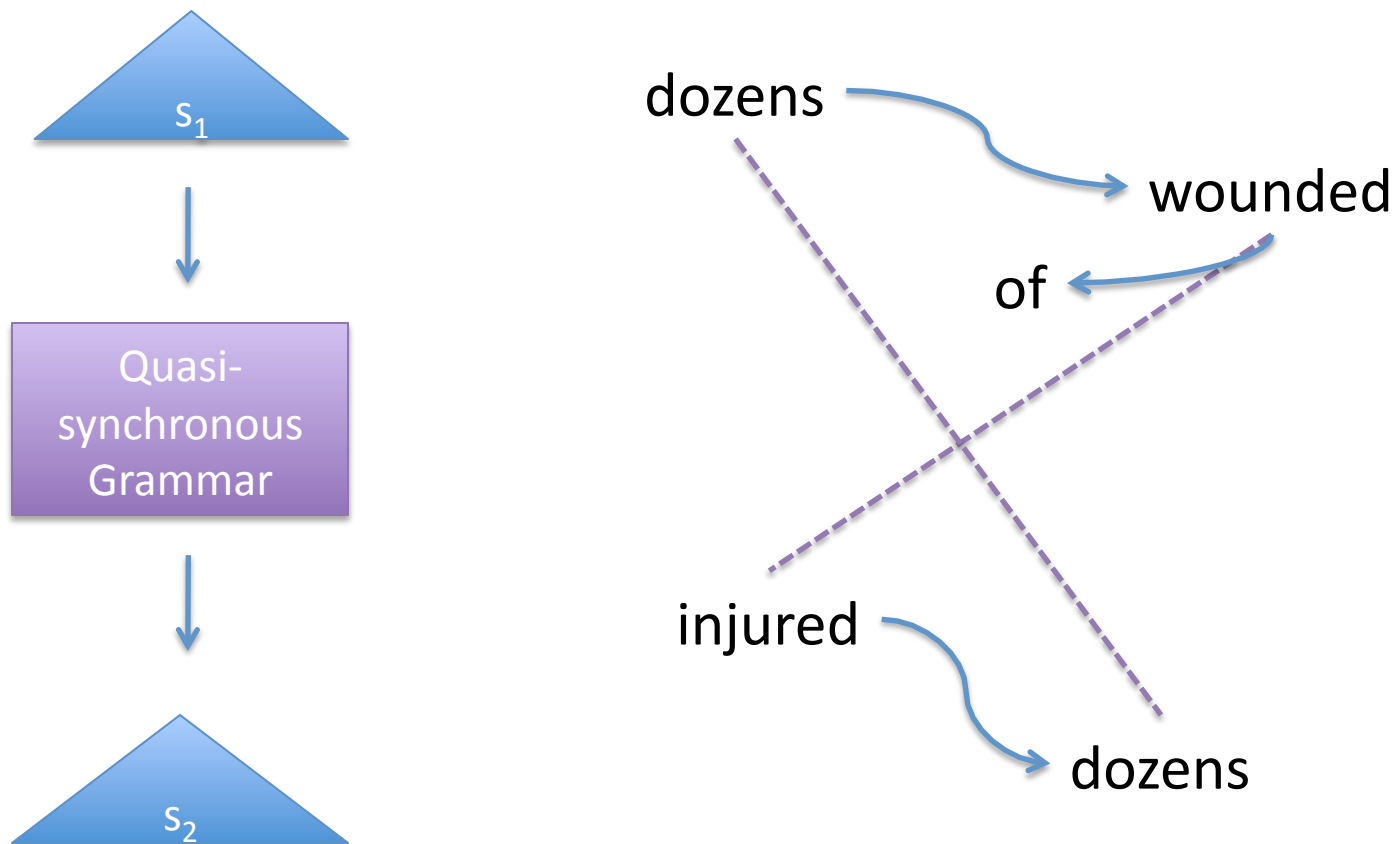
Alignment



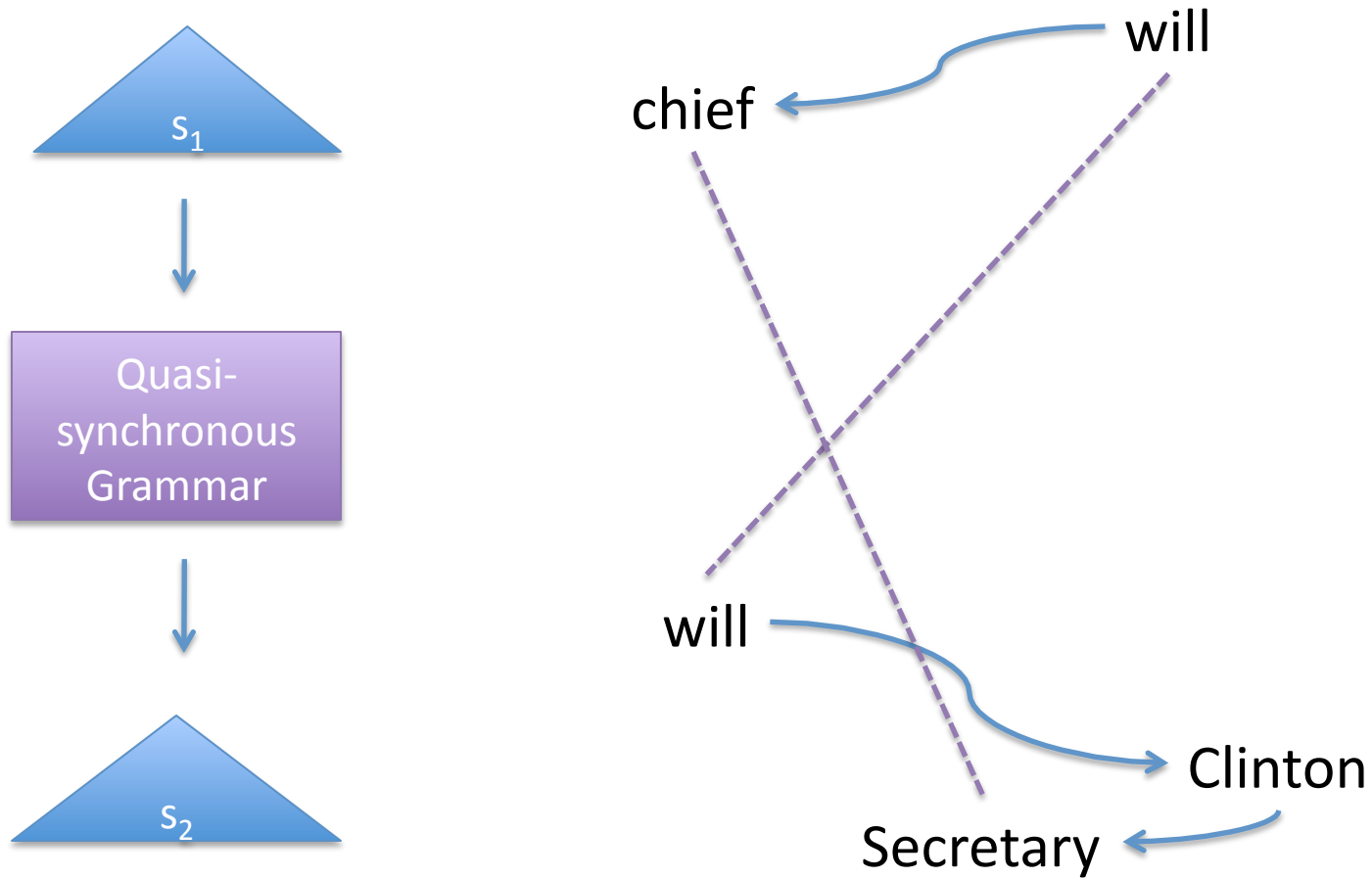
Parent-Child Configuration



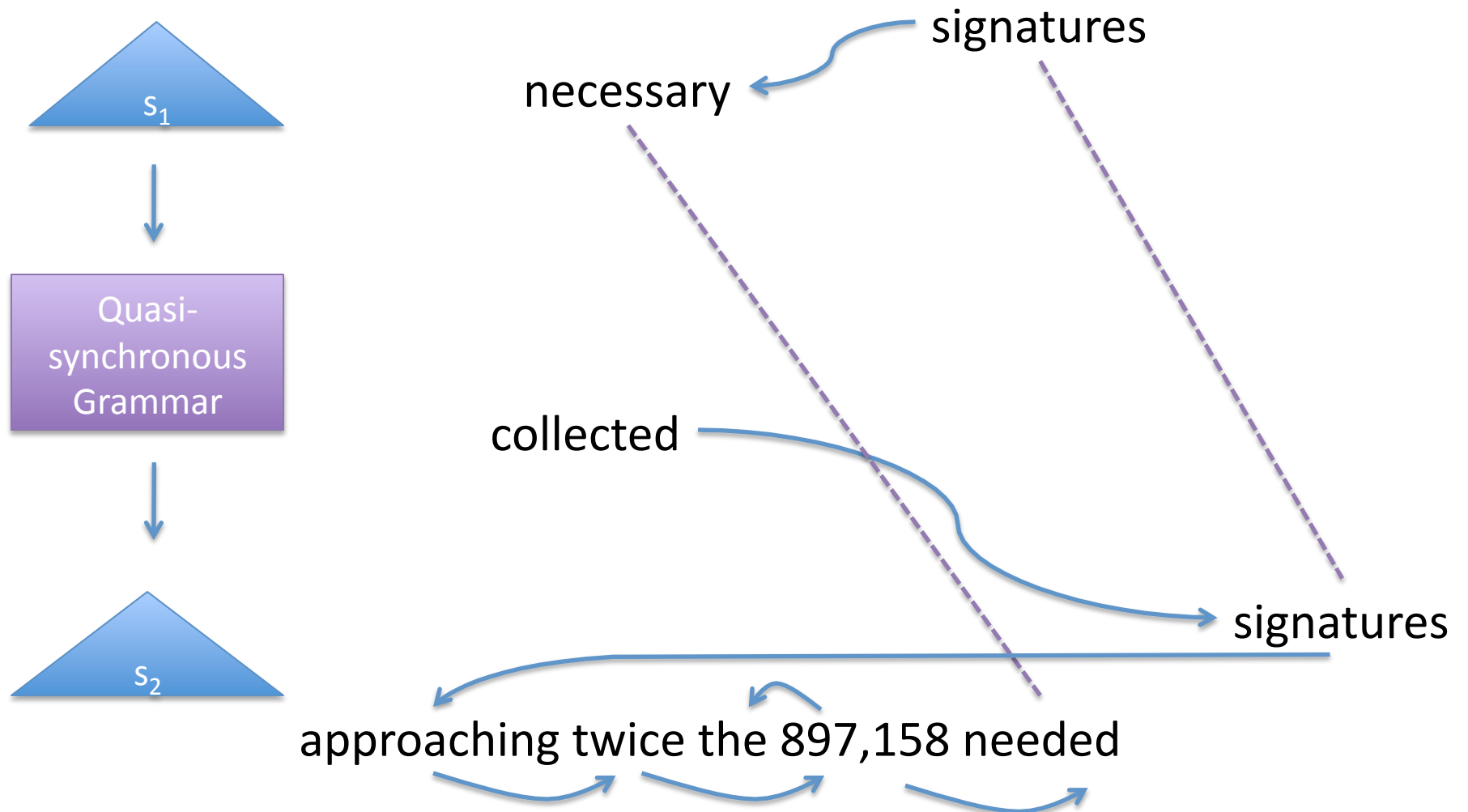
Child-Parent Configuration



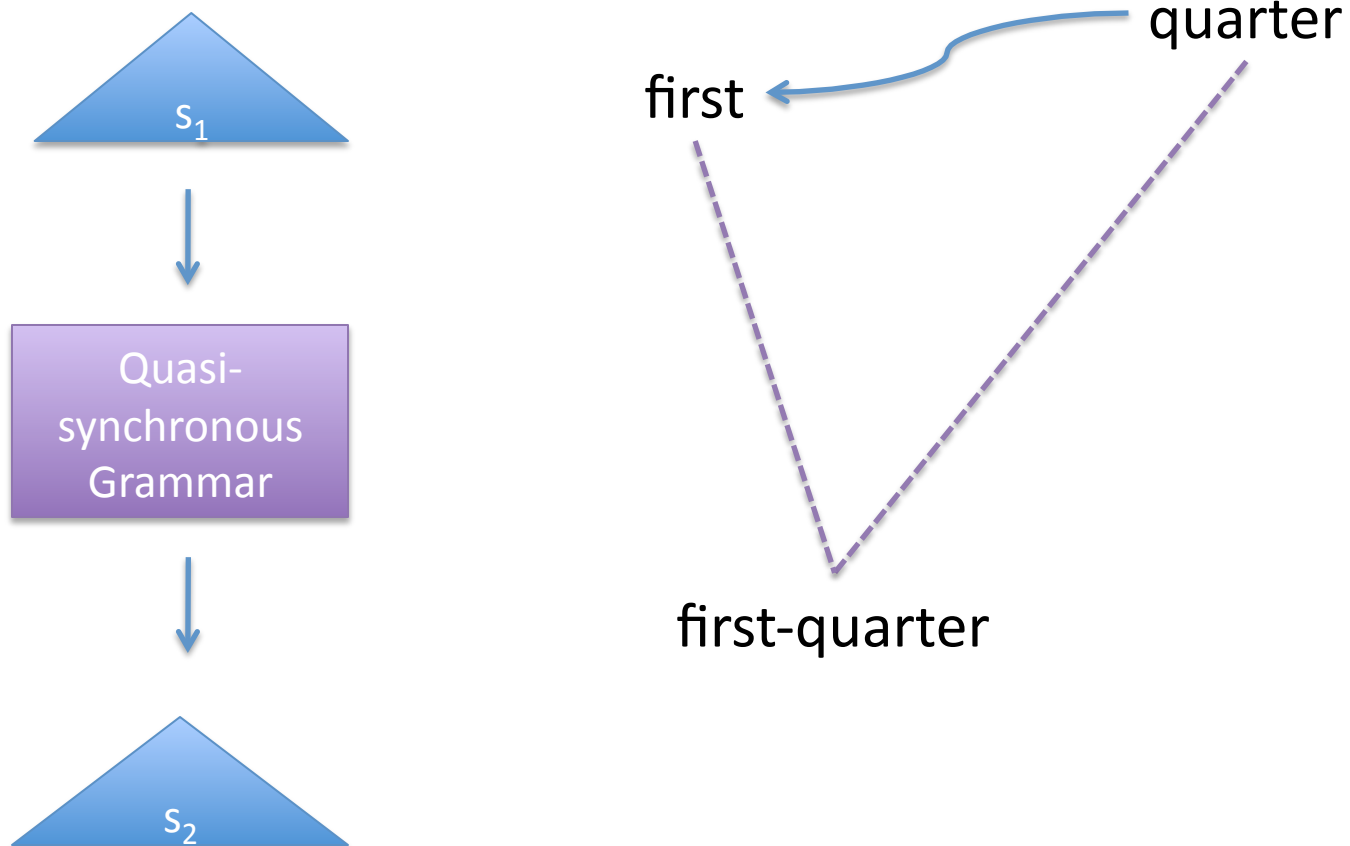
Grandparent-Child Configuration



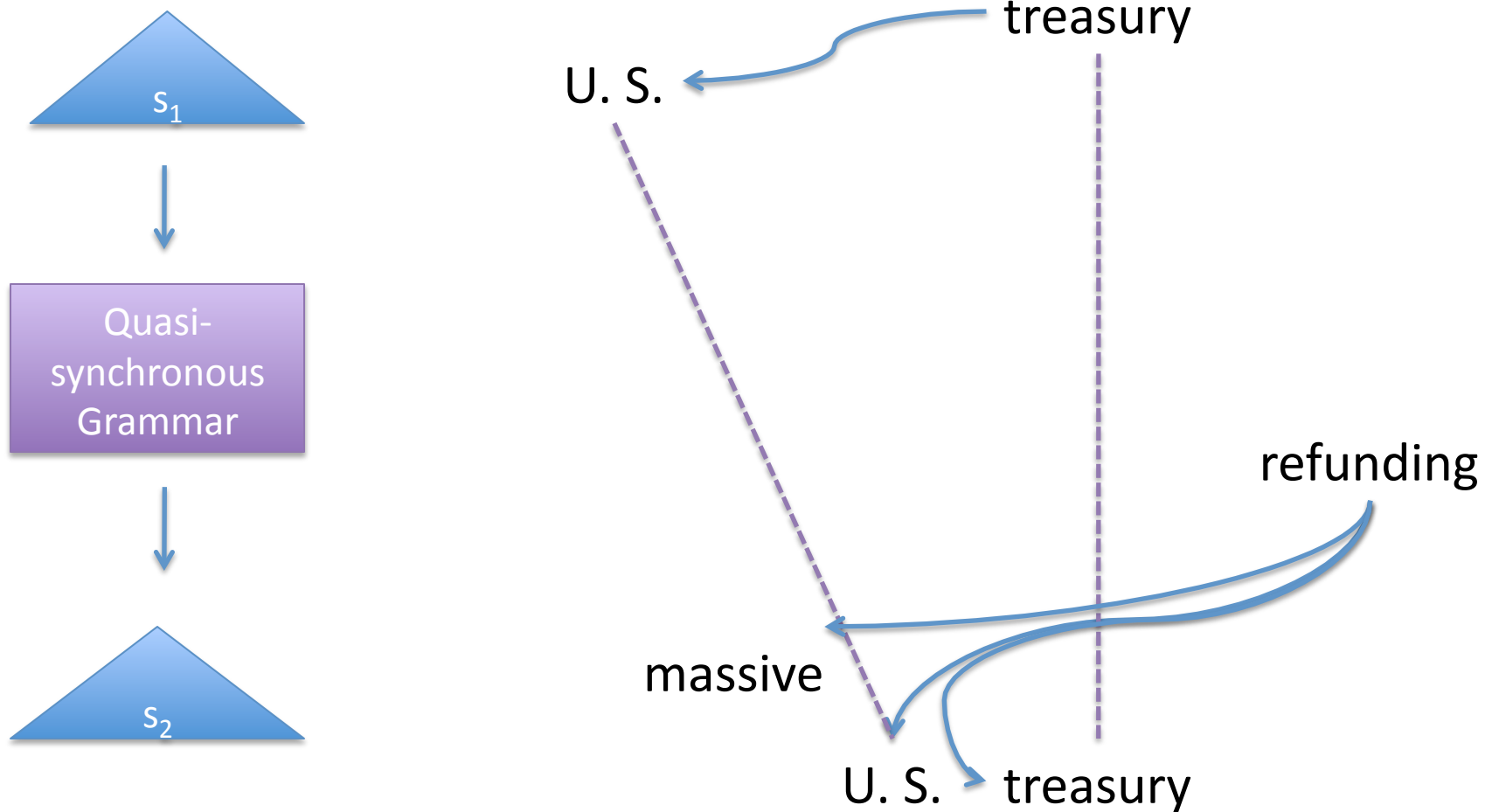
C-Command Configuration



Same Node Configuration



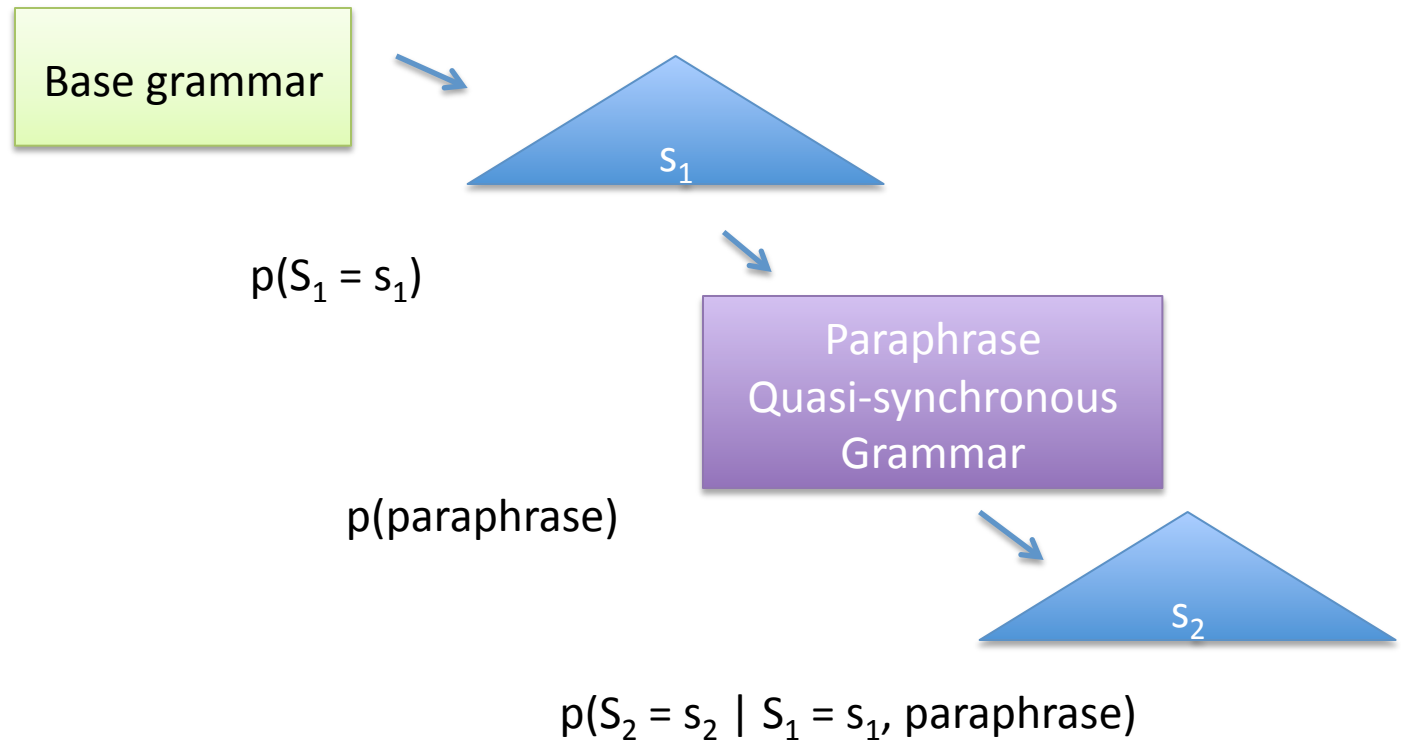
Sibling Configuration



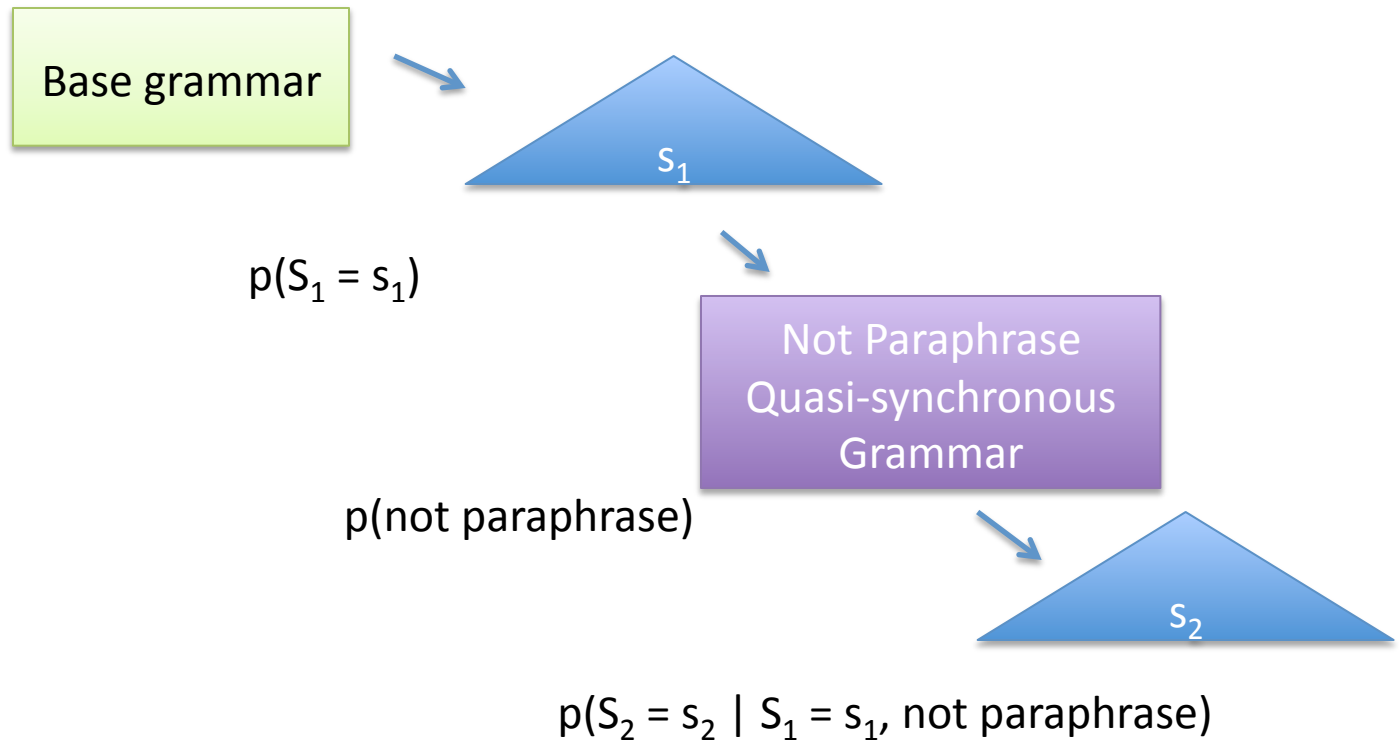
Probabilistic QG

- Probabilistic grammars – well known from parsing.
- From “parallel data,” we can learn:
 - relative frequencies of different configurations for different words
 - includes basic syntax (POS, dependency labels)
- We can also incorporate:
 - lexical semantics features that notice synonyms, hypernyms, etc.
 - named entity chunking

Generative Story (Paraphrase)



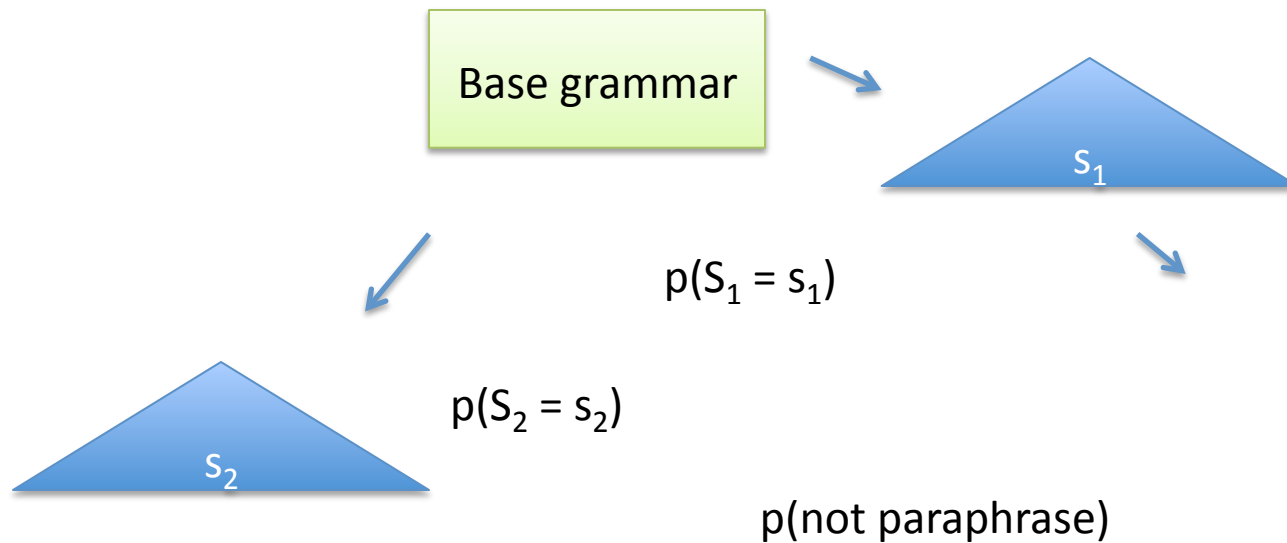
Generative Story (Not Paraphrase)



“Not Paraphrase” Grammar?

- This is the result of opting for a fully generative story to explain an unnatural dataset.
 - See David Chen and Bill Dolan’s (ACL 2011) approach to building a better dataset!
- We must account, probabilistically, for the event that two sentences are generated that are not paraphrases.
 - (Because it happens in the data!)
 - Generating twice from the base grammar didn’t work; in the data, “non paraphrases” look much more alike than you would expect by chance.

“Not Paraphrase” Model We Didn’t Use



Notes on the Model

- Although it is generative, we train it *discriminatively* (like a CRF).
- The correspondences (alignment) between the two sentences is treated as a **hidden variable**.
 - We sum it out during inference; this means all possible alignments are considered at once.
 - This is the main difference with other work based on overlap features.

But Overlap Features are Good!

- Much is explained by simple overlap features that don't easily fit the grammatical formalism (Finch et al., 2005; Wan et al., 2006; Corley and Mihalcea, 2005).
- Statistical modeling with a *product of experts* (i.e., two models that can veto each other) allowed us to incorporate shallow features, too.
- **We should not have to choose between two good, complementary representations!**
 - We just might have to *pay* for it.

Paraphrase Identification Experiments

- Test set: $N = 1,725$

Model	Accuracy	p-Precision	p-Recall
all paraphrase	66.49	66.49	100.00
Wan et al. SVM (reported)	75.63	77.00	90.00
Wan et al. SVM (replication on our test set)	75.42	76.88	90.14
Wan-like model	75.36	78.12	87.74
QG model	73.33	74.48	91.10
PoE (QG with Wan-like model)	76.06	79.57	86.05
Oracle PoE	83.19	100.00	95.29

Comments

- From a modeling point of view, this system is rather complicated.
 - Lots of components!
 - Training latent-variable CRFs is not for everyone.
- I'd like to see more elegant ways of putting together the building blocks (syntax, lexical semantics, hidden alignments, shallow overlap) within a single, discriminative model.

Jeopardy! Model

THIS CBC RADIO
SHOW, HOSTED BY
NORA YOUNG,
INTERVIEWED
COMPUTER SCIENTIST
NOAH SMITH IN 2011



QG for QA

- Essentially the same model works quite well for an **answer selection** task.
 - (I have the same misgivings about the data.)
- Briefly: learn $p(\text{question} \mid \text{answer})$ as a QG from question-answer data.
 - Then **rank** candidates.
- Full details in Wang, Mitamura, and Smith (EMNLP 2007).

Question-Answer Data

- Setup from Shen and Klakow (2006):
 - Rank answer candidates
- TREC dataset of just a few hundred questions with about 20 answers each; we manually judged which answers were correct (around 3 per question).
- Very small dataset!
 - We explored adding in noisily annotated data, but got no benefit.

Answer Selection Experiments

- Test set: $N = 100$

	No Lexical Semantics		With WordNet	
Model	MAP	MRR	MAP	MRR
TreeMatch	38.14	44.62	41.89	49.39
Cui et al. (2005)	43.50	55.69	42.71	52.59
QG model	48.28	55.71	60.29	68.52

QG: Summary

- QG is an elegant and attractive modeling component.
 - Really nice results on an answer selection task.
 - Okay results on a paraphrase identification task.
- Frustrations:
 - Integrating representations should be easier.
 - Is the model intuitive?

Outline

- ✓ Quasi-synchronous grammars
- 2. Tree edit models
- 3. A foray into text-to-text generation



A Different Approach: Tree Edits



I'm Mike Heilman, and I think those quasi-synchronous models are more complicated than they need to be.

- Full details in Heilman and Smith (NAACL 2010).

Tree Edit Models

- There are many algorithms for aligning trees or minimizing various **tree edit distances** (Klein, 1989; Zhang and Shasha, 1989).
- These allow deletion, insertion, and relabeling operations.
 - Simple, intuitive operations that transform the sentence incrementally.
- As noted in the QG work, **movement** is also desirable.
 - You can't have that and stay efficient.

An Example from Entailment

Pierce built the home for his daughter off
Rossville Blvd., as he lives nearby.

An Example from Entailment

Pierce built the home for his daughter off
Rossville Blvd., as he lives **nearby**.

relabel node

An Example from Entailment

Pierce built the home for his daughter off
Rossville Blvd., as he lives near.

An Example from Entailment

Pierce built the home for his daughter off
Rossville Blvd., as he lives near.

move node

An Example from Entailment

Pierce built the home for his daughter off,
as he lives near Rossville Blvd.

An Example from Entailment

move node

Pierce built the home for his daughter off,
as he lives near Rossville Blvd.

An Example from Entailment

built the home for his daughter off,
as Pierce he lives near Rossville Blvd.

An Example from Entailment

built the home for his daughter off,
as Pierce **he** lives near Rossville Blvd.

delete node

An Example from Entailment

built the home for his daughter off,
as Pierce lives near Rossville Blvd.

An Example from Entailment

built the home for his daughter off,
as Pierce lives near Rossville Blvd.

new root

An Example from Entailment

built the home for his daughter off,
as Pierce lives near Rossville Blvd.

An Example from Entailment

built the home for his daughter off,
as Pierce lives near Rossville Blvd.

delete node

An Example from Entailment

Pierce built the home for his daughter off
Rossville Blvd., as he lives nearby.

Pierce lives near Rossville Blvd.

Sketch of the Approach

1. Find a tree edit sequence for the sentences, allowing all the operations we want.
 - We use greedy heuristic search.
 - Don't worry about whether it's the “right” one.
2. Calculate features on the tree edit sequence.
3. Use a logistic regression model to classify the relationship.

Operations on Dependency Trees

- Insert child.
- Insert parent.
- Delete leaf.
- Delete and merge.
- Relabel node.
- Relabel edge.

- Move subtree.
- New root.
- Move sibling.

Heuristic Search

- Greedy best-first search (Pearl, 1984).
- Heuristic: 1 - Collins and Duffy's (2001) tree kernel, normalized.
 - Completely different context!
 - We use it as a similarity function from the candidate transformed sentence to the true output.
 - Our kernel is based on Moschitti (2006) and Zelenko et al. (2003).
- Constraints (in brief): don't insert elements not in the target; new edges take the most frequent label for the child POS.
- Maximum number of iterations (about 5 seconds per sentence pair). Fails less than 0.1% of the time.

33 Features of Edit Sequences

- Number of edits total, and by type
- Number of unedited nodes: total, verbs, nouns, numbers, proper nouns
- Relabel: same POS, same lemma, noun-to-pronoun, change of proper noun, numeric change greater than 5%
- Insert: noun-or-verb, proper noun
- Remove: noun-or-verb, proper noun, subject, object, verb complement, root edge
- Relabel (from or to): subject, object, verb complement, root edge
- Search failure

Experimental Notes

- Direction:
 - For entailment, from premise to hypothesis.
 - For paraphrase, both directions (double the features).
 - For answer selection, answer to question.

RTE-3 Experiments

Model	Accuracy	Precision	Recall
Harmeling (2007) - less general operations	59.5	66.49	100.00
de Marneffe et al. (2006) – align and classify	60.5	61.8	60.2
MacCartney and Manning (2008) – natural logic	59.4	70.1	36.1
MacCartney and Manning (2008) – hybrid	64.3	65.5	63.9
Tree edit model	62.8	61.9	71.2

Paraphrase Identification Experiments

- Test set: $N = 1,725$

Model	Accuracy	p-Precision	p-Recall
all paraphrase	66.49	66.49	100.00
Wan et al. SVM (reported)	75.63	77.00	90.00
Wan et al. SVM (replication on our test set)	75.42	76.88	90.14
Wan-like model	75.36	78.12	87.74
QG model	73.33	74.48	91.10
PoE (QG with Wan-like model)	76.06	79.57	86.05
Tree edit model	73.3	76.2	87.0

Answer Selection Experiments

- Test set: $N = 100$

Model	MAP	MRR
TreeMatch	38.14	44.62
with WordNet	41.89	49.39
Cui et al. (2005)	43.50	55.69
with WordNet	42.71	52.59
QG model with lex. sem. ablated	48.28	55.71
QG model, full	60.29	68.52
Tree edit model	60.91	69.17

Advantages of Tree Edit Model

- Very, very simple.
 - No lexical semantics, Bleu scores, hidden variable modeling, ...
 - ... but could be extended with these things.
- Learned models for the three tasks were highly similar.
- Intuitive way of breaking down the problem

Toward Generation

- Both quasi-synchronous grammar and tree edit models suggest ways of going about **generating** output.
 - QG: take input, build grammar, “parse Σ^* .”
This is sort of what we aim for in MT.
 - TE: search for high scoring transformations.
Totally untested idea.

Outline

- ✓ Quasi-synchronous grammars
 - ✓ Tree edit models
3. A foray into text-to-text generation



Finally, Generation

Two cases:

- Heilman and Smith (NAACL 2010) and Heilman (2011): factual question generation
- In brief: Martins and Smith (ILP Workshop 2009): sentence extraction and summarization

- in summarization paper, note the difficulty of not having a single dataset. scarce gains. could revisit this with better inference techniques (Lagrangian relaxation, etc.)
 - mention new Berkeley work that does this better?
- in question generation research: reliance on human judgments. still may be better than trying to build annotated data up front. play up that many errors resulted from parsing/analysis problems. could not get coref to help.

Question Generation

- Our formulation of the task:
- Given a document, generate questions that could be used to check comprehension.
 - Imagine a teacher who wants students to get reading practice on material of their choice, or current events. Can we help the teacher write a quiz?

(Historical aside: this developed out of an undergrad course project on question *answering*!)

How It Works

1. Extract sentences.
2. Nondeterministic rule-based answer-to-question transformations.
3. Statistical ranking learned from human judgments of sentence quality.

Example

- Monrovia was named after James Monroe, who was president of the United States in 1922.
- Monrovia was named after James Monroe.
- Was Monrovia named after James Monroe.
- Was Monrovia named after who.
- Who was Monrovia named after?

extract a simplified statement

subject-auxiliary inversion

answer to question phrase

WH movement

1. Sentence Extraction

- Related to textual entailment, except we're generating entailments.
- Preprocessing: parsing, supersense tagging, and coreference.
- Examples of operations:
 - removing discourse markers and adjuncts
 - splitting conjunctions
 - extract presupposed statements from certain well-catalogued constructions (Levinson, 1983)
 - pronouns replaced by more informative NPs (like Nenkova, 2008); using coreference

2. Question Formation

- Largely driven by syntax.
 - Robust, general rules written in a clean formalism, tregex.
 - Some semantic effects missed; overgenerates.
- Steps:
 1. Mark NPs, PPs, and subordinate clauses that can't be answer phrases
 2. Pick an answer phrase, generate question phrase
 3. Verb decomposition, aux.-inversion
 4. Substitute question phrase for answer phrase

Linguistics Helps!

- If you took a GB-oriented syntax class, you could be forgiven for thinking linguistics was the study of questions.
 - *What does Chris like the woman who wears?
 - *What does Dipanjan wonder where Mike went to buy?
 - *Who do you believe that came to my talk?
- Novel (to my knowledge): formulating these constraints in Tregex (Levy and Andrew, 2006).

3. Ranking

- Gathered human scores (1-5) of question quality.
 - In earlier work, we had them mark different kinds of errors; this was not helpful for the overall system and was more expensive.
- Learn a regression model from feature representation of question-answer to human acceptability scores.
- Rank on acceptability prediction.

Source	Question	Mean Rating
In 1924 the site was chosen to serve as the capital of the new Tajik Autonomous S. S. R. ..., and rapid industrial and population growth followed.	What followed?	1.00
Parliament offered the Crown not to James's eldest son ... but to William and Mary as joint Sovereigns.	Who did Parliament offer the Crown not to?	2.00
The National Archives has exhibits of historical documents.	What does the National Archives have exhibits of?	3.00
After the People's Republic of China took control of Tibet in 1950 and suppressed a Tibetan uprising in 1959, the passes into Sikkim became a conduit for refugees from Tibet.	What did the People's Republic of China take control of in 1950?	3.67
Asmara was a village of the Tigre people until the late 19 th century.	What was Asmara until the late 19 th century?	4.00
Each year the banquet for the winners of the Nobel Prizes is held in City Hall.	Where is the banquet for the winners of the Nobel Prizes held?	4.67

Agreement

- Fleiss κ in the 0.5-0.6 range, for various ways of making the comparison.
- Moderately difficult task.
- We were more careful in choosing annotators for the *test* set.

179 Question Quality Features

- Length of question, answer phrase, source
- Which WH word?
- Negation?
- Language model probability
- Grammatical phrase types in the answer
- Tense of the main verb
- Main verb is a form of *be*?
- Which sentence transformations were applied in stage 1?
- WH is subject?
- “Vagueness” features (e.g., pronouns, common nouns without modification)

Intrinsic Evaluation

- We considered sentence-level and document level tasks, and a few different datasets (encyclopedic text, elementary version, and Wikipedia).
- Precision at 5 is about 49% on a document-level evaluation.
- Michael's thesis quantifies the benefits of different components.
 - Ranking is very important.
- See thesis for lots of error analysis.
 - Punchline: keep working on core NLP.

User Study

- 17 teachers were given articles and asked to write quizzes (three questions).
 - Encyclopedia Britannica, history textbook, and U.S. Department of Energy materials for schoolchildren.
- In one condition they got to use a tool that suggested questions. Control: no suggestions.
- Measured time, self-reported mental effort, question acceptability.

1. Mouse over a sentence in the article to see suggested questions for that sentence (click to “lock” the sentence).

2. View ranked lists of suggested questions here.

3. Click on questions to select them (i.e., to add them to your quiz).

8. Click on shortcuts to see questions for the whole article (not just for 1 sentence).

7. Search for questions containing certain keywords.

4. View your list of selected questions at the bottom of the screen.

5. Add your own questions by clicking on “add your own” and typing them in.

6. Revise questions by clicking on the question or answer textboxes. Your revisions are saved automatically.

The screenshot shows a web application titled "A Tool for Generating Factual Questions". It features a main article on the left, a "Questions" panel in the center, and a "Quick Search" panel on the right. At the bottom, there are sections for "Selected Questions" and "Answers".

Annotations with red arrows point to the following elements:

- 1:** Points to a sentence in the article: "Rich oil fields were discovered in Abu Dhabi in 1958. Commercial production of oil began in 1962. The opening of Port Zayed in the early 1970s encouraged the city's economic development. Its primary exports are petroleum and petroleum products."
- 2:** Points to the "Questions" panel, which displays a list of suggested questions, with the first one highlighted: "When were Rich oil fields discovered in Abu Dhabi?".
- 3:** Points to a question in the "Questions" panel.
- 4:** Points to the "Selected Questions" section at the bottom, which shows the selected question: "When were Rich oil fields discovered in Abu Dhabi?".
- 5:** Points to the "add your own" button in the "Selected Questions" section.
- 6:** Points to the "Answers" section, which shows the answer: "in 1958".
- 7:** Points to the "Quick Search" panel, which includes a search bar and a "Search" button.
- 8:** Points to the "click to show all questions" link in the "Quick Search" panel.

Results

- Time: 5.0 minutes reduced to 3.8.
- Small, significant reduction in self-reported mental effort.
- Small, insignificant drop in acceptability rate; major change in distribution of questions.
 - Suggested questions led teachers to make easier quizzes.
- 10/17 would use the tool; 16/17 found it easy to use.

A Second Foray

- (Going beyond single sentences.)
- Martins and Smith (ILP workshop 2009): a joint model of sentence selection and sentence compression for extractive summarization.
 - This is hard. We used integer linear programming to solve the problem jointly, but learned two separate models on two separate datasets.
 - See more recent work by Berg-Kirkpatrick (ACL 2011) that overcomes the data problem.

Conclusions (1)

- Statistical models are building blocks, not black boxes.
 - We can put them together in naïve ways or sophisticated ways.
 - They don't require us to forgo good linguistic representations.
 - Talk with the parsing and machine learning people!

Conclusions (2)

- Small, noisy, imperfect data scenarios:
need more knowledge in the model.
 - I talked about formalisms, features, informed overgeneration, ...
 - We should also think about: priors, exploiting raw data, ...
- But building new datasets is honest work.

Conclusions (3)

- Predictive tasks as a useful abstraction that help us design better models that work for a range of problems.
- But let's not get stuck on the same tasks!

Acknowledgments

Student collaborators:

- Dipanjan Das
- Kevin Gimpel
- Michael Heilman
- André Martins
- Mengqiu Wang
(Stanford)

Sponsors: ICTI, NSF, Google,
DARPA

