A TAG formalism for Parsing and Translation

Xavier Carreras

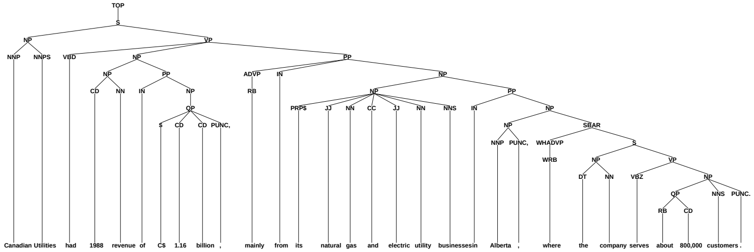
UPC

Joint work with Michael Collins, Terry Koo

Problem 1: Parsing

- Data: a treebank with pairs of sentences and parse trees
- Goal: learn a model that can predict the parse tree of a sentence

Canadian Utilities had 1988 revenue of C$ 1.16 billion, mainly from its natural gas and electric utility businesses in Alberta, where the company serves about 800,000 customers.



Statistical Machine Translation

- Data: a bilingual parallel corpus

Wiederaufnahme der Sitzungsperiode.	Resumption of the session.
Gibt es Einwände?	Are there any comments?
Wissenschaftlich betrachtet haben Sie recht.	Scientifically you are right.
Sie sind äußerst wichtig.	They are extremely important.
Das Wort hat Herr Simpson.	Mr Simpson has the floor.
Bedauerlicherweise wurde dies nicht eingehalten.	Sadly, that has not been the case.
Vielen Dank, Herr Simpson.	Thank you very much, Mr Simpson.

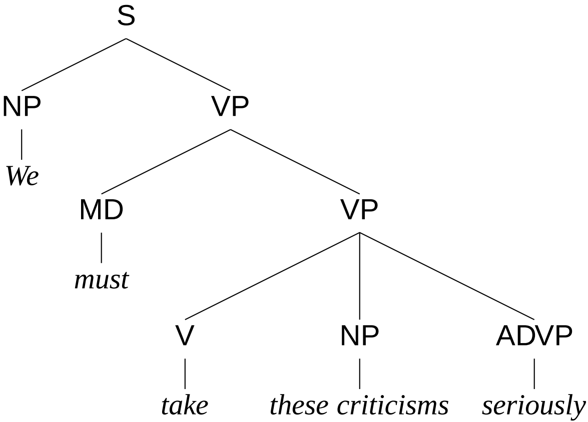
- Goal: learn a model that can predict an English translation given a German sentence

Problem 2: Translation as Parsing

wir müssen diese kritik ernst nehmen

(we must these criticisms seriously take)

⇓



Grammar Formalisms

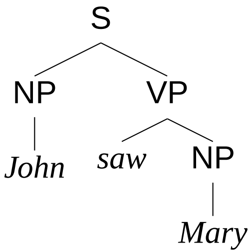
- The choice of grammar formalism implies a decomposition of parse trees into smaller units
- This choice is critical to:
  1. Representations that can be used
  2. Computational efficiency of underlying algorithms

Probabilistic Context-Free Grammars (PCFG)

A simple CFG:

```
S -> NP VP
...
NP -> John
NP -> Mary
...
VP -> slept
VP -> saw NP
...
```

A parse tree:



$$P(\mathit{Tree}) = P(\texttt{S -> NP VP} \mid \texttt{S}) \times P(\texttt{NP -> John} \mid \texttt{NP}) \times$$

$$P(\texttt{VP -> saw NP} \mid \texttt{VP}) \times P(\texttt{NP -> Mary} \mid \texttt{NP})$$

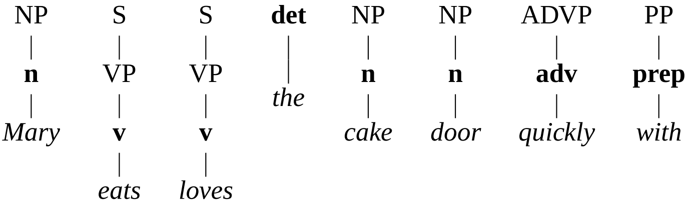
Outline

- A Tree Adjoining Grammar (TAG) formalism
- A TAG-based discriminative parser
- A TAG-based translation model

A TAG-Style Formalism

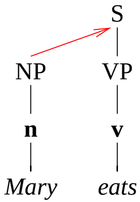
(Carreras, Collins, and Koo, 2008)

- In Tree Adjoining Grammar (TAG, Joshi, 1985) the grammar is defined by *a set of elementary trees*.
- Our elementary trees are Spines (See also Shen and Joshi, 2005):



A Combination Operation: *Sister Adjunction*

Sister adjunctions are used to combine spines to form trees.

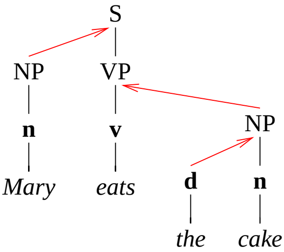


An adjunction operation attaches:

- A modifier spine
- To some position of a head spine

A Combination Operation: *Sister Adjunction*

Sister adjunctions are used to combine spines to form trees.

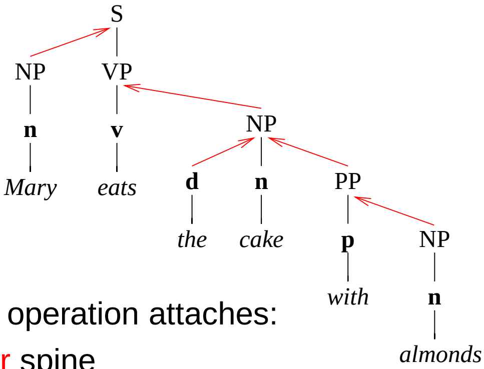


An adjunction operation attaches:

- A modifier spine
- To some position of a head spine

A Combination Operation: *Sister Adjunction*

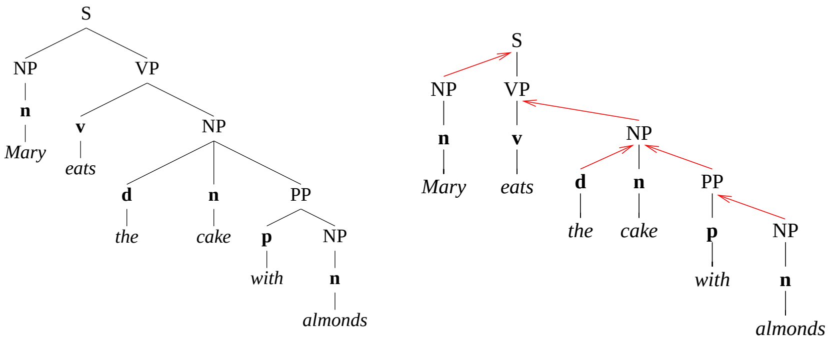
Sister adjunctions are used to combine spines to form trees.



An adjunction operation attaches:

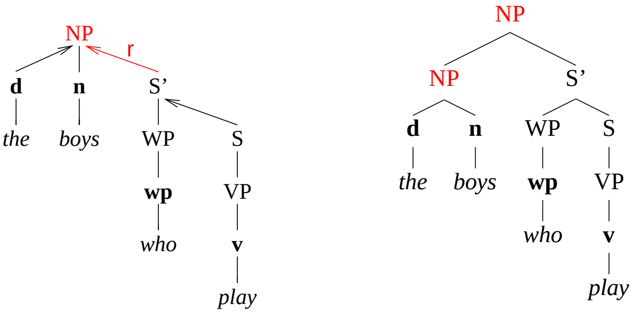
- A modifier spine
- To some position of a head spine

The Decomposition into Spines and Adjunctions



Another Operation: *Regular Adjunction*

Regular adjunctions add one level to the syntactic constituent they point to.

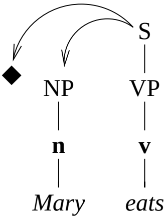


N.B.: This operation is simpler than adjunctions in classic TAG, resulting in more efficient parsing costs.

A Little More Formally....

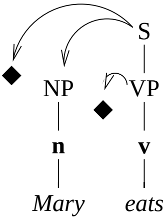
- Each spine has a separate left/right weighted finite-state automaton (HMM) at each level of the tree (in this case S, VP)
- The automata generate sequences of modifier spines at each level of the tree
- Parsing complexity: $O(n^3G)$ where n is the length of the string, G is a grammar constant (Eisner 2000)

A Little More Formally....



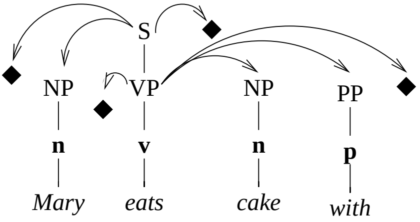
- Each spine has a separate left/right weighted finite-state automaton (HMM) at each level of the tree (in this case S, VP)
- The automata generate sequences of modifier spines at each level of the tree
- Parsing complexity: $O(n^3 G)$ where n is the length of the string, G is a grammar constant (Eisner 2000)

A Little More Formally....



- Each spine has a separate left/right weighted finite-state automaton (HMM) at each level of the tree (in this case S, VP)
- The automata generate sequences of modifier spines at each level of the tree
- Parsing complexity: $O(n^3G)$ where n is the length of the string, G is a grammar constant (Eisner 2000)

A Little More Formally....

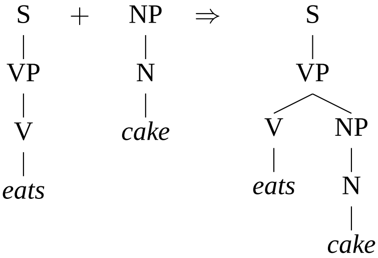


- Each spine has a separate left/right weighted finite-state automaton (HMM) at each level of the tree (in this case S, VP)
- The automata generate sequences of modifier spines at each level of the tree
- Parsing complexity: $O(n^3 G)$ where n is the length of the string, G is a grammar constant (Eisner 2000)

Advantages of TAG

- Lexical entries naturally capture constraints associated with lexical items

- Probabilities/costs can be associated with combination operations:



Outline

- A Tree Adjoining Grammar (TAG) formalism
- A TAG-based discriminative parser
- A TAG-based translation model

Structured Prediction Models for Parsing

- Conditional random fields (CRFs), and other discriminative models, are a powerful alternative to HMMs
  - A key strength: flexible representations
- **Can we generalize CRF-style models to parsing?**

Conditional Random Fields

(Lafferty, McCallum, and Pereira, 2001)

- Goal: learn a function from $\mathbf{x}$ to $\mathbf{y}$ where
  - $\mathbf{x} = x_1 x_2 \ldots x_n$ is an input sequence (e.g., a sequence of words)
  - $\mathbf{y} = y_1 y_2 \ldots y_n$ is an output sequence (e.g., a sequence of underlying states)

The Building Blocks for CRFs: Feature Vectors

$\mathbf{y} =$	N	V	D	N	P	N
$\mathbf{x} =$	Mary	eats	the	cake	with	almonds

- $\mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i)$ is a *feature vector* representing the transition $y_{i-1} \rightarrow y_i$ at position i in the sentence
- e.g., $i = 4$, $y_{i-1} =$ D, $y_i =$ N

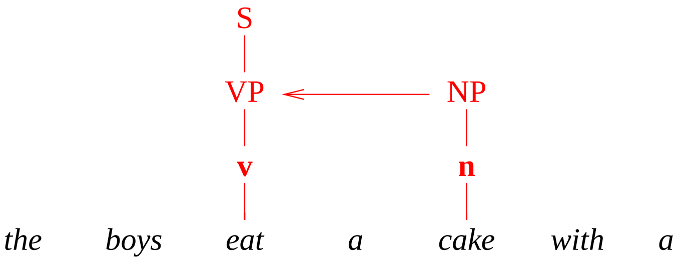
Conditional Random Fields

- Model form:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i)$$

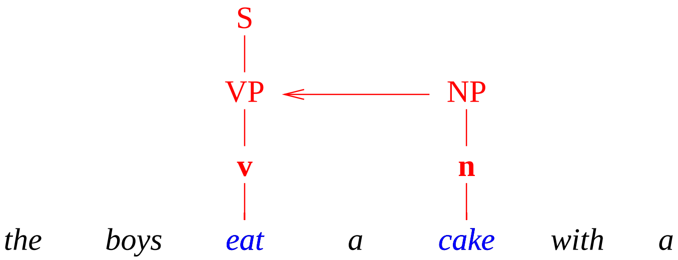
  - $\mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i)$ is a feature vector, $\mathbf{w}$ is a parameter vector
  - $\mathbf{w} \cdot \mathbf{f}(\mathbf{x}, i, y_{i-1}, y_i)$ is a measure of the plausibility/probability of state y_{i-1} being followed by state y_i at position i in the sentence $\mathbf{x}$
  - Can find $\mathbf{y}^*$ using the Viterbi algorithm

Features on Adjunctions



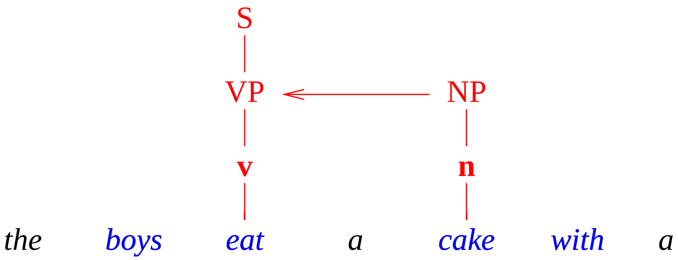
- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Features on Adjunctions



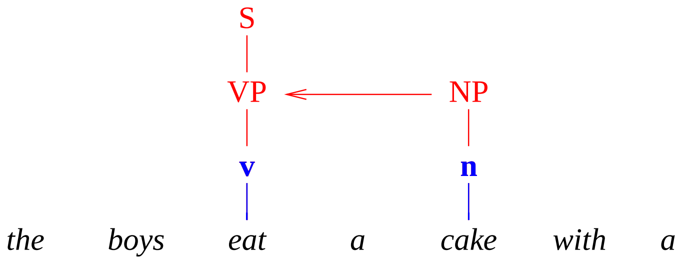
- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Features on Adjunctions



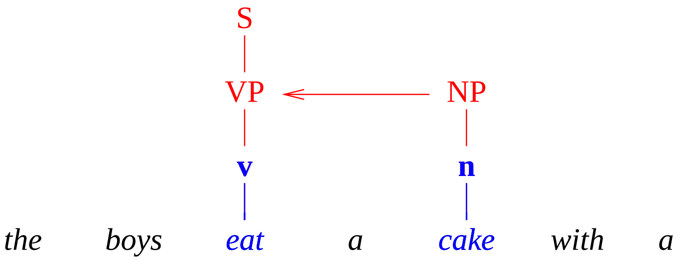
- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Features on Adjunctions



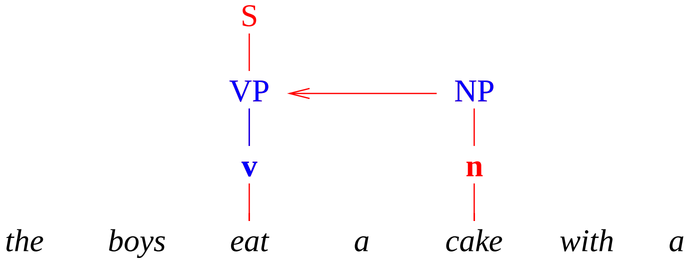
- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Features on Adjunctions



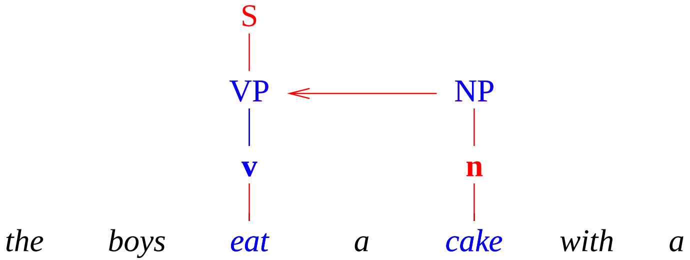
- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Features on Adjunctions



- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Features on Adjunctions

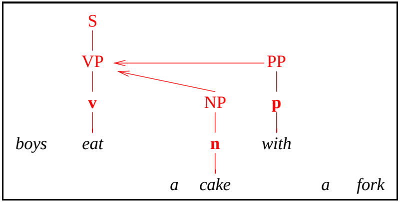


- Feature vectors $\mathbf{f}(\mathbf{x}, h, m, \sigma_h, \sigma_m, \text{POS})$ where
  - $\mathbf{x}$ is the sentence
  - $h = 3$ (index of head word), $m = 5$ (index of modifier word)
  - σ_h and σ_m are the head and modifier spines
  - POS is the position being adjoined into (e.g., VP)

Higher-Order Features on Adjunctions

We can extend the model with higher-order feature functions:

siblings

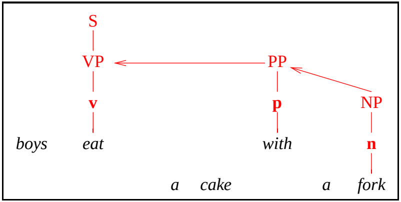


$O(n^3G)$

[Eisner 2000]

[McDonald & Pereira, 2006]

grandchildren



$O(n^4G)$

[Carreras, 2007]

[Koo & Collins, 2010]

A TAG-Based Model

- Goal: map an input sentence $\mathbf{x}$ to a parse tree $\mathbf{y}$
- Model form:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{r \in \mathbf{y}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, r)$$

  where each r is a tuple $\langle h, m, \sigma_h, \sigma_m, \text{POS} \rangle$ representing a combination of two spines in $\mathbf{y}$

- Parameter estimation: we used the averaged perceptron
- The inference problem: How to compute $\mathbf{y}^*$?
  Dynamic Programming + Coarse-to-fine strategy

Test results on WSJ data

FULL PARSERS	precision	recall	F_1
PCFG	·	·	$\sim$65
PCFG + parent annotations	·	·	$\sim$80
PCFG + head annotations	·	·	$\sim$88
Petrov et al. 2007	·	·	88.3
Finkel et al. 2008	88.2	87.8	88.0
Charniak 2000	89.5	89.6	89.6
Petrov & Klein 2007	90.2	89.9	90.1
this work	**91.4**	**90.7**	**91.1**

RERANKERS	precision	recall	F_1
Collins 2000	89.9	89.6	89.8
Charniak & Johnson 2005	·	·	91.4

Outline

- A Tree Adjoining Grammar (TAG) formalism
- A TAG-based discriminative parser
- A TAG-based translation model

Phrase-based Systems: Derivations

In wenigen Tagen finden Parlamentswahlen in Slowenien statt

- Translation involves:
  1. Segmenting the input into phrases, and choosing a translation for each phrase
  2. Choosing an ordering of the resulting English phrases

Phrase-based Systems: Derivations

[In wenigen] [Tagen] [finden] [Parlamentswahlen] [in Slowenien] [statt]
[In a few] [days] [take] [elections] [in Slovenia] [place]

- Translation involves:
  1. Segmenting the input into phrases, and choosing a translation for each phrase
  2. Choosing an ordering of the resulting English phrases

Phrase-based Systems: Derivations

[In wenigen] [Tagen] [finden] [Parlamentswahlen] [in Slowenien] [statt]
[In a few] [days] [take] [elections] [in Slovenia] [place]

$\Downarrow$

[In a few] [days] [elections] [take] [place] [in Slovenia]

- Translation involves:
  1. Segmenting the input into phrases, and choosing a translation for each phrase
  2. Choosing an ordering of the resulting English phrases

Phrase-base Systems: a Phrase Table

auch	⇒	also
auf nationaler und	⇒	at national and
bereits	⇒	already
dass	⇒	that
der kommission	⇒	the commission
des besitzstandes	⇒	of the acquis
die wichtigste	⇒	the most important
gemeinschaftspolitiken	⇒	community policies
im dezember in nizza	⇒	in december in nice
in diesem bericht enthaltenen	⇒	contained in this report
ist notwendig und	⇒	is necessary and
menschenrechte	⇒	human rights
oppositionsparteien und	⇒	opposition parties and
positiven auswirkungen der	⇒	positive effects of
trennlinie	⇒	dividing line
umsetzung der menschenrechte	⇒	implementation of human rights
und die	⇒	and the
wird schrittweise	⇒	should be gradually
zu beachten haben	⇒	to bear in mind

Phrase-base Systems: a Phrase Table

auch	$\Rightarrow$	also	(0.73)
auf nationaler und	$\Rightarrow$	at national and	(0.34)
bereits	$\Rightarrow$	already	(0.65)
dass	$\Rightarrow$	that	(0.92)
der kommission	$\Rightarrow$	the commission	(0.85)
des besitzstandes	$\Rightarrow$	of the acquis	(0.56)
die wichtigste	$\Rightarrow$	the most important	(0.44)
gemeinschaftspolitiken	$\Rightarrow$	community policies	(0.31)
im dezember in nizza	$\Rightarrow$	in december in nice	(0.37)
in diesem bericht enthaltenen	$\Rightarrow$	contained in this report	(0.81)
ist notwendig und	$\Rightarrow$	is necessary and	(0.94)
menschenrechte	$\Rightarrow$	human rights	(0.78)
oppositionsparteien und	$\Rightarrow$	opposition parties and	(0.53)
positiven auswirkungen der	$\Rightarrow$	positive effects of	(0.58)
trennlinie	$\Rightarrow$	dividing line	(0.67)
umsetzung der menschenrechte	$\Rightarrow$	implementation of human rights	(0.96)
und die	$\Rightarrow$	and the	(0.89)
wird schrittweise	$\Rightarrow$	should be gradually	(0.85)
zu beachten haben	$\Rightarrow$	to bear in mind	(0.44)

Trigram Language Models

$$score_{LM}(\text{In a few days elections take place in Slovenia})$$

$$\begin{aligned} = \; & P(\text{In}|\text{* *}) \times P(\text{a}|\text{* In}) \times P(\text{few}|\text{In a}) \times \\ & P(\text{days}|\text{a few}) \times P(\text{elections}|\text{few days}) \times \\ & P(\text{take}|\text{days elections}) \times P(\text{place}|\text{elections take}) \times \\ & P(\text{in}|\text{take place}) \times P(\text{Slovenia}|\text{place in}) \end{aligned}$$

Word-order Differences

bei all diesen problemen beschränkt sich der bericht brok darauf, von anpassung oder reformen zu sprechen.

$$\Downarrow$$

on all these subjects, the brok report confines itself to discussing adaptation and reform.

Word-order Differences

bei all diesen problemen beschränkt sich der bericht brok darauf, von anpassung oder reformen zu sprechen.

Paraphrase: on all these subjects confines itself the report brok on adaptation and reform to speak

$$\Downarrow$$

on all these subjects, the brok report confines itself to discussing adaptation and reform.

Word-order Differences

bei all diesen problemen beschränkt sich der bericht brok darauf, von anpassung oder reformen zu sprechen.

Paraphrase: on all these subjects confines itself the report brok on adaptation and reform to speak

$$\Downarrow$$

on all these subjects, the brok report confines itself to discussing adaptation and reform.

Translation: with all these problems is limited to the report brok to talk about reform or adjustment.

Word Order Differences

English: the dog has eaten the bone on Wednesday
German: the dog has the bone on Wednesday eaten
German: on Wednesday has the dog the bone eaten
German: the bone has the dog on Wednesday eaten

English: the president of the United States made the speech
Arabic: made the president of the United States the speech
Japanese: the president of the United States the speech made

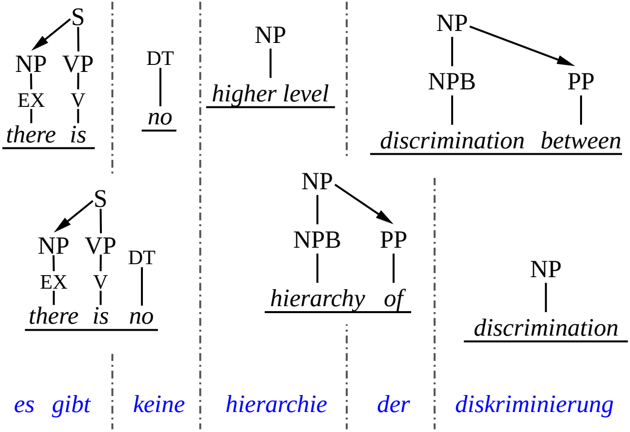
Word Order Differences

English:	the dog has eaten the bone on Wednesday
German:	the dog has the bone on Wednesday eaten
German:	on Wednesday has the dog the bone eaten
German:	the bone has the dog on Wednesday eaten

English:	Mary says that the dog has eaten the bone on Wednesday
German:	Mary says that the dog the bone on Wednesday eaten has

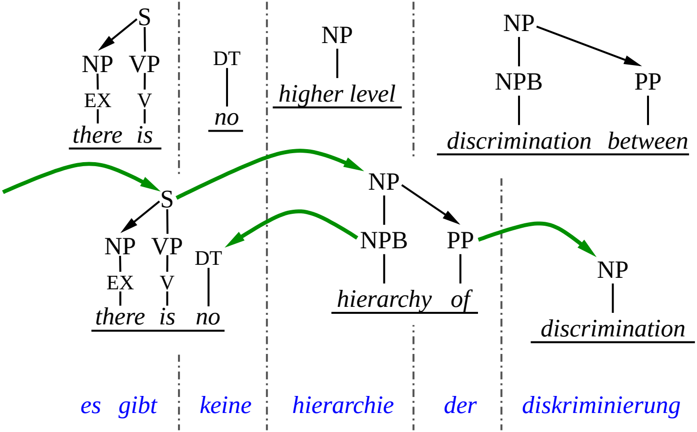
English:	the president of the United States made the speech
Arabic:	made the president of the United States the speech
Japanese:	the president of the United States the speech made

Phrase-based Translation with TAG operations



segmentation + s-phrase selection + adjunctions

Phrase-based Translation with TAG operations



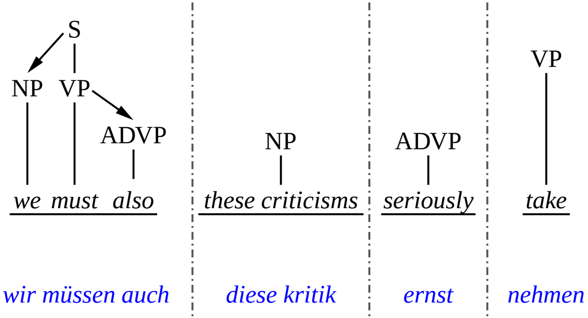
segmentation + s-phrase selection + adjunctions

Phrase-based Translation with TAG operations

A TAG-based syntactic translation model. Properties:

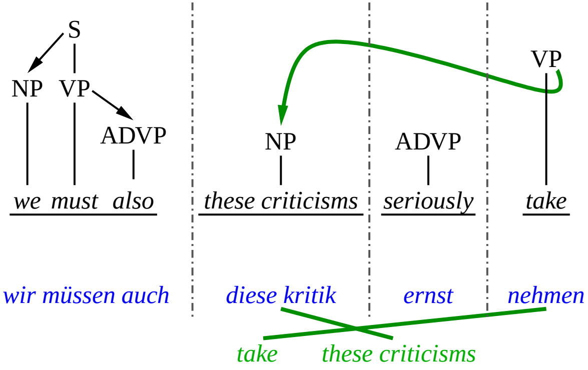
- Retains the full set of lexical entries of a phrase-based system
- Straightforward integration of a syntactic language model

Reordering via Non-Projective Operations



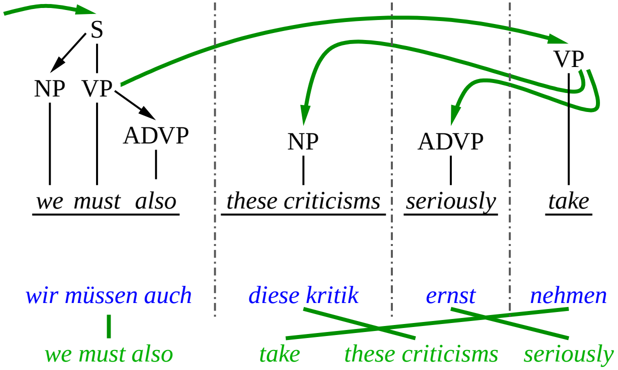
segmentation + s-phrase selection + non-projective adjunctions

Reordering via Non-Projective Operations



segmentation + s-phrase selection + non-projective adjunctions

Reordering via Non-Projective Operations



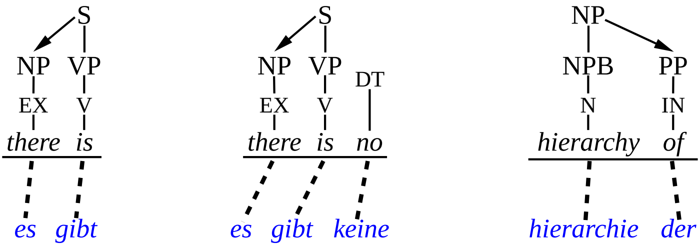
segmentation + s-phrase selection + non-projective adjunctions

Reordering via Non-Projective Operations

We model reordering with flexible non-projective adjunctions.

- How to control reorderings?
  - A discriminative model inspired by work in dependency parsing (e.g. [McDonald et al. 05])
  - Hard constraints
- How to decode efficiently?
  - A novel beam-search algorithm

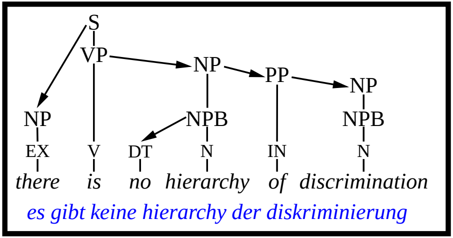
S-phrases: Syntactic Phrase-entries for Translation



An s-phrase consists of:

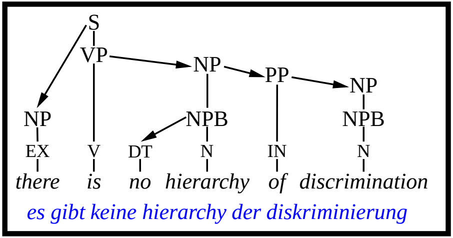
- Foreign words
- English words
- A syntactic structure
- An alignment

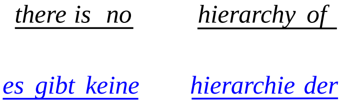
Extraction of S-phrases



- Training example = source sentence + English sentence + English parse tree
- We use phrasal entries from a standard phrase-based approach

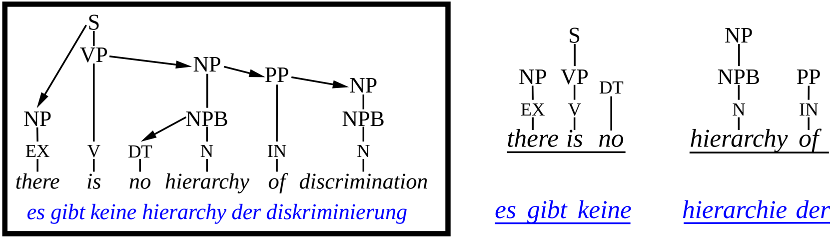
Extraction of S-phrases





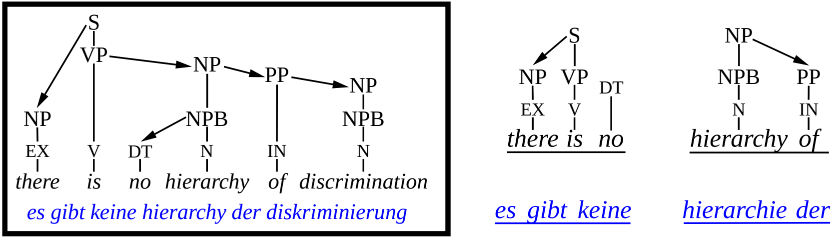
- Training example = source sentence + English sentence + English parse tree
- We use phrasal entries from a standard phrase-based approach

Extraction of S-phrases



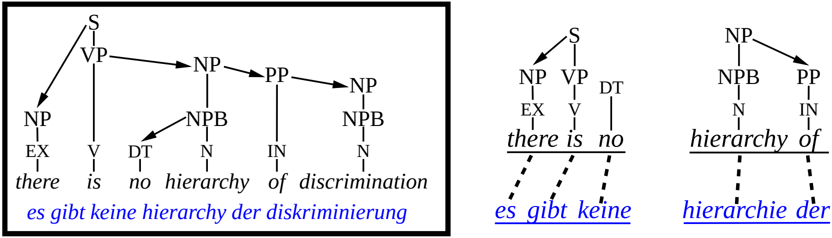
- Training example = source sentence + English sentence + English parse tree
- We use phrasal entries from a standard phrase-based approach

Extraction of S-phrases



- Training example = source sentence + English sentence + English parse tree
- We use phrasal entries from a standard phrase-based approach

Extraction of S-phrases



- Training example = source sentence + English sentence
  + English parse tree

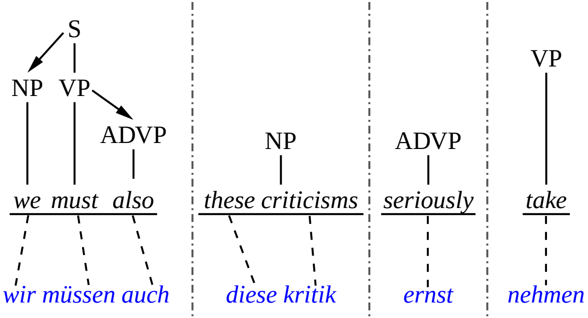
- We use phrasal entries from a standard phrase-based approach

Derivations

wir müssen auch *diese kritik* *ernst* *nehmen*

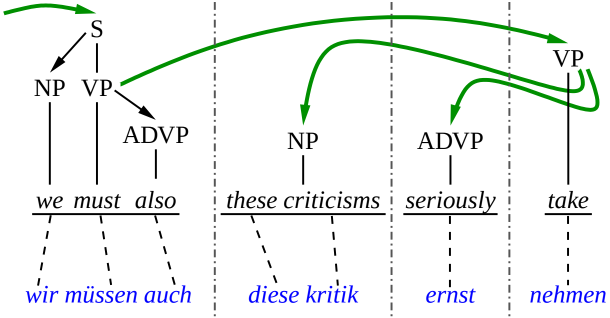
- A derivation:
  - Step 1: segment the input sentence,
    and choose an s-phrase for each segment
  - Step 2: connect s-phrases with adjunctions

Derivations



- A derivation:
  - Step 1: segment the input sentence,
    and choose an s-phrase for each segment
  - Step 2: connect s-phrases with adjunctions

Derivations



- A derivation:
  - Step 1: segment the input sentence,
    and choose an s-phrase for each segment
  - Step 2: connect s-phrases with adjunctions

Model

- Model score for a derivation d:

$$\begin{aligned} score(d) &= score_{LM}(d) + score_{P}(d) \\ &+ score_{SYN}(d) + score_{R}(d) \end{aligned}$$

where

- $score_{LM}$ is a trigram language model
- $score_{P}$ is a sum of standard phrase-based scores
- $score_{SYN}$ is a syntactic language model [Charniak et al. 03] [Shen et al. 08] (probabilities are associated with adjunctions)
- $score_{R}$ is a sum of discriminative adjunction scores

Model

- Model score for a derivation d:

$$\begin{aligned} score(d) &= score_{LM}(d) + score_{P}(d) \\ &+ score_{SYN}(d) + score_{R}(d) \end{aligned}$$

where

- $score_{LM}$ is a trigram language model
- $score_{P}$ is a sum of standard phrase-based scores
- $score_{SYN}$ is a syntactic language model [Charniak et al. 03] [Shen et al. 08] (probabilities are associated with adjunctions)
- $score_{R}$ is a sum of discriminative adjunction scores

Model

- Model score for a derivation d:

$$\begin{aligned} score(d) &= score_{LM}(d) + score_{P}(d) \\ &+ score_{SYN}(d) + score_{R}(d) \end{aligned}$$

where

- $score_{LM}$ is a trigram language model
- $score_{P}$ is a sum of standard phrase-based scores
- $score_{SYN}$ is a syntactic language model [Charniak et al. 03] [Shen et al. 08] (probabilities are associated with adjunctions)
- $score_{R}$ is a sum of discriminative adjunction scores

Model

- Model score for a derivation d:

$$\begin{aligned} score(d) &= score_{LM}(d) + score_{P}(d) \\ &+ score_{SYN}(d) + score_{R}(d) \end{aligned}$$

where

- $score_{LM}$ is a trigram language model
- $score_{P}$ is a sum of standard phrase-based scores
- $score_{SYN}$ is a syntactic language model [Charniak et al. 03] [Shen et al. 08] (probabilities are associated with adjunctions)
- $score_{R}$ is a sum of discriminative adjunction scores

Model

- Model score for a derivation d:

$$\begin{aligned} score(d) &= score_{LM}(d) + score_{P}(d) \\ &+ score_{SYN}(d) + score_{R}(d) \end{aligned}$$

where

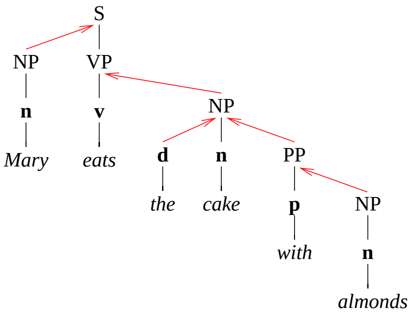
- $score_{LM}$ is a trigram language model
- $score_{P}$ is a sum of standard phrase-based scores
- $score_{SYN}$ is a syntactic language model [Charniak et al. 03] [Shen et al. 08] (probabilities are associated with adjunctions)
- $score_{R}$ is a sum of discriminative adjunction scores

Trigram Language Models

$$score_{LM}(\text{In a few days elections take place in Slovenia})$$

$$\begin{aligned} = \; & P(\text{In}|\text{* *}) \times P(\text{a}|\text{* In}) \times P(\text{few}|\text{In a}) \times \\ & P(\text{days}|\text{a few}) \times P(\text{elections}|\text{few days}) \times \\ & P(\text{take}|\text{days elections}) \times P(\text{place}|\text{elections take}) \times \\ & P(\text{in}|\text{take place}) \times P(\text{Slovenia}|\text{place in}) \end{aligned}$$

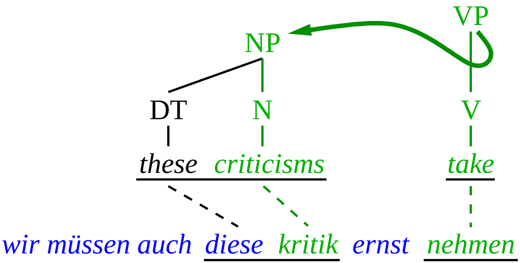
Syntactic Language Models



$P(\text{tree, sentence}) =$

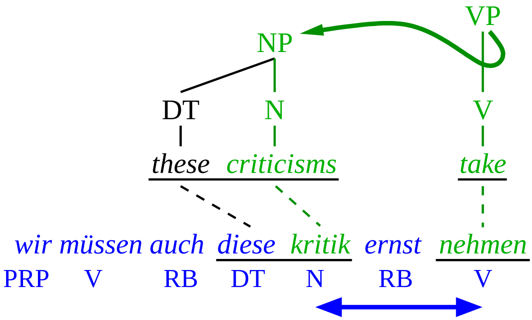
$$P(\text{S-VP-v-eats}|\text{ROOT}) \times P(\text{NP-n-Mary}|\text{S, eats, LEFT}) \times$$
$$P(\text{NP-n-cake}|\text{VP, eats, RIGHT}) \times P(\text{d-the}|\text{NP, cake, LEFT})$$
$$\times \ldots$$

$score_R$: A Discriminative Dependency Model



$score_R(d)$ is a **discriminative dependency model** (related to work in dependency parsing (e.g. [McDonald et al. 05]))

$score_R$: A Discriminative Dependency Model



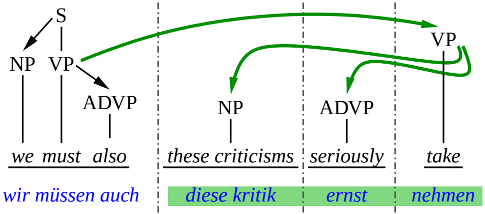
$score_R(d)$ is a **discriminative dependency model** (related to work in dependency parsing (e.g. [McDonald et al. 05]))

π-constituent constraint

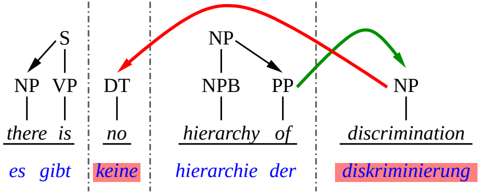
Define π-constituent: a head spine with all its descendants

Constraint any π-constituent must be aligned to a contiguous substring in the source sentence

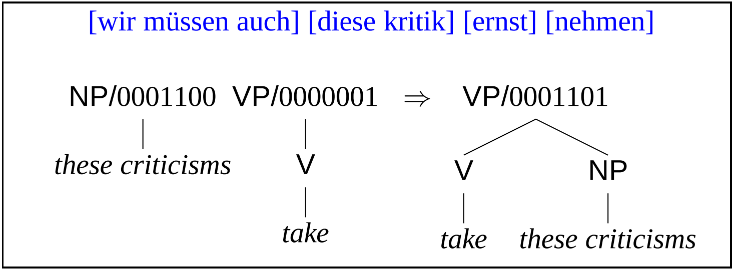
Satisfied:



Violated:



Decoding as Parsing



- Projective parsing: each constituent has an associated **span**
- A generalization: each constituent has a **bit-string** recording which foreign words have been translated
- Beam search strategy: ensures that the top N analyses for each foreign word are explored at each stage

Experiments

German to English using Europarl data (750K training sentences)

Test:

System	BLEU score
Phrase-based system (Pharaoh)	24.58
Syntax-based system	25.04 (+0.46)

significant ($p = 0.021$) under paired bootstrap resampling [Koehn 04]
close to significant ($p = 0.058$) under the sign test [Collins et al. 05]

Human Evaluations

Ref: `Now, however, we are seeing that president Putin is pursuing a policy of openness towards the west.`

`Now, however, we see that mr president Putin is pursuing a policy of openness towards the west.`

`We are, however, now that president Putin a policy of openness to the west out of blackmail.`

	Syntax	PB	=	Total
Syntax	51	3	7	61
PB	1	25	11	37
=	21	14	67	102
Total	73	42	85	200

both results are significant with $p < 0.05$ under the sign test

Human Evaluations

Ref: `Now, however, we are seeing that president Putin is pursuing a policy of openness towards the west.`

Syn: `Now, however, we see that mr president Putin is pursuing a policy of openness towards the west.`

PB : `We are, however, now that president Putin a policy of openness to the west out of blackmail.`

	Syntax	PB	=	Total
Syntax	51	3	7	61
PB	1	25	11	37
=	21	14	67	102
Total	73	42	85	200

both results are significant with $p < 0.05$ under the sign test

Translation Examples

Reference: on all these subjects, the brok report confines itself to discussing adaptation and reform.

Phrase-based: in all these issues is limited to the brok report, adjustment or reforms to speak.

Syntax: the brok report is limited to speak of adjustment or reforms in all these issues.

Translation Examples

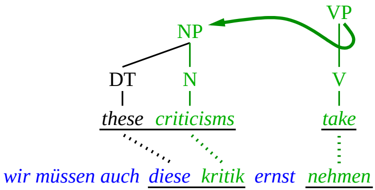
Reference: i believe that deferring the issue would be the worst possible option, both for the citizens of europe and for the citizens of the candidate countries.

Phrase-based: i believe, however, that postpone a decision would be the worst possible both for the citizens of europe , as well as for the citizens of the candidate countries.

Syntax: i believe, however, that a postponement would be the worst possible choice both for the citizens of the union and for the citizens of the candidate countries.

Future Work

A TAG-based syntactic translation model



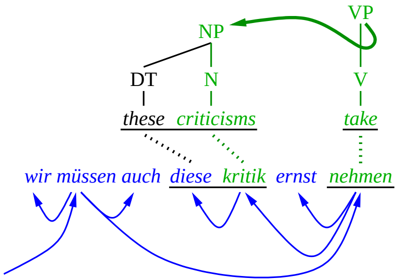
Non-projective adjunctions for reordering:

- Arbitrary reorderings
- Discriminative dependency model

Future work: Condition on syntactic structure of the source string

Future Work

A TAG-based syntactic translation model



Non-projective adjunctions for reordering:

- Arbitrary reorderings
- Discriminative dependency model

Future work: Condition on syntactic structure of the source string

Summary

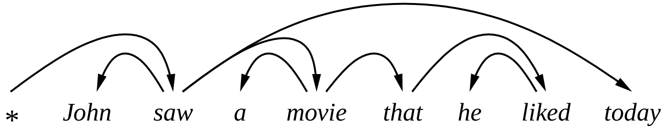
- A TAG-based formalism. Key points:
  - Combines dependency and constituency based representations
  - Allows relatively efficient parsing algorithms
- A TAG-based discriminative parser.
  Key points: feature-vector representations of TAG adjunctions, coarse-to-fine inference
- A TAG-based translation model. Key points:
  non-projective parsing operations, a discriminative dependency model

Extra Slides

Inference: Key Points

- Dynamic programming algorithms can be applied to the TAG grammars
- Exact inference is still very expensive
- A solution: *coarse-to-fine* dynamic programming (e.g., (Charniak, 1997; Charniak and Johnson, 2005))
  - Use a first-pass, simple, computationally-cheap model to restrict the search space of the full model

Dependency Structures

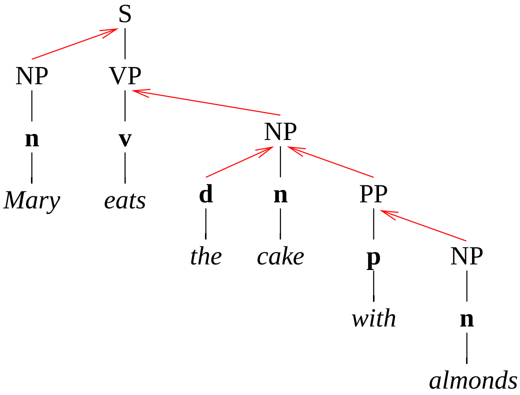


- Directed arcs represent *dependencies* between a *head word* and a *modifier word*.
- Dependency parsing models of McDonald et al. (2005, 2006):

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{r \in \mathbf{y}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, r)$$

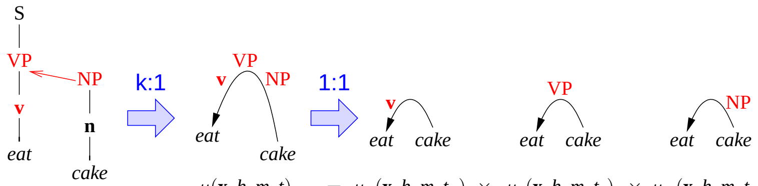
  where each r is a tuple $\langle h, m \rangle$ representing a dependency from modifier m to head h
- Can be parsed with DP in $O(Gn^3)$ time

TAG Parses and Dependency Structures



- A dependency structure augmented with *spines*, and *attachment positions*

Coarse-to-fine Dynamic Programming



$$\mu(\mathbf{x}, h, m, t) = \mu_{\mathrm{H}}(\mathbf{x}, h, m, t_{\mathrm{H}}) \times \mu_{\mathrm{P}}(\mathbf{x}, h, m, t_{\mathrm{P}}) \times \mu_{\mathrm{M}}(\mathbf{x}, h, m, t_{\mathrm{M}})$$

- Coarse-to-fine approach: we only allow the full TAG model to consider dependencies that have high probability under a (simple) dependency model
- The simple model estimates dependency probabilities in $O(n^3G)$ time, where $G \approx 60$ is the number of non-terminals (i.e., *VP*, *NP*, *S*, etc.)

Effect of the Beam (Validation Data)

	1st stage		2nd stage		
α	active	cov.	orac.	speed	F_1 error
10^{-4}	0.07	97.7	97.0	5:15	8.9
10^{-5}	0.16	98.5	97.9	11:45	8.4
10^{-6}	0.34	99.0	98.5	21:50	8.0

We can discard 99.6% of the possible adjunctions
and retain 98.5% of the correct syntactic constituents

Beam Search Decoding

0. Data structures: $\mathcal{Q}_i$ for $i = 1 \ldots n$ is a set of hypotheses for each length i, $\mathcal{S}$ is a set of chart entries
1. $\mathcal{S} \leftarrow \emptyset$
2. Initialize $\mathcal{Q}_1 \ldots \mathcal{Q}_n$ with basic chart entries derived from phrase entries
3. **For** $i = 1 \ldots n$
4.   **For** any $A \in$ BEAM$(\mathcal{Q}_i)$
5.   **If** $\mathcal{S}$ contains a chart entry with the same signature as A, and which has a higher inside score,
6.   **continue**
7.   **Else**
8.   Add A to $\mathcal{S}$
9.   For any chart entry C that can be derived from A together with another chart entry $B \in \mathcal{S}$, add C to the set $\mathcal{Q}_j$ where $j =$ length(C)
10. **Return** $\mathcal{Q}_n$, a set of items of length n

The Definition of BEAM

(BEAM) Given $\mathcal{Q}_i$, define $\mathcal{Q}_{i,j}$ for $j = 1 \ldots n$ to be the subset of items in $\mathcal{Q}_i$ which have their j'th bit equal to one (i.e., have the j'th source language word translated). Define $\mathcal{Q}'_{i,j}$ to be the N highest scoring elements in $\mathcal{Q}_{i,j}$. Then $\text{BEAM}(\mathcal{Q}_i) = \cup_{j=1}^{n} \mathcal{Q}'_{i,j}$.