A TAG formalism for Parsing and Translation

Xavier Carreras

UPC

Joint work with Michael Collins, Terry Koo
Syntactic Structures

S
  /\     /
 /  \   /  \
NP   VP  \
  /\    /\  \\
 N   V NP PP
   /\   /\  \\
  John saw Mary in Helsinki
Syntactic Structures and Grammatical Relations

⇒ *John* is the subject of *saw*
Syntactic Structures and Grammatical Relations

⇒ Mary is the object of saw
Syntactic Structures and Grammatical Relations

⇒ In Helsinki is a prepositional-phrase (PP) modifier to saw
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

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

- **Goal:** learn a model that can predict an English translation given a German sentence
In a strategic assessment of the Afghanistan conflict given to Robert Gates August 30, revealed by The Washington Post Monday, General McChrystal warned that without an increase in military resources in Afghanistan, the coalition might undergo a “failure.”
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:

\[
S \rightarrow NP \ VP \\
\ldots \\
NP \rightarrow John \\
NP \rightarrow Mary \\
\ldots \\
VP \rightarrow slept \\
VP \rightarrow saw NP \\
\ldots \\
\]

A parse tree:

\[
P(Tree) = P(S \rightarrow NP \ VP \mid S) \times P(NP \rightarrow John \mid NP) \times P(VP \rightarrow saw NP \mid VP) \times P(NP \rightarrow Mary \mid NP)
\]
Outline

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

```
  S
  NP  VP
    n    v
    Mary  eats
    the  cake
```

```
  N
  eat
    n
    the
    cake
```
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
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)
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

```
S
  | VP
  | V
  | saw
```

- Probabilities/costs can be associated with combination operations:

```
S + NP ⇒ S
  | VP
  | N
  | cake
  | V
  | eats
```

```
V NP
  | eats
  | cake
```
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 $x$ to $y$ where
  - $x = x_1 x_2 \ldots x_n$ is an input sequence
    (e.g., a sequence of words)
  - $y = y_1 y_2 \ldots y_n$ is an output sequence
    (e.g., a sequence of underlying states)
The Building Blocks for CRFs: Feature Vectors

\[ y = \text{N V D N P N} \]

\[ x = \text{Mary eats the cake with almonds} \]

- \( f(x, i, y_{i-1}, y_i) \) is a feature vector representing the transition \( y_{i-1} \rightarrow y_i \) at position \( i \) in the sentence

- e.g., \( i = 4, y_{i-1} = \text{D}, y_i = \text{N} \)
Conditional Random Fields

▶ Model form:

\[ y^* = \arg \max_y \sum_{i=1}^{n} w \cdot f(x, i, y_{i-1}, y_i) \]

▶ \( f(x, i, y_{i-1}, y_i) \) is a feature vector, \( w \) is a parameter vector

▶ \( w \cdot f(x, i, y_{i-1}, y_i) \) is a measure of the plausibility/probability of state \( y_{i-1} \) being followed by state \( y_i \) at position \( i \) in the sentence \( x \)

▶ Can find \( y^* \) using the Viterbi algorithm
Features on Adjunctions

Feature vectors $f(x, h, m, \sigma_h, \sigma_m, \text{POS})$ where

- $x$ is the sentence
- $h = 3$ (index of head word), $m = 5$ (index of modifier word)
- $\sigma_h$ and $\sigma_m$ are the head and modifier spines
- POS is the position being adjoined into (e.g., VP)
Features on Adjunctions

Feature vectors \( f(x, h, m, \sigma_h, \sigma_m, \text{POS}) \) where

- \( x \) is the sentence
- \( h = 3 \) (index of head word), \( m = 5 \) (index of modifier word)
- \( \sigma_h \) and \( \sigma_m \) are the head and modifier spines
- \( \text{POS} \) is the position being adjoined into (e.g., VP)
Features on Adjunctions

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

Feature vectors $f(x, h, m, \sigma_h, \sigma_m, \text{POS})$ where

- $x$ is the sentence
- $h = 3$ (index of head word), $m = 5$ (index of modifier word)
- $\sigma_h$ and $\sigma_m$ are the head and modifier spines
- POS is the position being adjoined into (e.g., VP)
Features on Adjunctions

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

Feature vectors $f(x, h, m, \sigma_h, \sigma_m, \text{POS})$ where

- $x$ is the sentence
- $h = 3$ (index of head word), $m = 5$ (index of modifier word)
- $\sigma_h$ and $\sigma_m$ are the head and modifier spines
- POS is the position being adjoined into (e.g., VP)
Features on Adjunctions

Feature vectors $f(x, h, m, \sigma_h, \sigma_m, \text{POS})$ where

- $x$ is the sentence
- $h = 3$ (index of head word), $m = 5$ (index of modifier word)
- $\sigma_h$ and $\sigma_m$ are the head and modifier spines
- POS is the position being adjoined into (e.g., VP)
A TAG-Based Model

- Goal: map an input sentence $x$ to a parse tree $y$

- Model form:

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

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

- Parameter estimation: we used the averaged perceptron

- The inference problem: How to compute $y^*$?
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):

\[ y^* = \arg \max_y \sum_{r \in y} w \cdot f(x, r) \]

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
A dependency structure augmented with *spines*, and
*attachment positions*
Coarse-to-fine Dynamic Programming

-$\mu(x, h, m, t) = \mu_H(x, h, m, t_H) \times \mu_P(x, h, m, t_P) \times \mu_M(x, h, m, t_M)$

- 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.)
## Test results on WSJ data

<table>
<thead>
<tr>
<th><strong>FULL PARSERS</strong></th>
<th>precision</th>
<th>recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFG</td>
<td>.</td>
<td>.</td>
<td>$\approx$65</td>
</tr>
<tr>
<td>PCFG + parent annotations</td>
<td>.</td>
<td>.</td>
<td>$\approx$80</td>
</tr>
<tr>
<td>PCFG + head annotations</td>
<td>.</td>
<td>.</td>
<td>$\approx$88</td>
</tr>
<tr>
<td>Petrov et al. 2007</td>
<td></td>
<td>.</td>
<td>88.3</td>
</tr>
<tr>
<td>Finkel et al. 2008</td>
<td>88.2</td>
<td>87.8</td>
<td>88.0</td>
</tr>
<tr>
<td>Charniak 2000</td>
<td>89.5</td>
<td>89.6</td>
<td>89.6</td>
</tr>
<tr>
<td>Petrov &amp; Klein 2007</td>
<td>90.2</td>
<td>89.9</td>
<td>90.1</td>
</tr>
<tr>
<td><strong>this work</strong></td>
<td>91.4</td>
<td>90.7</td>
<td>91.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>RERANKERS</strong></th>
<th>precision</th>
<th>recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collins 2000</td>
<td>89.9</td>
<td>89.6</td>
<td>89.8</td>
</tr>
<tr>
<td>Charniak &amp; Johnson 2005</td>
<td>.</td>
<td>.</td>
<td>91.4</td>
</tr>
</tbody>
</table>
Effect of the Beam (Validation Data)

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>1st stage</th>
<th>2nd stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>active</td>
<td>orac.</td>
</tr>
<tr>
<td></td>
<td>cov.</td>
<td>speed</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>0.07</td>
<td>97.0</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>0.16</td>
<td>97.9</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>0.34</td>
<td>98.5</td>
</tr>
</tbody>
</table>

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

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

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
<table>
<thead>
<tr>
<th>German Phrase</th>
<th>Possible English Translation</th>
<th>Confidence Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>auch</td>
<td>also</td>
<td>0.73</td>
</tr>
<tr>
<td>auf nationaler und</td>
<td>at national and</td>
<td>0.34</td>
</tr>
<tr>
<td>bereits</td>
<td>already</td>
<td>0.65</td>
</tr>
<tr>
<td>dass</td>
<td>that</td>
<td>0.92</td>
</tr>
<tr>
<td>der kommission</td>
<td>the commission</td>
<td>0.85</td>
</tr>
<tr>
<td>des besitzstandes</td>
<td>of the acquis</td>
<td>0.56</td>
</tr>
<tr>
<td>die wichtigste</td>
<td>the most important</td>
<td>0.44</td>
</tr>
<tr>
<td>gemeinschaftspolitiken</td>
<td>community policies</td>
<td>0.31</td>
</tr>
<tr>
<td>im dezember in nizza</td>
<td>in december in nice</td>
<td>0.37</td>
</tr>
<tr>
<td>in diesem bericht enthalten</td>
<td>contained in this report</td>
<td>0.81</td>
</tr>
<tr>
<td>ist notwendig und</td>
<td>is necessary and</td>
<td>0.94</td>
</tr>
<tr>
<td>menschenrechte</td>
<td>human rights</td>
<td>0.78</td>
</tr>
<tr>
<td>oppositionsparteien und</td>
<td>opposition parties and</td>
<td>0.53</td>
</tr>
<tr>
<td>positiven auswirkungen der trennlinie</td>
<td>positive effects of dividing line</td>
<td>0.58</td>
</tr>
<tr>
<td>umsetzung der menschenrechte</td>
<td>implementation of human rights</td>
<td>0.96</td>
</tr>
<tr>
<td>und die</td>
<td>and the</td>
<td>0.89</td>
</tr>
<tr>
<td>wird schrittweise</td>
<td>should be gradually</td>
<td>0.85</td>
</tr>
<tr>
<td>zu beachten haben</td>
<td>to bear in mind</td>
<td>0.44</td>
</tr>
</tbody>
</table>
Trigram Language Models

\[
score_{LM}(\text{In a few days elections take place in Slovenia}) = P(\text{In}|^*^*) \times P(\text{a}|^*\text{ In}) \times P(\text{few}|\text{In a}) \times \\
P(\text{days}|\text{a few}) \times P(\text{elections}|\text{few days}) \times \\
P(\text{take}|\text{days elections}) \times P(\text{place}|\text{elections take}) \times \\
P(\text{in}|\text{take place}) \times P(\text{Slovenia}|\text{place in})
\]
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

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

Translation: With all these problems is limited to the report Brok to talk about reform or adjustment.
Reference: consequently proposals are submitted to parliament under the assent procedure, meaning that parliament can no longer table amendments, as directives in this area were adopted as single market legislation under the codecision procedure on the basis of art.100a tec.

Translation: consequently, the proposals parliament after the assent procedure, the tabled amendments for offers no possibility of community directives, because as part of the internal market legislation on the basis of article 100a of the treaty in the codecision procedure have been adopted.
Reference: consequently proposals are submitted to parliament under the assent procedure, meaning that parliament can no longer table amendments, as directives in this area were adopted as single market legislation under the codecision procedure on the basis of art.100a tec.

Translation: consequently, the proposals parliament after the assent procedure, the tabled amendments for offers no possibility of community directives, because as part of the internal market legislation on the basis of article 100a of the treaty in the codecision procedure have been adopted.
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

*there is no* higher level of hierarchy discrimination between

*es gibt keine hierarchie der diskriminierung*

segmentation + s-phrase selection + adjunctions
Phrase-based Translation with TAG operations

There is no higher level of hierarchy discrimination between discrimination. There is no discrimination. Es gibt keine hierarchie der diskriminierung.

segmentation + s-phrase selection + adjunctions
Contributions (I)

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

we must also take these criticisms seriously
diese kritik ernst nehmen

wir müssen auch

segmentation + s-phrase selection + non-projective adjuncions
Reordering via Non-Projective Operations

- we must also take these criticisms seriously take
- wir müssen auch diese kritik ernst nehmen take these criticisms

segmentation + s-phrase selection + non-projective adjunctions
Reordering via Non-Projective Operations

we must also take these criticisms seriously

wir müssen auch diese kritik ernst nehmen

take these criticisms seriously

segmentation + s-phrase selection + non-projective adjunctions
Reordering via Non-Projective Operations

Contributions (II)

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
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
A derivation:

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

- Step 1: segment the input sentence, and choose an s-phrase for each segment
- Step 2: connect s-phrases with adjuncions
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 $d$:

$$
score(d) = score_{LM}(d) + score_{P}(d) + score_{SYN}(d) + score_{R}(d)
$$

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 $d$:

$$
\text{score}(d) = \text{score}_{LM}(d) + \text{score}_P(d) + \text{score}_{SYN}(d) + \text{score}_R(d)
$$

where

- $\text{score}_{LM}$ is a trigram language model
- $\text{score}_P$ is a sum of standard phrase-based scores
- $\text{score}_{SYN}$ is a syntactic language model [Charniak et al. 03] [Shen et al. 08] (probabilities are associated with adjunctions)
- $\text{score}_R$ is a sum of discriminative adjunction scores
Model

- Model score for a derivation $d$:

$$
score(d) = score_{LM}(d) + score_P(d) + score_{SYN}(d) + score_R(d)$$

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 $d$:
  \[
  score(d) = score_{LM}(d) + score_{P}(d) + score_{SYN}(d) + score_{R}(d)
  \]

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 \cite{Charniak2003, Shen2008}, (probabilities are associated with adjunctions)
- $score_{R}$ is a sum of discriminative adjunction scores
Model

Model score for a derivation $d$:

$$
score(d) = score_{LM}(d) + score_{P}(d) + score_{SYN}(d) + score_{R}(d)
$$

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 adjuctions)
- $score_{R}$ is a sum of discriminative adjunction scores
Trigram Language Models

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

\[
= P(\text{In} | \ast \ast) \times P(\text{a} | \ast \text{ In}) \times P(\text{few} | \text{In a}) \times
\]

\[
P(\text{days} | \text{a few}) \times P(\text{elections} | \text{few days}) \times
\]

\[
P(\text{take} | \text{days elections}) \times P(\text{place} | \text{elections take}) \times
\]

\[
P(\text{in} | \text{take place}) \times P(\text{Slovenia} | \text{place in})
\]
Syntactic Language Models

\[
P(\text{tree, sentence}) =
\]

\[
P(\text{S-VP-v-eats|ROOT}) \times P(\text{NP-n-Mary|S, eats, LEFT}) \times
\]

\[
P(\text{NP-n-cake|VP, eats, RIGHT}) \times P(\text{d-the|NP, cake, LEFT}) \times \ldots
\]
Two challenges

All permutations of s-phrases are possible.

Two challenges:

1. Constraining reorderings
2. Search
A Discriminative Dependency Model

$score_R$: $score_R(d)$ 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:**

```
we must also
wir müssen auch
diese kritik ernst nehmen
```

**Violated:**

```
there is no
es gibt keine hierarchie der diskriminierung
```
Decoding as Parsing

[wir müssen auch] [diese kritik] [ernst] [nehmen]

NP/0001100  VP/0000001  ⇒  VP/0001101

these criticisms  V  take  V  NP

these criticisms  take

- 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
Beam Search Decoding

0. Data structures: $Q_i$ for $i = 1 \ldots n$ is a set of hypotheses for each length $i$, $S$ is a set of chart entries

1. $S \leftarrow \emptyset$

2. Initialize $Q_1 \ldots Q_n$ with basic chart entries derived from phrase entries

3. For $i = 1 \ldots n$

4. For any $A \in \text{BEAM}(Q_i)$

5. If $S$ contains a chart entry with the same signature as $A$, and which has a higher inside score, continue

6. Else

7. Add $A$ to $S$

8. For any chart entry $C$ that can be derived from $A$ together with another chart entry $B \in S$, add $C$ to the set $Q_j$ where $j = \text{length}(C)$

9. Return $Q_n$, a set of items of length $n$
(BEAM) Given $Q_i$, define $Q_{i,j}$ for $j = 1 \ldots n$ to be the subset of items in $Q_i$ which have their $j$’th bit equal to one (i.e., have the $j$’th source language word translated). Define $Q'_{i,j}$ to be the $N$ highest scoring elements in $Q_{i,j}$. Then $\text{BEAM}(Q_i) = \bigcup_{j=1}^{n} Q'_{i,j}$. 

The Definition of BEAM
Experiments
German to English using Europarl data (750K training sentences)

Development:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax-based</td>
<td>25.2</td>
</tr>
<tr>
<td>Syntax (no disc. model)</td>
<td>23.7 (-1.5)</td>
</tr>
<tr>
<td>Syntax (no $\pi$-c constraint)</td>
<td>24.4 (-0.8)</td>
</tr>
</tbody>
</table>

Test:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-based system (Pharaoh)</td>
<td>24.58</td>
</tr>
<tr>
<td>Syntax-based system</td>
<td>25.04 (+0.46)</td>
</tr>
</tbody>
</table>

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

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

Development:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax-based</td>
<td>25.2</td>
</tr>
<tr>
<td>Syntax (no disc. model)</td>
<td>23.7 (-1.5)</td>
</tr>
<tr>
<td>Syntax (no $\pi$-c constraint)</td>
<td>24.4 (-0.8)</td>
</tr>
</tbody>
</table>

Test:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-based system (Pharaoh)</td>
<td>24.58</td>
</tr>
<tr>
<td>Syntax-based system</td>
<td>25.04 (+0.46)</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>Syntax</th>
<th>PB</th>
<th>=</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>51</td>
<td>3</td>
<td>7</td>
<td>61</td>
</tr>
<tr>
<td>PB</td>
<td>1</td>
<td>25</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td>=</td>
<td>21</td>
<td>14</td>
<td>67</td>
<td>102</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>42</td>
<td>85</td>
<td>200</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>Syntax</th>
<th>PB</th>
<th>=</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>51</td>
<td>3</td>
<td>7</td>
<td>61</td>
</tr>
<tr>
<td>PB</td>
<td>1</td>
<td>25</td>
<td>11</td>
<td>37</td>
</tr>
<tr>
<td>=</td>
<td>21</td>
<td>14</td>
<td>67</td>
<td>102</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>42</td>
<td>85</td>
<td>200</td>
</tr>
</tbody>
</table>

both results are significant with $p < 0.05$ under the sign test
Reference: now, however, we are seeing that president putin is pursuing a policy of openness towards the west.

Phrase-based: we are, however, now that president putin a policy of openness to the west out of blackmail

Syntax: now, however, we see that mr president putin is pursuing a policy of openness towards the west.
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 postponing 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.
Syntactified BLEU Scores

- Regular BLEU scores:
  
  \[
  \text{SYNTAX} = 24.96, 59.6/31.5/19.0/11.8 \quad (\text{BP}=0.980)
  \]
  
  \[
  \text{PHRASEB} = 24.48, 59.7/30.7/18.3/11.3 \quad (\text{BP}=0.986)
  \]

- Adding syntactic annotations:

  parliament: \text{S+VP+NP} \quad \text{agrees:ROOT+ROOT+S}

- Syntactified BLEU scores:

  \[
  \text{SYNTAX} = 18.83, 46.5/23.7/14.1/8.8 \quad (\text{BP}=0.980)
  \]
  
  \[
  \text{PHRASEB} = 16.93, 43.4/21.4/12.4/7.5 \quad (\text{BP}=0.986)
  \]
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