Applications of Tree Automata Theory
Lecture IV: Machine Translation — Basics

Andreas Maletti

Institute of Computer Science
Universität Leipzig, Germany

on leave from: Institute for Natural Language Processing
Universität Stuttgart, Germany

maletti@ims.uni-stuttgart.de

Yekaterinburg — August 24, 2014

Roadmap

1. Theory of Tree Automata
2. Parsing — Basics and Evaluation
3. Parsing — Advanced Topics
4. Machine Translation — Basics and Evaluation
5. Theory of Tree Transducers
6. Machine Translation — Advanced Topics

Always ask questions right away!

Statistical Machine Translation

Foundations

Statistical Machine Translation

Definition

A statistical machine translation system is a (usually fully automatic) computer system that translates based on statistical models learnt from parallel corpora.

Parallel Corpus

Definition

A parallel corpus is a (linguistic) resource, in which text in one language is presented together with sentence-by-sentence translations into another language

Notes

- parallel corpus $=$ sentence-aligned bi-text
- parallel corpus $\neq$ comparable corpus

Parallel Corpus

Definition (Sentence alignment)

A sentence alignment is an injective (partial) mapping
$f : \mathbb{N} \to \mathbb{N}$ (between sentence numbers)

Parallel Corpus

Definition (Sentence alignment)

A sentence alignment is an injective (partial) mapping
$f : \mathbb{N} \to \mathbb{N}$ (between sentence numbers)

Definition (Parallel corpus [formal])

A parallel corpus consists of three (partial) mappings:

- $s : \mathbb{N} \to \Sigma^*$ (source sentences)
- sentence alignment $f : \mathbb{N} \to \mathbb{N}$
- $t : \mathbb{N} \to \Delta^*$ (target sentences)

Parallel Corpus

Example (Aligned sentences)

1. nice example:
   - “We can help countries catch up, but not by putting their neighbors on hold”
   - “Wir können Ländern beim Aufholen helfen, aber nicht, indem wir ihre Nachbarn in den Wartesaal schicken”

Parallel Corpus

Example (Aligned sentences)

1. nice example:
   - “We can help countries catch up, but not by putting their neighbors on hold”
   - “Wir können Ländern beim Aufholen helfen, aber nicht, indem wir ihre Nachbarn in den Wartesaal schicken”

2. questionable example:
   - “We must bear in mind the Community as a whole”
   - “Wir müssen uns davor hüten, alles vergemeinschaften zu wollen”

Parallel Corpus

English-Russian example

- “Indeed, Republican lawyers identified only 300 cases of electoral fraud in the United States in a decade.”
- “К тому же юристы республиканцев насчитали только 300 случаев электоральных фальсификаций в Соединенных Штатах за десять лет.”

Sentences

Definition

A sentence is a sequence of tokens

(the result of tokenization)

Sentences

Definition

A sentence is a sequence of tokens

(the result of tokenization)

Notes

- normally the atomic unit is the word (dictionary entry)
- notable exceptions: Japanese, Chinese, etc.

Sentences

Example

- Sentence:

  We must bear in mind the Community as a whole

- Token sequence: length: 10

We	must	bear	in	mind	the	Community	as	a	whole

Sentences

Definition

Given a sentence S of length n, we write

- $\text{pos}(S) = \{1, \ldots, n\}$ positions in S
- $S[i]$ with $i \in \text{pos}(S)$ for its i-th token
- $S^{-1}[w] = \{i \in \text{pos}(S) \mid S[i] = w\}$ its inverse
- $S[i, j]$ with $i, j \in \text{pos}(S)$, $i \leq j$ for its $[i, j]$-span

$$S[i, j] = S[i]S[i + 1] \cdots S[j]$$

Sentences

Example

$S =$ We must bear in mind the Community as a whole

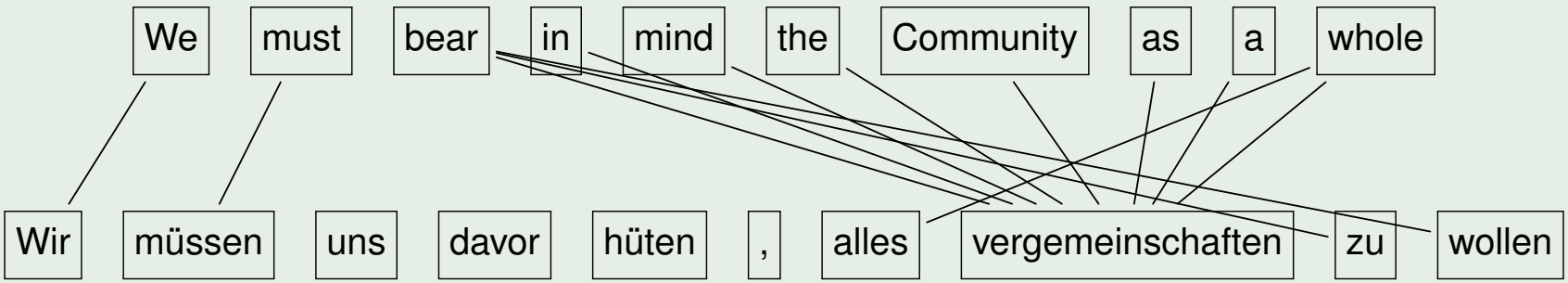
- $\text{pos}(S) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
- $S[4] =$ in
- $S^{-1}[\text{must}] = \{2\}$
- $S[6, 8] =$ the Community as

Word Alignment

Definition

A word alignment for two sentences S_1 and S_2 is a relation $\rho \subseteq \text{pos}(S_1) \times \text{pos}(S_2)$ on its positions

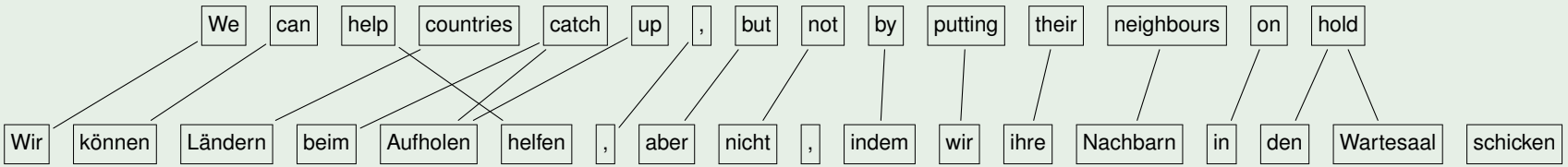
Example



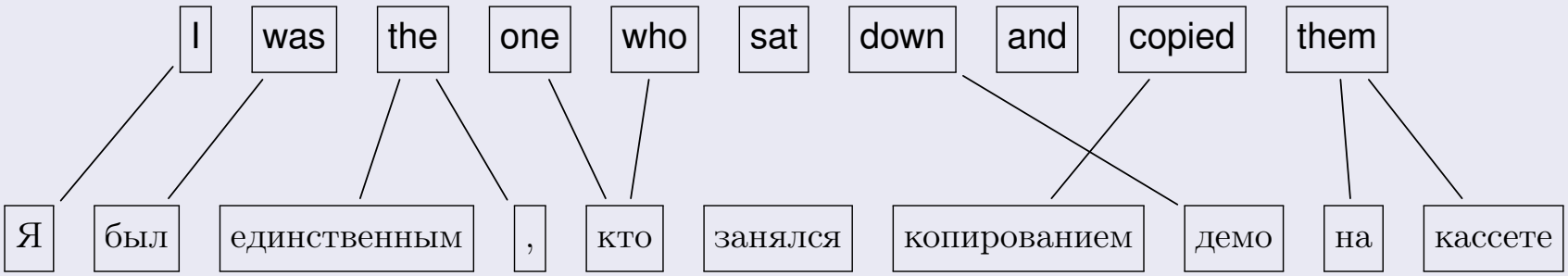
1–1 2–2 3–8 3–9 3–10 4–8 5–8 6–8 7–8 8–8 9–8 10–7 10–8

Word Alignment

Nicer example



English-Russian example



1-1 2-2 3-3 3-4 4-5 5-5 9-7 7-8 10-9 10-10

Parallel Corpus

EUROPARL German-English parallel corpus

- 1,920,209 parallel sentences
- 44,548,491 words in German
- 47,818,827 words in English
- sentence-aligned, but not word-aligned
- from parliament proceedings

Parallel Corpus

MULTIUN Chinese-English parallel corpus

- $9,564,315$ parallel sentences
- $256,720,000$ words in English
- sentence-aligned, but not word-aligned
- from official UN documents

YANDEX parallel corpus

- $\approx$ 1M parallel sentences
- Russian-English

Statistical Machine Translation

Phrase-based Models

Phrase-based Models

Key points [Och, Ney, 2004]

- phrase as basic translation unit
- dominant model for many language pairs
  incl. Russian-English
- can deal with (arbitrary) many-to-many alignments
- words internally move together

Phrase-based Models

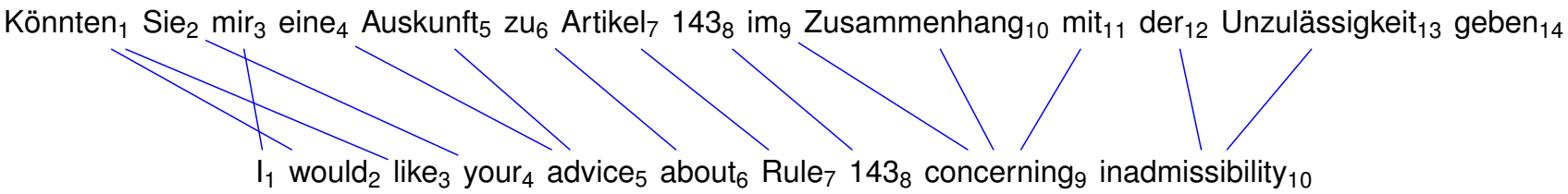
Key points [Och, Ney, 2004]

- phrase as basic translation unit
- dominant model for many language pairs
  incl. Russian-English
- can deal with (arbitrary) many-to-many alignments
- words internally move together

Definition (Phrase)

phrase is a contiguous sequence of words
(usually given by span $[i, i']$)

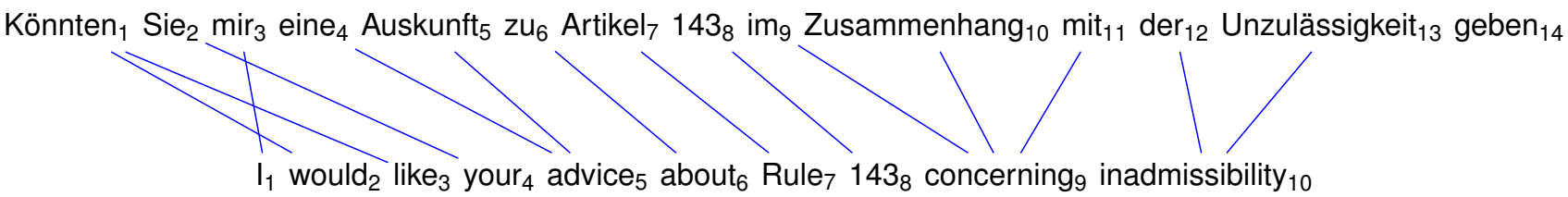
Phrase-based Models



Algorithm

1. phrase pair $([j, j'], [i, i'])$ consistently aligned if
   - $\ell' \in [i, i']$ for all $\ell \in [j, j']$ and $(\ell, \ell') \in A$
   - $\ell \in [j, j']$ for all $\ell' \in [i, i']$ and $(\ell, \ell') \in A$

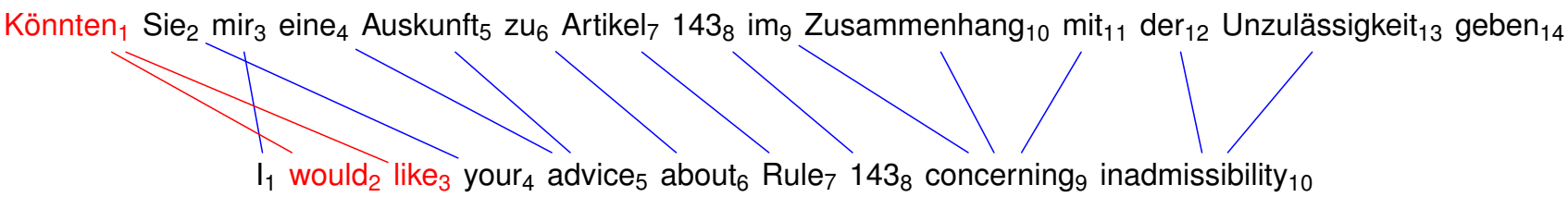
Phrase-based Models



Algorithm

1. phrase pair $([j, j'], [i, i'])$ consistently aligned if
   - $\ell' \in [i, i']$ for all $\ell \in [j, j']$ and $(\ell, \ell') \in A$
   - $\ell \in [j, j']$ for all $\ell' \in [i, i']$ and $(\ell, \ell') \in A$
2. extract all consistently aligned phrase pairs
3. (restrict length of phrases based on corpus size)

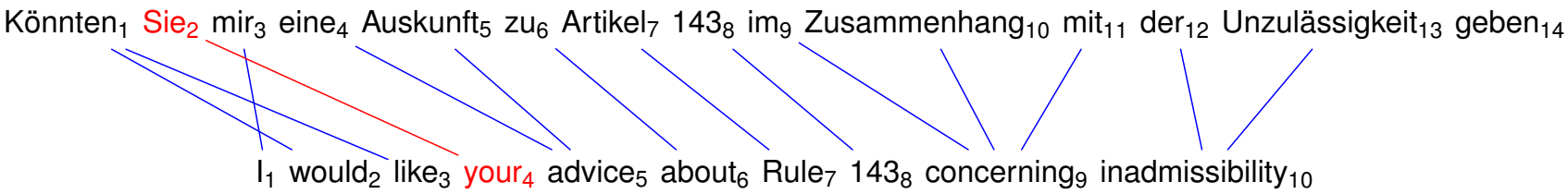
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

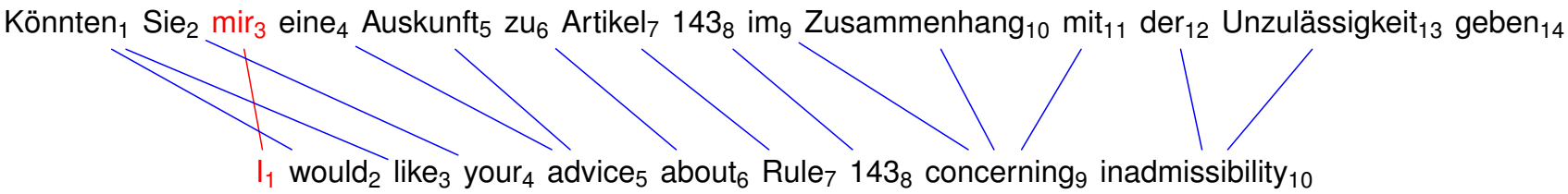
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

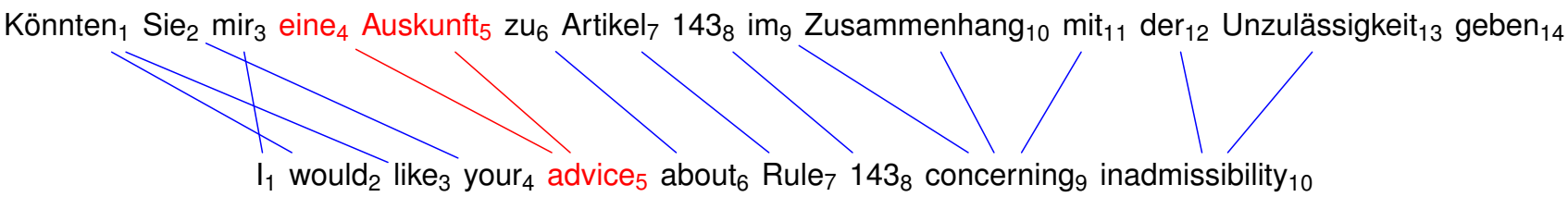
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

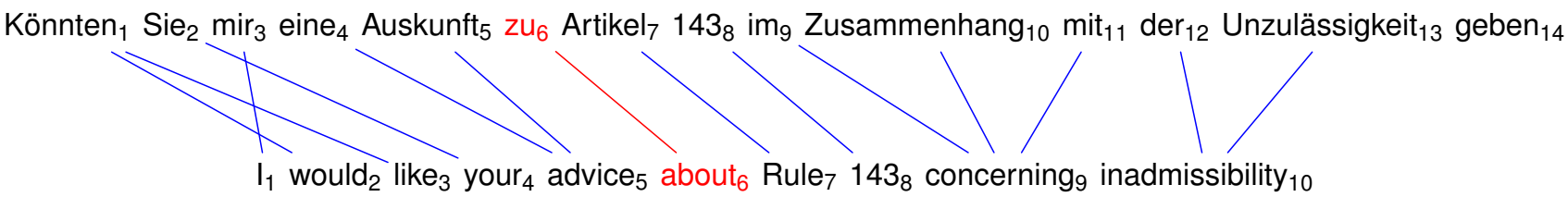
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

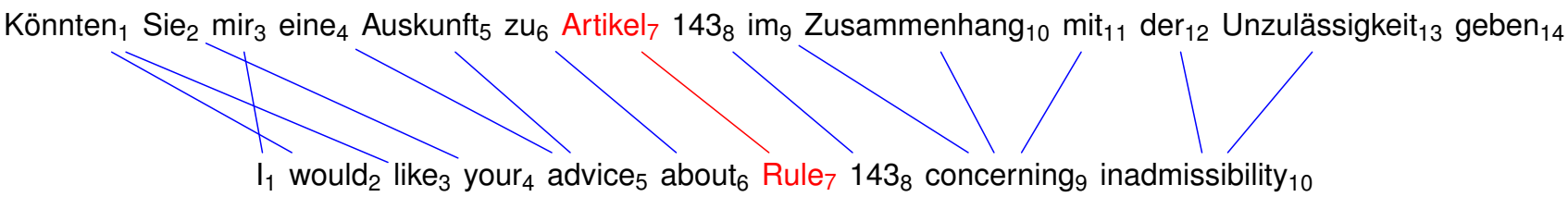
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

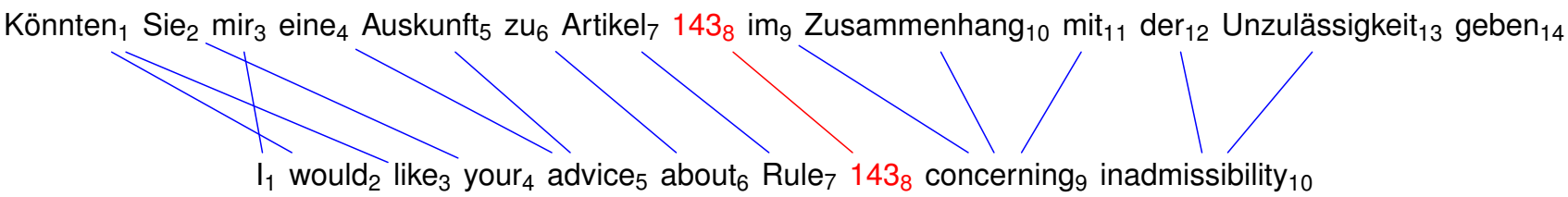
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

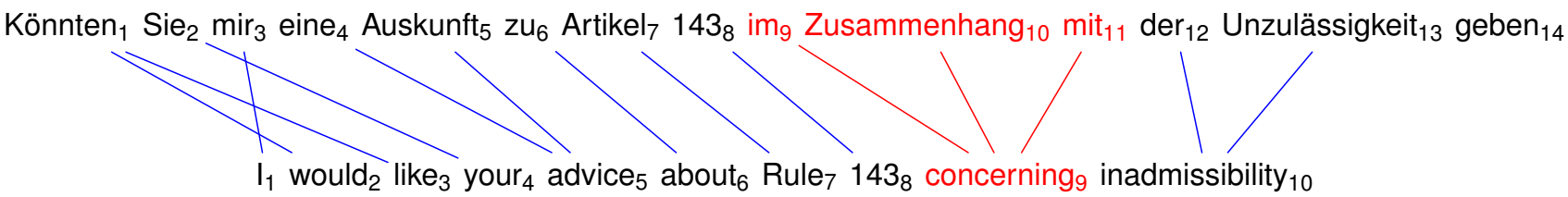
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

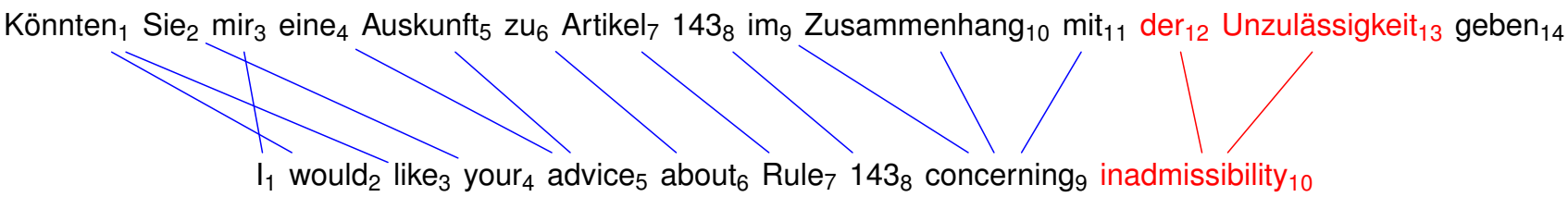
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

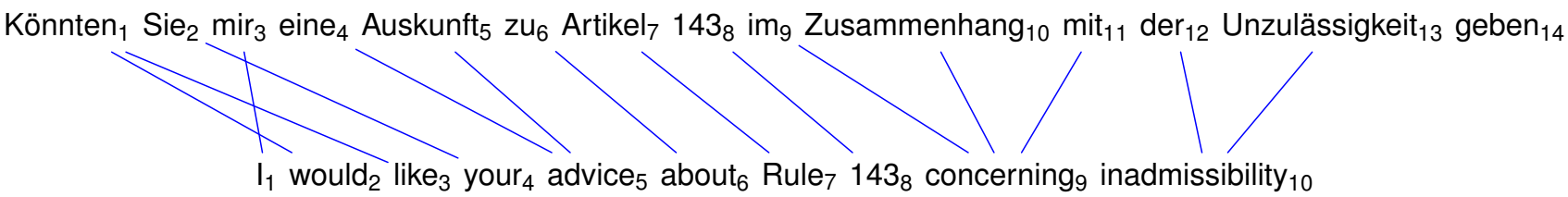
Phrase-based Models



Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

Phrase-based Models



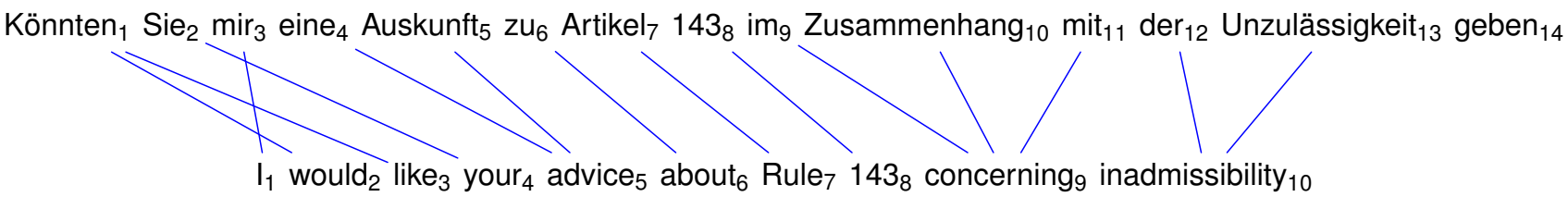
Formally:

([1,1], [2,3])	([2,2], [4,4])	([3,3], [1,1])
([4,5], [5,5])	([6,6], [6,6])	([7,7], [7,7])
([8,8], [8,8])	([9,11], [9,9])	([12,13], [10,10])

For better readability:

Könnten — would like	Sie — your	mir — I
eine Auskunft — advice	zu — about	Artikel — Rule
143 — 143	im Zusammenhang mit — concerning	der Unzulässigkeit — inadmissibility

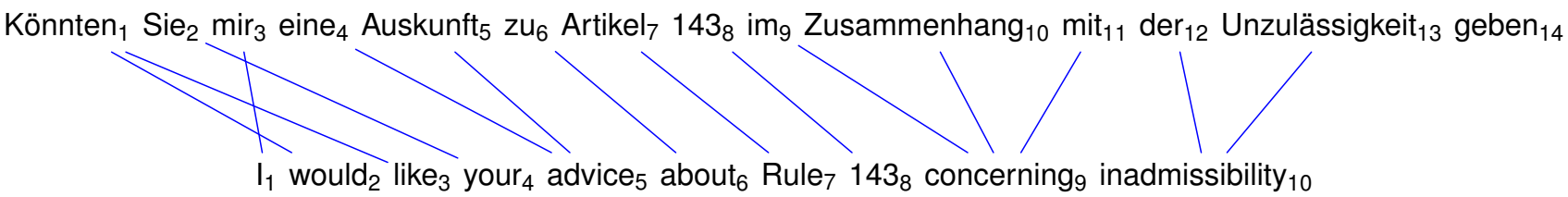
Phrase-based Models



Notes

- these were only minimal phrase pairs
- extract all (sensible) combinations of these
- e.g., $([1,1],[2,3])$ and $([2,2],[4,4])$ yield $([1,2],[2,4])$

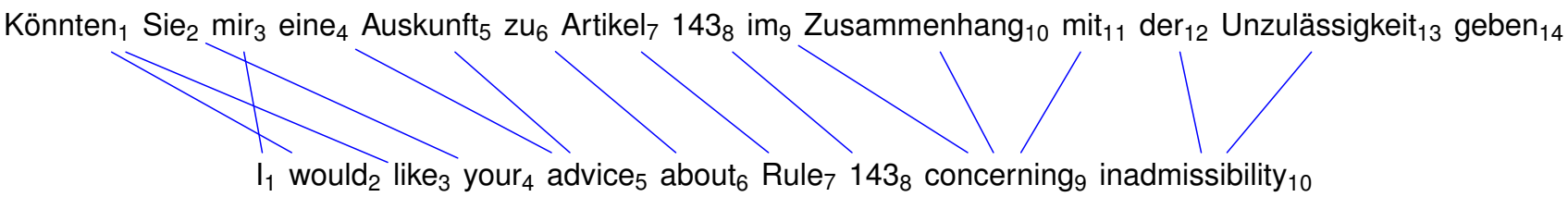
Phrase-based Models



Notes

- these were only minimal phrase pairs
- extract all (sensible) combinations of these
- e.g., $([1,1],[2,3])$ and $([2,2],[4,4])$ yield $([1,2],[2,4])$
- unaligned words can be added to neighboring phrases
- e.g., $([12,13],[10,10])$ extends to $([12,14],[10,10])$

Phrase-based Models

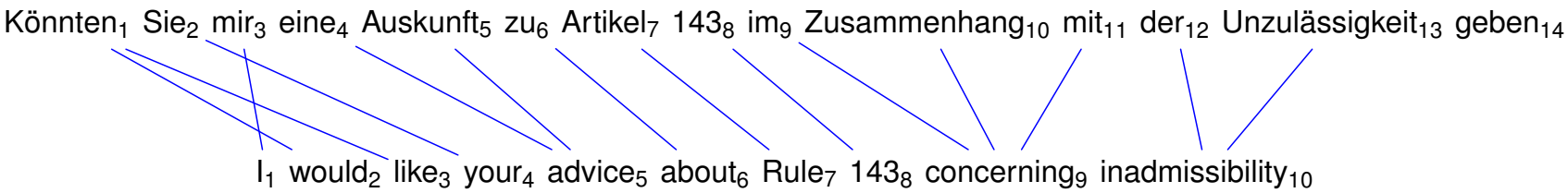


Notes

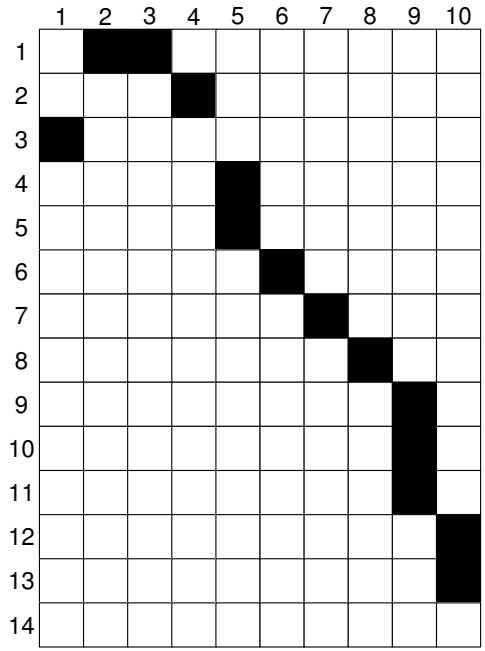
- these were only minimal phrase pairs
- extract all (sensible) combinations of these
- e.g., $([1,1],[2,3])$ and $([2,2],[4,4])$ yield $([1,2],[2,4])$
- unaligned words can be added to neighboring phrases
- e.g., $([12,13],[10,10])$ extends to $([12,14],[10,10])$

Könnten Sie — would like your der Unzulässigkeit geben — inadmissibility

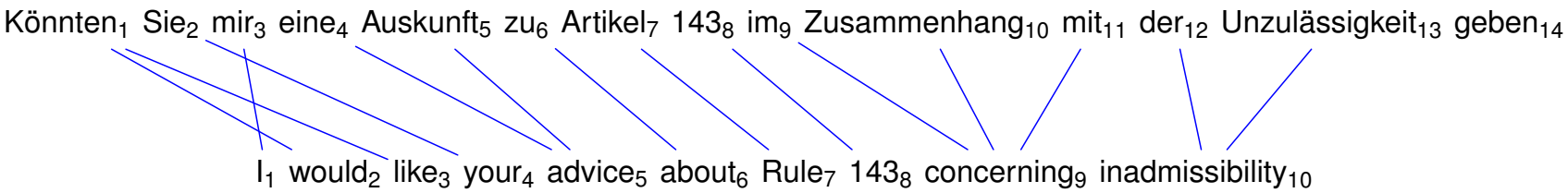
Phrase-based Models



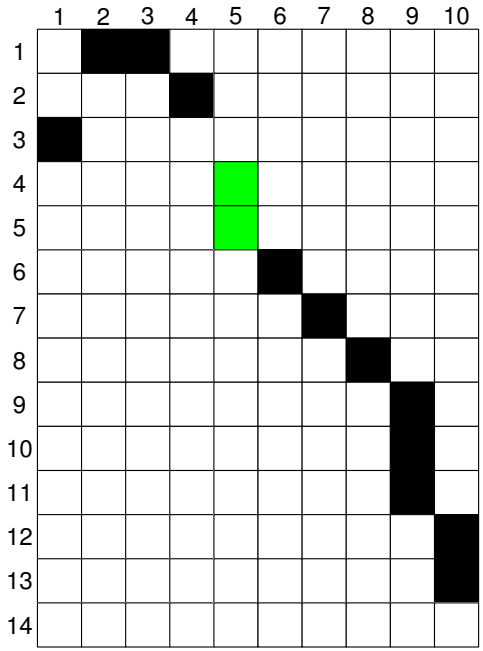
Alternative representation (rectangles):



Phrase-based Models



Alternative representation (rectangles):

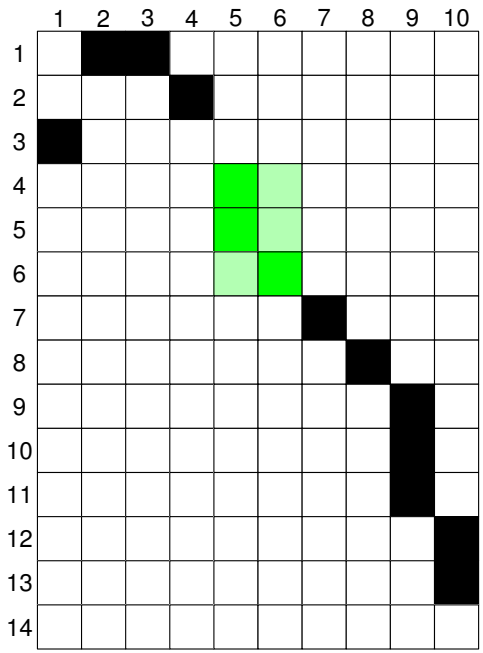


Phrase-based Models

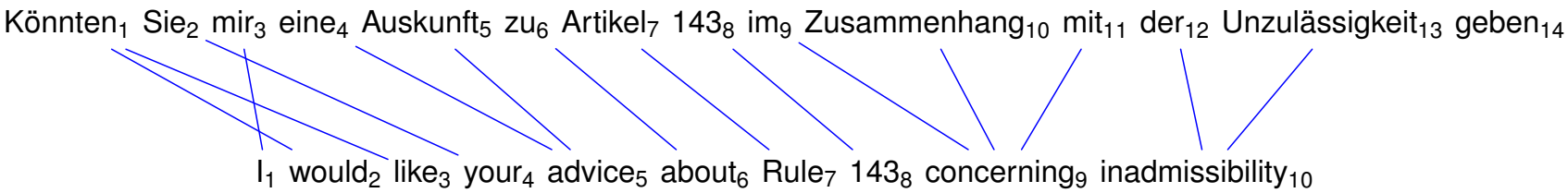
Könnten$_1$ Sie$_2$ mir$_3$ eine$_4$ Auskunft$_5$ zu$_6$ Artikel$_7$ 143$_8$ im$_9$ Zusammenhang$_{10}$ mit$_{11}$ der$_{12}$ Unzulässigkeit$_{13}$ geben$_{14}$

I$_1$ would$_2$ like$_3$ your$_4$ advice$_5$ about$_6$ Rule$_7$ 143$_8$ concerning$_9$ inadmissibility$_{10}$

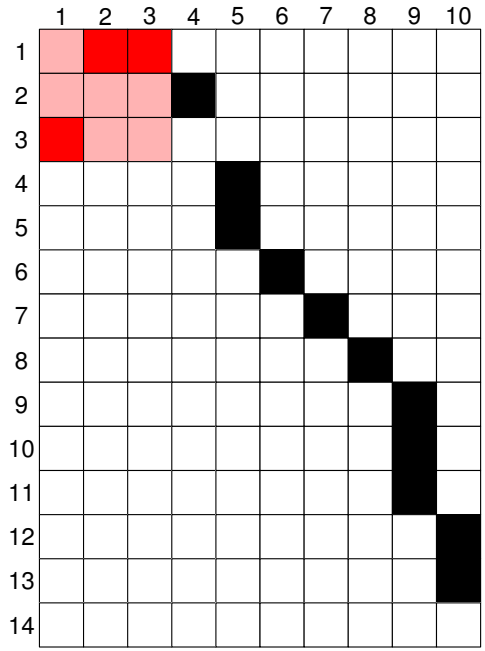
Alternative representation (rectangles):



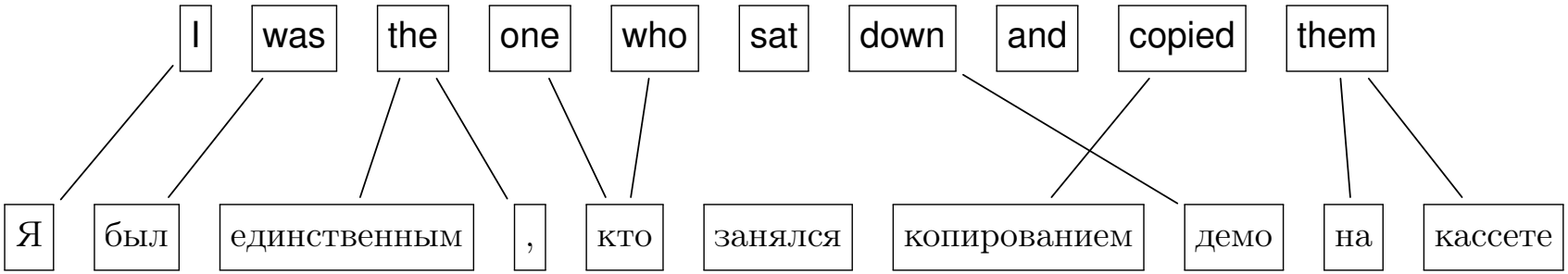
Phrase-based Models



Alternative representation (rectangles):



Phrase-based Models



Extractable phrase pairs

I was — Я был

down and copied them — копированием демо на кассете

the one who sat — единственным, кто занялся

Phrase-based Models

Notes

- phrases are not linguistic phrases!

  (noun phrases, verb phrases, etc.)

- non-linguistic phrase pair: Sinne der — keeping with the

Phrase-based Models

Notes

- phrases are not linguistic phrases!

  (noun phrases, verb phrases, etc.)
- non-linguistic phrase pair: Sinne der — keeping with the
- having only linguistic phrases lowers translation quality

Phrase-based Models

Notes

- phrases are not linguistic phrases!

  (noun phrases, verb phrases, etc.)
- non-linguistic phrase pair: Sinne der — keeping with the
- having only linguistic phrases lowers translation quality
- phrase translation table typically huge

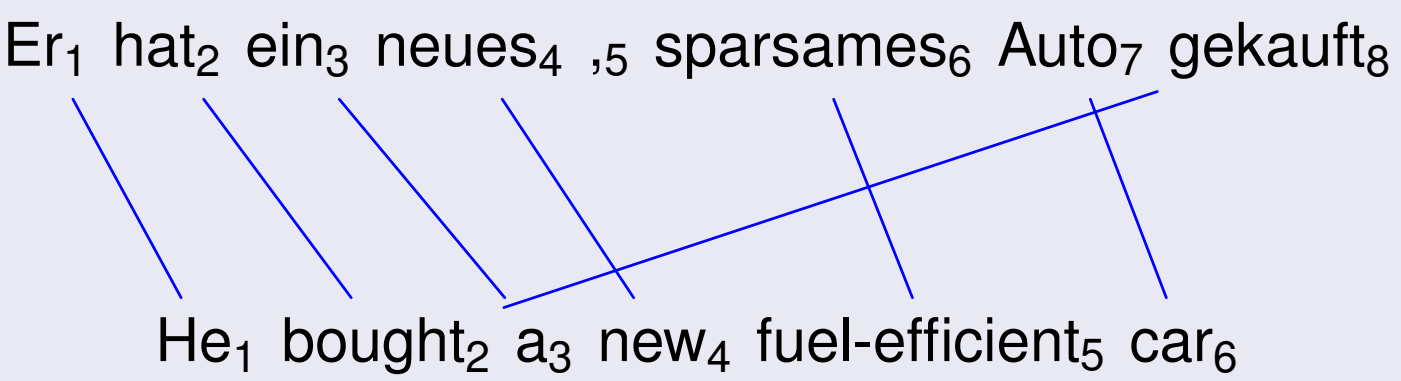
  (much larger than parallel corpus)

Statistical Machine Translation

Hierarchical Phrase-based Models

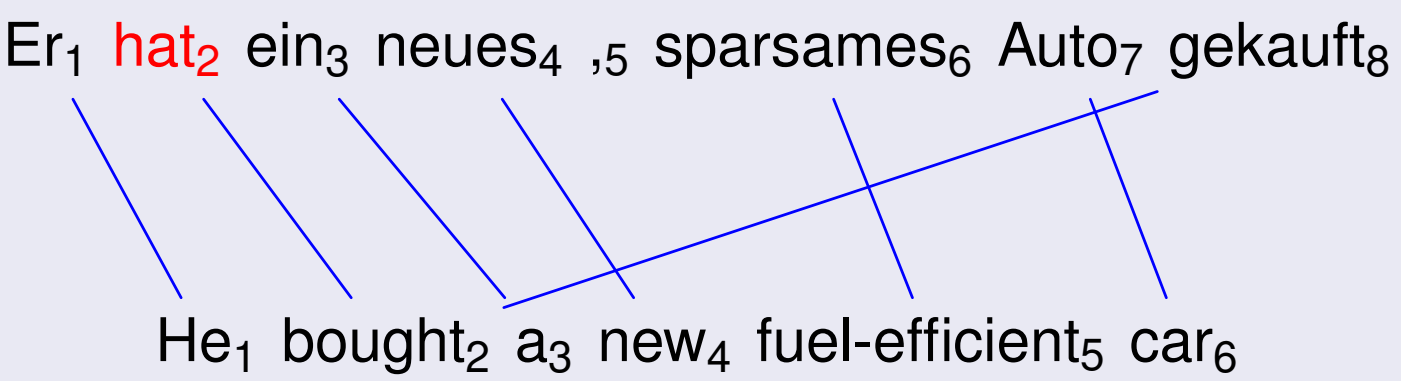
Phrase-based Models

Input



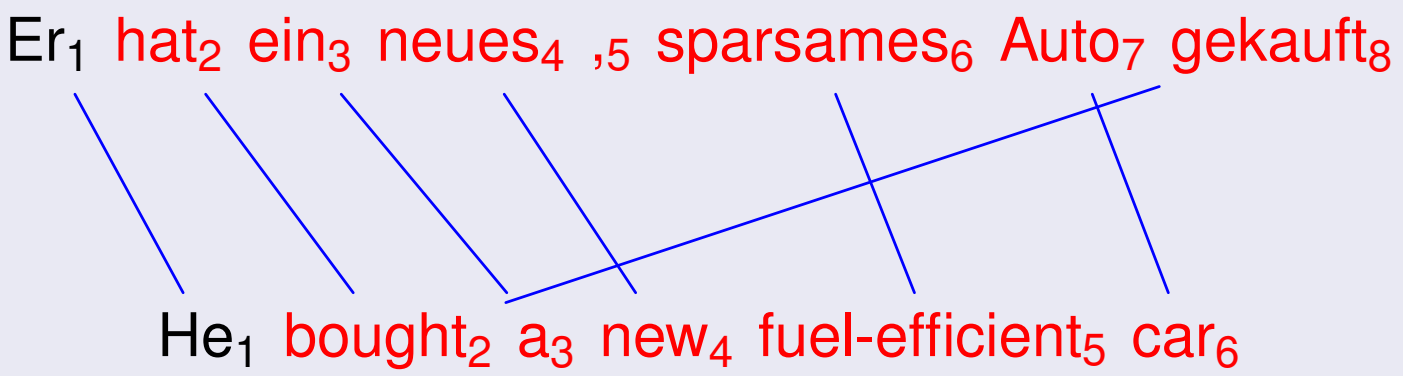
Phrase-based Models

Input



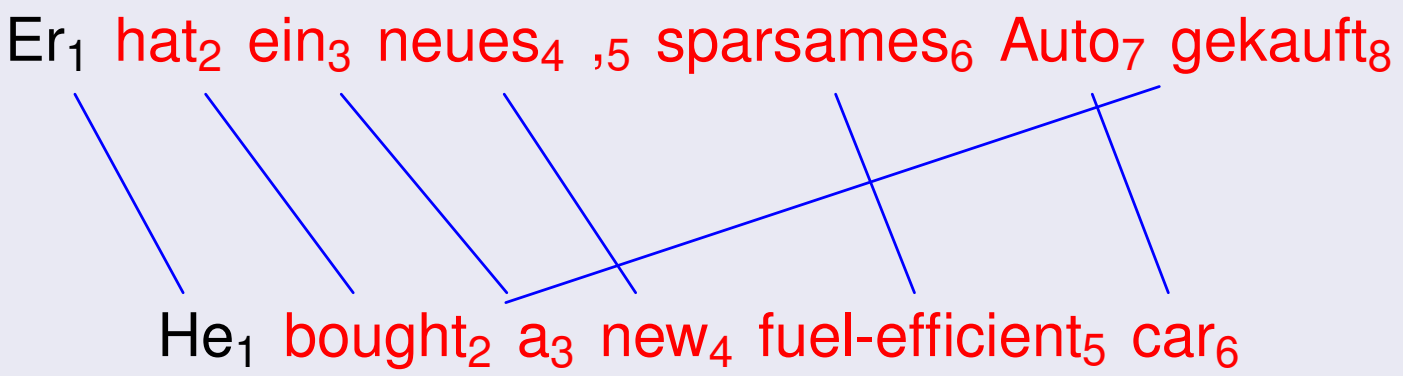
Phrase-based Models

Input



Phrase-based Models

Input



Problem

Only rule translating hat or gekauft has very limited use:

hat ein neues , sparsames Auto gekauft
— bought a new fuel-efficient car

Hierarchical Phrase-based Models

Analysis

- the restriction to phrases (i.e., contiguous segments) yields undesirably long phrase pairs

hat ein neues , sparsames Auto gekauft

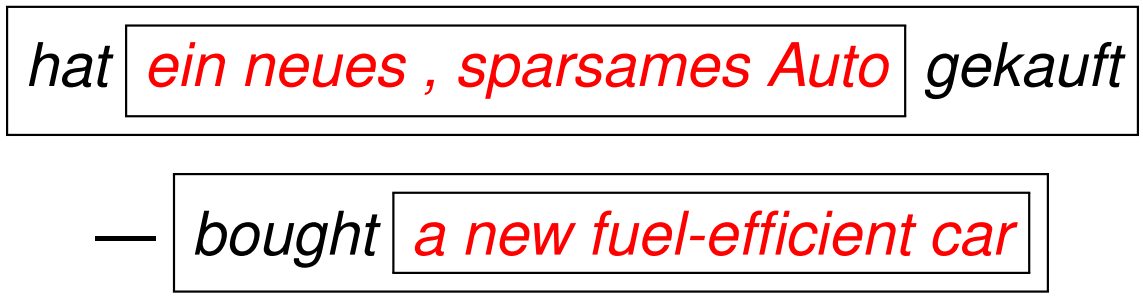
— *bought a new fuel-efficient car*

Hierarchical Phrase-based Models

Analysis

- the restriction to phrases (i.e., contiguous segments) yields undesirably long phrase pairs
- it would be better to allow phrases inside phrases

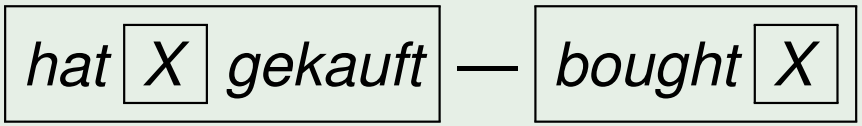
→ hierarchical phrases [CHIANG, 2007]



Hierarchical Phrase-based Models

Hierarchical phrases

- phrases with gaps (written as *X*)



Hierarchical Phrase-based Models

Hierarchical phrases

- phrases with gaps (written as *X*)

$$\boxed{\textit{hat}\ \boxed{X}\ \textit{gekauft}} \quad — \quad \boxed{\textit{bought}\ \boxed{X}}$$

- gaps filled by other hierarchical phrases (maybe gaps)

$$\boxed{\textit{ein neues , sparsames Auto}} \quad — \quad \boxed{\textit{a new fuel-efficient car}}$$

$$\boxed{\textit{ein}\ \boxed{X}\ \textit{Auto}} \quad — \quad \boxed{\textit{a}\ \boxed{X}\ \textit{car}}$$

Hierarchical Phrase-based Models

Hierarchical phrases

- phrases with gaps (written as *X*)

$$\boxed{\textit{hat}\ \boxed{X}\ \textit{gekauft}} \quad — \quad \boxed{\textit{bought}\ \boxed{X}}$$

- gaps filled by other hierarchical phrases (maybe gaps)

$$\boxed{\textit{ein neues , sparsames Auto}} \quad — \quad \boxed{\textit{a new fuel-efficient car}}$$

$$\boxed{\textit{ein}\ \boxed{X}\ \textit{Auto}} \quad — \quad \boxed{\textit{a}\ \boxed{X}\ \textit{car}}$$

- put together we have

$$\boxed{\textit{hat}\ \boxed{\textit{ein}\ \boxed{X}\ \textit{Auto}}\ \textit{gekauft}} \quad — \quad \boxed{\textit{bought}\ \boxed{\textit{a}\ \boxed{X}\ \textit{car}}}$$

Hierarchical Phrase-based Models

Restrictions

- must contain lexical item in source side

illegal: [X] — [*the* X]

Hierarchical Phrase-based Models

Restrictions

- must contain lexical item in source side

  illegal: $[X] \text{ — } [\textit{the}\ [X]]$

- at most 5 gaps (for efficiency: at most 2 gaps)

  illegal: $[X_1]\ \textit{hat}\ [X_2]\ \textit{des}\ [X_3] \text{ — } [X_2]\ \textit{of}\ [X_3]\ \textit{was}\ [X_1]$

Hierarchical Phrase-based Models

Restrictions

- must contain lexical item in source side

  illegal: $[X] \text{ — } [\textit{the}\ [X]]$

- at most 5 gaps (for efficiency: at most 2 gaps)

  illegal: $[X_1]\ \textit{hat}\ [X_2]\ \textit{des}\ [X_3] \text{ — } [X_2]\ \textit{of}\ [X_3]\ \textit{was}\ [X_1]$

- no adjacent gaps in source side

  illegal: $\textit{ihm}\ [X_1]\ [X_2] \text{ — } \textit{the}\ [X_2]\ \textit{of}\ [X_1]\ \textit{to him}$

Hierarchical Phrase-based Models

Algorithm

1. Extract phrase-pairs as usual

hat ein neues , sparsames Auto gekauft — *bought a new fuel-efficient car*

ein neues , sparsames Auto — *a new fuel-efficient car*

. . . — . . .

Hierarchical Phrase-based Models

Algorithm

1. Extract phrase-pairs as usual

hat ein neues , sparsames Auto gekauft	—	*bought a new fuel-efficient car*
ein neues , sparsames Auto	—	*a new fuel-efficient car*
...	—	...

2. Find phrase-pair that contains another phrase-pair

Hierarchical Phrase-based Models

Algorithm

1. Extract phrase-pairs as usual

hat ein neues , sparsames Auto gekauft — *bought a new fuel-efficient car*

ein neues , sparsames Auto — *a new fuel-efficient car*

. . . — . . .

2. Find phrase-pair that contains another phrase-pair
3. Remove inner phrase-pair and leave new X_i (if possible)

hat X_1 *gekauft* — *bought* X_1

Hierarchical Phrase-based Models

Algorithm

1. Extract phrase-pairs as usual

hat ein neues , sparsames Auto gekauft — *bought a new fuel-efficient car*

ein neues , sparsames Auto — *a new fuel-efficient car*

. . . — . . .

2. Find phrase-pair that contains another phrase-pair
3. Remove inner phrase-pair and leave new X_i (if possible)

hat X_1 *gekauft* — *bought* X_1

4. if there were changes, then go to 2

Statistical Machine translation

Statistical Machine translation

Translation

- Input:

  Республиканская стратегия сопротивления повторному избранию обамы

Statistical Machine translation

Translation

- Input:
  Республиканская стратегия сопротивления повторному избранию обамы
- Hierarchical phrase-based (YANDEX corpus):
  Republican strategy resistance renewal elect obama

Statistical Machine translation

Translation

- Input:
  Республиканская стратегия сопротивления повторному избранию обамы
- Hierarchical phrase-based (YANDEX corpus):
  Republican strategy resistance renewal elect obama
- GOOGLE Translate:
  The Republican strategy of resistance to the re-election of Obama

Statistical Machine translation

Translation

- Input:
  Республиканская стратегия сопротивления повторному избранию обамы
- Hierarchical phrase-based (YANDEX corpus):
  Republican strategy resistance renewal elect obama
- GOOGLE Translate:
  The Republican strategy of resistance to the re-election of Obama
- YANDEX Translate:
  The Republican strategy of resistance for the re-election of Obama

Statistical Machine translation

Translation

- **Input:**

  Кроме того, эти законы сокращают период досрочного голосования, упраздняют пра во регистрации избирателя в день волеизъявления и отнимают право голоса у граждан, имеющих судимость.

Statistical Machine translation

Translation

- **Input:**
  Кроме того, эти законы сокращают период досрочного голосования, упраздняют пра во регистрации избирателя в день волеизъявления и отнимают право голоса у граждан, имеющих судимость.
- **Hierarchical phrase-based** (YANDEX corpus):
  In addition, these laws reduce period early voting, упраздняют right registering voters in the day of voting and take the citizens have the right to vote, have a criminal record.

Statistical Machine translation

Translation

- **Input:**
  Кроме того, эти законы сокращают период досрочного голосования, упраздняют пра во регистрации избирателя в день волеизъявления и отнимают право голоса у граждан, имеющих судимость.
- **Hierarchical phrase-based** (YANDEX corpus):
  In addition, these laws reduce period early voting, упраздняют right registering voters in the day of voting and take the citizens have the right to vote, have a criminal record.
- **GOOGLE Translate:**
  In addition, these laws reduce the early voting period, abolish the right-in voter registration on the day of expression and the right to take away votes from citizens who have a criminal record.

Statistical Machine translation

Translation

- **Input:**
  Кроме того, эти законы сокращают период досрочного голосования, упраздняют пра во регистрации избирателя в день волеизъявления и отнимают право голоса у граждан, имеющих судимость.
- **Hierarchical phrase-based** (YANDEX corpus):
  In addition, these laws reduce period early voting, упраздняют right registering voters in the day of voting and take the citizens have the right to vote, have a criminal record.
- **GOOGLE Translate:**
  In addition, these laws reduce the early voting period, abolish the right-in voter registration on the day of expression and the right to take away votes from citizens who have a criminal record.
- **YANDEX Translate:**
  In addition, these laws reduce the early voting period will void the law of registration of the voter on the day of will and take away the right to vote of citizens with criminal records.

BLEU — Bilingual Evaluation Understudy

Approach [Papineni et al., 2002]

- compare translation output T to reference translation R
- count n-gram matches (recall)
- normalize by count of all possible n-gram matches

$$\text{prec}_i = \frac{\text{match}_i(T, R)}{|T| - i + 1}$$

BLEU — Bilingual Evaluation Understudy

Approach [Papineni et al., 2002]

- compare translation output T to reference translation R
- count n-gram matches (recall)
- normalize by count of all possible n-gram matches

$$\text{prec}_i = \frac{\text{match}_i(T, R)}{|T| - i + 1}$$

- BLEU-n = brev-pen $\cdot \sqrt[n]{\prod_{i=1}^{n} \text{prec}_i}$ (geometric mean)

BLEU — Bilingual Evaluation Understudy

Approach [Papineni et al., 2002]

- compare translation output T to reference translation R
- count n-gram matches (recall)
- normalize by count of all possible n-gram matches

$$\text{prec}_i = \frac{\text{match}_i(T, R)}{|T| - i + 1}$$

- BLEU-n $= \text{brev-pen} \cdot \sqrt[n]{\prod_{i=1}^{n} \text{prec}_i}$ (geometric mean)
- adjustment factor brevity penalty (punishes short output)

$$\text{brev-pen} = \min(1, \frac{|T|}{|R|})$$

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
1 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
1 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
2 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
3 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
4 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
4 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
4 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
5 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
5 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
6 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
6 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
6 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
7 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
7 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
8 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
9 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
10 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
11 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
11 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
12 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
13 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
14 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
15 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
16 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
17 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
17 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
18 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
19 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	1 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	2 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	3 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	4 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	5 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	6 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	7 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	8 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	9 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	1 of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	2 of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	3 of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	4 of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	5 of 27	of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	5 of 27	1 of 26

Evaluation — BLEU Scoring

Translation: (Google Translate)

The rotated with miniature cameras mounted on helmets recordings are checked at the airbase in Kandahar and then sent to London , where they are on the Internet .

Reference: The images , taken with miniature cameras attached to troop helmets , are monitored by a command centre in Kandahar and then transferred to London from where they are uploaded onto the Internet .

1-grams	2-grams	3-grams	4-grams
20 of 29	10 of 28	5 of 27	1 of 26

$$\text{BLEU-4} = \frac{29}{34} \cdot \sqrt[4]{\frac{20 \cdot 10 \cdot 5 \cdot 1}{29 \cdot 28 \cdot 27 \cdot 26}} = 17.46\%$$

Evaluation — BLEU Scoring

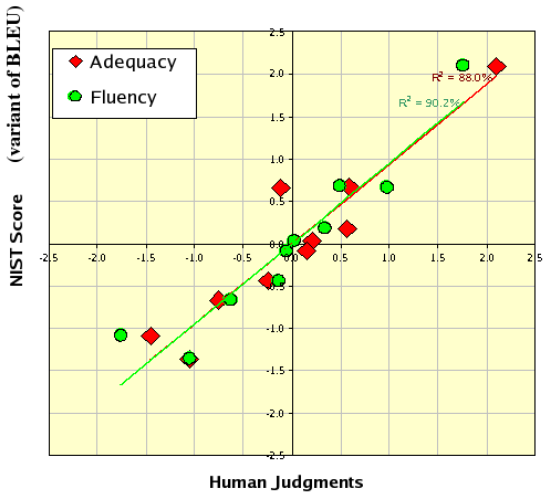
Notes

- should be used with multiple references
- should be used for documents (not sentences)
- primary evaluation for MT systems
- correlates reasonably well with human judgements

Evaluation — BLEU Scoring

Notes

- should be used with multiple references
- should be used for documents (not sentences)
- primary evaluation for MT systems
- correlates reasonably well with human judgements



(Figure from [Koehn, 2010])

Evaluation

Russian-to-English

System	BLEU-4
WMT '13 winner [PINO et al., 2013] (hierarchical phrase-based)	25.9
hierarchical phrase-based (vanilla)	21.9
MBOT string-to-tree (vanilla)	20.7
string-to-tree (vanilla)	19.8

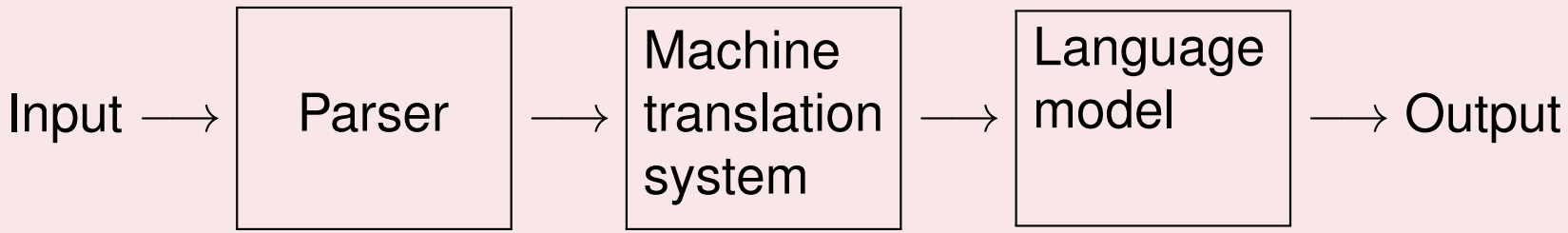
WMT '13 winner: Hierarchical phrase-based translation with wFSTs.
Pre-processing: STANFORD CoreNLP, Morfessor, Stem+POStag (TreeTagger)
Post-processing: lattices 5-gram-LM rescored and union of lattices rescored via Lattice Minimum Bayes Risk.

Statistical Machine translation

Syntax-based Models

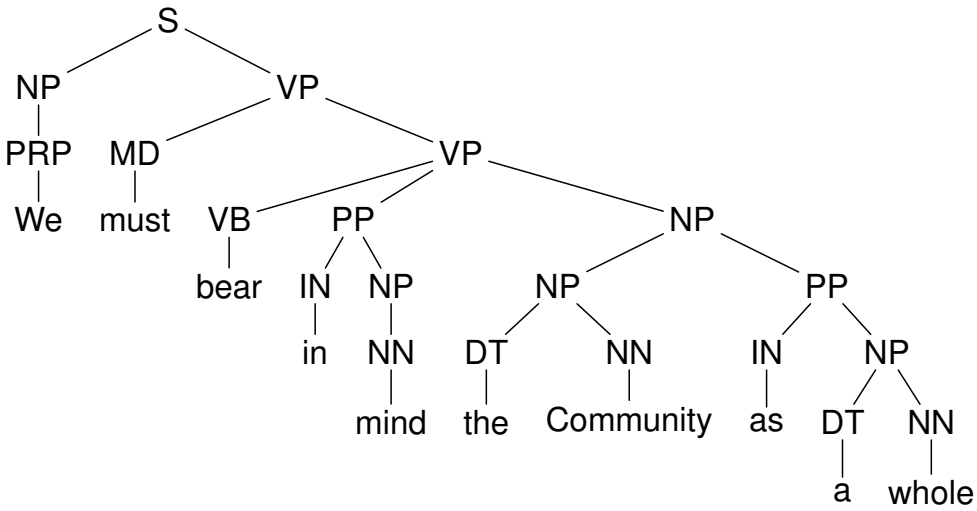
Syntax-based Machine Translation

Syntax-based systems



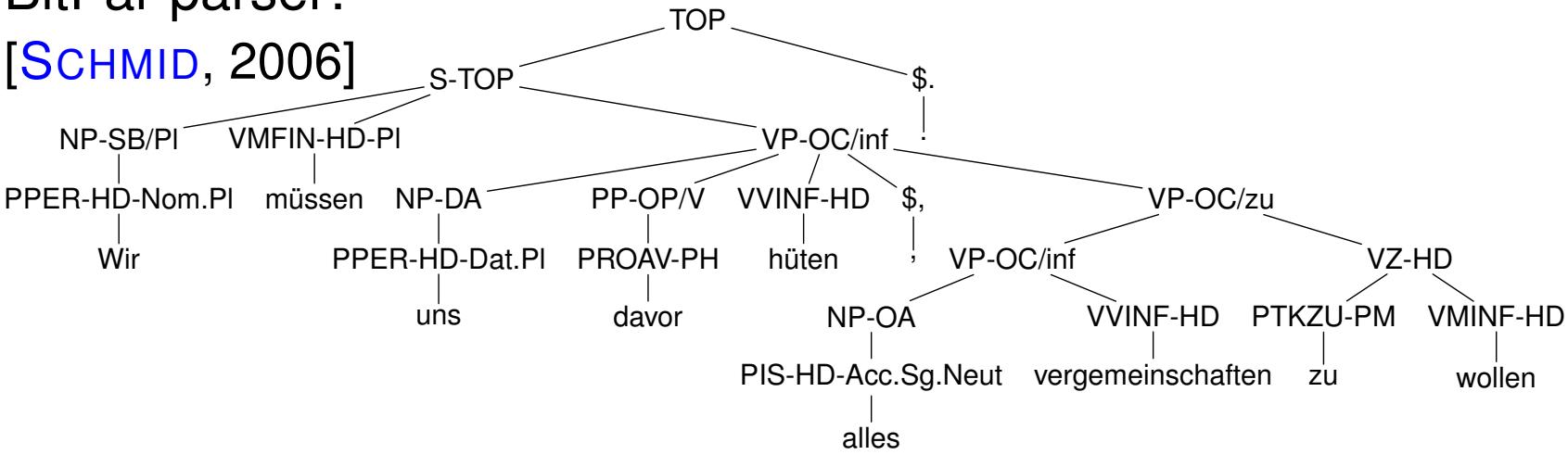
Syntax-based Machine Translation

CHARNIAK parser: [CHARNIAK, JOHNSON, 2005]



BitPar parser:

[SCHMID, 2006]



Syntax-based Machine Translation

Arabic-English

Yugoslav President Voislav signed for Serbia.

و تولى التوقيع عن صربيا الرئيس اليوغوسلافي فويسلاف

Translit.: w twlY AltwqyE En SrbyA Alr}ys Alywgwslafy fwyslAf.

And then the matter was decided, and everything was put in place.

ف كان ان تم الحسم و وضعت الأمور في نصاب ها

Translit.: f kAn An tm AlHsm w wDEt Al>mwr fy nSAb hA.

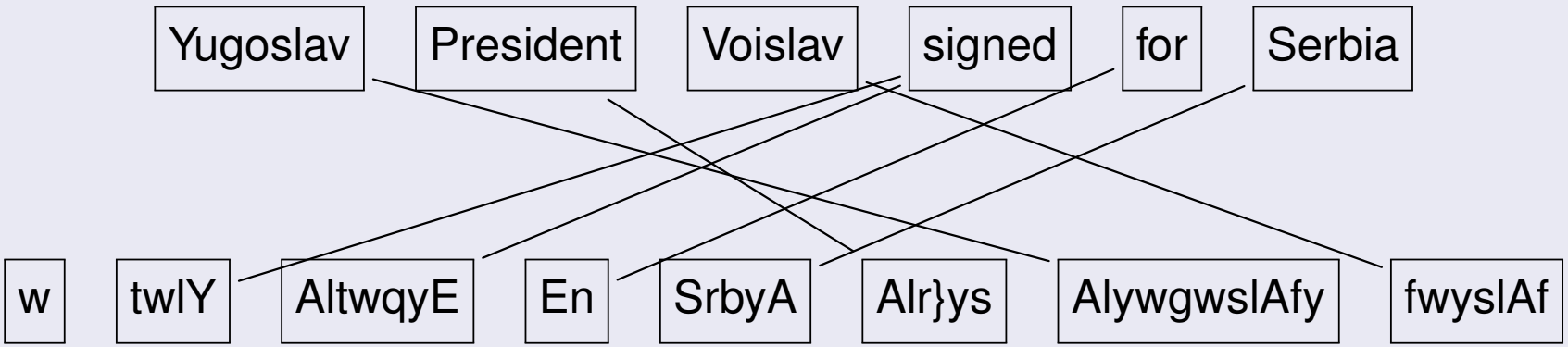
Below are the male and female winners in the different categories.

و هنا الأوائل و الأوليات في مختلف الفئات

Translit.: w hnA Al>wA}l w Al>wlyAt fy mxtlf Alf}At.

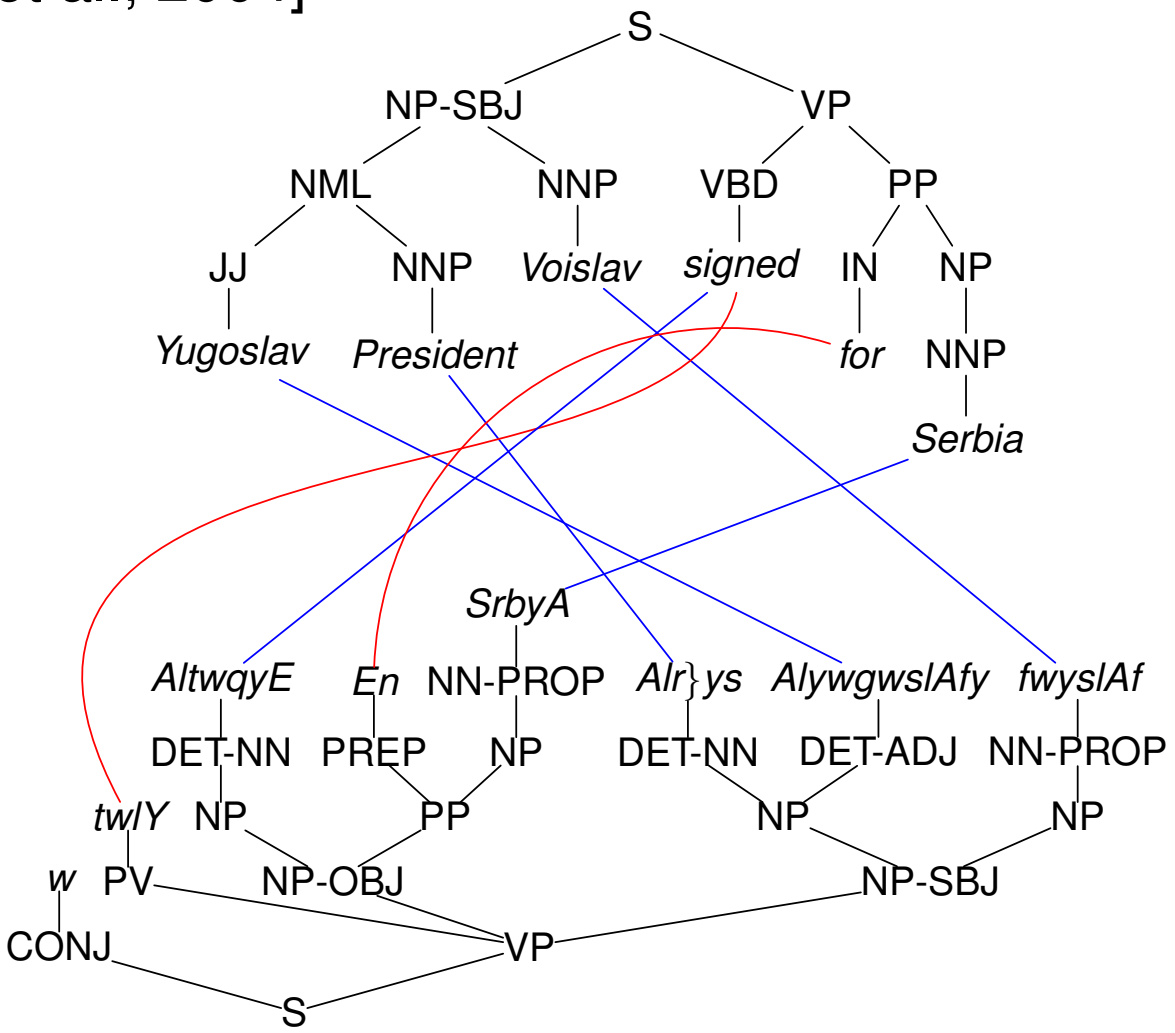
Syntax-based Machine Translation

Alignment

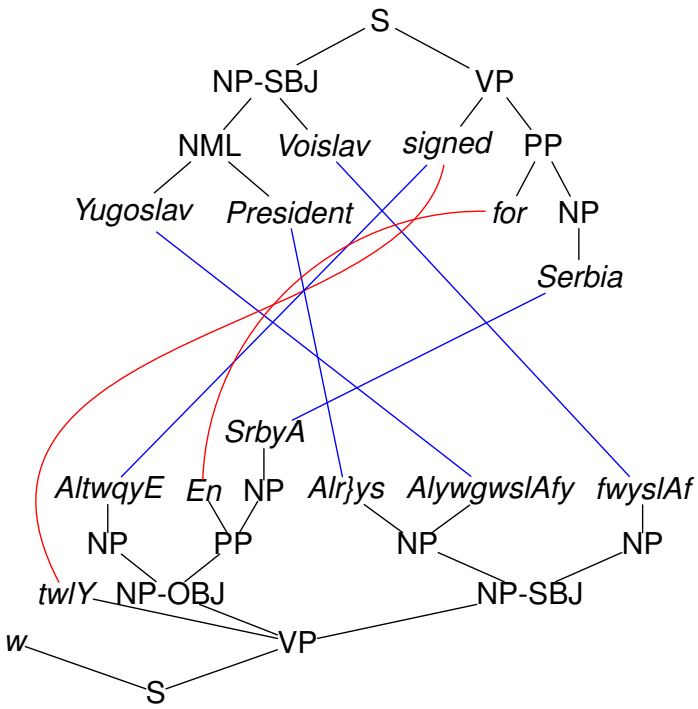


Syntax-based Machine Translation

[GALLEY et al., 2004]

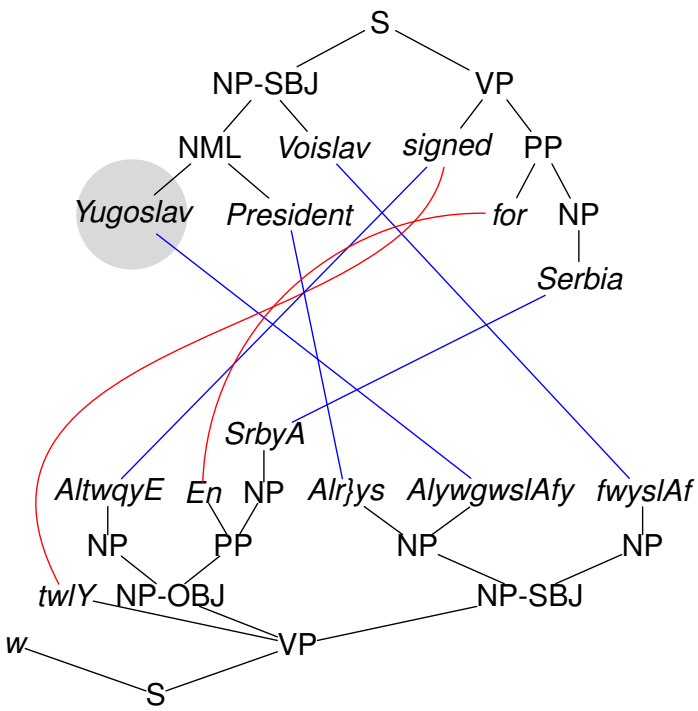


Syntax-based Machine Translation



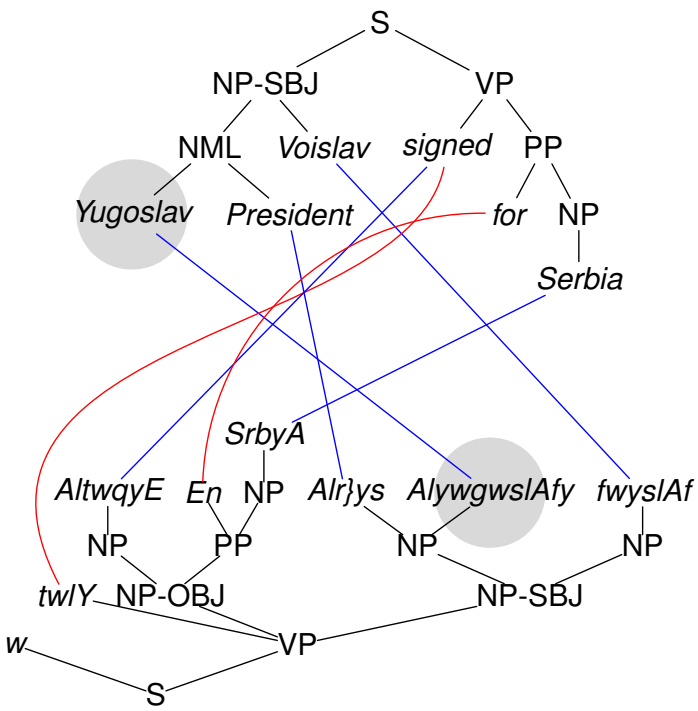
- Select next node bottom-up

Syntax-based Machine Translation



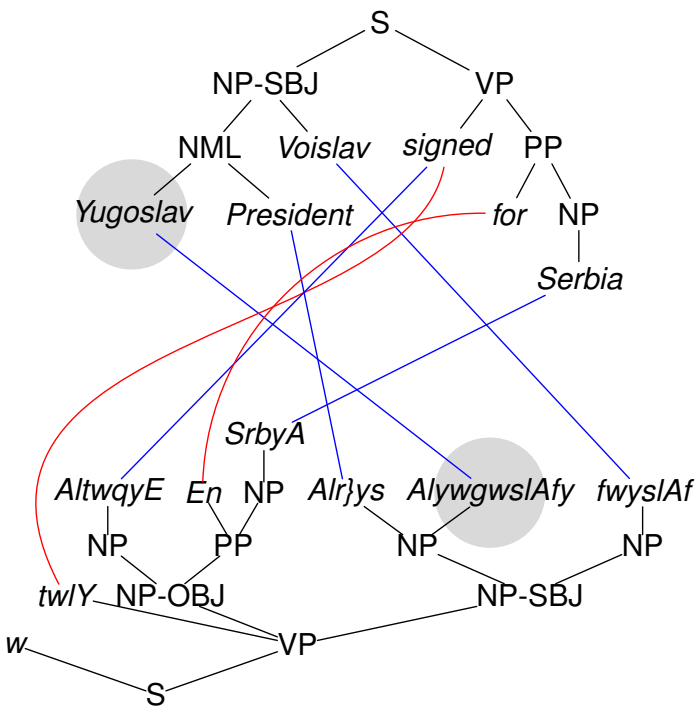
- Select next node bottom-up
- Identify maximal subtree of aligned nodes

Syntax-based Machine Translation



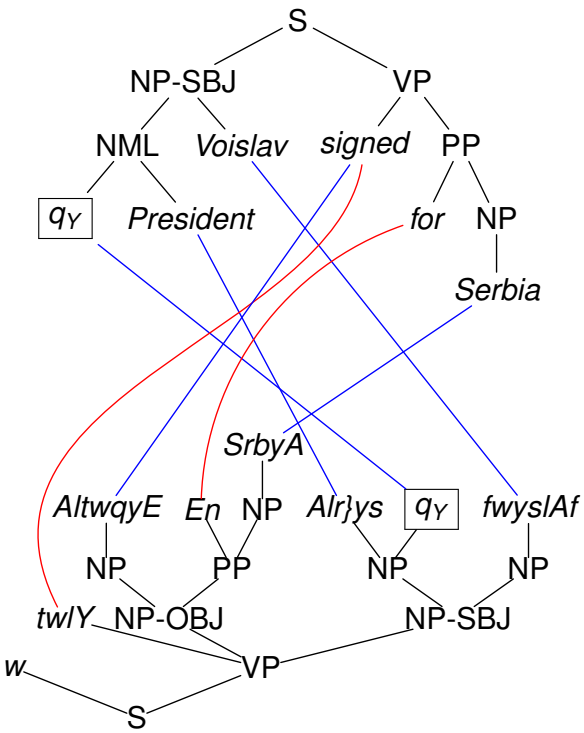
- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state

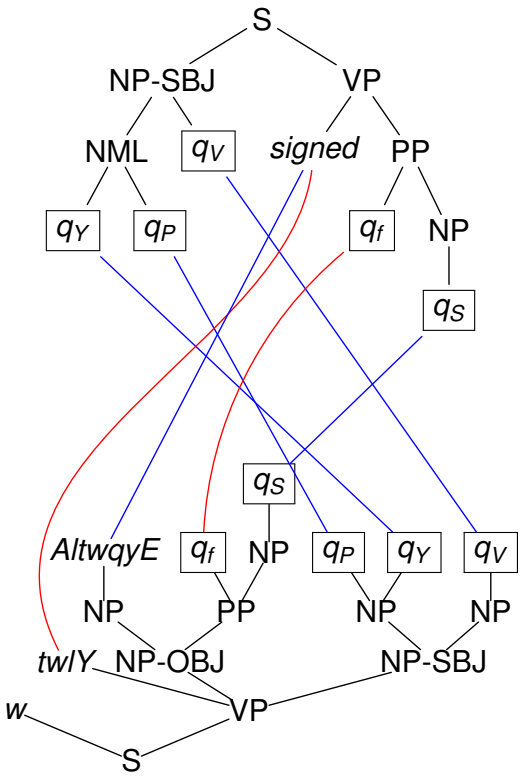
Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

  Yugoslav $\frac{q_Y}{}$ *AlywgwslAfy*

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

$$\textit{Yugoslav} \xrightarrow{q_Y} \textit{AlywgwslAfy}$$

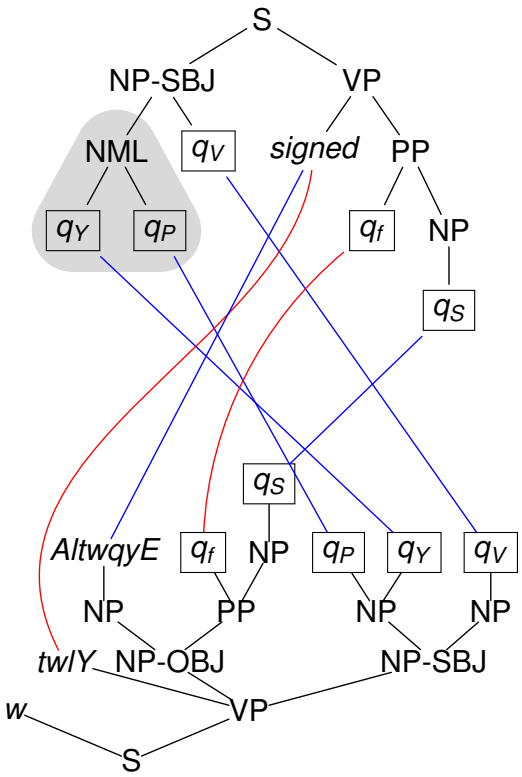
$$\textit{President} \xrightarrow{q_P} \textit{Alr}\}\textit{ys}$$

$$\textit{Voislav} \xrightarrow{q_V} \textit{fwyslAf}$$

$$\textit{for} \xrightarrow{q_f} \textit{En}$$

$$\textit{Serbia} \xrightarrow{q_S} \textit{SrbyA}$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

$$\textit{Yugoslav} \xrightarrow{q_Y} \textit{AlywgwslAfy}$$

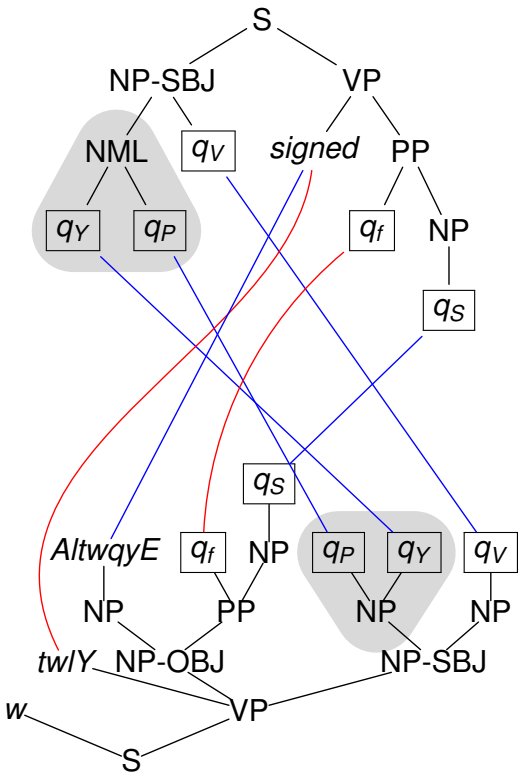
$$\textit{President} \xrightarrow{q_P} \textit{Alr}\}\textit{ys}$$

$$\textit{Voislav} \xrightarrow{q_V} \textit{fwyslAf}$$

$$\textit{for} \xrightarrow{q_f} \textit{En}$$

$$\textit{Serbia} \xrightarrow{q_S} \textit{SrbyA}$$

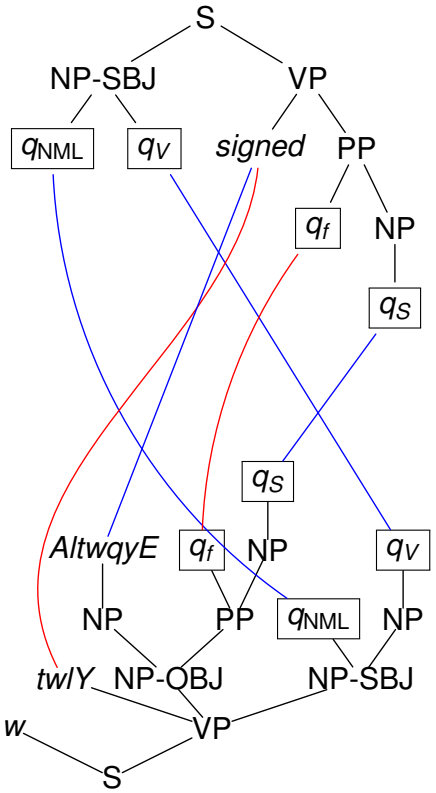
Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

$$\mathrm{NML}(q_Y, q_P) \xrightarrow{q_{\mathrm{NML}}} \mathrm{NP}(q_P, q_Y)$$

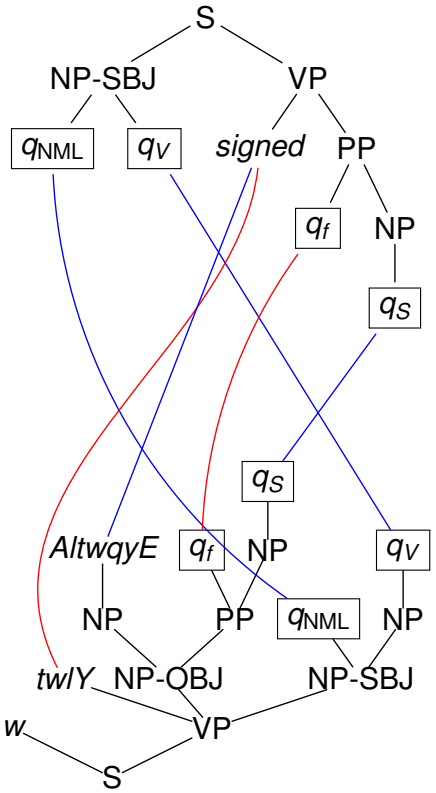
Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

$$\text{NML}(q_Y, q_P) \overset{q_{\text{NML}}}{\underline{\quad}} \text{NP}(q_P, q_Y)$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

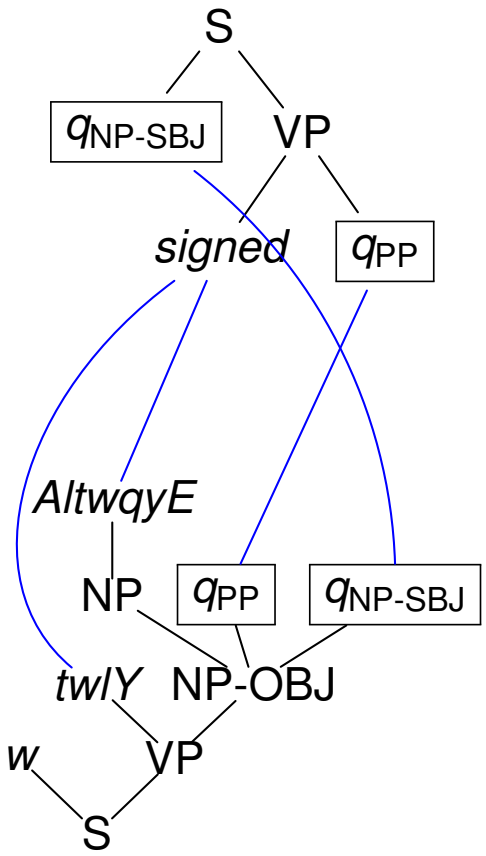
$$\text{NML}(q_Y, q_P) \overset{q_{\text{NML}}}{\underline{\quad}} \text{NP}(q_P, q_Y)$$

$$\text{NP}(q_S) \overset{q_{\text{NP}}}{\underline{\quad}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \overset{q_{\text{PP}}}{\underline{\quad}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \overset{q_{\text{NP-SBJ}}}{\underline{\quad}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

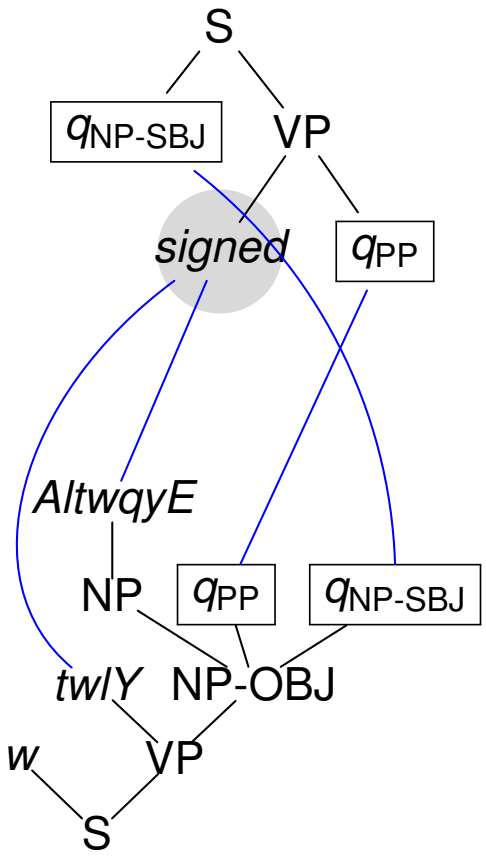
$$\text{NML}(q_Y, q_P) \overset{q_{\text{NML}}}{\underline{\hspace{1em}}} \text{NP}(q_P, q_Y)$$

$$\text{NP}(q_S) \overset{q_{\text{NP}}}{\underline{\hspace{1em}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \overset{q_{\text{PP}}}{\underline{\hspace{1em}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \overset{q_{\text{NP-SBJ}}}{\underline{\hspace{1em}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

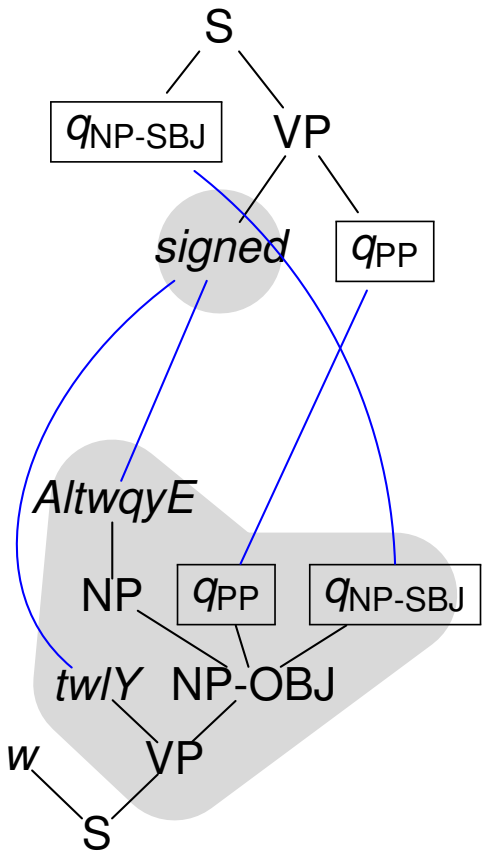
$$\text{NML}(q_Y, q_P) \xrightarrow{q_{\text{NML}}} \text{NP}(q_P, q_Y)$$

$$\text{NP}(q_S) \xrightarrow{q_{\text{NP}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \xrightarrow{q_{\text{PP}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \xrightarrow{q_{\text{NP-SBJ}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

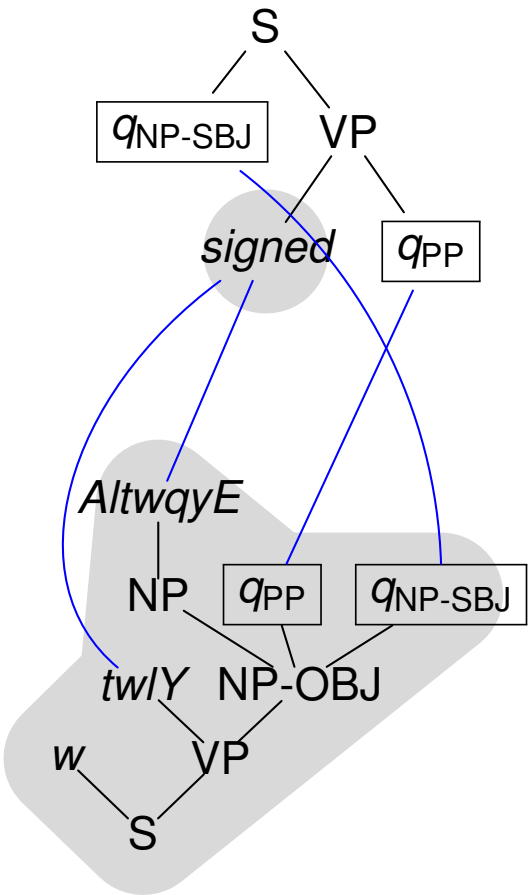
$$\text{NML}(q_Y, q_P) \overset{q_{\text{NML}}}{\underline{\hspace{1em}}} \text{NP}(q_P, q_Y)$$

$$\text{NP}(q_S) \overset{q_{\text{NP}}}{\underline{\hspace{1em}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \overset{q_{\text{PP}}}{\underline{\hspace{1em}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \overset{q_{\text{NP-SBJ}}}{\underline{\hspace{1em}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

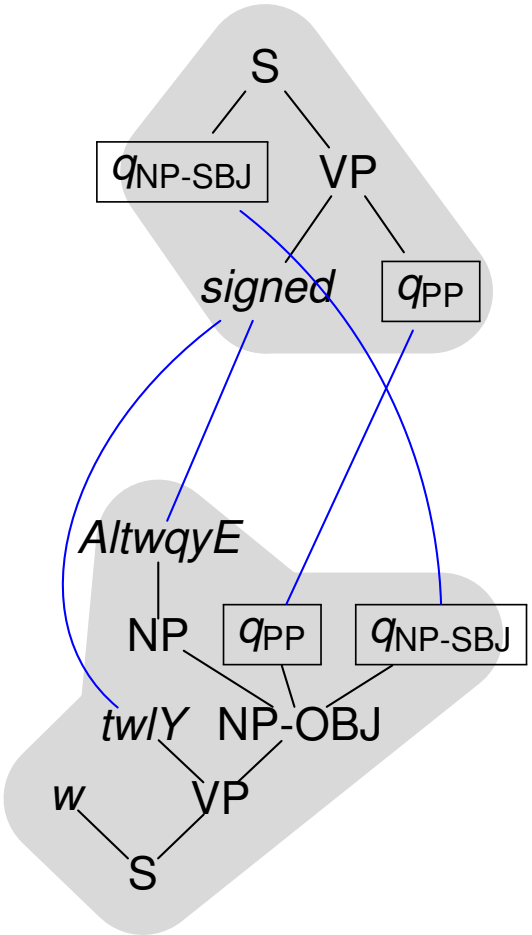
$$\text{NML}(q_Y, q_P) \overset{q_{\text{NML}}}{\underline{\hspace{1em}}} \text{NP}(q_P, q_Y)$$

$$\text{NP}(q_S) \overset{q_{\text{NP}}}{\underline{\hspace{1em}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \overset{q_{\text{PP}}}{\underline{\hspace{1em}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \overset{q_{\text{NP-SBJ}}}{\underline{\hspace{1em}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation



- Select next node bottom-up
- Identify maximal subtree of aligned nodes
- Identify subtree of nodes aligned to aligned nodes, etc.
- Extract rule and leave state
- Repeat

$$\text{NML}(q_Y, q_P) \overset{q_{\text{NML}}}{\underline{\hspace{1em}}} \text{NP}(q_P, q_Y)$$

$$\text{NP}(q_S) \overset{q_{\text{NP}}}{\underline{\hspace{1em}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \overset{q_{\text{PP}}}{\underline{\hspace{1em}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \overset{q_{\text{NP-SBJ}}}{\underline{\hspace{1em}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation

Rules

$$\textit{Yugoslav} \xrightarrow{q_Y} \textit{AlywgwslAfy} \qquad \textit{President} \xrightarrow{q_P} \textit{Alr\}ys}$$

$$\textit{Voislav} \xrightarrow{q_V} \textit{fwyslAf} \qquad \textit{for} \xrightarrow{q_f} \textit{En}$$

$$\textit{Serbia} \xrightarrow{q_S} \textit{SrbyA}$$

$$\text{NML}(q_Y, q_P) \xrightarrow{q_{\text{NML}}} \text{NP}(q_P, q_Y) \qquad \text{NP}(q_S) \xrightarrow{q_{\text{NP}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \xrightarrow{q_{\text{PP}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \xrightarrow{q_{\text{NP-SBJ}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

Syntax-based Machine Translation

Rules

$$\textit{Yugoslav} \xrightarrow{q_Y} \textit{AlywgwslAfy} \qquad \textit{President} \xrightarrow{q_P} \textit{Alr\}ys}$$

$$\textit{Voislav} \xrightarrow{q_V} \textit{fwyslAf} \qquad \textit{for} \xrightarrow{q_f} \textit{En}$$

$$\textit{Serbia} \xrightarrow{q_S} \textit{SrbyA}$$

$$\text{NML}(q_Y, q_P) \xrightarrow{q_{\text{NML}}} \text{NP}(q_P, q_Y) \qquad \text{NP}(q_S) \xrightarrow{q_{\text{NP}}} \text{NP}(q_S)$$

$$\text{PP}(q_f, q_{\text{NP}}) \xrightarrow{q_{\text{PP}}} \text{PP}(q_f, q_{\text{NP}})$$

$$\text{NP-SBJ}(q_{\text{NML}}, q_V) \xrightarrow{q_{\text{NP-SBJ}}} \text{NP-SBJ}(q_{\text{NML}}, \text{NP}(q_V))$$

$\rightarrow$ Rules of an Extended Top-down Tree Transducer

Extended Top-down Tree Transducer

Advantages

- ✓ simple and natural model
- ✓ easy to train (from linguistic resources)
  [GRAEHL et al., 2008]
- ✓ symmetric

Extended Top-down Tree Transducer

Advantages

- ✓ simple and natural model
- ✓ easy to train (from linguistic resources)
  [GRAEHL et al., 2008]
- ✓ symmetric

Generic implementation

- TIBURON [MAY, KNIGHT, 2006]

Extended Top-down Tree Transducer

Disadvantages (also of STSG)

- ✗ no discontinuities
- ✗ not binarizable
  [Aho, Ullman, 1972; Zhang et al., 2006]
- ✗ inefficient input/output restriction
  [∼, Satta, 2010]
- ✗ not composable
  [Arnold, Dauchet, 1982]

Literature

Selected references

CHIANG: *Hierarchical Phrase-Based Translation*
Comput. Linguist. 33, 2007

GALLEY, HOPKINS, KNIGHT, MARCU
What's in a Translation Rule? Proc. NAACL, 2004

KOEHN: *Statistical Machine Translation*
Cambridge University Press, 2010

OCH, NEY: *The Alignment Template Approach to Statistical Machine Translation.* Comput. Linguist. 30, 2004

PAPINENI, ROUKOS, WARD, ZHU: *BLEU: A Method for Automatic Evaluation of Machine Translation.* Proc. 40th ACL, 2002