Challenges in Statistical Machine Translation

Philipp Koehn

koehn@csail.mit.edu

Computer Science and Artificial Intelligence Lab

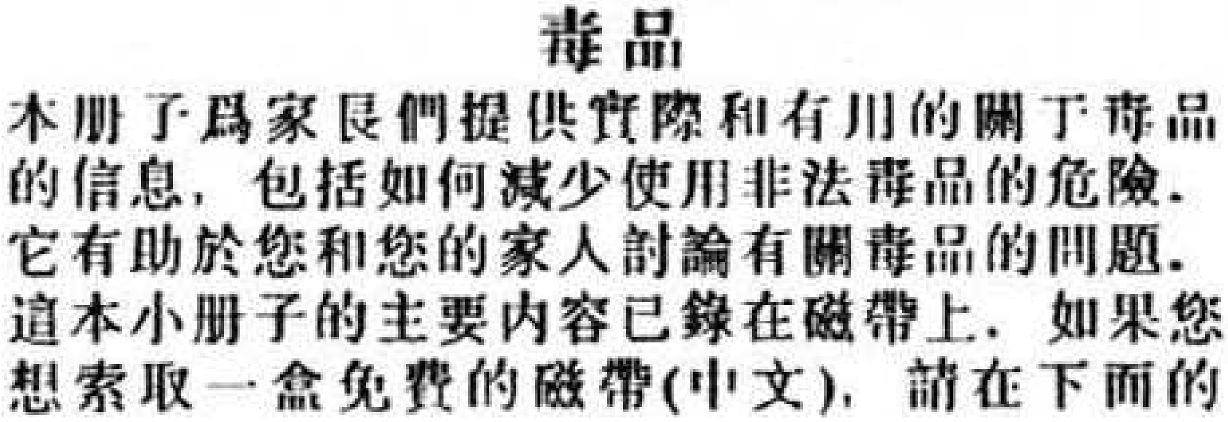
Massachusetts Institute of Technology

Outline

- **Statistical Machine Translation**
- What is wrong with MT?
- Divide and Conquer: Noun Phrase Translation
- Syntactic Transformations
- Discriminative Training

Machine Translation

- Task: Make sense of foreign text like



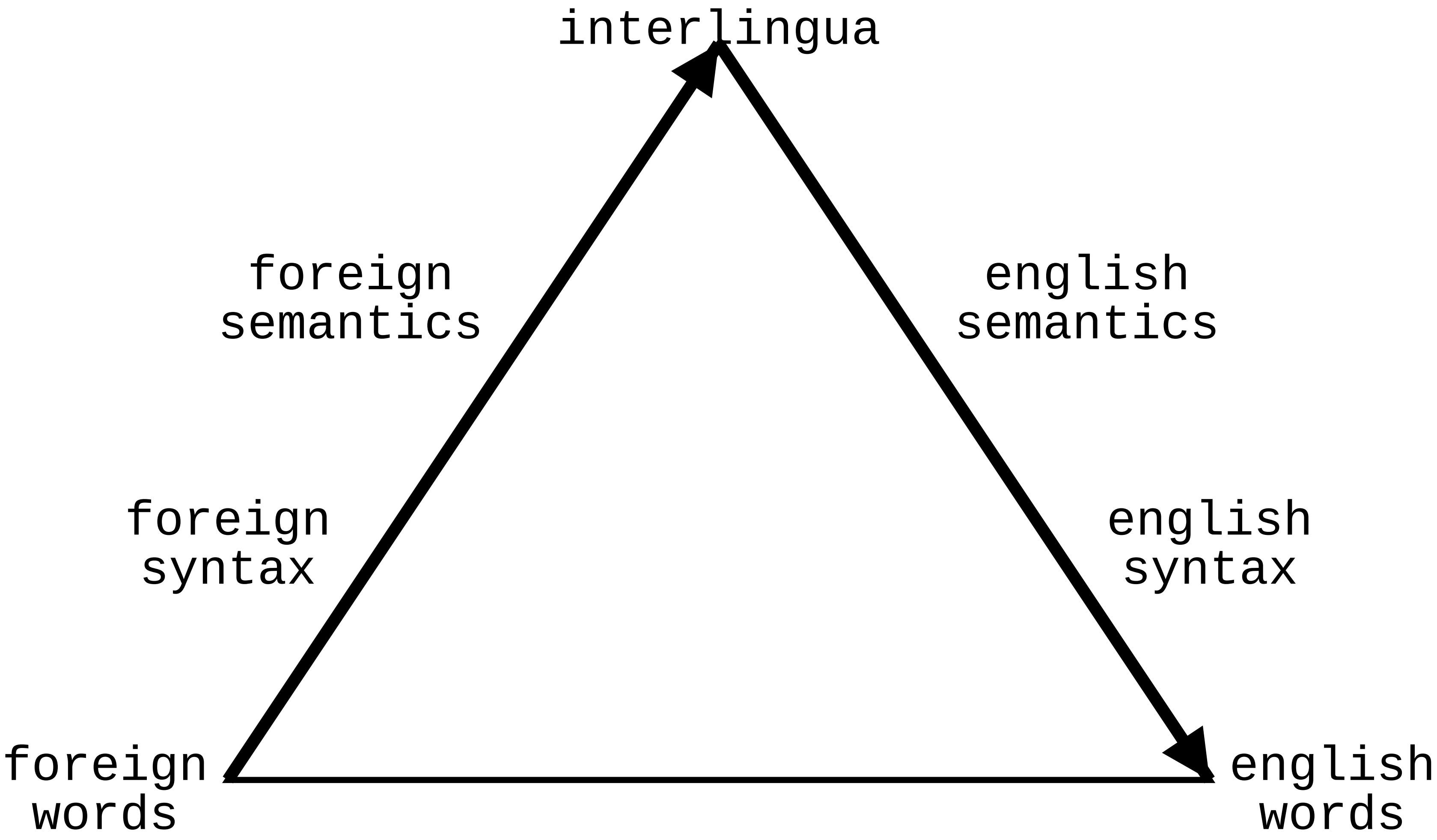
毒品

本冊子爲家長們提供實際和有用的關于毒品
的信息，包括如何減少使用非法毒品的危險．
它有助於您和您的家人討論有關毒品的問題．
這本小冊子的主要內容已錄在磁帶上． 如果您
想索取一盒免費的磁帶(中文)，請在下面的

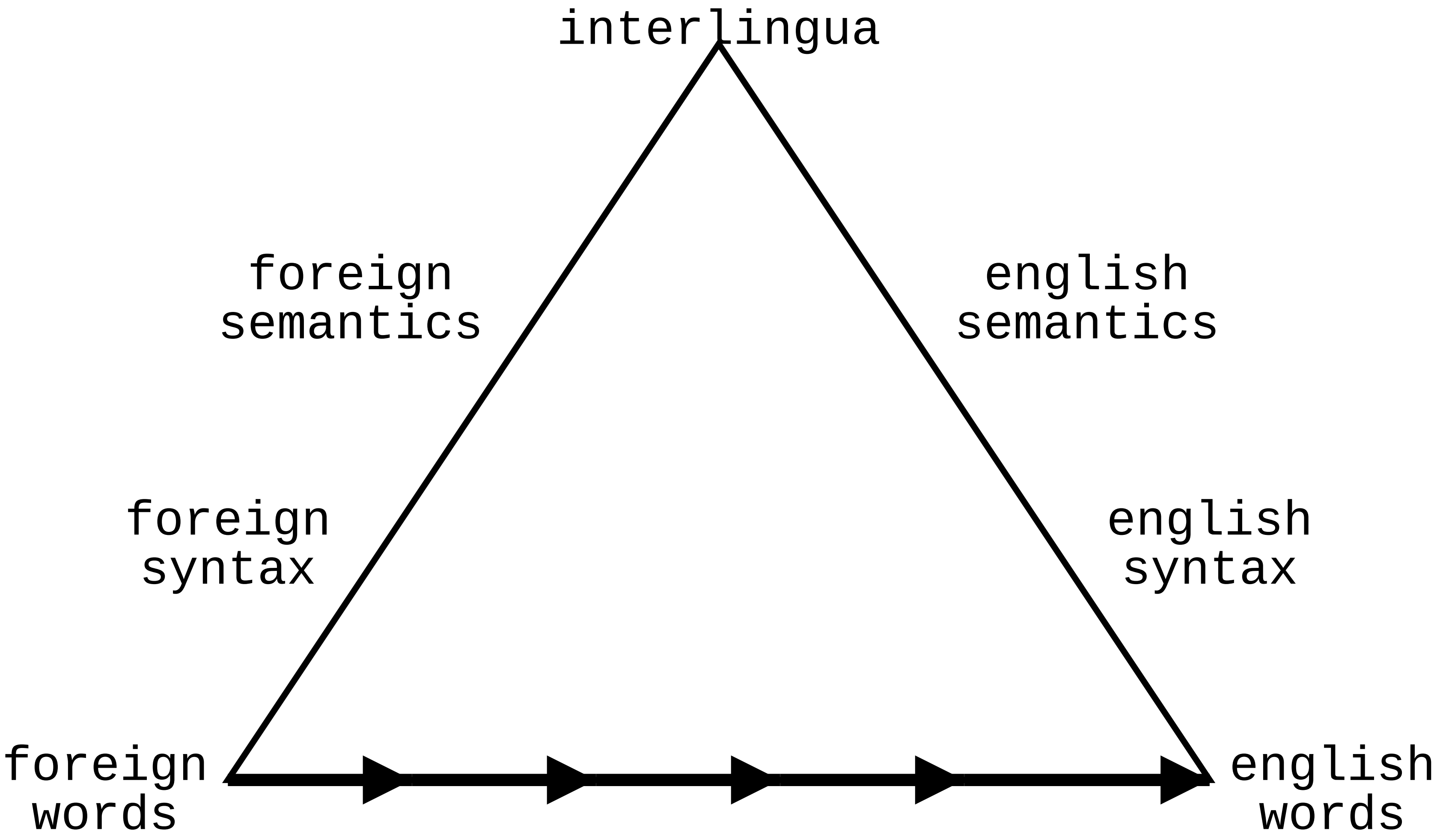
- One of the oldest problems in Artificial Intelligence

- AI-hard: reasoning and world knowledge required

The Machine Translation Pyramid



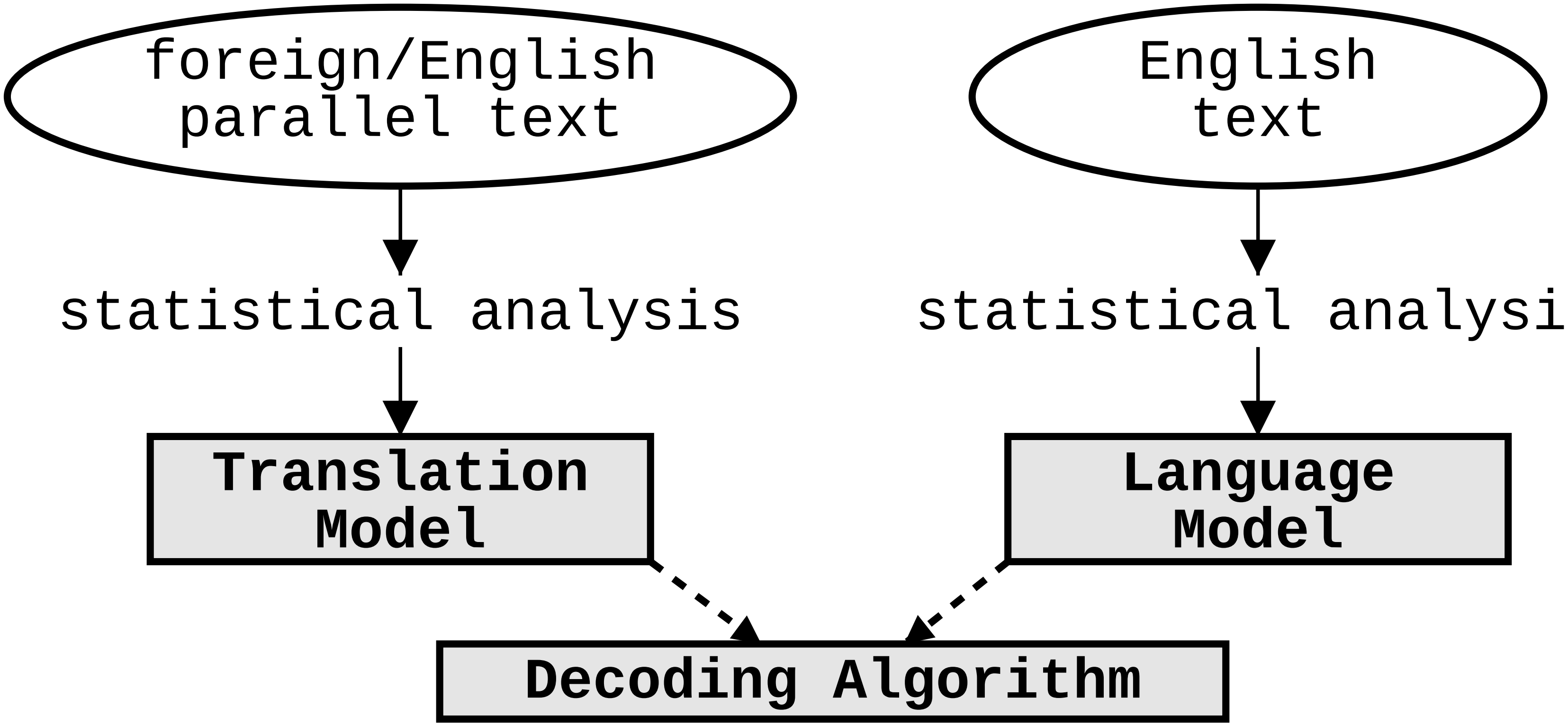
The Machine Translation Pyramid



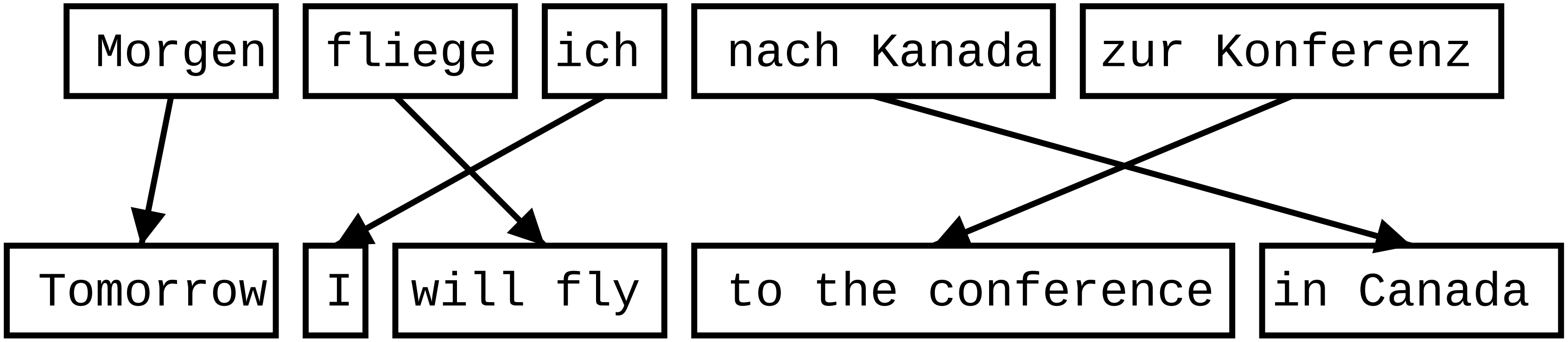
however, the currently best performing statistical machine translation systems are still crawling at the bottom.

Statistical Machine Translation Models

- Components: Translation model, language model, decoder



Phrase-Based Translation



- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered
- See [Koehn et al., NAACL2003] as introduction

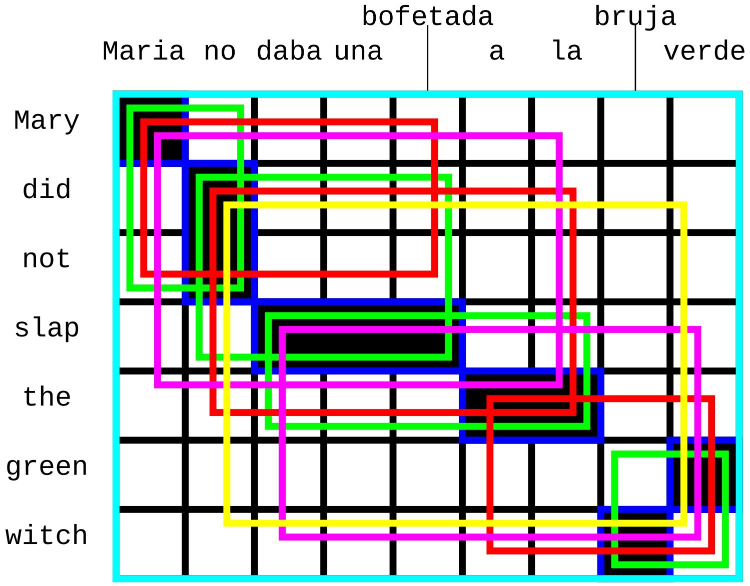
How to Learn the Phrase Translation Table?

- Start with the word alignment:

	Maria	no	daba	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap			■	■	■				
the						■	■		
green									■
witch								■	

- Collect all phrase pairs that are consistent with the word alignment

Collect Phrase Pairs



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch), (Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch), (Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Probability Distribution of Phrase Pairs

- We need a probability distribution over the collected phrase pairs

$\Rightarrow$ Possible choices

- relative frequency of collected phrases:
  $\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})}$
- or, conversely $\phi(\bar{e}|\bar{f})$
- use lexical translation probabilities

Phrase Translation Table

- Phrase Translations for “den Vorschlag”:

English	$\phi(e\|f)$	**English**	$\phi(e\|f)$
the proposal	0.6227	the suggestions	0.0114
’s proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Phrase-Based MT: Do it yourself

- Phrase-based MT has currently best performance
- Corpora available at LDC, ISI, other places

  e.g., Europarl: `http://www.isi.edu/∼koehn/europarl/`
- Giza++ toolkit available at RWTH Aachen

  `http://www-i6.informatik.rwth-aachen.de/web/Software/GIZA++.html`
- Language model available at SRI

  `http://www.speech.sri.com/projects/srilm/`
- Pharaoh decoder available at ISI

  `http://www.isi.edu/licensed-sw/pharaoh/`

Outline

- Statistical Machine Translation
- **What is wrong with MT?**
- Divide and Conquer: Noun Phrase Translation
- Syntactic Transformations
- Discriminative Training

Output of Chinese System

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

MIT statistical MT system, tested on NIST Eval 2002, about 100 million words training data

Output of Chinese System

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Partially excellent translations

Output of Chinese System

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Mangled grammar

Output of Chinese System

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Wrong word choices

Output of Chinese System

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Name Translation

Output of Arabic System

Canada send forces to join war

Ottawa 8 October / Xinhua / The minister of defence announced AlkndY aart Ayjltwn in a press conference Canada here today ships and aircraft send military forces and the special adhering to the war against terrorism, led by the United States. He said that the Ayjltwn six ships of the navy and air force aircraft and soldiers of the special forces will become involved in this process, which will include more than 2,000 fighters . He said that ”this Ayjltwn battle would be different from any other battle xDnAhA.” He added that ”all the important role in this battle. Each country is determined to put a stop to terrorism something that can make a difference. He said that the process is scheduled to take up to six months.

Unknown words

Output of German System

German-English system trained on the Europarl corpus

We need to Biarritz, it is also a little more foresight.

⇒ What we need in Biarritz is also a little more foresight.

⇒ We neeed in Biarritz also a little more foresight.

President Prodi, I believe that this sound recommendations made.

⇒ President Prodi, I believe that these are sound recommendations.

With regard to the events in recent times strikes me as a matter of gas prices particularly noteworthy.

⇒ With regard to the events in recent times, the matter of gas prices strikes me as particularly noteworthy.

Problem: Syntactic transformations

Solution

- More training data
- Better modeling
- Syntax, Semantics, Pragmatics

Outline

- Statistical Machine Translation
- What is wrong with MT?
- **Divide and Conquer: Noun Phrase Translation**
- Syntactic Transformations
- Discriminative Training

Motivation [Koehn and Knight, 2003]

- More knowledge for statistical machine translation
  - current methods use little to none syntactic or semantic knowledge
- Vision: divide and conquer
  - break up translation process into smaller modules
  - enables dedicated modeling

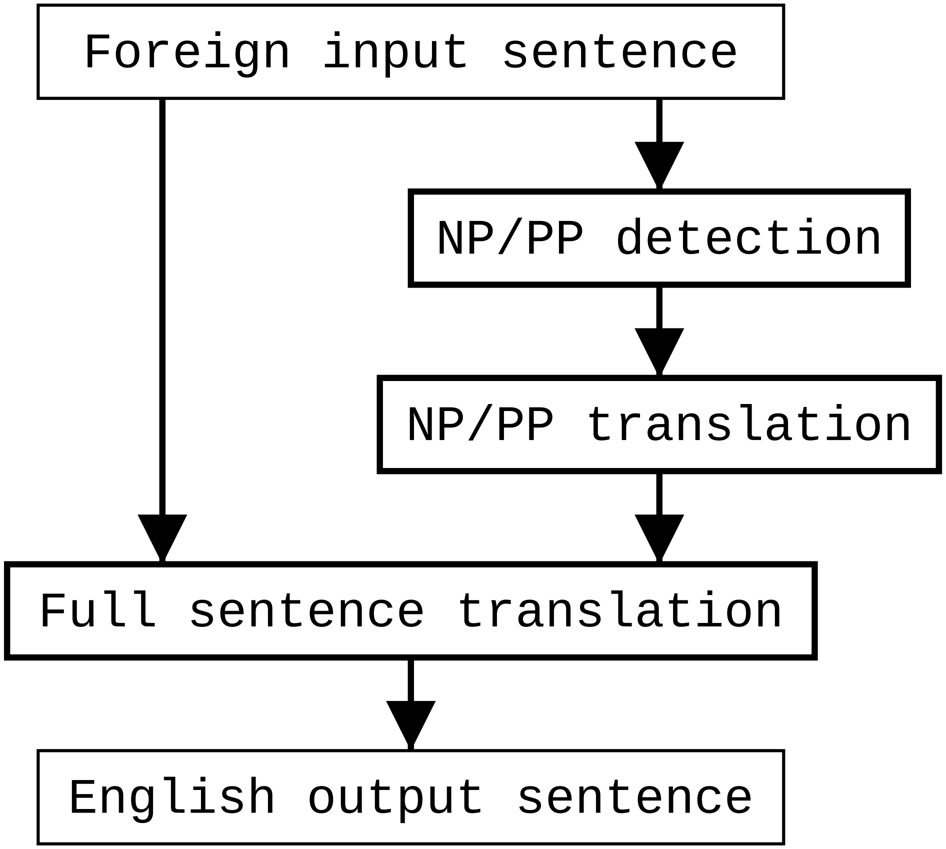
⇒ Translation of noun phrases

  - can be translated in isolation
  - more expensive features and methods can be used

Definition

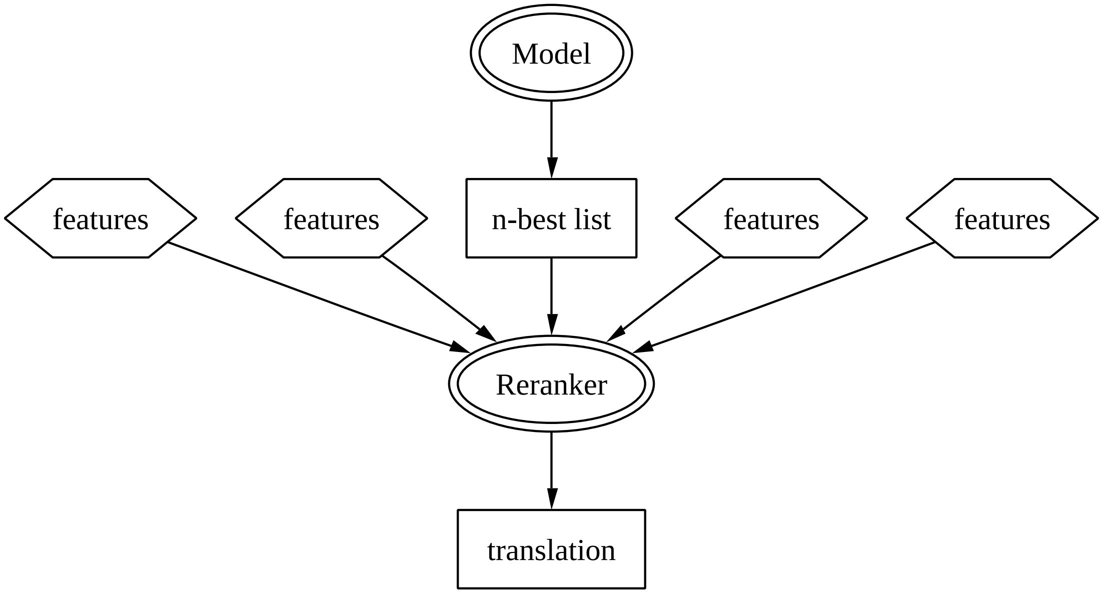
- Definition NP/PP:
  - the maximal noun phrases that are attached at the clause level
  - not contain relative clauses
  - not just baseNP
  - also includes prepositional phrases
- Are NP/PPs translated as NP/PPs?
  - German-English: 75% are translated, 98% can be

Framework



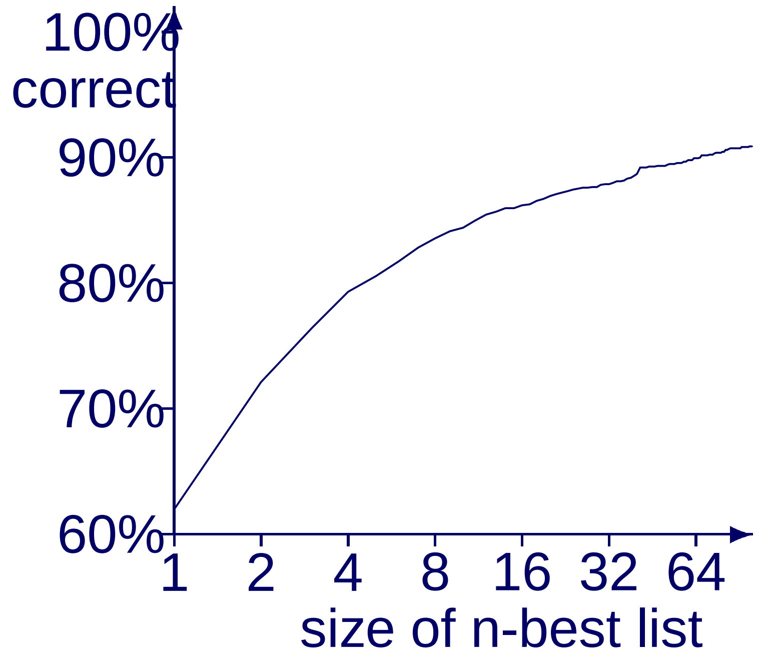
- NP/PPs translated by modular subsystem

Translation as Reranking



- Base model proposes candidate
- Reranking with additional features
  - maximum entropy
  - similar to [Och and Ney, 2002]

Translation as Reranking: Why Possible?

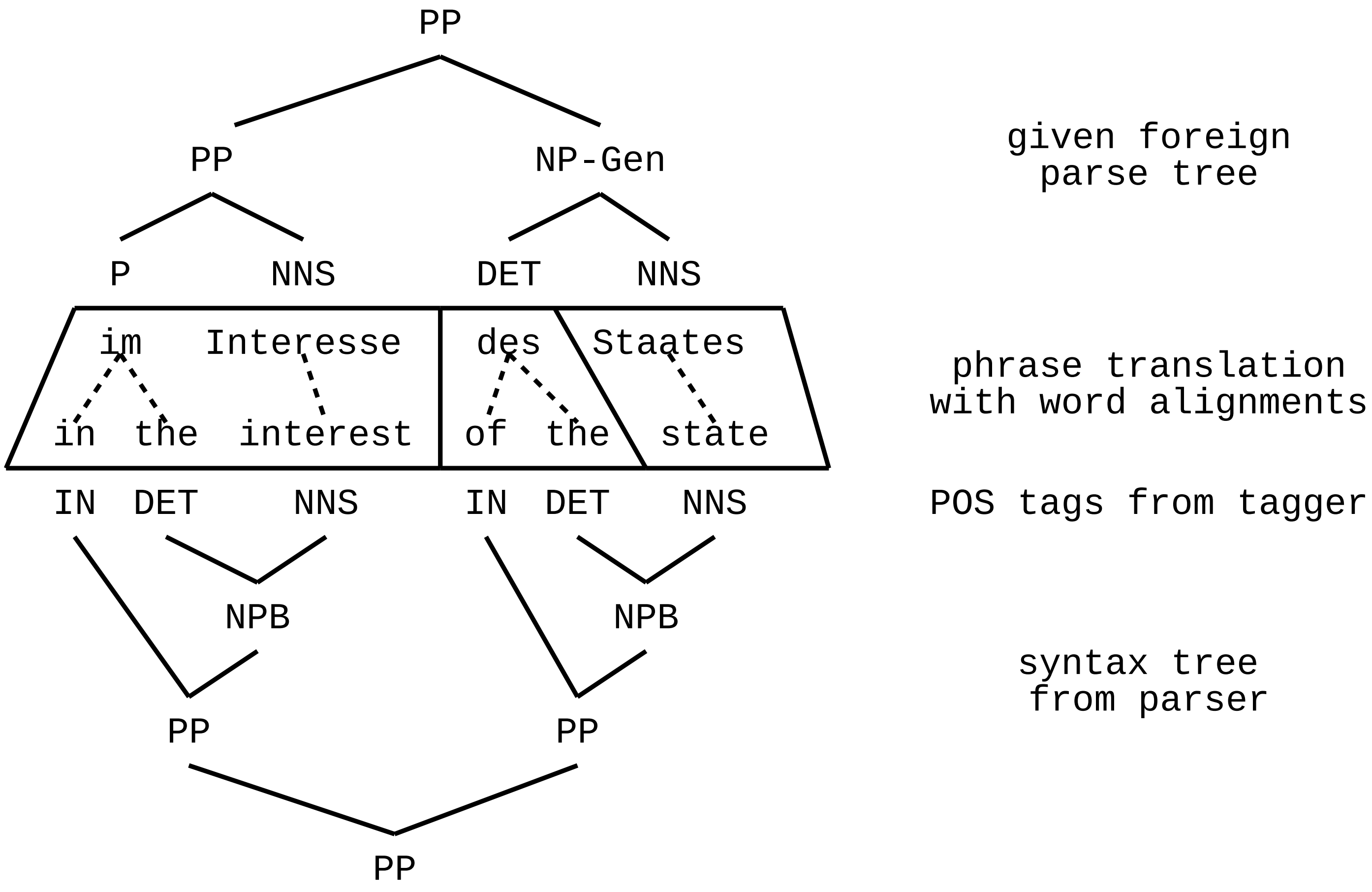


- 60% of NP/PPs translated correctly
- 90% of NP/PPs have correct translation in 100-best list
- Advantage of reranking: global features

Special Modeling for NP/PP Translation

- Compound splitting
- Web n-Grams
- Syntactic features

Syntactic Features



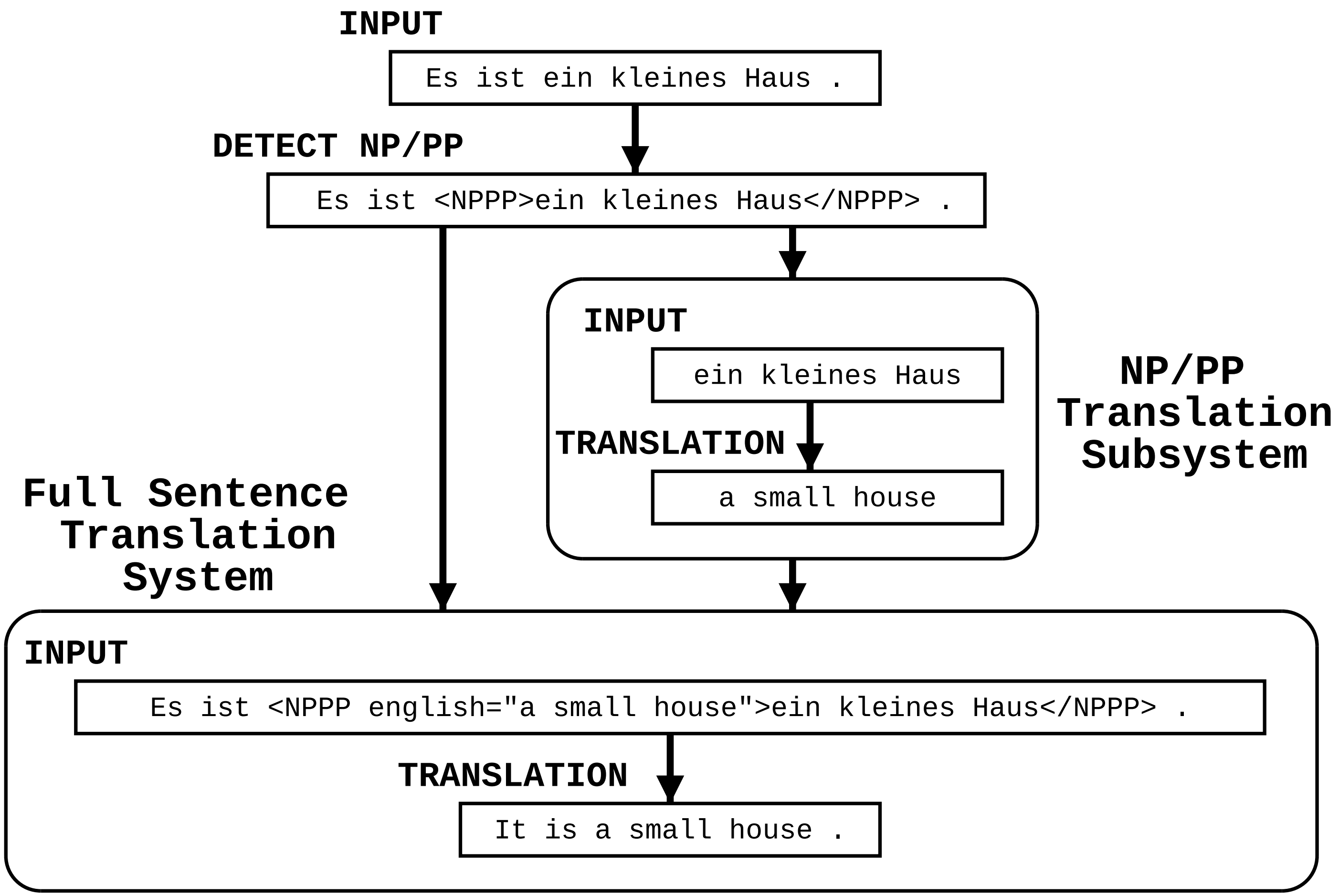
- Keep foreign syntactic parse tree
- Annotate English candidate translation with syntax

Accuracy (Human Judgment)

System	**NP/PP Correct**	
Word-Based Model	724	53.2%
Phrase-Based Model	800	58.7%
Compound Splitting	838	61.5%
Re-Estimated Parameters	858	63.0%
Web Count Features	881	64.7%
Syntactic Features	892	65.5%

- Overall +12.3% improvement
- 95% Statistical significance interval 2.5%

Integration


INPUT
Es ist ein kleines Haus .
DETECT NP/PP
Es ist <NPPP>ein kleines Haus</NPPP> .
INPUT
ein kleines Haus
TRANSLATION
a small house
NP/PP Translation Subsystem
Full Sentence Translation System
INPUT
Es ist <NPPP english="a small house">ein kleines Haus</NPPP> .
TRANSLATION
It is a small house .

- Translations passed to full sentence translation system
  - using XML markup
  - allow passing of reranked list (with probabilities)

Evaluation of Integration

- Performance on full-sentence translation (BLEU score)

System	Word-Based MT	Phrase-Based MT
baseline system	17.6%	22.0%
with NP/PP subsystem	19.9%	22.4%

- Why little improvement for phrase-based MT?
  - cuts around NP/PP disable overlapping phrase translations
  - parsing errors force hard decisions

Conclusions on NP/PP Translation

- It is possible to separate out NP/PP translation
- Improved NP/PP translation performance
- Improved overall sentence translation performance
  - still needs better integration
  - still needs better conditioning on sentence context

Outline

- Statistical Machine Translation
- What is wrong with MT?
- Divide and Conquer: Noun Phrase Translation
- **Syntactic Transformations**
- Discriminative Training

Weaknesses of Phrase-Based Models

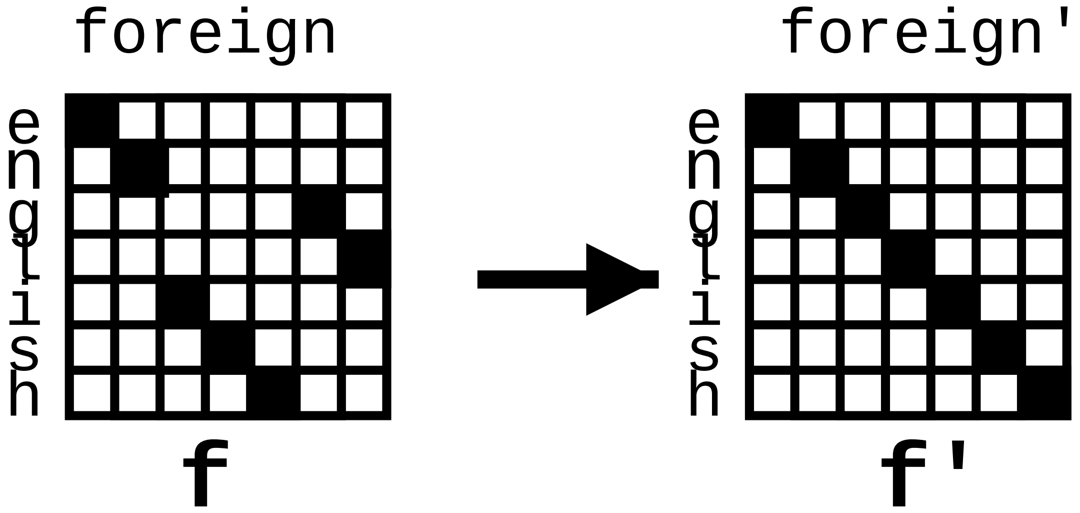
- Phrase-based SMT is pretty good at
  - word choices
  - ideomatic expressions
  - local restructuring
- ... but bad at
  - large-scale reordering
  - add, drop, change of function words for non-local reasons
  - correct syntax on sentence level

German Verb Movement

- Ongoing work with Ivona Kucera
- Example
  - `OBJ V SBJ` ⇒ `SBJ V OBJ`
  - `NP AUX NP NP NP V` ⇒ `NP AUX V NP NP NP`
- Preliminary results on rules for verb movement
  - deterministic preprocessing on test and train

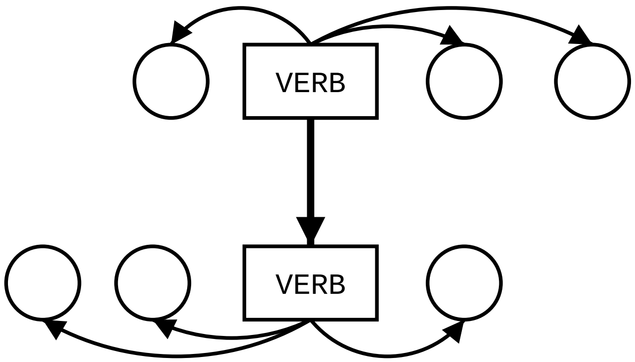
  ⇒ improvement in BLEU

Data-Driven Transformation Model

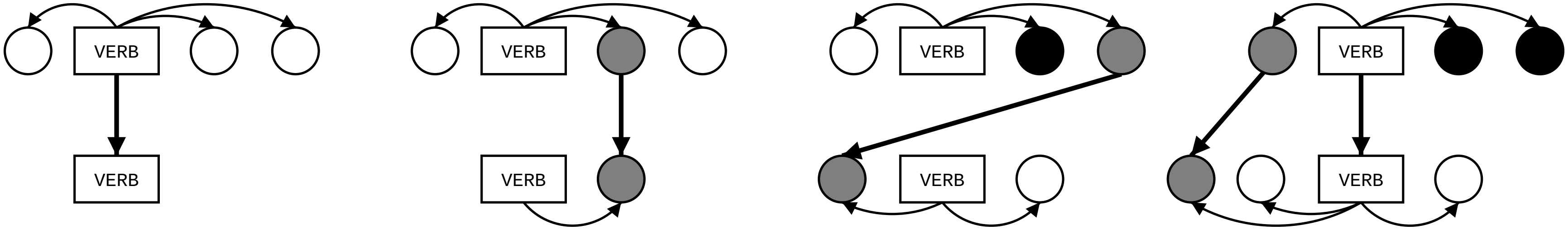


- Definition of the reordering task
  - reorder foreign to be more similar to English word order
  - can be learned from parallel corpus (supervised data)
  - error metric: number/length of discontinuities
  - one possible model: $\prod_i p(\text{move}|\text{word}_i, \text{additional features})$

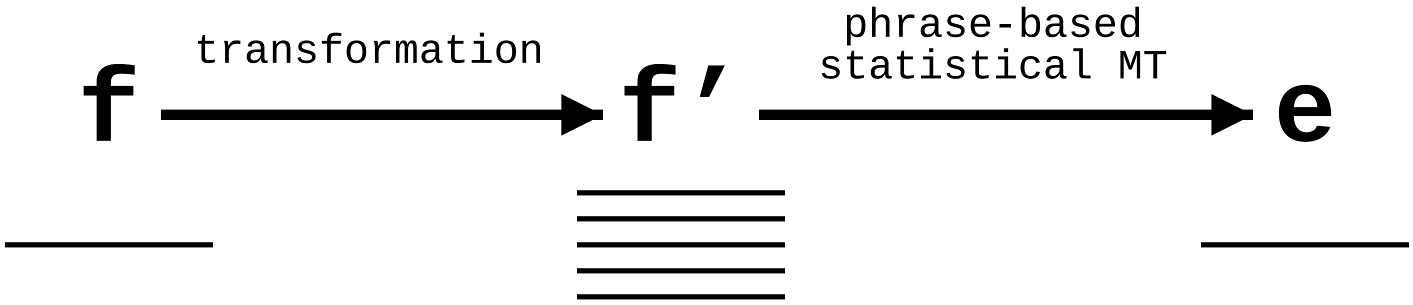
Verbal Argument Structure



- Another model: verb-focused syntax model
  - flat tree on clause structure
  - first map verb, then restructure arguments



Integration



f

transformation

f'

phrase-based statistical MT

e

- Transform f into f' with our methods
- Translate n-best restructurings with phrase-based MT
  - uses both transformation score and translation/language model score
  - if no restructuring → baseline performance
- Transformation does not need to be perfect
  - phrase-based model may still reorder

Outline

- Statistical Machine Translation
- What is wrong with MT?
- Divide and Conquer: Noun Phrase Translation
- Syntactic Transformations
- **Discriminative Training**

Knowledge Sources

- Many different knowledge sources useful
  - language model
  - reordering (distortion) model
  - phrase translation model
  - word translation model
  - word penalty
  - additional language models
  - additional features

Components in 2004 NIST Eval System

- reordering model
- language model trained on all data
- language model trained on news data
- phrase translation model f→e
- phrase translation model e→f
- word translation model f→e
- word translation model e→f
- word penalty
- phrase penalty

Log-Linear Models

- IBM Models provided mathematical justification for factoring components together

$$p_{LM} \times p_{TM} \times p_D$$

- These may be weighted

$$p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D}$$

- Many components p_i with weights λ_i

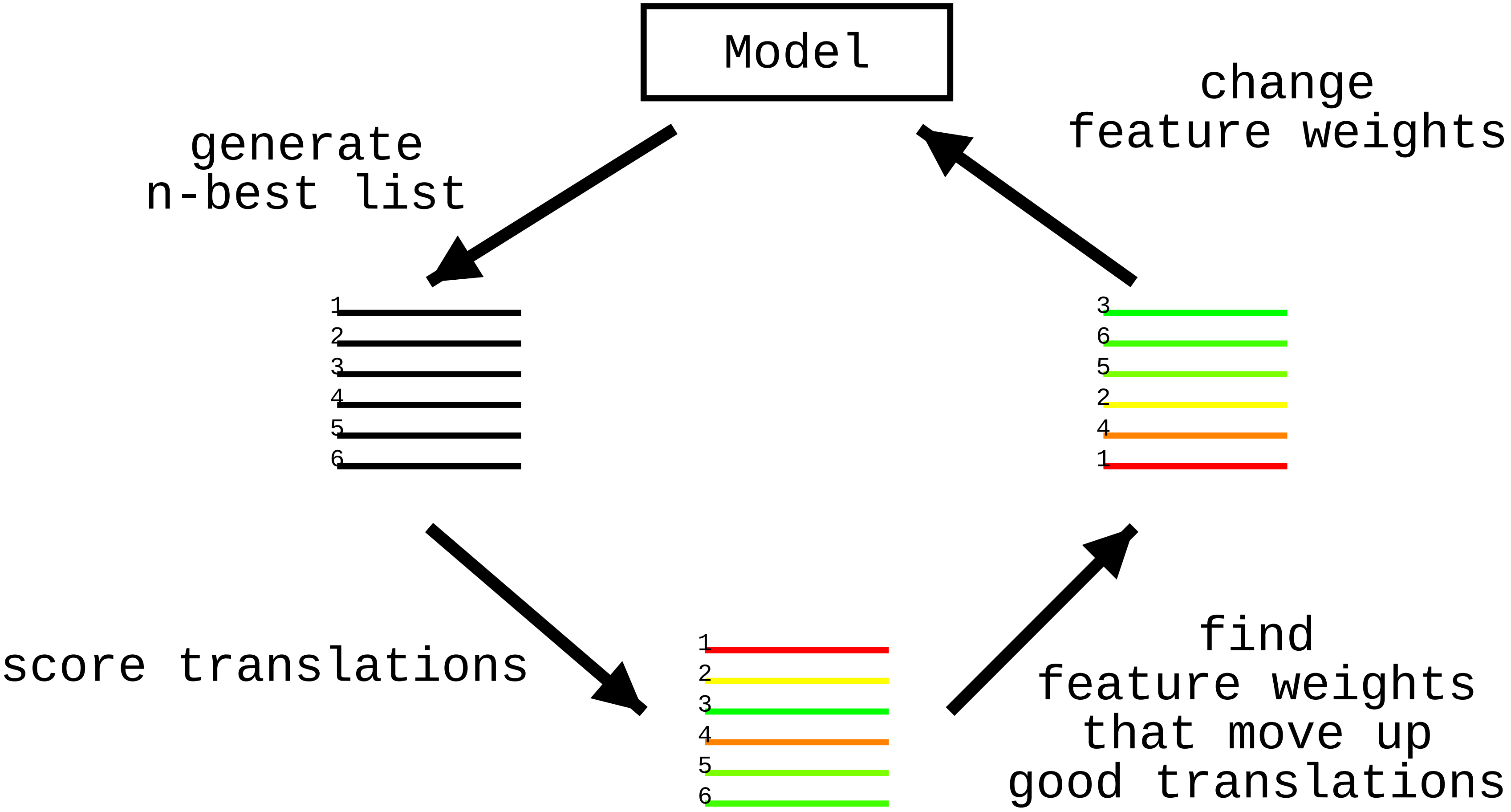
$\Rightarrow \prod_i p_i^{\lambda_i} = exp(\sum_i \lambda_i log(p_i))$

$\Rightarrow log \prod_i p_i^{\lambda_i} = \sum_i \lambda_i log(p_i)$

Set Feature Weights

- Contribution of components p_i determined by weight λ_i
- Methods
  - manual setting of weights: try a few, take best
  - automate this process
- Learn weights
  - set aside a development corpus
  - set the weights, so that optimal translation performance on this development corpus is achieved
  - requires automatic scoring method (e.g., BLEU)

Learn Feature Weights



Discriminative vs. Generative Models

- Generative models
  - translation process is broken down to steps
  - each step is modeled by a probability distribution
  - each probability distribution is estimated from the data by maximum likelihood
- Discriminative models
  - model consist of a number of features (e.g. the language model score)
  - each feature has a weight, measuring its value for judging a translation as correct
  - feature weights are optimized on training data, so that the system output matches correct translations as close as possible

Discriminative Training (2)

- Training set (“development set”)
  - different from original training set
  - small (maybe 1000 sentences)
  - must be different from test set
- Current model translates this development set
  - n-best list of translations (n=100, 10000)
  - translations in n-best list can be scored
- Feature weights are adjusted
- N-Best list generation and feature weight adjustment repeated for a number of iterations

Learning Task

- Task: find weights, so that feature vector of the correct translations scores best

  methods differ in what is meant by **find**, **correct**, **translations**, and **best**

	TRANSLATION	**LM**	**TM**	**WP**	**SER**
1	Mary not give slap witch green .	-17.2	-5.2	-7	1
2	Mary not slap the witch green .	-16.3	-5.7	-7	1
3	Mary not give slap of the green witch .	-18.1	-4.9	-9	1
4	Mary not give of green witch .	-16.5	-5.1	-8	1
5	Mary did not slap the witch green .	-20.1	-4.7	-8	1
6	Mary did not slap green witch .	-15.5	-3.2	-7	1
7	Mary not slap of the witch green .	-19.2	-5.3	-8	1
8	Mary did not give slap of witch green .	-23.2	-5.0	-9	1
9	Mary did not give slap of the green witch .	-21.8	-4.4	-10	1
10	Mary did slap the witch green .	-15.5	-6.9	-7	1
11	**Mary did not slap the green witch .**	**-17.4**	**-5.3**	**-8**	**0**
12	Mary did slap witch green .	-16.9	-6.9	-6	1
13	Mary did slap the green witch .	-14.3	-7.1	-7	1
14	Mary did not slap the of green witch .	-24.2	-5.3	-9	1
15	Mary did not give slap the witch green .	-25.2	-5.5	-9	1
rank	**translation**	**feature vector**			

Previous Work

- System tuning:
  - small development set
  - few features
- Approaches
  - maximum entropy [Och and Ney, ACL2002]

    also used for noun phrase translation reranking [Koehn and Knight, 2003]
  - minimum error rate training [Och, ACL2003]
  - ordinal regression [Shen et al., NAACL2004]

Ongoing Work

- Ongoing work with Michael Collins, Luke Zettlemoyer, and Brooke Cowan
  - training over entire training corpus
- Define likelyhood of good translations
  - compare reference translation to system output
  - or: loss function that assigns partial credit to n-best
- Algorithms
  - various gradient descent methods

Thank You!

- Questions?