

# SYNTAX FOR STATISTICAL MACHINE TRANSLATION

Franz Josef Och

och@isi.edu

**Franz Josef Och – USC/ISI**

**Daniel Gildea – UPenn**

**Anoop Sarkar – SFU**

**Kenji Yamada – XRCE**

**Sanjeev Khudanpur – JHU**

**Alex Fraser – USC/ISI**

**Shankar Kumar – JHU**

**David Smith – JHU**

**Libin Shen – UPenn**

**Viren Jain – UPenn**

**Katherine Eng – Stanford**

**Zhen Jin – Mt. Holyoke**

**+ Assoc.: Dragomir Radev – Univ. of Michigan**

major tasks:

- **MODELING:** introducing structures into the probabilistic dependencies

$$(I_1 | J_1) = \lambda_1^M (I_1 | J_1)$$

- **TRAINING:** estimation of free parameters using training data  $(f_1^S \ e_1^S)$

$$\hat{\lambda}_1^M = \operatorname{argmax}_{\lambda_1^M} \text{SOME-CRITERION}(f_1^S \ e_1^S \ \lambda_1^M)$$

- **SEARCH:** finding 'best' translation

$$\hat{e} = \operatorname{argmax}_e \hat{\lambda}_1^M(e|f)$$

additional important tasks:

- **data acquisition:** collection of parallel training data
- **pre-/postprocessing:** tokenization, normalization, ...
- **evaluation:** assessing the quality of MT output

Approach to model (e|f): decompose into simpler dependencies

- $m \left( \begin{smallmatrix} I \\ 1 \end{smallmatrix} \mid \begin{smallmatrix} J \\ 1 \end{smallmatrix} \right)$  with  $\lambda_m = 1$  : feature functions
- $m$  with  $\lambda_m = 1$  : interpolation parameter

Log-linear models:

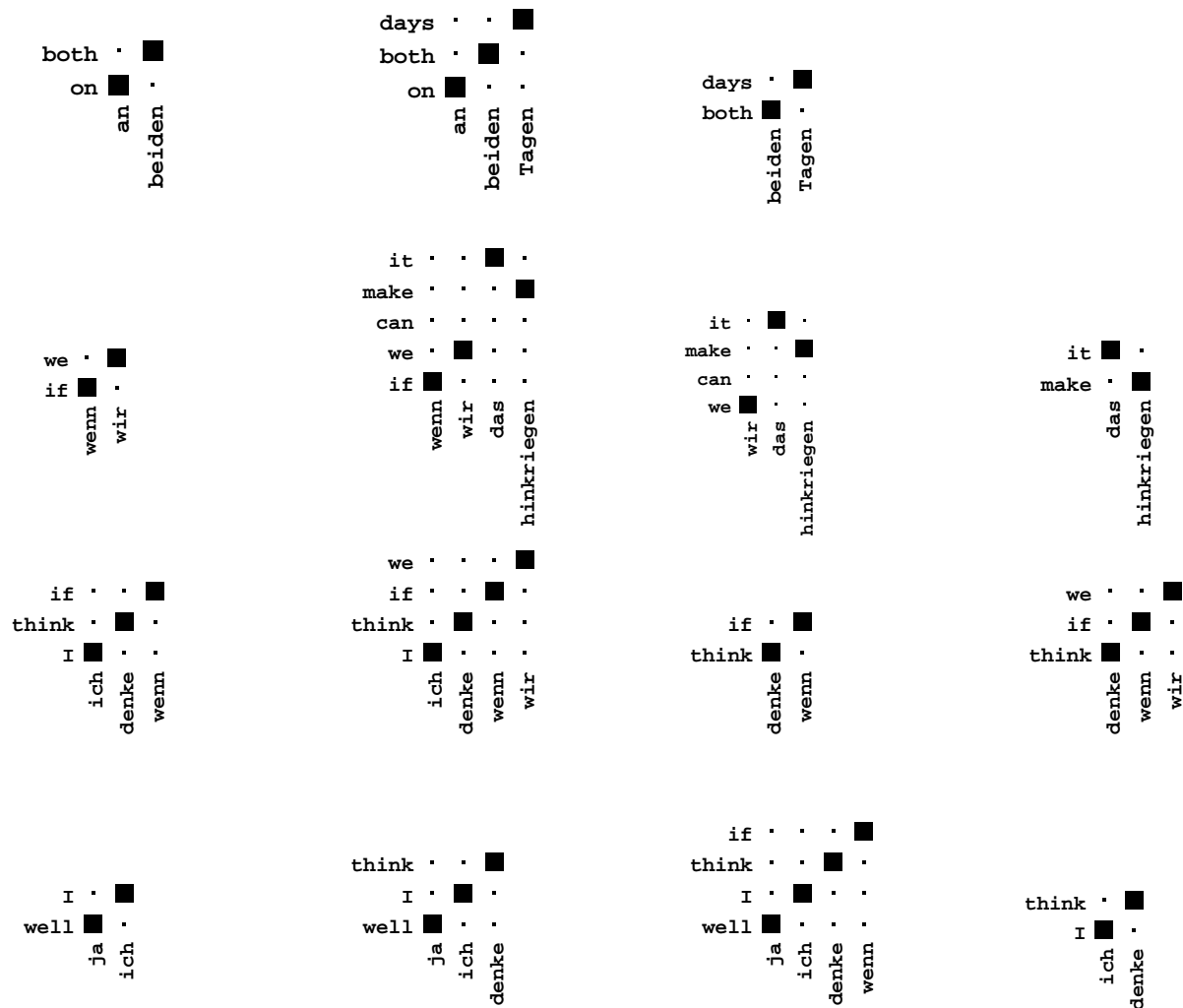
$$\begin{aligned} p \left( \begin{smallmatrix} I \\ 1 \end{smallmatrix} \mid \begin{smallmatrix} J \\ 1 \end{smallmatrix} \right) &= \lambda_1^M \left( \begin{smallmatrix} I \\ 1 \end{smallmatrix} \mid \begin{smallmatrix} J \\ 1 \end{smallmatrix} \right) \\ &= \frac{\exp \left[ \sum_{m=1}^M m \left( \begin{smallmatrix} I \\ 1 \end{smallmatrix} \mid \begin{smallmatrix} J \\ 1 \end{smallmatrix} \right) \cdot \lambda_m \right]}{\sum_{e_1^I} \exp \left[ \sum_{m=1}^M m \left( \begin{smallmatrix} I \\ 1 \end{smallmatrix} \mid \begin{smallmatrix} J \\ 1 \end{smallmatrix} \right) \cdot \lambda_m \right]} \end{aligned}$$

Very simple decision rule (with zero-one loss function):

$$\hat{e}_1^I = \operatorname{argmax}_{e_1^I} \left\{ \sum_{m=1}^M m \left( \begin{smallmatrix} I \\ 1 \end{smallmatrix} \mid \begin{smallmatrix} J \\ 1 \end{smallmatrix} \right) \cdot \lambda_m \right\}$$

What are the relevant dependencies?





hidden variables of alignment template translation model:

- $\frac{K}{1}$ : sequence of alignment templates
- $\frac{K}{1}$ : alignment of alignment templates

Feature functions:

alignment template selection:

$$\rightarrow \mathbf{AT} \left( \frac{I}{1} \quad \frac{J}{1} \quad \frac{K}{1} \quad \frac{K}{1} \right) = \log \left( \frac{K}{1} \mid \frac{J}{1} \right) = \log \prod_{k=1}^K \left( k \mid \tilde{k} \right)$$

(phrase) alignment:

$$\rightarrow \mathbf{AL} \left( \frac{I}{1} \quad \frac{J}{1} \quad \frac{K}{1} \quad \frac{K}{1} \right) = \log \left( \frac{K}{1} \mid \frac{J}{1} \quad \frac{K}{1} \right) =$$

word selection:

$$\rightarrow \mathbf{WRD} \left( \frac{I}{1} \quad \frac{J}{1} \quad \frac{K}{1} \quad \frac{K}{1} \right) = \log \prod_{i=1}^I \left( i \mid \text{'aligned French words'} \right)$$

word-based trigram language model feature function:

$$\rightarrow \mathbf{LM} \left( \frac{I}{1} \quad \frac{J}{1} \quad \frac{K}{1} \quad \frac{K}{1} \right) = \log \prod_{i=1}^I \left( i \mid i-2 \quad i-1 \right)$$

number of produced words (word penalty):

number of produced phrases (alignment template penalty):

...

**(Discriminative) Training Criteria:**

- **standard training criterion: maximum posterior probability**

$$\hat{\lambda}_1^M = \operatorname{argmax}_{\lambda_1^M} \left\{ \sum_{s=1}^S \log \lambda_1^M(e_s | f_s) \right\}$$

- **direct optimization of BLEU score / NIST score [Och03]**

**Search:**

- **DP beam-search producing sentence from left to right**  
baseline feature functions can be easily decomposed for left-to-right search
- **restrict possible reordering to a maximal jump distance of 10 words**
- **N-best list extraction: (optimal) A\* search**

---

**CHINESE-ENGLISH DARPA/NIST MT EVALUATION (LARGE DATA TRACK)**

<b>System</b>	<b>2002</b>	<b>2003</b>
<b>alignment template approach</b>	<b>7.6 (RWTH)</b>	<b>9.0 (ISI)</b>
<b>best competing research systems</b>	<b>7.3</b>	<b>7.9</b>
<b>best of six commercial off-the-shelf systems</b>	<b>6.1</b>	<b>6.6</b>

**note: 95%-confidence intervals: +/- 0.15**

**good results, but: system makes often 'stupid' errors**

- The resolution urged the to the ceasefire and demands ...

**BETTER:** The resolution urges Israel and the Palestinians to cease fire and demands ...

- Indonesia that oppose the presence of foreign troops.

**BETTER:** Indonesia reiterated its opposition to foreign military presence.

- in brief ) ( South Korea will be on high-ranking official to visit North Korea

**BETTER:** (News in Brief) South Korean high-ranking officials to visit North Korea in April

- Condemns US interference in its internal affairs.

**BETTER:** Ukraine condemns US interference in its internal affairs.

- Japan to freeze Rusia to provide humanitarian aid

**BETTER:** Japan to freeze humanitarian assistance to Russia

- ... if the west further sanctions against zimbabwe ...  
BETTER: ... if western countries impose further sanctions against zimbabwe ...
- ... he is fully able to activate team .  
BETTER: ... he is fully able to activate the team .
- ... , particularly those who cheat the audience the players.  
BETTER: ... , particularly those players who cheat the audience.
- ... the trial of the outcome ...  
BETTER: ... the outcome of the trial ...

---

idea: add additional feature functions that 'explicitly' look at syntactic well-formedness of produced translation

syntactic framework: statistical parser for source and target language

feature functions depend also on parse trees:  $( \quad ) \rightarrow ( \quad e \quad f )$

feature functions: simple  $\rightarrow$  complicated

- verb present in both sentences
- verb takes same number of arguments
- ...
- relationships between phrases should be transferred
- ...
- projecting one tree: constituent reordering
- aligning tree pairs: elementary tree pairs

**1. Error Analysis / Feature Hunting**

- (a) step 1: detect systematic error by contrasting produced/oracle/reference translations**
- (b) step 2: develop feature function for that error**
- (c) step 3: perform disc. training with that feature**
- (d) step 4: evaluate system performance and keep feature if performance improves**
- (e) goto 1**

**2. Development of Feature Functions**

→ ...

**3. Diskriminative Training Techniques**

- **maximum entropy training: YASMET toolkit**
- **directly optimizing BLEU scores: opt-nbest**
- **SVM reranking**

**4. Search:**

- **-best rescoring**
- **minimum Bayes risk search with syntactically motivated loss functions**

- **Framework: Large Data Track Chinese–English → 150M words (per language)**
  1. **Parallel training data: Chinese treebank, FBIS, Xinhua News, Hongkong News, UN, Hongkong Hansards, Sinorama, Hongkong Laws**
  2. **Chinese treebank for parser training**
  3. **Development data: Evalset-01, Devset-03-1 Devset-03-2 → 175K words**
  4. **Test data: Evalset-02 → 25K words**
  5. **Blind test data: Evalset-03 → 25K words**
  6. **16384-best lists for development and test corpora produced by baseline alignment template system**

<b>-best size</b>	<b>1</b>	<b>1024</b>	<b>16384</b>
<b>BLEU[%]</b>	<b>32</b>	<b>45%</b>	<b>50%</b>

7. **Large parts of the data are tagged/chunked/parsed**
- **Ready-to-use tools**
    1. **Syntax tools: English/Chinese POS tagger, chunker, parser + conv. tools**
    2. **Discriminative training: maximum entropy training (YASMET), maximum BLEU training, maximum NIST training**
    3. **Feature functions: tree-to-tree/tree-to-string alignment models**

How do we define success?

We are successful, if

- we achieve significant improvement of BLEU/NIST scores
  - absolute improvement of BLEU: 1%: statistically significant
  - 1 – 2%: ok
  - 2 – 3%: good
  - 3 – 4%: very good
  - 4 – 6%: fantastic (more is unrealistic)
- potential problem: BLEU/NIST are probably not sensitive enough to measure improvements in syntactic wellformedness
- human evaluation judges week-5 system significantly better (with respect to syntax) than initial system
  - plan: subjective evaluation performed at Butler Hill Group or at LDC with five human evaluators for 900 sentences
- ( we show that explicit models of syntax are useless in MT )

- **week 1:**
  - **one-click-program for:**  
computing FF, training of parameters on development corpus, BLEU on test corpus  
goal: very quick development training/test cycle of under sixty minutes (for simple FF's)
  - **integrate tree-to-tree, tree-to-string alignment models**
  - **contrastive error analysis**
  - **plan development of new feature functions**
- **week 2/3/4/5/6:**
  - **error analysis / feature function development**
  - **end of week: evaluate progress on blind test corpus**
- **end of week 5:**
  - **build system for subjective evaluation**

**Summary:**

- **Starting point: best existing Chinese-English MT system: alignment template system from USC/ISI**
- **Contrastive error analysis**
- **Develop specific syntactic feature functions that try to 'fix' the errors**
- **Perform discriminative training of feature function weights**