Forced Derivations for Hierarchical Machine Translation

Stephan Peitz, Arne Mauser, Joern Wuebker and Hermann Ney
Human Language Technology and Pattern Recognition, RWTH Aachen University

RWTH AACHEN UNIVERSITY

Motivation

- heuristic hierarchical rule extraction causes two problems
  - translation probabilities depend on simple counts from a word alignment
  - large number of extracted rules
- employ forced derivation procedure on parallel training data
  - learn better rule probabilities with an EM-inspired algorithm
  - apply more consistent pruning regarding the translation process

Overview

- efficient framework to estimate translation probabilities
- perform an EM-inspired algorithm on parallel training data
  - expectation step: calculate expected counts for each applied rule
  - maximization step: update the translation probabilities
- during the forced derivation step
  - two-parse algorithm [Dyer, HLT-NAACL 2010]
  - inside-outside algorithm [Čmejrek et al., IWSLT 2009]
  - leave-one-out [Wuebker et al., ACL 2010]
  - log-linear combination of all features used in the translation process
- after the forced derivation procedure
  - threshold pruning to reduce rule set size using expected counts
- experimental results on following Europarl task from the WMT 2012
  - German→English
  - French→English
- open-source translation toolkit Jane [Wuebker et al., CoLing 2012]
  - `http://www.hltpr.rwth-aachen.de/jane`

Forced Derivation Step

- goal: calculate expected counts for each applied rule
- all possible synchronous derivations are needed
- two-parse algorithm reduces average run-time
- for a given sentence pair (f_1^J, e_1^I)
  - parse f_1^J, extract applied rules
  - annotate rules with the source span
  - parse e_1^I with annotated rules
  - perform inside-outside algorithm on target parse tree
  - calculate expected count using inside and outside probabilities
- expected counts for a rule are summed up over all sentence pairs

Rule Annotation

f_1^5 = Und zwar sollen derartige Strafen

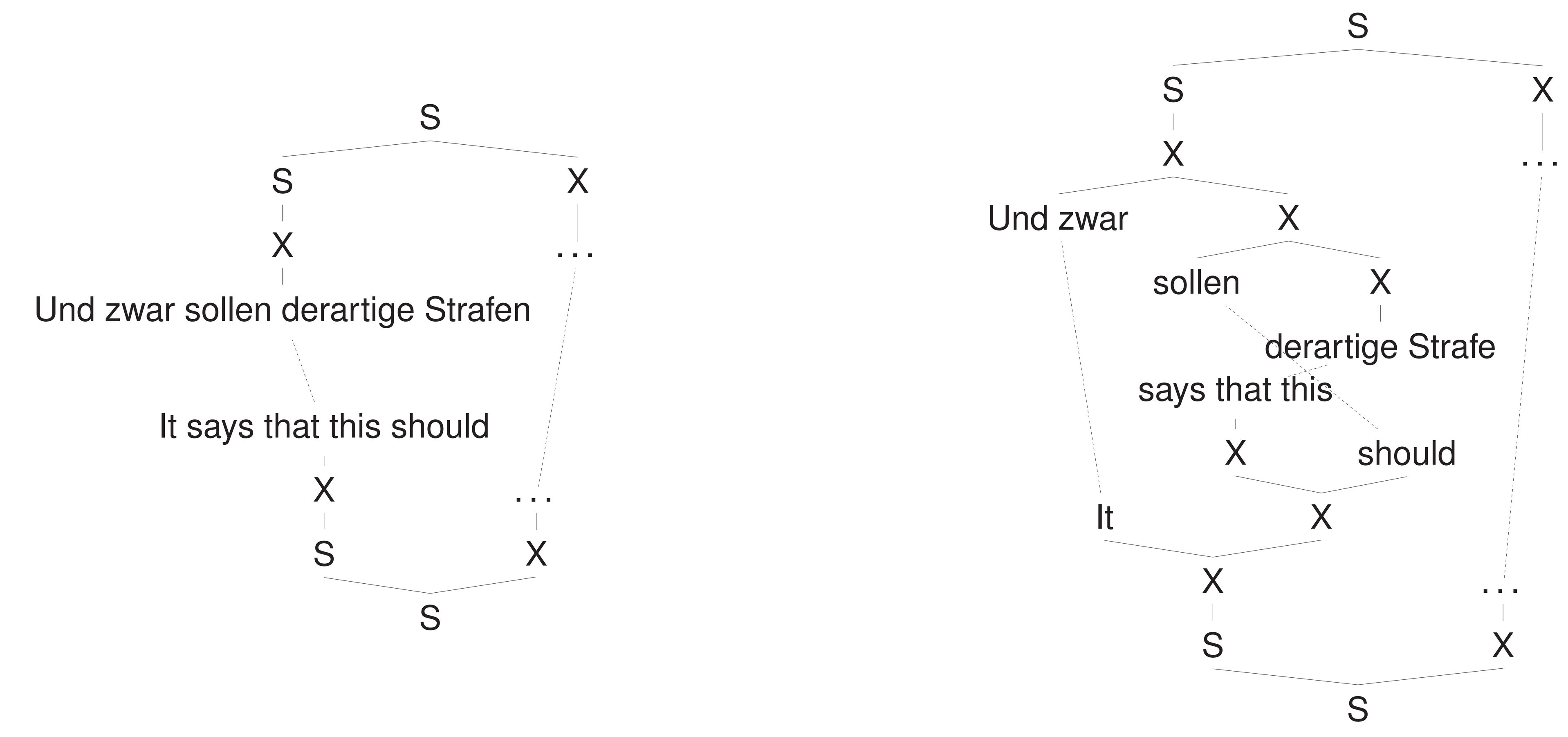
$$\Downarrow$$

$$X \rightarrow \langle \text{sollen } X, X \text{ should} \rangle$$

$$\Downarrow$$

$$X_3^5 \rightarrow \langle \text{sollen } X, X_4^5 \text{ should} \rangle$$

Leave-one-out



- Derivation example without and with leave-one-out

Experimental Results

- parallel training data: around 2M sentences
- initial rule set heuristically extracted
- parsing of 2000 sentences in 2.5 hours on a single machine (on average)
- preliminary experiments on the development set of the German→English task

	dev BLEU	avg. # applied glue rules /sent.
without l1o	20.3	0.7
length-based l1o	**21.0**	**5.7**
baseline	20.8	3.4

cutoff threshold	dev BLEU	% of full rule set
0.2	21.0	3.2
0.15	21.4	3.9
0.1	**21.4**	**4.9**
0.01	21.2	13.2
0.001	21.1	23.4
full	21.0	92.0

- in addition: log-linear interpolation
  - intersect learned rule set with initial rule set
  - interpolation weight ω was adjusted on the development set
- reduction of the rule set size by more than 95%
- improvements on the test set of the German→English and French→English tasks

setup	German→English		French→English	
	BLEU	TER	BLEU	TER
baseline	19.1	63.4	24.6	57.2
forced derivation +*l1o* +*cutoff*	19.5	63.1	25.0	57.2
interpolation $\omega = 0.2$	**19.8**	**62.6**	**25.6**	**56.3**