The HDU Discriminative SMT System for Constrained Data PatentMT at NTCIR10

Institute for Computational Linguistics, Heidelberg University, Germany
Patrick Simianer, Gesa Stupperich, Laura Jehl, Katharina Wäschle, Artem Sokolov, Stefan Riezler
{simianer,stupperich,jehl,waeschle,sokolov,riezler}@cl.uni-heidelberg.de

Patents are easy to translate, they contain lots of repetitive and formulaic text ⇒ train a model of **sparse lexicalized features** on a large data set using **multi-task learning**; incorporate ℓ_1/ℓ_2 **regularization** to find most important features

Sparse, lexicalized features attached to SCFG rules

(1) $X \rightarrow$ X_1 要件 の X_2 | X_2 of X_1 requirements
(2) $X \rightarrow$ この とき 、 X_1 は | this time , the X_1 is
(3) $X \rightarrow$ テキスト メモリ 41 に X_1 | X_1 in the text memory 41

Rule identifiers: unique rule identifier

Rule *n*-grams: bigrams in source and target side of a rule, e.g. `of` X_1, X_1 `requirements`

Rule shape: 39 patterns identifying location of sequences of terminal and non-terminal symbols, e.g. (for rule (1))

`NT, term*, NT | NT, term*, NT, term*`

There is a very large number of potential features (≫ than the number of rules in the grammar)

Pairwise-ranking model

$$\begin{aligned} g(x_1) > g(x_2) &\Leftrightarrow f(x_1) > f(x_2) \\ &\Leftrightarrow f(x_1) - f(x_2) > 0 \\ &\Leftrightarrow w \cdot x_1 - w \cdot x_2 > 0 \quad (1) \\ &\Leftrightarrow w \cdot \underbrace{(x_1 - x_2)}_{=\bar{x}_i} > 0 \end{aligned}$$

$x_{1,2}$ feature representations of translations
$g(\cdot)$ (per-sentence) BLEU score
$f(\cdot)$ model score of the decoder
w weight vector (model/decoder)
$x \cdot y$ vector dot product

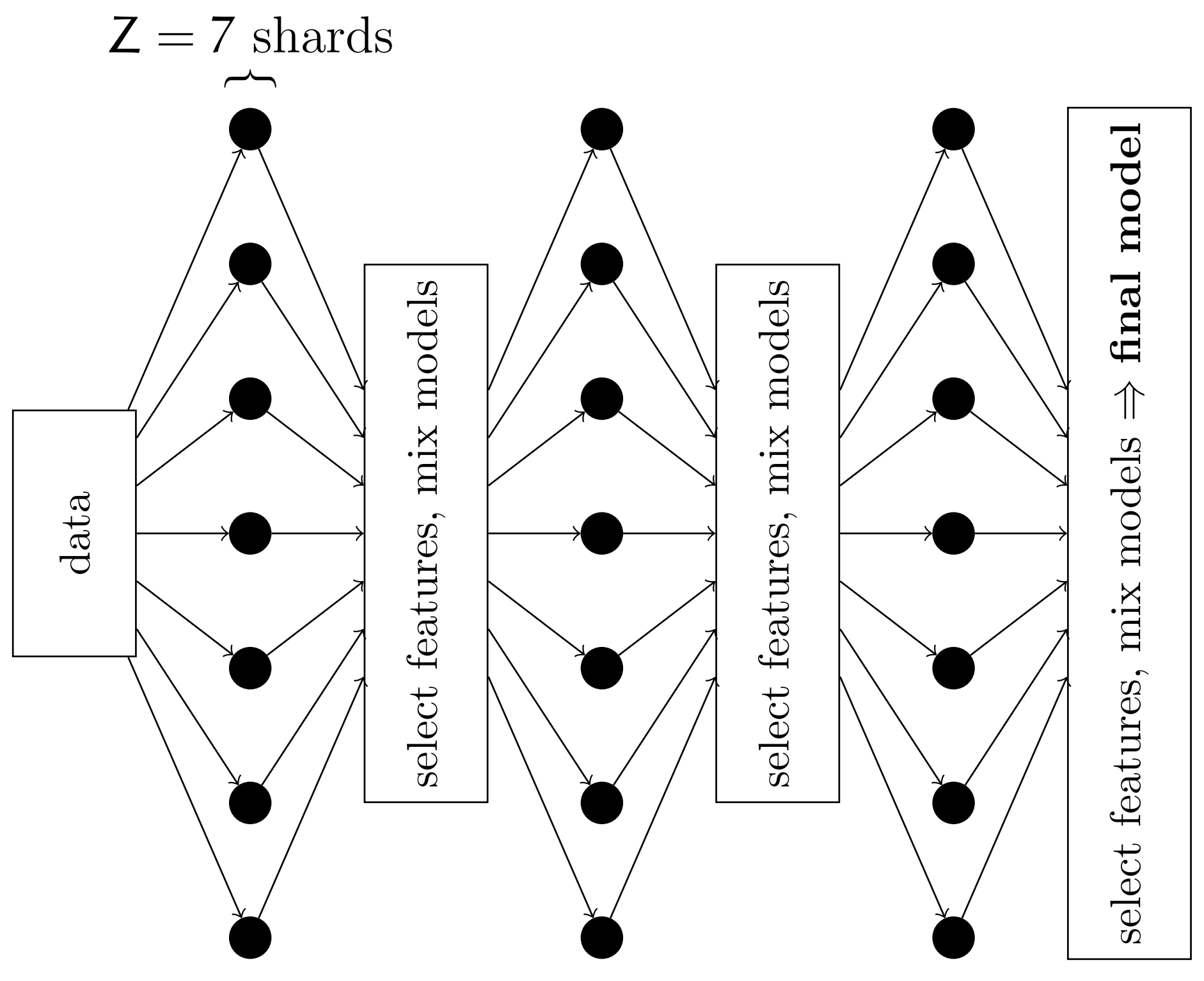
Hinge loss for a stochastic pairwise-ranking perceptron

$$L_i(w) = \max(0, -w \cdot \bar{x}_i) \quad (2)$$

$$\nabla L_i = \begin{cases} -\bar{x}_i \text{ if } w \cdot \bar{x}_i \leqslant 0, \\ 0 \qquad \text{otherwise.} \end{cases} \quad (3)$$

Gold standard ranking: BLEU+1 scores of translations of *k*best lists

Multi-task learning, ℓ_1/ℓ_2 regularization and parallelization


Z = 7 shards
data
select features, mix models
select features, mix models
select features, mix models ⇒ final model

(a) Parallelization strategy

	f_1	f_2	f_3	f_4	f_5
w_1	[5	4	3	4	0]
w_2	[2	0	4	1	1]
w_3	[4	0	0	3	0]
ℓ_2 norms	9	4	5	5	1
sort	f_1	f_3	f_4	f_2	f_5
select K = 3	f_1	f_3	f_4		
mix	11/3	7/3	7/3		

(b) Feature selection

Figure 1 : Multi-task learning algorithm

- Randomly split data into **Z** shards
- Select top **K** feature columns that have highest ℓ_2 norm over shards (or equivalently, by setting a threshold λ)
- Average weights of selected features over shards
- Resend reduced weight vector to shards for new epoch

devtest results	**tuning set**			
tuning method	*dev1*	*dev2*	*dev3*	*dev1,2,3*
baseline (MERT)	27.85	27.63	27.6	27.76
single dev, dense features	27.83	–	–	–
single dev, sparse features	28.84	28.08	28.71	29.03
multi-task, sparse features	–	–	–	28.92

Preprocessing (JP-EN only)

- JP: Full-width-latin characters converted to their standard UTF-8 equivalents
- JP: **Katakana term splitting** (RWTH NTCIR9) w/ compound splitter (Koehn/Knight, 2003)
- EN: Customized tokenizer (avoid splitting of `FIG.` or `PAT.` ...)
- both: **Consistent tokenization** (BBN NTCIR9): training data aligned using regular expressions; for test/dev sources applied the most common variants

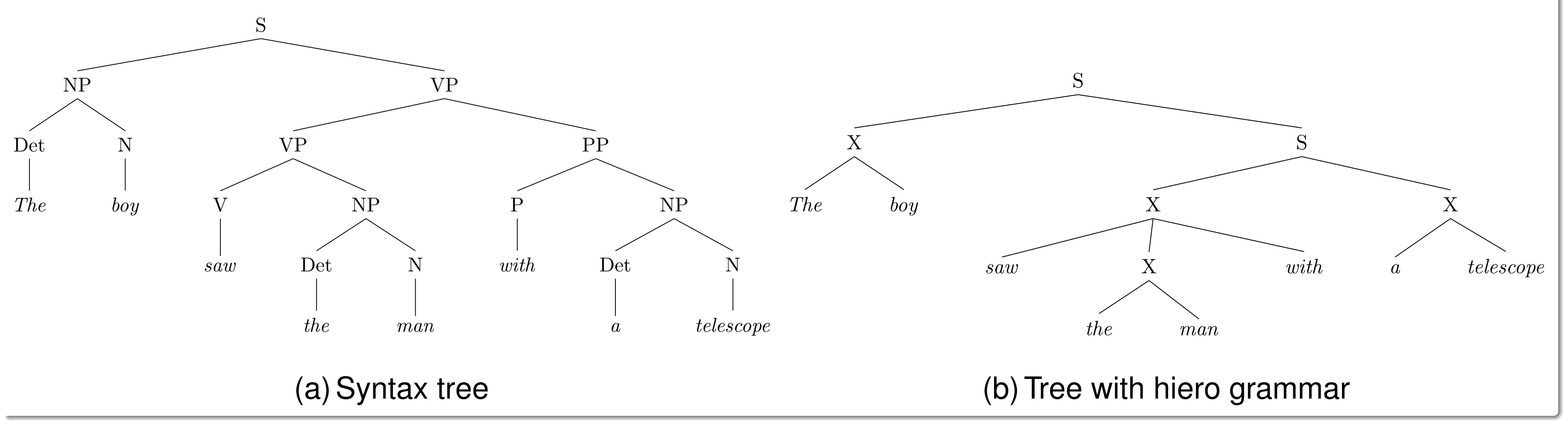
SMT setup: `cdec` SCFG decoder (Dyer, 2010); Hiero grammars (2 non-terminals max., ...) built w/ impl. of the suffix array extraction technique of (Lopez, 2007); 5gram modified Kneser-Ney smoothed LM built w/ SRILM; lowercased models; high values for cube pruning pop limit (500) and span size limit (100) at test time; Chinese segmentation w/ Stanford Segmenter, Japanese w/ MeCab; parses w/ Stanford Parser; English tokenizing/recasing/truecasing w/ `moses tools`

Patents are hard to translate, long sentences and an unusual jargon are common ⇒ enable **soft-syntactic constraints** in a SCFG/Hiero model to deal with long distance dependencies

Parsematch rescoring feature (Vilar et al, 2010)

- Introduce a quantity, $m(i,j)$ which records the distance (penalized exponentially) of a span i,j to its closest syntactic label
- For matching we only consider single sentence pairs (original work used all data)

No improvements on dev: 34.06 (baseline) → 34.07



(a) Syntax tree (b) Tree with hiero grammar

Marton & Resnik's (2008) soft-syntactic constraints

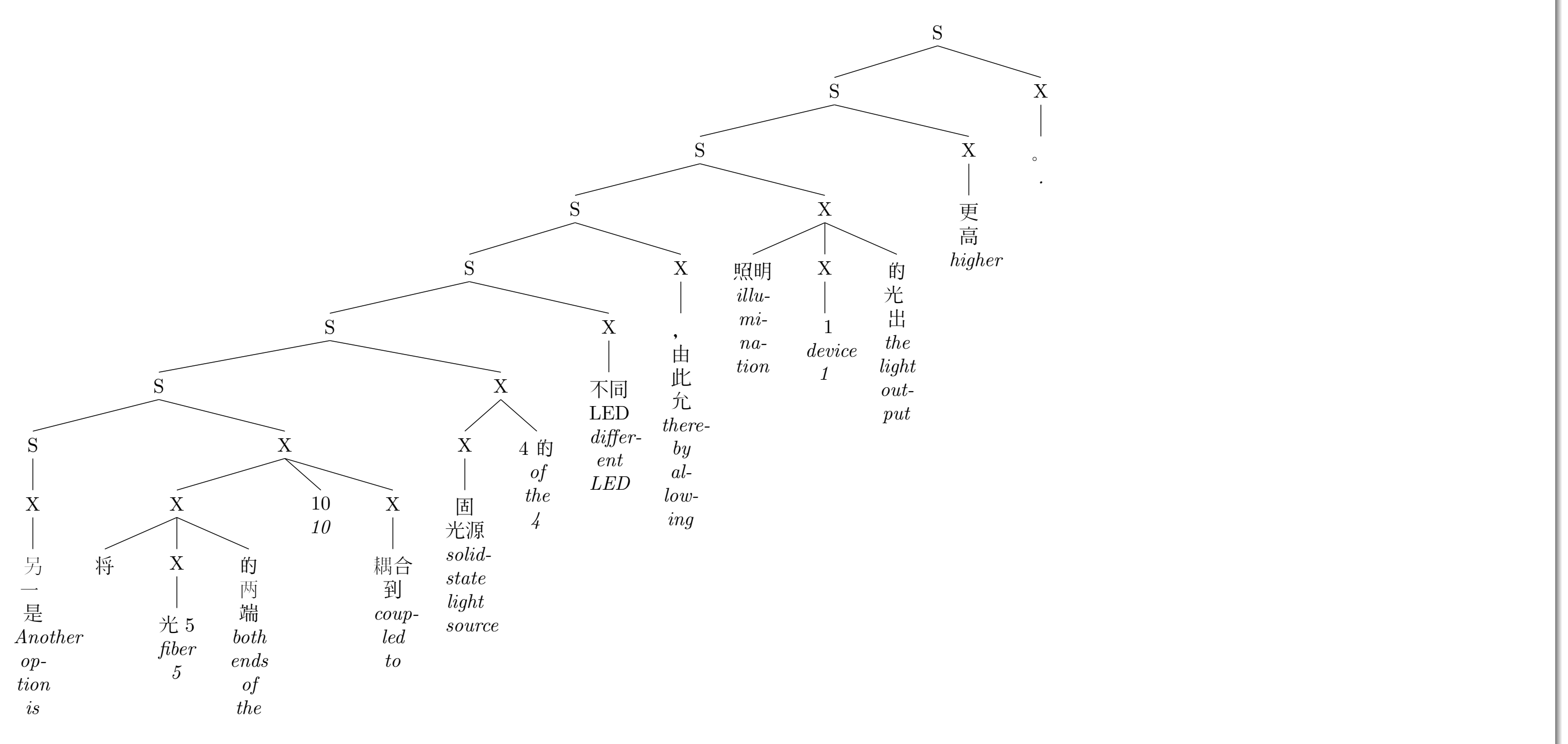
$\{$ADJP,ADVP,CP,DNP,IP,LCP,NP,PP,QP,VP$\} \times \{$=,+$\}$

- Indicate if spans in decoder derivations **match =** or **cross +** constituents of syntactic trees
- In contrast to (Chiang, 2005) these features do include the actual phrase labels
- Weights may be tied (marker: '2') or set independently (marker: '_')
- *IP2 VP2 NP_* (5 features, NP tied, IP/VP independent); *XP2* (20 features)

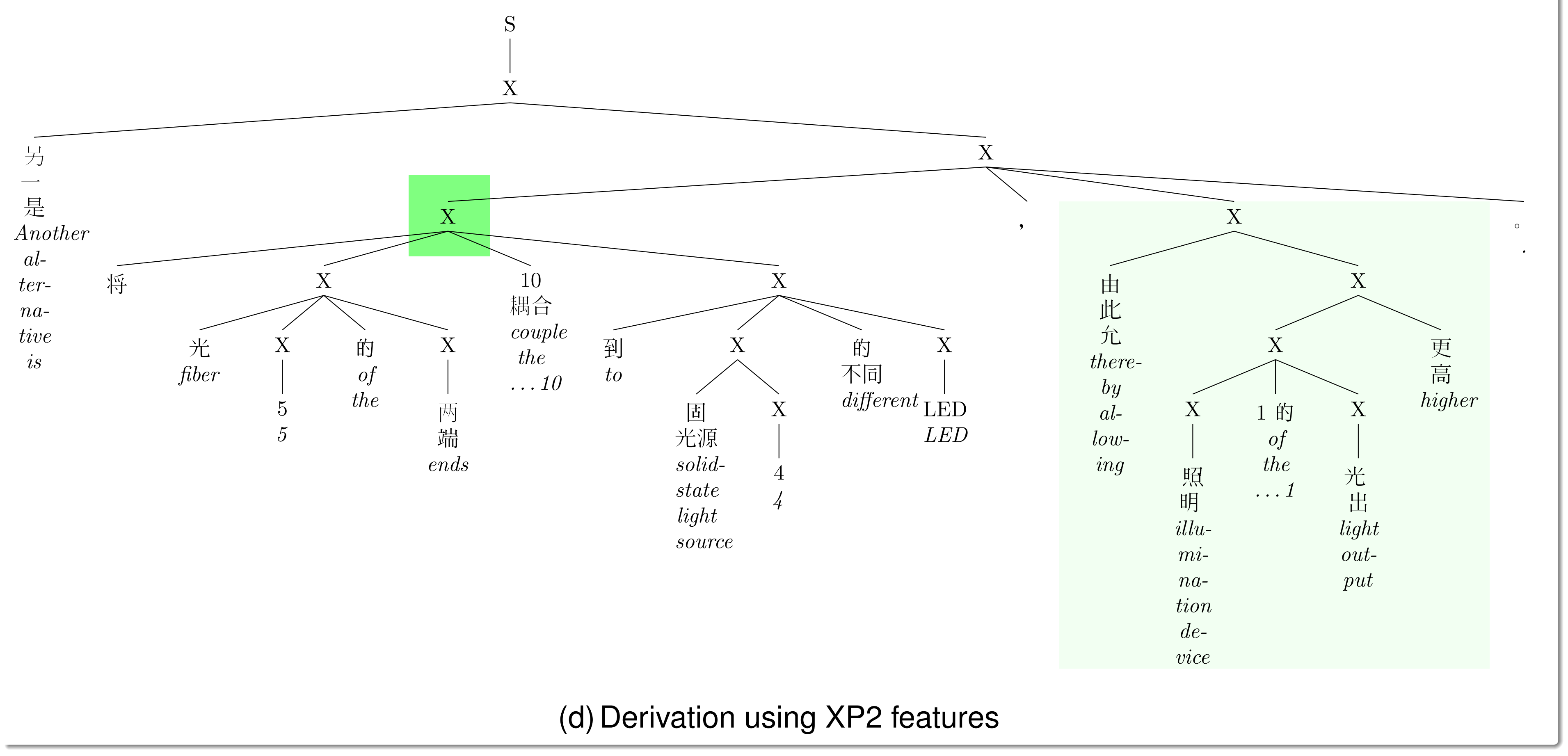
Results on dev: 34.06 (baseline) → 34.57 (*IP2 VP2 NP_*) → 34.84 (*XP2*)

Effects of soft-syntactic constraints

baseline	Another option is coupled to both ends of ..., thereby allowing ...
XP2	Another alternative is to couple the ends of ..., thereby allowing ...
reference	A further option is to optically couple both ends 10 of ..., thus allowing ...



(c) Baseline derivation



(d) Derivation using XP2 features

Systems & results:

Constrained setup for both JP-EN and ZH-EN subtasks: using only provided parallel data

Japanese-to-English subtask

HDU-1 Multi-task training with sparse features combining all four available dev sets
HDU-2 Identical to HDU-1 but training stopped early

Rank #5 and #6 in terms of **BLEU** on the *IE* test set (#2/#3 considering constrained systems), **#8 IE adequacy**, **#6 IE acceptability**

Chinese-to-English subtask

HDU-1 Marton & Resnik's soft-syntactic features (*XP2* configuration), tuned w/ single dev set
HDU-2 System as JP-EN with sparse rule features, but model learned on a single dev set

Rank #9 and #10 in terms of BLEU on *IE* test set (constrained #3/#4), **#4 IE adequacy**, **#4 IE acceptability**