Analyzing the Potential of Source Sentence Reordering in Statistical Machine Translation

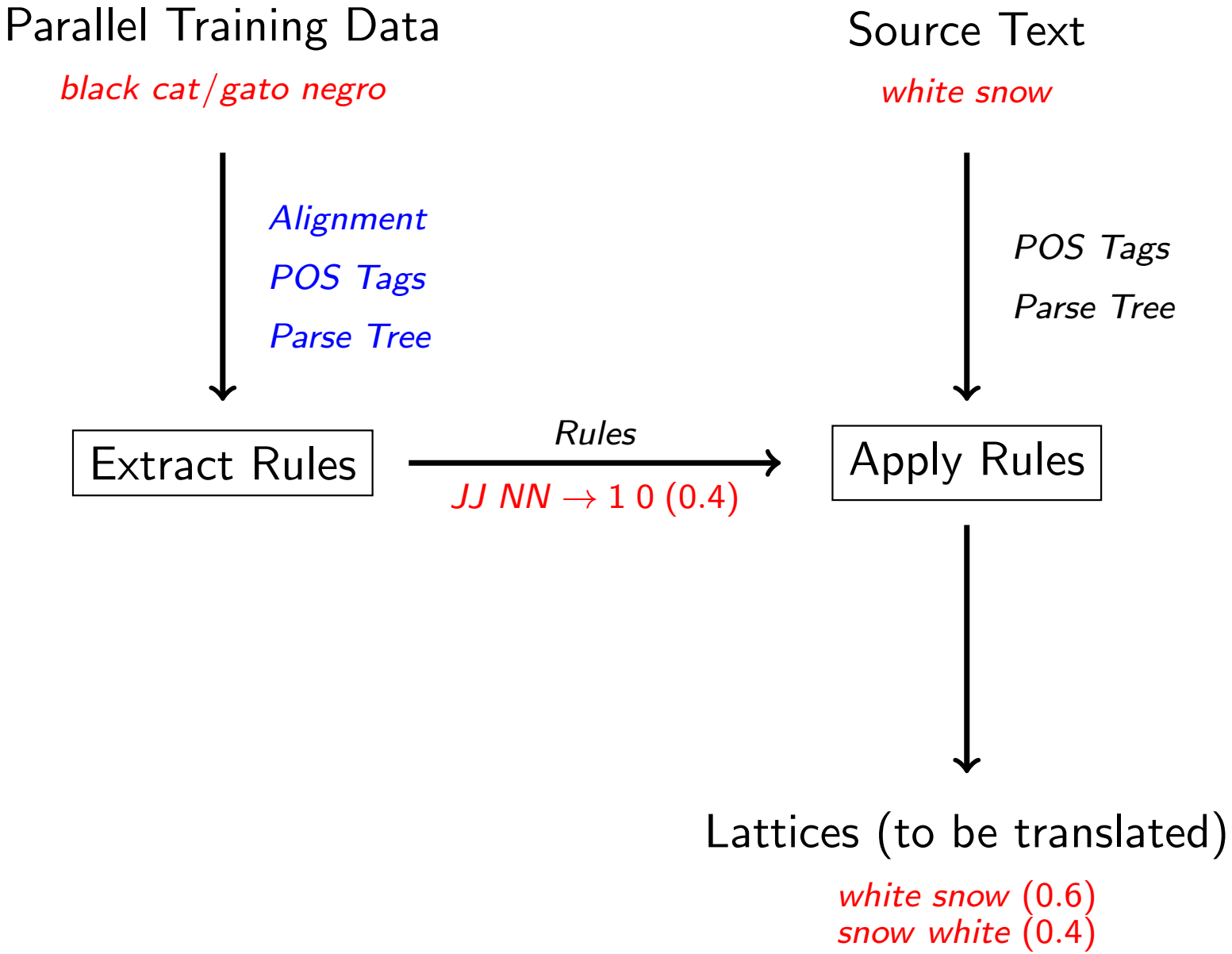
Ge Wu

May 13, 2014

Illustration

English	The black cat climbed to the tree top.
Spanish	The cat black climbed to the top tree.
Japanese	The black cat the tree top to climbed.

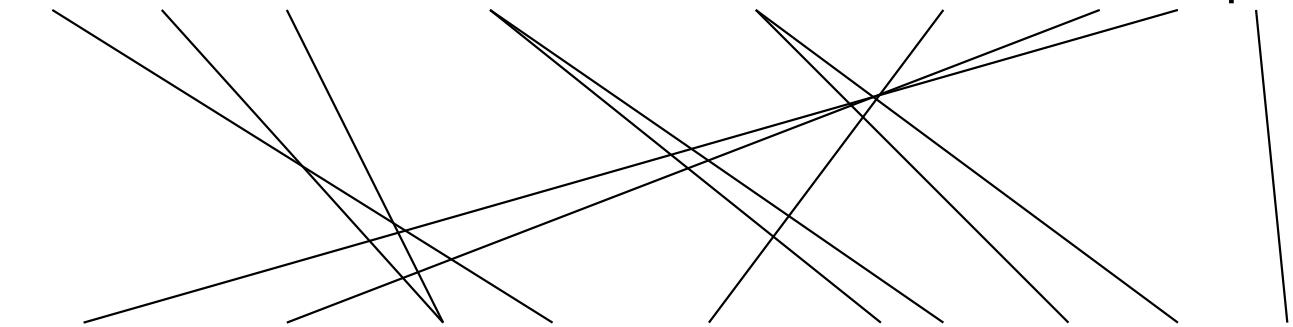
Illustration



Alignment

1 2 3 4 5 6 7 8 9 10 11 12

On the 24th , a series of attacks occurred in Iraq .



伊拉克 境内 二十四 号 发生 了 多 起 袭击 事件 。

1 2 3 4 5 6 7 8 9 10 11

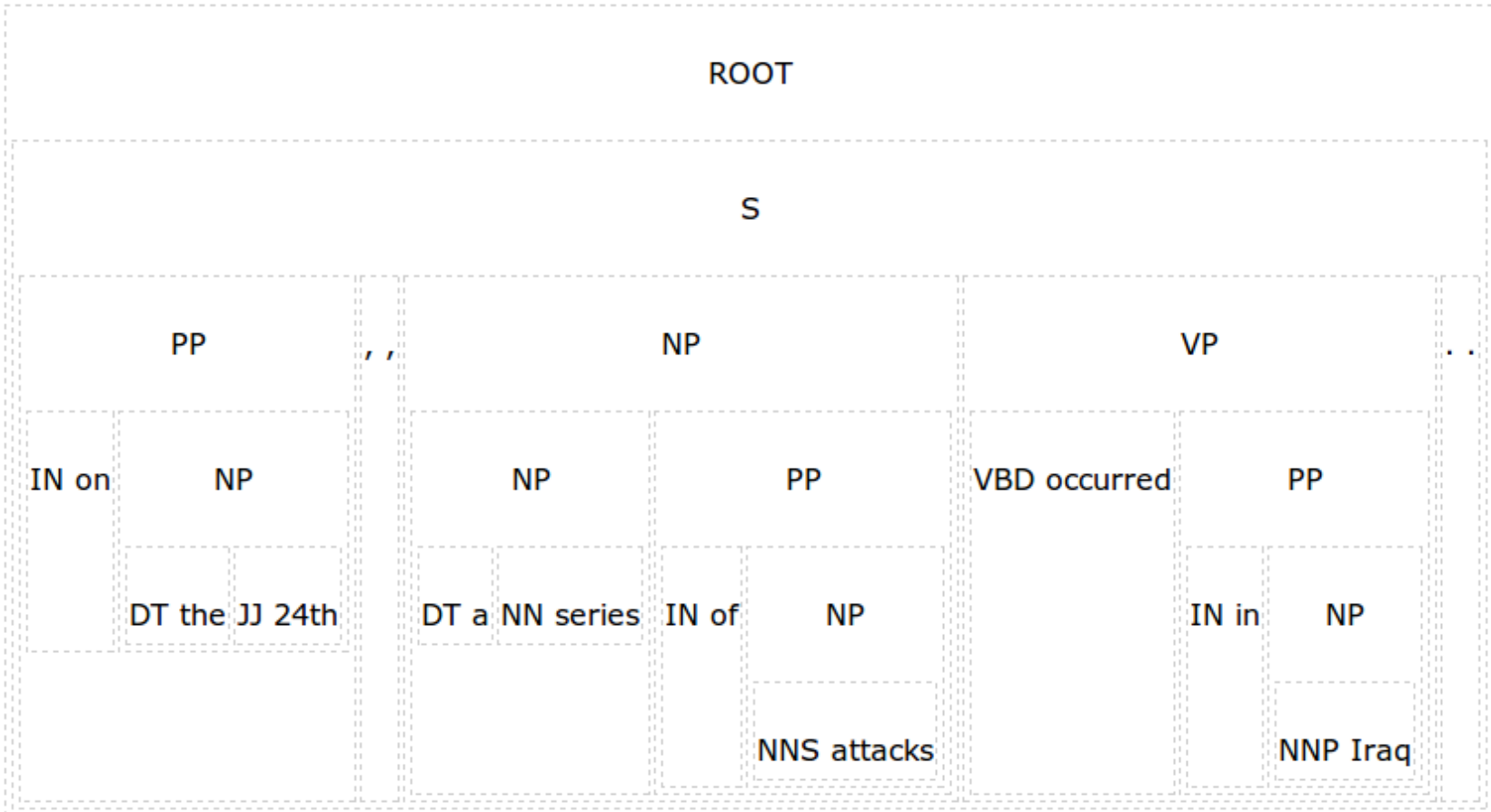
1-4 2-3 3-3 6-7 6-8 8-9 8-10 9-5 10-2 11-1 12-11

POS Tags

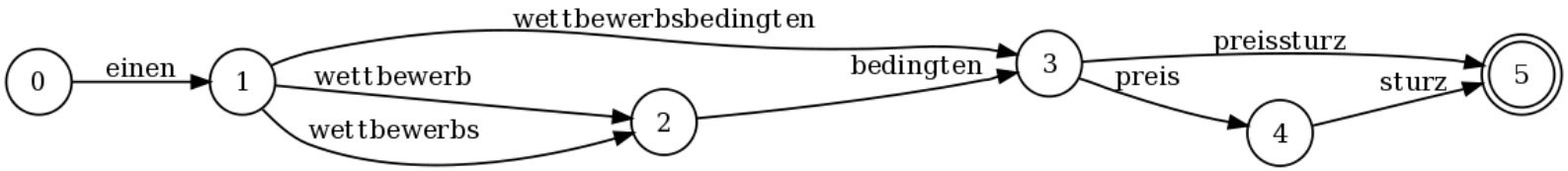
On the 24th , a series of attacks occurred in Iraq .

```
IN DT  JJ , DT NN IN  NNS    VVN IN NP SENT
```

Parse Tree


ROOT
S
PP
, ,
NP
VP
. .
IN on
NP
NP
PP
VBD occurred
PP
DT the
JJ 24th
DT a
NN series
IN of
NP
IN in
NP
NNS attacks
NNP Iraq

Lattices


0
einen
1
wettbewerb
wettbewerbs
2
wettbewerbsbedingten
bedingten
3
preis
preissturz
sturz
4
5

Rule Types

- Short Rules

  JJ NN $\rightarrow$ 1 0 — 0.573
- Long Rules

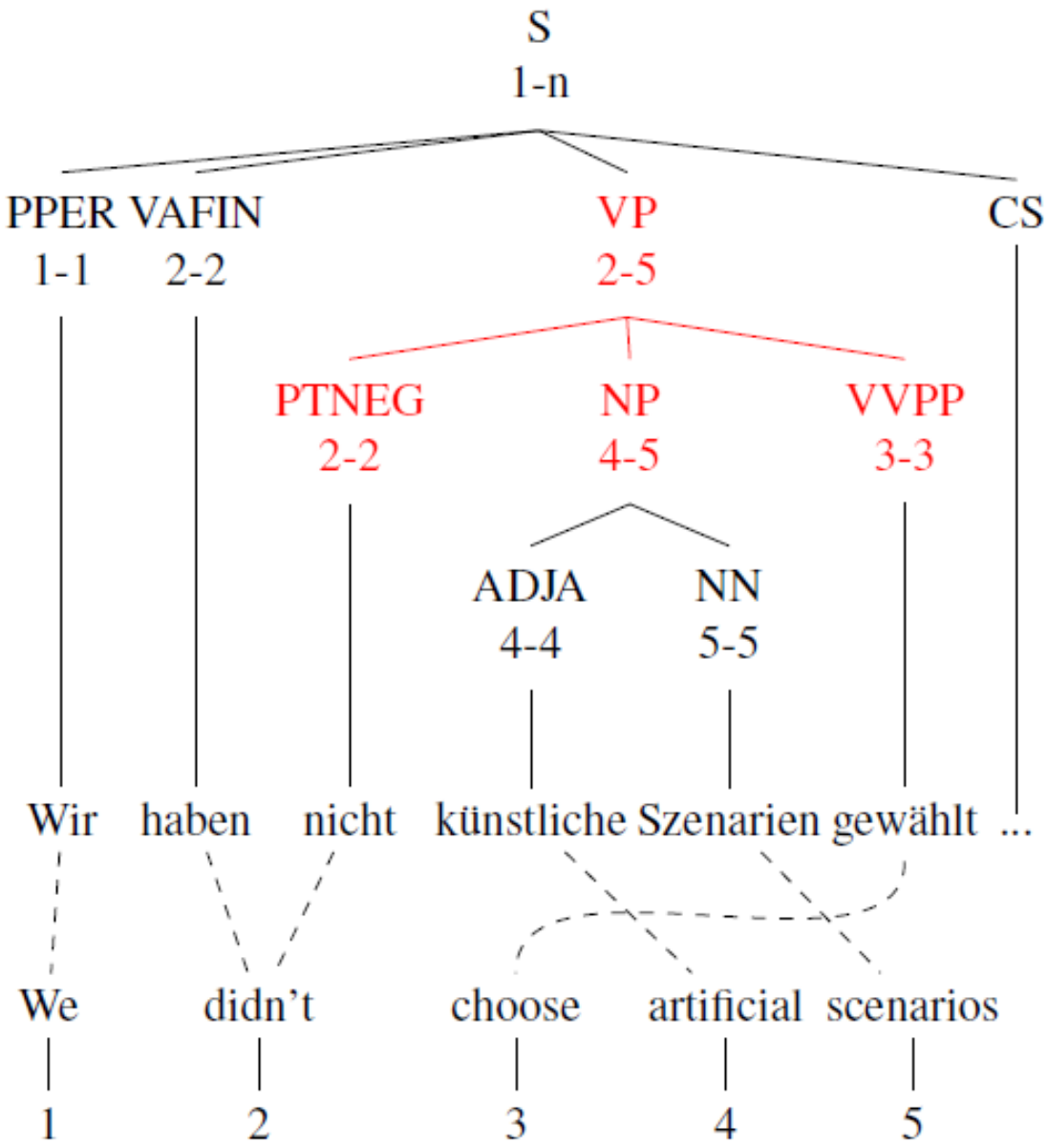
  NN ADV * VAFIN $\rightarrow$ 0 3 1 2 — 0.18
- Tree Rules

  NP (ADJP NNS) $\rightarrow$ 1 0 — 0.103279
- Multilayer Tree Rules

  ADJP (JJ PP (IN NP)) $\rightarrow$ 0 2 1 — 0.02988506

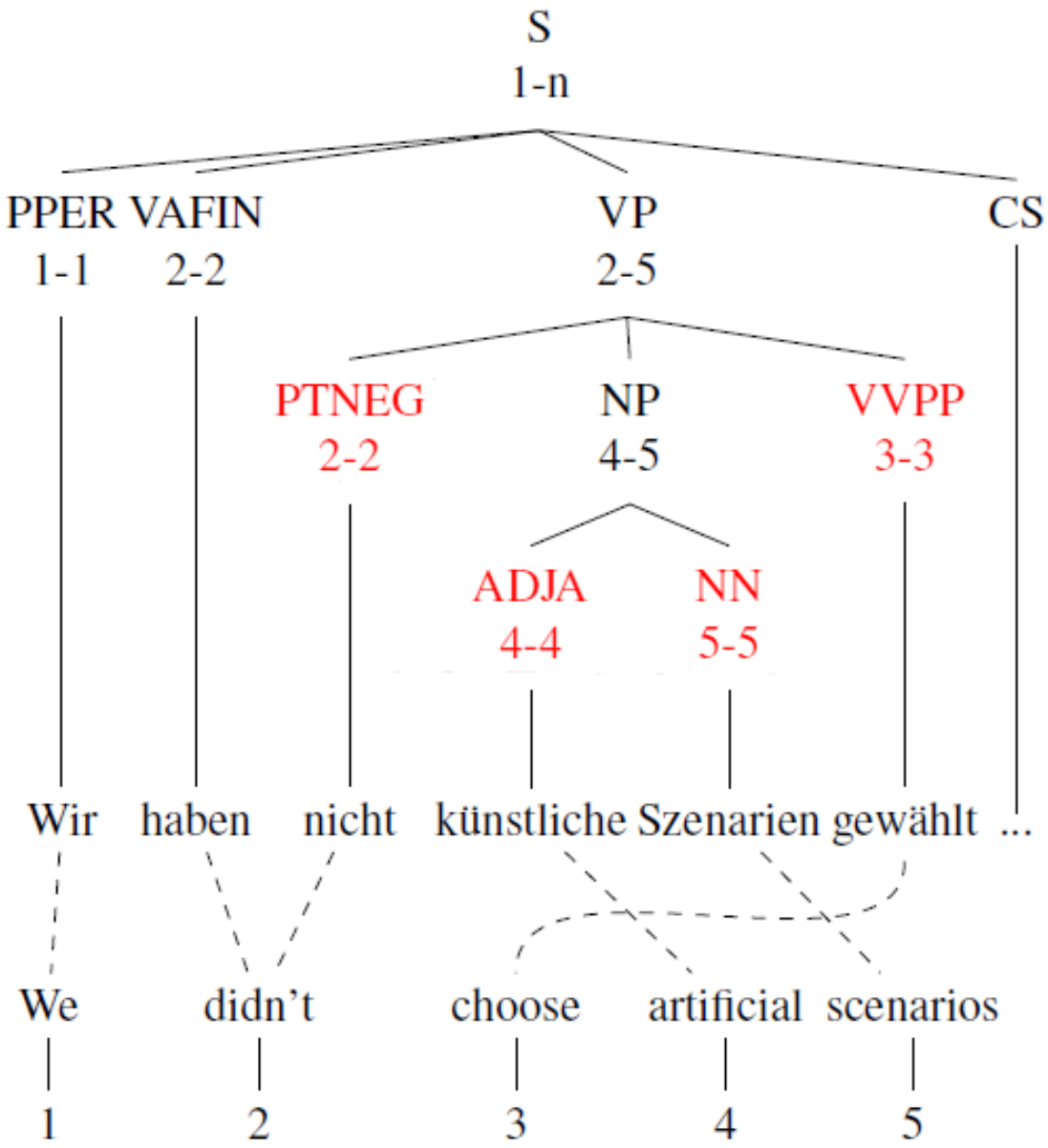
Extracting Rules

VP (PTNEG NP VVPP) → 0 2 1



Extracting Rules

VP (PTNEG NP (ADJA NN) VVPP) → 0 3 1 2



Evaluation

System	Rule #	BLEU Score
Baseline		8.09
Short Rules	95584	8.52
Long Rules	12306	8.57
Tree Rules	1573	8.79
Combined	109456	8.93
Multilayer Tree Rules (layers = 2)	4029	9.04

Training Data: English $\rightarrow$ Chinese, 75873 lines, 12.4MB

Vector Representation as Feature

Corpus

↓ word2vec (Neural Network)

```
...
spain     -0.274035 -0.153813 -0.073527 -0.080498 ...
larger    -0.000062  0.030795 -0.152307 -0.286478 ...
products -0.348087 -0.112983  0.120410 -0.176838 ...
parties   -0.259261 -0.040402 -0.047077 -0.312133 ...
night      0.096195  0.019403  0.063992  0.248290 ...
...
```

Vector Representation as Feature

```
Enter word or sentence (EXIT to break): translation

                  Word              Cosine distance
------------------------------------------------------------
          translations                    0.652776
                 bible                    0.620422
            translated                    0.591767
                  text                    0.564570
            septuagint                    0.559373
            dictionary                    0.557046
               tyndale                    0.546539
               vulgate                    0.542804
            translator                    0.542666
             apocrypha                    0.537319
           translators                    0.526699
                   ...                    ...
```