Enriching SCFG Rules Directly from Efficient Bilingual Chart Parsing

Martin Čmejrek, Bowen Zhou, Bing Xiang
{martin.cmejrek,bzhou,bxiang}@us.ibm.com
IBM, T. J. Watson Research Center, Yorktown Heights
Europarl: German-English, hierarchical MT

Why is the rule r5 missing from the model?
- Heuristic approaches to rule extraction (max phrase pair length, min span of non-terminals, etc.)

Let's define rule arithmetic and propose r5 := r6 + r7

We like r5 better for its contextual restriction.
Outline

- Baseline: Formally syntax-based system (ForSyn)
- Bilingual chart parsing with SCFG
- Estimating rule probabilities from the parse forest using EM
- Improving grammar coverage
- Proposing new rules with rule arithmetic
- Experiments and results
Enriching SCFG Rules Directly from Efficient Bilingual Chart Parsing

Selection of promising rules

Baseline Training word alignments + rule extraction

Hierarchical Rules

Sentence Pair

Bilingual Chart Parsing + EM

Select promising rules

Rule Arithmetics

Temporal Ruleset

Proposed Rules with Expected Counts

Filtering
Formally syntax-based system

- **Synchronous Context-Free Grammar (SCFG)**
  A synchronous rewriting system generating source and target pairs simultaneously
  \[ X \rightarrow \langle \gamma, \alpha, \sim \rangle \]
  only one nterm. \( X, \gamma, \alpha \) are term./nterm. strings, \( \sim \) is coindexation.

- (David Chiang. 2007. Hierarchical phrase-based translation)

- Decoder: ForSyn (Bowen Zhou et al., 2008. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels)

Parallel corpus

<table>
<thead>
<tr>
<th>Alignments</th>
<th>wiederaufnahme der sitzungsperiode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>resumption of the session</td>
</tr>
</tbody>
</table>

**Phrase pairs**

- \( X \sqsubset \langle \text{wiederaufnahme, resumption} \rangle \)
- \( X \sqsubset \langle \text{der, of the} \rangle \)
- \( X \sqsubset \langle \text{sitzungsperiode, session} \rangle \)
- …

**Abstract rules**

- \( X \sqsubset \langle \text{wiederaufnahme } X_1, \text{ resumption } X_1 \rangle \)
- \( X \sqsubset \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle \)
- …

**Glue**

- \( X \sqsubset \langle X_1 X_2, X_1 X_2 \rangle \)
Modeling

- **Baseline: a log-linear framework including 9 features**
  - conditional rule probabilities in both directions: $P(\alpha|\gamma)$ and $P(\gamma|\alpha)$
  - lexical weights in both directions: $P_w(\alpha|\gamma)$ and $P_w(\gamma|\alpha)$
  - other 5 features: LM, word bonus, abstraction, rule, and glue penalties

Decoding implemented as searching for the optimal derivation

$$P(D) \propto P_{LM}(e)^{\lambda_{LM}} \times \prod_i \prod_{X \rightarrow \langle \gamma, \alpha \rangle \in D} \phi_i(X \rightarrow \langle \gamma, \alpha \rangle)^{\lambda_i},$$
... Modeling

- **Problem with** $P(\alpha|\gamma)$ and $P(\gamma|\alpha)$
  - phrasal rules features can be estimated by maximum likelihood
  - abstract rules counts are not so reliable, eg:
    - $X \triangleleft <\text{wiederaufnahme X1, resumption X1}>$
    - was generated from both
    - wiederaufnahme der sitzungsperiode – resumption of the session
    - and
    - wiederaufnahme der – resumption of the

- **Solution: Use EM to estimate SCFG rule probabilities!**
  - Use Bilingual Chart Parsing to obtain parse forest
  - Compute inside, outside probabilities, and expected counts
  - Re-estimate joint probabilities $P(\alpha,\gamma)$, then compute $P(\alpha|\gamma)$ and $P(\gamma|\alpha)$
  - S. Huang and B. Zhou, “An EM algorithm for SCFG…”, ICASSP’09
Baseline Training word alignments + rule extraction → Hierarchical Rules → Bilingual Chart Parsing + EM → Select promising rules → Rule Arithmetics

Sentence Pair

Filtering → Proposed Rules with Expected Counts → Temporary Ruleset

Proposed Rules with Expected Counts
Bilingual Chart Parsing (CKY)

1: wiederaufnahme 2: der 3: sitzungsperiode
1: resumption 2: of 3: session

chart cell rule
[i, j, k, l]

\[
[1,1,1,1] \ X \triangleleft \ <\text{wiederaufnahme},\ \text{resumption}> \\
[3,3,4,4] \ X \triangleleft \ <\text{sitzungsperiode},\ \text{session}> \\
[2,2,2,3] \ X \triangleleft \ <\text{der},\ \text{of the}> \\
[1,2,1,3] \ X \triangleleft \ <\text{wiederaufnahme} \ X_1,\ \text{resumption} \ X_1> \ [2,2,2,3] \\
\quad X \triangleleft \ <X_1\ X_2,\ X_1\ X_2>\ [[1,1,1,1],[2,2,2,3]] \\
[2,3,2,4] \ X \triangleleft \ <X_1\ X_2,\ X_1\ X_2>\ [[2,2,2,3],[3,3,4,4]] \\
[1,3,1,4] \ X \triangleleft \ <\text{wiederaufnahme} \ X_1,\ \text{resumption} \ X_1>\ [2,3,2,4] \\
\quad X \triangleleft \ <X_1\ \text{der} \ X_2,\ \text{X} \ of \ the \ \text{X}_2>\ [[1,1,1,1],[3,3,4,4]] \\
\quad X \triangleleft \ <X_1\ X_2,\ X_1\ X_2>\ [[1,1,1,1],[2,3,2,4]],[[1,2,1,3],[3,3,4,4]]
\]

\[\text{Success!}\]
Computing inside probabilities

• Inside probability – probability of generating a parallel sequence from $X$

$$\beta_{ijkl}(X) = P(X \Rightarrow^* e_i^j; f_k^l)$$

• Can be defined recursively (there is only one non-terminal, so we can omit $X$)

$$\beta_{ijkl} = \sum_{r \in t_{ijkl}} P(r, r) \prod_{(i'j'k'l') \in r, bp} \beta_{i'j'k'l'}$$

• And computed dynamically while parsing...

$$\beta_{1,1,1,1} = P(X \triangleleft \langle \text{wiederaufnahme, resumption} \rangle)$$

$$\ldots$$

$$\beta_{1,3,1,4} = P(X \triangleleft \langle \text{wiederaufnahme } X_1, \text{ resumption } X_1 \rangle) \beta_{2,3,2,4}$$

$$+ P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) \beta_{1,1,1,1} \beta_{3,3,4,4}$$

$$+ P(X \triangleleft \langle X_1, X_2, X_1 X_2 \rangle) \beta_{1,1,1,1} \beta_{2,3,2,4}$$

$$+ P(X \triangleleft \langle X_1 X_2, X_1 X_2 \rangle) \beta_{1,2,1,3} \beta_{3,3,4,4}$$
Computing outside probabilities

- Outside probability – probability of generating the parallel sequence outside of $X$.

$$\alpha_{ijkl}(X) = P(S \Rightarrow \ast \epsilon_1^{i-1}, X, \epsilon_{j+1}^M; f_1^{k-1}, X, f_{l+1}^N)$$

- Can be computed by iterating the chart in top-down ordering, starting from the root cell:

$$\alpha_{1,M,1,N} := 1$$

- And propagating the probability mass to backpointed cells for rules with 1 non-terminal:

$$\alpha_{\rho, bp1} + = P(\rho, r)\alpha_{ijkl}$$

$$\alpha_{\rho, bp1} + = P(\rho, r)\alpha_{ijkl}\beta_{\rho, bp2}$$

$$\alpha_{\rho, bp2} + = P(\rho, r)\alpha_{ijkl}\beta_{\rho, bp1}$$

So that in the root cell $[1,3,1,4]$ the updates would look like:

$$\alpha_{2,3,2,4} + = P(X \triangleleft \langle \text{wiederaufnahme } X_1, \text{resumption } X_1 \rangle) \ast 1$$

$$\alpha_{1,1,1,1} + = P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) \ast 1 \ast \beta_{3,3,4,4}$$

$$\alpha_{3,3,4,4} + = P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) \ast 1 \ast \beta_{1,1,1,1}$$

$$\alpha_{1,1,1,1} + = P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ of the } X_2 \rangle) \ast 1 \ast \beta_{2,3,2,4}$$

$$\alpha_{2,3,2,4} + = P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ der } X_2 \rangle) \ast 1 \ast \beta_{1,1,1,1}$$

$$\alpha_{1,2,1,3} + = P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ der } X_2 \rangle) \ast 1 \ast \beta_{3,3,4,4}$$

$$\alpha_{3,3,4,4} + = P(X \triangleleft \langle X_1 \text{ der } X_2, X_1 \text{ der } X_2 \rangle) \ast 1 \ast \beta_{1,2,1,3}$$
Computing expected counts

- Contributions to rule expected counts can be computed by iterating the chart in any ordering

\[ c(\rho, r) = \frac{P(\rho, r)\alpha_{ijkl} \prod_{i=1}^{\beta_{n,1}} \beta_{\rho, bp_i}}{\beta_{1, M, 1, N}}, \]

- Collect counts from the whole training corpus.

M-step

- Finally, joint probabilities are re-estimated as

\[ P(r) = \frac{c(r)}{\sum_{r' \in R : L(r') = L(r)} c(r')} \]

L(r) means the left-hand side of the rule r. (Always X, trivial)

- Conditional probabilities \( P(\alpha|\gamma) \) and \( P(\gamma|\alpha) \) are computed as normalized probabilities of rules with the same source and target side, respectively.

\( P(r) \) for scoring

- The joint probability of a rule is normalized for length by

\[ c_{size}(s)^{s-1} \]

- Combination of all rule features is used (lexical weights are important too)
Improving the grammar coverage

- Many sentences cannot be parsed
  - On Europarl data, 70% for *union*, 20% for *grow-diag-final*
  - Structural complexity
    
    \[
    \text{ich dann erneut verweisen werde} \\
    \text{i will then raise again}
    \]

    Every parse needs one of the following rules:
    \[
    X \triangleleft \text{erneut } X_1, X_1 \text{ again} \\
    X \triangleleft \text{X}_1 \text{ verweisen , raise X}_1>
    \]

  - Low frequency words and asymmetric translation pairs

    \[
    \text{nach monatelangen und weltweiten konsultationen wird nun im donaldson …} \\
    \text{and after consulting world wide for many months the donaldson report …}
    \]

    As a consequence, either the whole sentence pair cannot be parsed (losing expected counts for other words), or another asymmetric rules (trying to fix weltweiten konsultationnen) are boosted

  - Add Swap Glue to increase parsability: \( X \triangleleft <X_1 X_2, X_2, X_1> \)
  - Insertion and deletion rules (ITG - like)
    \[
    X \triangleleft <X_1, X_1 e>, X \triangleleft <X_1, e X_1>, X \triangleleft <X_1 f, X_1>, X \triangleleft <f X_1, X_1>
    \]
Enriching SCFG Rules Directly from Efficient Bilingual Chart Parsing

Sentence Pair

Baseline Training word alignments + rule extraction

Hierarchical Rules

Filtering

Bilingual Chart Parsing + EM

Proposed Rules with Expected Counts

Temporary Ruleset

Select promising rules

Rule Arithmetics
Proposing new rules

- Parse the sentence pair, estimate expected counts

- Select the “most promising” rule usages
  - Currently selected as productions $\rho$ with the highest contributions to expected counts

- Use Rule arithmetic to combine rules
  1) create span projections – for both rules
  2) merge span projections
  3) collect rules

- Parse again, this time also using new proposed rules, estimate expected counts
die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

```
x \otimes <diese, the>
x \otimes <systeme, system>
x \otimes <diese systeme, the system>
```

“Rule arithmetic Hello world” Combining phrasal rules
The challenge is to make the system the very best.

Combining phrasal rules (contiguous projections required)
die herausforderung besteht darin **diese** **systeme** **zu** **den besten der welt** **zu machen**

the challenge is **to make** **the system** **the very best**

\[
X \vartriangleleft <zu machen, to make>
\]

\[
+ X \vartriangleleft <X_1X_2X_2X_1>
\]

\[
= X \vartriangleleft <X_1 zu machen, to make X_1>
\]

Combining phrasal rule and swap glue.
die herausforderung besteht darin diese systeme zu den besten der welt zu machen

the challenge is to make the system the very best

\[
X \triangleleft \langle \text{besteht darin, is} \rangle \\
+ \\
X \triangleleft \langle X_1 \text{ zu } X_2, \text{ to } X_2 X_1 \rangle \\
= \\
X \triangleleft \langle \text{besteht darin } X_1 \text{ zu } X_2, \text{ is to } X_2 X_1 \rangle
\]

Combining phrasal rule and abstract rule with 2 non-terminals.
die herausforderung besteht darin **diese systeme** zu den besten der welt zu machen

the challenge is **to make** the **system** the **very best**

\[
\begin{align*}
X \trianglelefteq & \ <\text{diese } X_1, \text{ the } X_1> \\
+ & \ <X_1 \text{ zu } X_2, \text{ to } X_2 X_1> \\
= & \ <\text{diese } X_1 \text{ zu } X_2, \text{ to } X_2 \text{ the } X_1>
\end{align*}
\]

Combining 2 abstract rules

Note that the span \( X_1 \) of the second rule was shortened
Experiment

- Presenting 3 sets of results (BLEU)
  - 11 Iterations of EM training
  - Proposing new rules after the 1st iteration (adding new rules to the baseline)
    - filtering (using only rules proposed at least from 2 sentence pairs)
  - + 1 iteration of EM with new rules
# Results (BLEU)

## German-English data from the Europarl

- 297k sentence pairs, lowercase, punctuation removed
- 1k dev set, 1 reference
- 1k test set, 1 reference
- 13M baseline rules
- 100k proposed rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>Ger-En dev</th>
<th>Ger-En test</th>
<th>Farsi-En dev</th>
<th>Farsi-En test</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>23.852</td>
<td>25.447</td>
<td>41.095</td>
<td>38.248</td>
</tr>
<tr>
<td>EM i0</td>
<td>24.394</td>
<td><strong>26.122</strong></td>
<td>40.764</td>
<td><strong>39.114</strong></td>
</tr>
<tr>
<td>i1</td>
<td>24.365</td>
<td>25.826</td>
<td>41.295</td>
<td>38.509</td>
</tr>
<tr>
<td>i2</td>
<td>24.433</td>
<td>25.936</td>
<td>41.424</td>
<td>38.238</td>
</tr>
<tr>
<td>i4</td>
<td>24.409</td>
<td>25.936</td>
<td>41.563</td>
<td>39.594</td>
</tr>
<tr>
<td>i5</td>
<td>24.300</td>
<td>26.259</td>
<td>41.557</td>
<td>39.242</td>
</tr>
<tr>
<td>i6</td>
<td>24.339</td>
<td>26.197</td>
<td>41.726</td>
<td>39.209</td>
</tr>
<tr>
<td>i7</td>
<td>23.985</td>
<td>25.827</td>
<td>41.542</td>
<td>39.445</td>
</tr>
<tr>
<td>i8</td>
<td>24.129</td>
<td><strong>26.305</strong></td>
<td>41.634</td>
<td>39.326</td>
</tr>
<tr>
<td>i9</td>
<td>23.988</td>
<td>25.940</td>
<td>41.422</td>
<td>39.583</td>
</tr>
<tr>
<td>i10</td>
<td>24.079</td>
<td>26.226</td>
<td>41.287</td>
<td><strong>39.650</strong></td>
</tr>
<tr>
<td>proposed i0</td>
<td>24.418</td>
<td><strong>26.122</strong></td>
<td>40.729</td>
<td>38.382</td>
</tr>
<tr>
<td>proposed i0 + EM</td>
<td>24.837</td>
<td><strong>26.408</strong></td>
<td>41.836</td>
<td><strong>40.246</strong></td>
</tr>
</tbody>
</table>

## Farsi-English conversational data

- 1.4k dev set, 1 reference
- 417 test set, 4 references
- 11.3M baseline rules
- 120k proposed rules
Conclusion

- Introduced algorithms for bilingual parsing, and estimation of rule probabilities
- Presented a new method for synthesizing new rules from the most confident rules within the parse forest
- The method is independent of bilingual word alignments and complementary to heuristic alignment-based approaches
- Presented results on conversational data from two different language pairs: +1 BLEU on German-English translations of Europarl data, +2.00 BLEU on Farsi.
Thank you
Example of rules proposed by rule arithmetic

1 ... um \( X_1 \), in order \( X_1 \)
natürlicher \( X_1 \), of course \( X_1 \)
deshalb \( X_1 \), this is why \( X_1 \)
\( X_1 \) zu koennen, to \( X_1 \)
\( X_1 \) ist, it is \( X_1 \)
nach der tagesordnung folgt die \( X_1 \), the next item is the \( X_1 \)
herr \( X_1 \) herr kommissar \( X_2 \), mr \( X_1 \) commissioner \( X_2 \)
die \( X_1 \) der \( X_2 \), \( X_1 \) the \( X_2 \)
im gegenteil \( X_1 \), on the contrary \( X_1 \)
nach der tagesordnung folgt \( X_1 \), the next item is \( X_1 \)
\( X_1 \) die \( X_2 \), the \( X_1 \) the \( X_2 \)
die \( X_1 \) die, the \( X_1 \)
ausserdem \( X_1 \), in addition \( X_1 \)
daher \( X_1 \), that is why \( X_1 \)
wir \( X_1 \) nicht \( X_2 \), we \( X_1 \) not \( X_2 \)
die \( X_1 \) der \( X_2 \), the \( X_2 \) \( X_1 \)
deshalb \( X_1 \), for this reason \( X_1 \)
um \( X_1 \) zu \( X_2 \), to \( X_2 \) \( X_1 \)
\( X_1 \) nicht \( X_2 \) werden, \( X_1 \) not be \( X_2 \)
Example of rules proposed by rule arithmetic

\[ X_1 \text{ nimmt } X_2 \text{ wird } X_1 \text{ zur } X_2 \]
\[ \text{parliament adopted the draft legislative} \]
\[ \text{it is very} \]
\[ \text{there } X_1 \text{ the} \]
\[ \text{the lisbon } X_1 \]
\[ \text{in this } X_1 \]
\[ \text{commissioner } X_1 \text{ would like to } X_2 \]
\[ \text{X}_1 \text{ approved the} \]
\[ \text{X}_1 \text{ cannot have} \]
\[ \text{for them} \]
\[ \text{also needs} \]
\[ \text{believe} \]
\[ \text{have proposed} \]
\[ \text{you } X_1 \]
\[ \text{they are} \]
\[ \text{voted in favour of these proposals} \]
\[ X_1 \text{ auditors} \]
Results – German 2 English

- **Data**
  - German-English data from the Europarl corpus.
    - 297k sentence pairs,
    - 1k dev set, 1 reference
    - 1k test set, 1 reference
    - 13M baseline rules
    - 100k proposed rules

<table>
<thead>
<tr>
<th></th>
<th>Ger-En dev</th>
<th>Ger-En test</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>23.852</td>
<td>25.447</td>
</tr>
<tr>
<td>i0</td>
<td>24.394</td>
<td>26.122</td>
</tr>
<tr>
<td>i1</td>
<td>24.365</td>
<td>25.826</td>
</tr>
<tr>
<td>i2</td>
<td>24.433</td>
<td>25.936</td>
</tr>
<tr>
<td>i3</td>
<td>24.375</td>
<td>26.047</td>
</tr>
<tr>
<td>i4</td>
<td>24.409</td>
<td>25.936</td>
</tr>
<tr>
<td>i5</td>
<td>24.300</td>
<td>26.259</td>
</tr>
<tr>
<td>i6</td>
<td>24.339</td>
<td>26.197</td>
</tr>
<tr>
<td>i7</td>
<td>23.985</td>
<td>25.827</td>
</tr>
<tr>
<td>i8</td>
<td>24.129</td>
<td>26.305</td>
</tr>
<tr>
<td>i9</td>
<td>23.988</td>
<td>25.940</td>
</tr>
<tr>
<td>i10</td>
<td>24.079</td>
<td>26.226</td>
</tr>
<tr>
<td>proposed i0</td>
<td>24.418</td>
<td>26.122</td>
</tr>
<tr>
<td>proposed i0 + EM</td>
<td>24.837</td>
<td>26.408</td>
</tr>
</tbody>
</table>
Results - Farsi

Data
- Farsi-English conversational data
  - 1439 dev set, 1 reference
  - 417 test set, 4 references

<table>
<thead>
<tr>
<th></th>
<th>Farsi-En dev</th>
<th>Farsi-En test</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>41.095</td>
<td>38.248</td>
</tr>
<tr>
<td>i0</td>
<td>40.764</td>
<td><strong>39.114</strong></td>
</tr>
<tr>
<td>i1</td>
<td>41.295</td>
<td>38.509</td>
</tr>
<tr>
<td>i2</td>
<td>41.424</td>
<td>38.238</td>
</tr>
<tr>
<td>i3</td>
<td>41.339</td>
<td>39.339</td>
</tr>
<tr>
<td>i4</td>
<td>41.563</td>
<td>39.594</td>
</tr>
<tr>
<td>i5</td>
<td>41.557</td>
<td>39.242</td>
</tr>
<tr>
<td>i6</td>
<td>41.726</td>
<td>39.209</td>
</tr>
<tr>
<td>i7</td>
<td>41.542</td>
<td>39.445</td>
</tr>
<tr>
<td>i8</td>
<td>41.634</td>
<td>39.326</td>
</tr>
<tr>
<td>i9</td>
<td>41.422</td>
<td>39.583</td>
</tr>
<tr>
<td>i10</td>
<td>41.287</td>
<td><strong>39.650</strong></td>
</tr>
<tr>
<td>proposed i0</td>
<td>40.729</td>
<td>38.382</td>
</tr>
<tr>
<td>proposed i0 + EM</td>
<td>41.836</td>
<td><strong>40.246</strong></td>
</tr>
</tbody>
</table>
Parsing efficiency

- The efficient implementation of $R_{Spans}(i,j,k,l)$ is very important.

- Create 2 prefix trees to encode all rule sides relevant to the sentence pair. And to store their spans, and spans of their non-terminals.

  $[1,2,1,3] \ X \triangleleft \ <\text{wiederaufnahme } X_1, \ \text{resumption } X_1> \ [2,2,2,3]

  $[1,3,1,4] \ X \triangleleft \ <\text{wiederaufnahme } X_1, \ \text{resumption } X_1> \ [2,3,2,4]

- The most time consuming is the glue rule, since there is no non-terminal between $X_1$ and $X_2$
Discussion, future directions

- **Speed**
  - ~5s/sentence
  - Pruning?
  - In literature, usually, parsing is done once, then 50 – 100 EM iterations, remembering the parse forest. Can we update the forest just for proposed rules?

- **Additional features, more linguistic information**

- **In monolingual parsing, rules are extracted from a treebank, then EM runs on a different data...**

- **Operation of subtraction?**
Acknowledgements

- **Bowen Zhu**
  - discussions of the BCP and rule arithmetics, ForSyn decoder

- **Bing Xiang**
  - discussions over the BCP

- **Songfang Huang**
  - initial EM implementation

- and to the whole S2S team!