Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation

Rico Sennrich

Institute for Language, Cognition and Computation
University of Edinburgh

September 19 2015
Problem: ungrammatical translation output

subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)

subcategorisation: *überraschen* is transitive
Problem: ungrammatical translation output

die Ergebnisse+pl der jüngsten Umfrage wird+sg für viele überraschen.

what’s wrong?

• subject-verb agreement: die Ergebnisse (pl) – wird (sg)
• subcategorisation: überraschen is transitive
Problem: ungrammatical translation output

what’s wrong?

- subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)
- subcategorisation: *überraschen* is transitive
Problem: ungrammatical translation output

what’s wrong?

- subject-verb agreement: *die Ergebnisse* (pl) – *wird* (sg)
- subcategorisation: *überraschen* is transitive

syntactic n-grams

- n-gram language models are sensitive to string distance
- dependency chains (rebranded **syntactic n-grams** [Sidorov et al., 2013]) are more robust
## Contribution

### Previous Work
- Large body of research on syntactic language models for SMT
  - Charniak et al., 2003, Och et al., 2004, Quirk et al., 2004, Post and Gildea, 2008, Cherry and Quirk, 2008, Shen et al., 2010
- Promising results with dependency language models

### Our Contribution
- Novel **relational** dependency language model
- Optimization of global SMT parameters on syntactic MT metric → Better appreciation of syntactic language models
Towards a relational dependency language model

### Previous Work

[Quirk et al., 2004, Shen et al., 2010]
- Unlabelled
- Varying degrees of word order modeling:
  - None [Quirk et al., 2004]
  - Heavy reliance on position [Shen et al., 2010]

### Our Model

- Relational: dependency labels as atomic elements
  - Use dependency labels as context
    - Verb must agree with subject, but not with object
  - Also predict dependency labels
    - Side-effect: models subcategorisation
- Sibling order is considered, but not relied on
Notation

- $S$: sequence of words
- $D$: sequence of dependency labels
- $T$: sequence of head positions (tree topology)

common approximation: $P(S) \approx P(S|T)$
Ergebnisse der jüngsten Umfrage wird für viele überraschen.
Dependency Language Model (DLM)

\[ P(S) = P(w_1, w_2, ..., w_n) \]
\[ \approx \prod_{i=1}^{n} P(w_i | h_s(i), h_a(i)) \]  
(1)

Markov assumption: use window of (closest) \( q \) siblings and \( r \) ancestors:

\[ P(S) \approx \prod_{i=1}^{n} P(w_i | h_s(i)^q, h_a(i)^r) \]  
(2)

\[ \text{sent} \]
\[ \text{vroot} \]
\[ \text{subj} \]
\[ \text{det} \quad \text{NN} \quad \text{gmod} \]
\[ \text{det} \quad \text{attr} \quad \text{NN} \]
\[ \text{det} \quad \text{ART} \quad \text{ADJA} \]
\[ \text{die} \quad \text{Ergebnisse} \quad \text{der} \quad \text{jüngsten} \quad \text{Umfrauge} \]
\[ \text{wird} \quad \text{für} \quad \text{viele} \quad \text{überraschen} \quad \text{.} \]
Relational Dependency Language Model (RDLM)

relational model predicts dependency labels, and is conditioned on ancestor/sibling labels:

\[
P(S, D) = P(D) \times P(S|D)
\]

\[
\approx \prod_{i=1}^{n} P_l(i) \times P_w(i)
\]

\[
P_l(i) = P(l_i \mid h_s(i)_{q1}, l_s(i)_{q1}, h_a(i)_{r1}, l_a(i)_{r1})
\]

\[
P_w(i) = P(w_i \mid h_s(i)_{q1}, l_s(i)_{q1}, h_a(i)_{r1}, l_a(i)_{r1}, l_i)
\]
final model generates all \( (m) \) nodes, including preterminals \( (<PT>) \) and virtual STOP nodes \( (<S>) \).

\[
P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} 
P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ 
P_l(i), & \text{otherwise} \end{cases}
\]

(4)
final model generates all \( (m) \) nodes, including preterminals \(<\text{PT}>\) and virtual STOP nodes \(<\text{S}>\).

\[
P(S, D, T) \approx \prod_{i=1}^{m} \left\{ \begin{array}{ll} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{array} \right. \tag{4}
\]
The final model generates all \( m \) nodes, including preterminals \(<\text{PT}>\) and virtual STOP nodes \(<\text{S}>\).

\[
P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} 
P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\
P_l(i), & \text{otherwise}
\end{cases}
\] (4)

[Diagram of tree structure]

<table>
<thead>
<tr>
<th>( N )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>subj</td>
<td>det</td>
<td>&lt;PT&gt;</td>
<td>&lt;S&gt;</td>
<td>&lt;PT&gt;</td>
<td>gmod</td>
<td>det</td>
<td>&lt;PT&gt;</td>
<td>&lt;S&gt;</td>
<td>attr</td>
<td>&lt;PT&gt;</td>
<td>&lt;S&gt;</td>
<td>&lt;PT&gt;</td>
<td>&lt;S&gt;</td>
<td>&lt;S&gt;</td>
</tr>
<tr>
<td>( S )</td>
<td>Ergebnisse</td>
<td>die</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
<td>Umfrage</td>
<td>der</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
<td>jüngsten</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
<td>( \epsilon )</td>
</tr>
<tr>
<td>( T )</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>
final model generates all \( (m) \) nodes, including preterminals \(<\text{PT}>\) and virtual STOP nodes \(<\text{S}>\).

\[
P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} 
P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ 
P_l(i), & \text{otherwise} \end{cases}
\] (4)
Predicting Tree Topology

The final model generates all \( m \) nodes, including preterminals \(<\text{PT}>\) and virtual STOP nodes \(<\text{S}>\).

\[
P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases}
\]  

(4)
final model generates all \( (m) \) nodes, including preterminals (\(<PT>\)) and virtual STOP nodes (\(<S>\)).

\[
P(S, D, T) \approx \prod_{i=1}^{m} \begin{cases} P_l(i) \times P_w(i), & \text{if } w_i \neq \epsilon \\ P_l(i), & \text{otherwise} \end{cases}
\]  

(4)
Neural Network Training

- feed-forward network architecture similar to [Vaswani et al., 2013]
- separate networks for $P_l$ and $P_w$
- one hidden layer
- big vocabulary: 500 000

Figure: Neural network architecture [Vaswani et al., 2013]
Decoding

Decoding with (R)DLM

- string-to-tree SMT decoder
  - decoder builds dependency trees
  - we score each hypothesis with (R)DLM
- decoding is bottom-up, but (R)DLM is top-down
  - dummy tokens for unavailable context
  - embedding of dummy token is weighted average of all words/labels
  - nodes are rescored as more context becomes available
A syntactic SMT metric for optimization and evaluation

Desideratum
- metric that rewards grammaticality beyond n-grams

Head-word chain metric (HWCM) [Liu and Gildea, 2005]
- precision-oriented reference-based metric (like BLEU)
- precision is estimated for dependency chains instead of n-grams

example chain: wird - Ergebnisse - Umfrage - der

Our contribution
- we use HWCM (f-score) for optimization of SMT parameters.
  → first use of (non-shallow) syntactic metric for tuning
Evaluation

### Metrics
- automatic SMT metrics
- agreement errors

### Data and methods
- English-German (and -Russian) data from WMT 2014
- 4.5 million sentence pairs parallel data; 120 million sentences monolingual data
- automatically parsed with ParZu [Sennrich et al., 2013]
- string-to-tree baseline as in [Williams et al., 2014]
- 3 runs of k-best batch MIRA optimization
- Moses toolkit
Evaluation: English → German (newstest2014)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU (tuned on BLEU)</th>
<th>BLEU (tuned on BLEU + HWCM_f)</th>
<th>HWCM_f (tuned on BLEU)</th>
<th>HWCM_f (tuned on BLEU + HWCM_f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-gram NNLM</td>
<td>22.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shen et al. (2010)</td>
<td>21.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DLM</td>
<td>22.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDLM</td>
<td>23.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The graph shows the performance of various methods on the newstest2014 dataset, with results measured in BLEU score (%). The methods include baseline, 5-gram NNLM, Shen et al. (2010), DLM, and RDLM, each with different models and tuning strategies.
<table>
<thead>
<tr>
<th>English → German</th>
<th>BLEU</th>
<th>HWCM$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>20.3</td>
<td>23.2</td>
</tr>
<tr>
<td>+RDLM</td>
<td>21.0</td>
<td>24.1</td>
</tr>
<tr>
<td>+HWCM tuning</td>
<td>21.6</td>
<td>24.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English → Russian</th>
<th>BLEU</th>
<th>HWCM$_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>25.9</td>
<td>23.9</td>
</tr>
<tr>
<td>+RDLM</td>
<td>26.6</td>
<td>26.5</td>
</tr>
<tr>
<td>+HWCM tuning</td>
<td>26.8</td>
<td>27.3</td>
</tr>
</tbody>
</table>
Evaluation: morphological agreement errors

- Baseline
- 5-gram NNLM
- Shen et al. (2010) DLM
- RDLM

Number of sentences with agreement error:
- Tuned on BLEU
- Tuned on BLEU + HWCM$_f$
Conclusions

Relational Dependency Language Model (RDLM)
- substantially improves fluency
  \( \text{BLEU}/\text{HWCM}_f \); agreement errors; ranked 1–2 (out of 16) @ WMT 15
- relational variant outperforms unlabelled model and related work

HWCM Tuning
- dependency-based metric suitable for tuning
  (see also: RED @ WMT15 tuning task)
- synergy effects between metric and model

Follow-Up Work
- A Joint Dependency Model of Morphological and Syntactic Structure for SMT
  come see my talk! (Mo, 13:45, room 1)
Thank you!

code

- RDLM/HWCM are integrated in Moses: http://statmt.org/moses/
- configs: https://github.com/rsennrich/wmt2014-scripts
Syntax-based language models for statistical machine translation.
In MT Summit IX, New Orleans, USA.

Discriminative, Syntactic Language Modeling through Latent SVMs.

Modeling Inflection and Word-Formation in SMT.

Syntactic Features for Evaluation of Machine Translation.

A Smorgasbord of Features for Statistical Machine Translation.

Parsers as language models for statistical machine translation.
Dependency Tree Translation: Syntactically Informed Phrasal SMT.

DEPFIX: A System for Automatic Correction of Czech MT Outputs.

Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis.

String-to-dependency Statistical Machine Translation.

Syntactic Dependency-based N-grams As Classification Features.

Decoding with Large-Scale Neural Language Models Improves Translation.
Edinburgh’s Syntax-Based Systems at WMT 2014.
### Table: Translation quality of English→Russian string-to-tree SMT system.

<table>
<thead>
<tr>
<th>MIRA objective</th>
<th>system</th>
<th>dev</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>HWCM&lt;sub&gt;f&lt;/sub&gt;</td>
<td>TER</td>
<td>BLEU</td>
<td>HWCM&lt;sub&gt;f&lt;/sub&gt;</td>
<td>TER</td>
<td>BLEU</td>
<td>HWCM&lt;sub&gt;f&lt;/sub&gt;</td>
<td>TER</td>
<td></td>
</tr>
<tr>
<td>BLEU</td>
<td>baseline</td>
<td>22.5</td>
<td>21.6</td>
<td>56.7</td>
<td>17.1</td>
<td>18.8</td>
<td>64.7</td>
<td>25.9</td>
<td>23.9</td>
<td>54.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DLM</td>
<td>23.3*</td>
<td>23.5</td>
<td>56.0</td>
<td>17.5</td>
<td>20.2</td>
<td>64.0</td>
<td>26.4</td>
<td>26.1</td>
<td>53.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RDLM</td>
<td>23.1</td>
<td>23.7</td>
<td>56.0</td>
<td>17.6</td>
<td>20.4</td>
<td>63.8</td>
<td>26.6</td>
<td>26.5</td>
<td>53.7</td>
<td></td>
</tr>
<tr>
<td>BLEU+HWCM&lt;sub&gt;f&lt;/sub&gt;</td>
<td>baseline</td>
<td>22.5</td>
<td>22.9*</td>
<td>56.1*</td>
<td>17.2</td>
<td>19.7*</td>
<td>63.9*</td>
<td>25.8</td>
<td>25.1*</td>
<td>54.1*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DLM</td>
<td>23.0</td>
<td>24.1*</td>
<td>55.6*</td>
<td>17.6</td>
<td>20.8*</td>
<td>63.2*</td>
<td>26.4</td>
<td>26.9*</td>
<td>53.3*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RDLM</td>
<td>23.1</td>
<td>24.4*</td>
<td>55.4*</td>
<td>17.6</td>
<td>20.9*</td>
<td>63.1*</td>
<td>26.8*</td>
<td>27.3*</td>
<td>53.0*</td>
<td></td>
</tr>
</tbody>
</table>

Rico Sennrich  
modelling/optimizing syntactic n-grams for SMT
### Evaluation: automatic SMT metrics

<table>
<thead>
<tr>
<th>MIRA objective</th>
<th>system</th>
<th>BLEU</th>
<th>HWCM$_f$</th>
<th>METEOR</th>
<th>TER</th>
<th>BLEU</th>
<th>HWCM$_f$</th>
<th>METEOR</th>
<th>TER</th>
<th>BLEU</th>
<th>HWCM$_f$</th>
<th>METEOR</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BLEU</strong></td>
<td>baseline</td>
<td>34.4</td>
<td>32.6</td>
<td>52.5</td>
<td>47.4</td>
<td>19.8</td>
<td>22.8</td>
<td>39.7</td>
<td>62.4</td>
<td>20.3</td>
<td>23.2</td>
<td>42.0</td>
<td>62.7</td>
</tr>
<tr>
<td></td>
<td>5-gram NNLM</td>
<td>35.3</td>
<td>33.1</td>
<td><strong>53.2</strong></td>
<td>46.4</td>
<td><strong>20.4</strong></td>
<td>23.2</td>
<td>40.2</td>
<td><strong>61.7</strong></td>
<td><strong>21.0</strong></td>
<td>23.5</td>
<td>42.5</td>
<td><strong>62.2</strong></td>
</tr>
<tr>
<td></td>
<td>[Shen et al., 2010]</td>
<td>34.4*</td>
<td>33.2</td>
<td>52.7*</td>
<td>46.9</td>
<td>20.0</td>
<td>23.2</td>
<td>40.0*</td>
<td>62.3</td>
<td>20.4</td>
<td>23.5</td>
<td>42.3</td>
<td>62.9</td>
</tr>
<tr>
<td></td>
<td>DLM</td>
<td>34.9*</td>
<td><strong>33.8</strong></td>
<td><strong>53.1</strong></td>
<td>46.8</td>
<td>20.3</td>
<td>23.6</td>
<td>40.1*</td>
<td><strong>61.7</strong></td>
<td>20.8</td>
<td>23.9</td>
<td>42.3</td>
<td><strong>62.2</strong></td>
</tr>
<tr>
<td></td>
<td>RDLM</td>
<td>35.0</td>
<td><strong>33.9</strong></td>
<td><strong>53.1</strong></td>
<td>46.7</td>
<td><strong>20.5</strong></td>
<td><strong>23.8</strong></td>
<td><strong>40.4</strong></td>
<td><strong>61.7</strong></td>
<td><strong>21.0</strong></td>
<td><strong>24.1</strong></td>
<td><strong>42.7</strong></td>
<td><strong>62.2</strong></td>
</tr>
<tr>
<td></td>
<td>5-gram + RDLM</td>
<td>35.5</td>
<td>34.0</td>
<td>53.4*</td>
<td>46.3</td>
<td>20.7</td>
<td>23.7</td>
<td>40.6*</td>
<td>61.5</td>
<td>21.4</td>
<td>24.1</td>
<td>42.9</td>
<td>61.7</td>
</tr>
<tr>
<td><strong>BLEU + HWCM$_f$</strong></td>
<td>baseline</td>
<td>34.4</td>
<td>33.0*</td>
<td>52.4</td>
<td>46.9*</td>
<td>20.0*</td>
<td>23.0*</td>
<td>39.6</td>
<td>61.9*</td>
<td>20.5*</td>
<td>23.3*</td>
<td>41.8</td>
<td>62.2*</td>
</tr>
<tr>
<td></td>
<td>5-gram NNLM</td>
<td>35.2</td>
<td>33.5*</td>
<td><strong>53.0</strong></td>
<td>46.0*</td>
<td>20.6*</td>
<td>23.4*</td>
<td>40.1</td>
<td>60.9*</td>
<td>21.1*</td>
<td>23.6</td>
<td>42.3</td>
<td>61.5*</td>
</tr>
<tr>
<td></td>
<td>[Shen et al., 2010]</td>
<td>34.2</td>
<td>33.8*</td>
<td>52.4</td>
<td>46.4*</td>
<td>20.2*</td>
<td>23.5*</td>
<td>39.8</td>
<td>61.8*</td>
<td>20.7*</td>
<td>23.7*</td>
<td>42.1</td>
<td>62.2*</td>
</tr>
<tr>
<td></td>
<td>DLM</td>
<td>34.8</td>
<td>34.3*</td>
<td>52.7</td>
<td><strong>45.9</strong></td>
<td>20.4</td>
<td>23.8*</td>
<td>39.8</td>
<td><strong>60.7</strong></td>
<td>21.4*</td>
<td>24.2*</td>
<td>42.0</td>
<td><strong>60.9</strong></td>
</tr>
<tr>
<td></td>
<td>RDLM</td>
<td>34.9</td>
<td><strong>34.5</strong></td>
<td><strong>53.0</strong></td>
<td><strong>45.8</strong></td>
<td><strong>20.9</strong></td>
<td><strong>24.2</strong></td>
<td><strong>40.3</strong></td>
<td><strong>60.7</strong></td>
<td><strong>21.6</strong></td>
<td><strong>24.5</strong></td>
<td><strong>42.5</strong></td>
<td><strong>60.8</strong></td>
</tr>
<tr>
<td></td>
<td>5-gram + RDLM</td>
<td>35.4</td>
<td>34.6*</td>
<td>53.2</td>
<td>45.4*</td>
<td>21.0*</td>
<td>24.1*</td>
<td>40.4</td>
<td>60.5*</td>
<td>21.8*</td>
<td>24.4*</td>
<td>42.7</td>
<td>60.6*</td>
</tr>
</tbody>
</table>

**Table**: Translation quality of English → German string-to-tree SMT system.

Rico Sennrich

modelling/optimizing syntactic n-grams for SMT
System-level rank correlation (Kendall’s $\tau$) between automatic metrics and number of agreement errors.
also the user manages his identity and can therefore be anonymous.

auch der Benutzer verwaltet seine Identität und können daher anonym sein.

auch der Benutzer verwaltet seine Identität und kann daher anonym sein.

darüber hinaus verwaltet der Inhaber seine Identität und kann somit anonym bleiben.

subject-verb agreement

baseline has singular subject, but plural verb
<table>
<thead>
<tr>
<th>source</th>
<th>how do you <strong>apply this definition</strong> to their daily life and social networks?</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>wie kann man <strong>diese Definition</strong> für ihr tägliches Leben und soziale Netzwerke <strong>gelten</strong>?</td>
</tr>
<tr>
<td>RDLM</td>
<td>wie kann man <strong>diese Definition</strong> auf ihren Alltag und sozialen Netzwerken <strong>anwenden</strong>?</td>
</tr>
<tr>
<td>ref</td>
<td>wie wird <strong>diese Definition</strong> auf seinen Alltag und die sozialen Netzwerke <strong>angewendet</strong>?</td>
</tr>
</tbody>
</table>

**subcategorisation**

*gelten* is intransitive.

*anwenden* is correct in transitive construction.

(hard-to-fix error for lemma-based SMT system with inflection prediction [Fraser et al., 2012] or post-correction approach [Rosa et al., 2012]).