BBN’s Systems for the Chinese-English Sub-task of the NTCIR-10 PatentMT Evaluation

Zhongqiang Huang, Jacob Devlin, Spyros Matsoukas, Rich Schwartz
{zhuang, jdevlin, smatsouk, schwartz}@bbn.com

Speech, Language, and Multimedia
Raytheon BBN Technologies
Cambridge, MA, U.S.A.
Overview

- Statistical machine translation framework
- Building patent machine translation systems
- Official evaluation results
- Summary
Part I:
Statistical Machine Translation Framework
Statistical Machine Translation (MT) Framework

Parallel Text → Rule Extraction → Chart Parsing → K-best Generation → Feature Extraction → Re-ranking → 1-Best Translation

Target Language Text → LM Estimation → Feature Extraction → Re-ranking → 1-Best Translation

Shared Forest of Derivations → K-Best Translations

Test Data → 1-Best Translation

Training

Decoding
String-to-Dependency Translation Model

- Modified version of Chiang’s Hiero algorithm
- Extract hierarchical rules with well-formed dependencies on the target side
  - Well-formed dependency structure:
    - Single rooted tree, with each child being a complete sub-tree
    - Sequence of siblings, each being a complete sub-tree
  - Use POS tag of head word as non-terminal labels on the target side

\[ X : X_1 \text{ 出发} \text{ 去} X_2 \rightarrow VB : NR_1 \text{ leaves for } NN_2 \]

- Extract all phrasal rules, ignoring dependency
- Features:
  - 10+ core features
  - ~50K sparse binary features
Part II:
Building Patent Machine Translation Systems
BBN Patent MT systems - Overview

- Data released by the NTCIR-10 organizers
  - Parallel data: 45M words of Chinese-English sentence pairs
  - Extra LM data: 14B words of US patents in English
  - Development data: 2K Chinese-English sentence pairs, split into tuning and test set

- Model training
  - Translation Model: trained on the 45M parallel corpus
  - Language Models:
    - 45M LM: trained on the target side of the 45M parallel corpus
    - 14B LM: trained on the 45M words plus the 14B US patent words

- Summary of results on the test set (development)
Review of Work for BBN NTCIR-9

• Consistent tokenization
  • Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”
Review of Work for BBN NTCIR-9

• Consistent tokenization
  • Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”

• Special token sharing
  • Replace special tokens with a common token for each type in translation and language model
    • Numbers: e.g., 2,596, -123.321
    • Patent IDs: e.g., No.5,400,788, No. 5,405,753
    • Math expressions: e.g., p=0.004, Sine(45)=0.7071
    • Material names: e.g., C15H23N2O5P, LiEt3BH
    • Labeled names: e.g., 3.05kg, 200ml
Review of Work for BBN NTCIR-9

- Consistent tokenization
  - Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”

- Special token sharing
  - Replace special tokens with a common token for each type in translation and language model
    - Numbers: e.g., 2,596, -123.321
    - Patent IDs: e.g., No.5,400,788, No. 5,405,753
    - Math expressions: e.g., p=0.004, Sine(45)=0.7071
    - Material names: e.g., C15H23N2O5P, LiEt3BH
    - Labeled names: e.g., 3.05kg, 200ml

- Patent case-LM
  - Re-trained on the 45M LM data
• Consistent tokenization
  • Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”
• Special token sharing
  • Replace special tokens with a common token for each type in translation and language model
    • Numbers: e.g., 2,596, -123.321
    • Patent IDs: e.g., No.5,400,788, No. 5,405,753
    • Math expressions: e.g., p=0.004, Sine(45)=0.7071
    • Material names: e.g., C15H23N2O5P, LiEt3BH
    • Labeled names: e.g., 3.05kg, 200ml
• Patent case-LM
  • Re-trained on the 45M LM data
• Optimized word segmenter
  • Re-optimized for patent translation
Review of Work for BBN NTCIR-9

• **Consistent tokenization**
  • Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”

• **Special token sharing**
  • Replace special tokens with a common token for each type in translation and language model
    • Numbers: e.g., 2,596, -123.321
    • Patent IDs: e.g., No.5,400,788, No. 5,405,753
    • Math expressions: e.g., p=0.004, Sine(45)=0.7071
    • Material names: e.g., C15H23N2O5P, LiEt3BH
    • Labeled names: e.g., 3.05kg, 200ml

• **Patent case-LM**
  • Re-trained on the 45M LM data

• **Optimized word segmenter**
  • Re-optimized for patent translation

• **Top 100 sparse features**
  • Due to the smaller tuning set, we use only the top 100 features of the highest weights in each category of the 50K sparse features
Review of Work for BBN NTCIR-9

• Consistent tokenization
  • Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”

• Special token sharing
  • Replace special tokens with a common token for each type in translation and language model
    • Numbers: e.g., 2,596, -123.321
    • Patent IDs: e.g., No.5,400,788, No. 5,405,753
    • Math expressions: e.g., p=0.004, Sine(45)=0.7071
    • Material names: e.g., C15H23N2O5P, LiEt3BH
    • Labeled names: e.g., 3.05kg, 200ml

• Patent case-LM
  • Re-trained on the 45M LM data

• Optimized word segmenter
  • Re-optimized for patent translation

• Top 100 sparse features
  • Due to the smaller tuning set, we use only the top 100 features of the highest weights in each category of the 50K sparse features

• Document-level LM adaptation
  • Find documents in monolingual English patent corpus that are similar to test document
  • Estimate a separate LM and interpolate with the general LM

\[ P_{LM}(s) = (1 - \alpha)P_{generalLM}(s) + \alpha P_{biasLM}(s) \]
Review of Work for BBN NTCIR-9

- **Consistent tokenization**
  - Fixed inconsistent tokenization of ASCII strings in the source and target sides, e.g., “IS-1000” vs. “IS – 1000”

- **Special token sharing**
  - Replace special tokens with a common token for each type in translation and language model
    - Numbers: e.g., 2,596, -123.321
    - Patent IDs: e.g., No.5,400,788, No. 5,405,753
    - Math expressions: e.g., $p=0.004$, $\text{Sine}(45)=0.7071$
    - Material names: e.g., C15H23N2O5P, LiEt3BH
    - Labeled names: e.g., 3.05kg, 200ml

- **Patent case-LM**
  - Re-trained on the 45M LM data

- **Optimized word segmenter**
  - Re-optimized for patent translation

- **Top 100 sparse features**
  - Due to the smaller tuning set, we use only the top 100 features of the highest weights in each category of the 50K sparse features

- **Document-level LM adaptation**
  - Find documents in monolingual English patent corpus that are similar to test document
  - Estimate a separate LM and interpolate with the general LM

\[
P_{\text{LM}}(s) = (1 - \alpha)P_{\text{generalLM}}(s) + \alpha P_{\text{biasLM}}(s)
\]

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBN Baseline with 45M LM</td>
<td><strong>34.01</strong></td>
</tr>
<tr>
<td>+ consistent tokenization</td>
<td>34.56</td>
</tr>
<tr>
<td>+ more token sharing</td>
<td>34.97</td>
</tr>
<tr>
<td>+ patent case-LM</td>
<td>36.47</td>
</tr>
<tr>
<td>+ optimized word segmenter</td>
<td>36.95</td>
</tr>
<tr>
<td>+ top 100 features</td>
<td><strong>37.71</strong></td>
</tr>
<tr>
<td>+ 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ document-level LM adaptation</td>
<td><strong>40.04</strong></td>
</tr>
</tbody>
</table>
Development for NTCIR-10 Evaluation

- Overview
  - Miscellaneous additional features
  - Sentence-level LM adaptation
  - Robust context dependent translation
  - Recurrent neural network LM
  - Translation-based caser
Miscellaneous Additional Features

• Bigram lexical translation model
  • Extension of context-based lexical probabilities to
    model joint likelihood of target bigrams given source
    context
    \[
    P(t_{s_i}, t_{s_{i-1}} | s_i, s_{i-1}, s_{i+1}, s_{i-2})
    \]
  • Apply chain rule and use simple back-off smoothing
  • Similarly for the backward direction
Miscellaneous Additional Features

• Bigram lexical translation model
  • Extension of context-based lexical probabilities to model joint likelihood of target bigrams given source context
    \[ P(t_{s_i}, t_{s_{i-1}} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) \]
  • Apply chain rule and use simple back-off smoothing
  • Similarly for the backward direction

• Trait features, e.g.,
  • Percent of NULL source content words
  • Percent of words that re-order
  • Percent of low-frequency n-grams
  • Source-to-target length ratio
Miscellaneous Additional Features

- Bigram lexical translation model
  - Extension of context-based lexical probabilities to model joint likelihood of target bigrams given source context
    \[ P(t_{s_i}, t_{s_{i-1}} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) \]
  - Apply chain rule and use simple back-off smoothing
  - Similarly for the backward direction
- Trait features, e.g.,
  - Percent of NULL source content words
  - Percent of words that re-order
  - Percent of low-frequency n-grams
  - Source-to-target length ratio
- Disable feature normalization
Miscellaneous Additional Features

- Bigram lexical translation model
  - Extension of context-based lexical probabilities to model joint likelihood of target bigrams given source context
  
  \[ P\left(t_{s_i}, t_{s_{i-1}} | s_i, s_{i-1}, s_{i+1}, s_{i-2}\right) \]

- Apply chain rule and use simple back-off smoothing
- Similarly for the backward direction

- Trait features, e.g.,
  - Percent of NULL source content words
  - Percent of words that re-order
  - Percent of low-frequency n-grams
  - Source-to-target length ratio

- Disable feature normalization

---

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR-9 system with 45M LM</td>
<td>37.71</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>38.06</td>
</tr>
<tr>
<td>NTCIR-9 system with 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>39.51</td>
</tr>
</tbody>
</table>
Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents
Sentence-level LM adaptation

• Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents

Percentage of source n-grams (tokens) in the test sentences that are observed in the parallel training for newswire (GALE) and patent (NTCIR-10)
Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents

Percentage of source n-grams (tokens) in the test sentences that are observed in the parallel training for newswire (GALE) and patent (NTCIR-10)

Percentage of target n-grams (tokens) in the patent test sentences that are also observed in the patent parallel corpus and the monolingual English patent corpus
Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents

![Graph 1: Percentage of source n-grams (tokens) in the test sentences that are observed in the parallel training for newswire (GALE) and patent (NTCIR-10)]

![Graph 2: Percentage of target n-grams (tokens) in the patent test sentences that are also observed in the patent parallel corpus and the monolingual English patent corpus]

Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents

![Graph 1: Percentage of source n-grams (tokens) in the test sentences that are observed in the parallel training for newswire (GALE) and patent (NTCIR-10)]

![Graph 2: Percentage of target n-grams (tokens) in the patent test sentences that are also observed in the patent parallel corpus and the monolingual English patent corpus]
Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents

Percentage of source n-grams (tokens) in the test sentences that are observed in the parallel training for newswire (GALE) and patent (NTCIR-10)

Percentage of target n-grams (tokens) in the patent test sentences that are also observed in the patent parallel corpus and the monolingual English patent corpus
Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents.

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR-9 system with 45M LM</td>
<td>37.71</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>38.06</td>
</tr>
<tr>
<td>NTCIR-9 system with 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>39.51</td>
</tr>
<tr>
<td>+ document-level LM adaptation</td>
<td>39.94</td>
</tr>
</tbody>
</table>
Sentence-level LM adaptation

- Patent documents tend to use well-structured sentence and re-use n-grams in other patent documents

**Graph 1:**
- Percentage of source n-grams (tokens) in the test sentences that are observed in the parallel training for newswire (GALE) and patent (NTCIR-10)

**Graph 2:**
- Percentage of target n-grams (tokens) in the patent test sentences that are also observed in the patent parallel corpus and the monolingual English patent corpus

**Table:**

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR-9 system with 45M LM</td>
<td>37.71</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>38.06</td>
</tr>
<tr>
<td>NTCIR-9 system with 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>39.51</td>
</tr>
<tr>
<td>+ document-level LM adaptation</td>
<td>39.04</td>
</tr>
<tr>
<td>+ sentence-level LM adaptation</td>
<td>40.95</td>
</tr>
</tbody>
</table>
Robust Context-Dependent Modeling

- High order context-dependent translation models may be very sparse

\[ P(t_{s_i}, t_{s_{i-1}} | s_i, s_{i-1}, s_{i+1}, s_{i-2}) \]
Robust Context-Dependent Modeling

- High order context-dependent translation models may be very sparse

\[ P(t_{s_i}, t_{s_{i-1}} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) \]

- Common solution
  - First, apply the chain rule
  \[ P(t_{s_i}, t_{s_{i-1}} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) = P(t_{s_i} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) \]
  - Back-off each probability independent
Robust Context-Dependent Modeling

- High order context-dependent translation models may be very sparse

\[
P(t_{s_i}, t_{s_{i-1}} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2})
\]

- Common solution
  - First, apply the chain rule
  \[
P(t_{s_i}, t_{s_{i-1}} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) = P(t_{s_i} \mid s_i, s_{i-1}, s_{i+1}, s_{i-2}) P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2})
\]
  - Back-off each probability independent

- But, unlike LM, there is no clear back-off ordering
  - Is \( P(t_{s_{i-1}} \mid t_{s_i}, s_{i-1}) \) “better” than \( P(t_{s_{i-1}} \mid s_i, s_{i-1}) \)?
Robust Context-Dependent Modeling

• Our solution: interpolate all possible back-off components
  • Sparse context types can be added independently of one another

\[
P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) = \omega_0 P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) + \omega_1 P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}) + \cdots + \omega_3 P(t_{s_{i-1}} \mid t_{s_i})
\]
Robust Context-Dependent Modeling

• Our solution: interpolate all possible back-off components
  • Sparse context types can be added independently of one another

\[
P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) = \omega_0 P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) + \omega_1 P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}) + \cdots + \omega_3 P(t_{s_{i-1}} \mid t_{s_i})
\]

• Each weight \( \omega \) is a function of the marginal count

\[
\omega_j P(t_{s_i} \mid s_i, s_{i-1}) = \frac{1}{Z} \alpha_j \log(C(s_i, s_{i-1})) \frac{C(t_{s_i}, s_i, s_{i-1})}{C(s_i, s_{i-1})}
\]
Robust Context-Dependent Modeling

- Our solution: interpolate all possible back-off components
  - Sparse context types can be added independently of one another

\[ P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) = \omega_0 P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}, s_{i-2}) + \omega_1 P(t_{s_{i-1}} \mid t_{s_i}, s_i, s_{i-1}, s_{i+1}) + \cdots + \omega_{30} P(t_{s_{i-1}} \mid t_{s_i}) \]

- Each weight \( \omega \) is a function of the marginal count

\[ \omega_j P(t_{s_i} \mid s_i, s_{i-1}) = \frac{1}{Z_j} \alpha_j \log(C(s_i, s_{i-1})) \frac{C(t_{s_i}, s_i, s_{i-1})}{C(s_i, s_{i-1})} \]

- Weights \( \alpha \) are optimized to maximize likelihood on a held-out set
  - Least useful components are thrown out for efficiency
Robust Context-Dependent Modeling

- Our solution: interpolate all possible back-off components
  - Sparse context types can be added independently of one another

\[
P(t_{s_{i-1}} \mid t_s, s_i, s_{i-1}, s_{i+1}, s_{i-2}) = \omega_0 P(t_{s_{i-1}} \mid t_s, s_i, s_{i-1}, s_{i+1}, s_{i-2}) + \omega_1 P(t_{s_{i-1}} \mid t_s, s_i, s_{i-1}, s_{i+1}) + \cdots + \omega_{30} P(t_{s_{i-1}} \mid t_s)
\]

- Each weight \( \omega \) is a function of the marginal count
  \[
  \omega_j P(t_s \mid s_i, s_{i-1}) = \frac{1}{Z} \alpha_j \log(C(s_i, s_{i-1})) \frac{C(t_s, s_i, s_{i-1})}{C(s_i, s_{i-1})}
  \]

- Weights \( \alpha \) are optimized to maximize likelihood on a held-out set
  - Least useful components are thrown out for efficiency

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR-9 system with 45M LM</td>
<td>37.71</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>38.06</td>
</tr>
<tr>
<td>+ <strong>robust context dependent translation</strong></td>
<td>38.72</td>
</tr>
<tr>
<td>NTCIR-9 system with 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>39.51</td>
</tr>
<tr>
<td>+ sentence-level LM adaptation</td>
<td>40.95</td>
</tr>
<tr>
<td>+ <strong>robust context dependent translation</strong></td>
<td>41.09</td>
</tr>
</tbody>
</table>
Neural Net LM

- Trained a recurrent neural net LM for rescoring
  - Mikolov’s toolkit:
    http://www.fit.vutbr.cz/~imikolov/rnnlm/
  - Interpolated with 5-gram KN Smoothing LM
Neural Net LM

• Trained a recurrent neural net LM for rescoring
  • Mikolov’s toolkit: http://www.fit.vutbr.cz/~imikolov/rnnlm/
  • Interpolated with 5-gram KN Smoothing LM

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR-9 system with 45M LM</td>
<td>37.71</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>38.06</td>
</tr>
<tr>
<td>+ robust context dependent translation</td>
<td>38.72</td>
</tr>
<tr>
<td>+ recurrent neural network LM</td>
<td>39.35</td>
</tr>
<tr>
<td>NTCIR-9 system with 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>39.51</td>
</tr>
<tr>
<td>+ document-level LM adaptation</td>
<td>39.94</td>
</tr>
<tr>
<td>+ sentence-level LM adaptation</td>
<td>40.95</td>
</tr>
<tr>
<td>+ robust context dependent translation</td>
<td>41.09</td>
</tr>
<tr>
<td>+ recurrent neural network LM</td>
<td>41.43</td>
</tr>
</tbody>
</table>
Translation-based Caser

- Treats casing as a translation problem
  - Similar to (Hassan, et al. 2006)’s MaTrEx system
  - Trained on 45M LM training data
  - Use rule probabilities, case LM probability, and sparse features, e.g., *Is the target word upper cased and does it follow a period? Is the target word upper cased and a proper noun?*
Translation-based Caser

- Treats casing as a translation problem
  - Similar to (Hassan, et al. 2006)’s MaTrEx system
  - Trained on 45M LM training data
  - Use rule probabilities, case LM probability, and sparse features, e.g., *Is the target word upper cased and does it follow a period? Is the target word upper cased and a proper noun?*

<table>
<thead>
<tr>
<th>System</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR-9 system with 45M LM</td>
<td>37.71</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>38.06</td>
</tr>
<tr>
<td>+ robust context dependent translation</td>
<td>38.72</td>
</tr>
<tr>
<td>+ recurrent neural network LM</td>
<td>39.35</td>
</tr>
<tr>
<td><strong>+ translation-based caser</strong></td>
<td><strong>40.02</strong></td>
</tr>
<tr>
<td>NTCIR-9 system with 14B LM</td>
<td>39.14</td>
</tr>
<tr>
<td>+ miscellaneous features</td>
<td>39.51</td>
</tr>
<tr>
<td>+ document-level LM adaptation</td>
<td>39.94</td>
</tr>
<tr>
<td>+ sentence-level LM adaptation</td>
<td>40.95</td>
</tr>
<tr>
<td>+ robust context dependent translation</td>
<td>41.09</td>
</tr>
<tr>
<td>+ recurrent neural network LM</td>
<td>41.43</td>
</tr>
<tr>
<td><strong>+ translation-based caser</strong></td>
<td><strong>42.13</strong></td>
</tr>
</tbody>
</table>
Part III:
Official Evaluation Results
Official Automatic (BLEU) Results

• The two BBN systems
  • BBN-1: the primary system, trained on 45M parallel corpus plus 14B English patent corpus
  • BBN-2: the secondary system, trained on 45M parallel corpus only

• NCTIR Official Baseline systems
  • Baseline1– Moses phrase-based hierarchical SMT system
  • Baseline2– Moses phrase-based SMT system
Official Automatic (BLEU) Results

- The two BBN systems
  - BBN-1: the primary system, trained on 45M parallel corpus plus 14B English patent corpus
  - BBN-2: the secondary system, trained on 45M parallel corpus only

- NCTIR Official Baseline systems
  - Baseline1—Moses phrase-based hierarchical SMT system
  - Baseline2—Moses phrase-based SMT system

<table>
<thead>
<tr>
<th>System</th>
<th>Intrinsic evaluation</th>
<th>Chronological evaluation</th>
<th>Multilingual evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBN-1</td>
<td>42.68</td>
<td>39.44</td>
<td>41.09</td>
</tr>
<tr>
<td>BBN-2</td>
<td>39.98</td>
<td>36.69</td>
<td>38.93</td>
</tr>
<tr>
<td>Baseline1</td>
<td>32.52</td>
<td>30.74</td>
<td>17.96</td>
</tr>
<tr>
<td>Baseline2</td>
<td>31.34</td>
<td>29.34</td>
<td>18.05</td>
</tr>
</tbody>
</table>

* → indicates the change in BLEU from NTCIR-9 evaluation to NTCIR-10 evaluation
Official Manual Evaluation Results

• Adequacy: scores from 5 (best) to 1 (worst)

<table>
<thead>
<tr>
<th>System</th>
<th>Average adequacy</th>
<th>Allocation of scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>BBN-1</td>
<td>42.68</td>
<td>156</td>
</tr>
<tr>
<td>Baseline1</td>
<td>32.52</td>
<td>46</td>
</tr>
<tr>
<td>Baseline2</td>
<td>31.34</td>
<td>38</td>
</tr>
</tbody>
</table>
Official Manual Evaluation Results

- Adequacy: scores from 5 (best) to 1 (worst)

<table>
<thead>
<tr>
<th>System</th>
<th>Average adequacy</th>
<th>Allocation of scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>BBN-1</td>
<td>42.68</td>
<td>156</td>
</tr>
<tr>
<td>Baseline1</td>
<td>32.52</td>
<td>46</td>
</tr>
<tr>
<td>Baseline2</td>
<td>31.34</td>
<td>38</td>
</tr>
</tbody>
</table>

- Acceptability: scores in AA (best), A, B, C, and F (worst)
- Pairwise acceptability: percentage of wins and ties when comparing acceptability score with other submissions

<table>
<thead>
<tr>
<th>System</th>
<th>Pairwise score</th>
<th>Allocation of scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AA</td>
</tr>
<tr>
<td>BBN-1</td>
<td>0.69</td>
<td>81</td>
</tr>
</tbody>
</table>
Official Manual Evaluation Results

- Adequacy: scores from 5 (best) to 1 (worst)
  
<table>
<thead>
<tr>
<th>System</th>
<th>Average adequacy</th>
<th>Allocation of scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>BBN-1</td>
<td>42.68</td>
<td>156</td>
</tr>
<tr>
<td>Baseline1</td>
<td>32.52</td>
<td>46</td>
</tr>
<tr>
<td>Baseline2</td>
<td>31.34</td>
<td>38</td>
</tr>
</tbody>
</table>

- Acceptability: scores in AA (best), A, B, C, and F (worst)
- Pairwise acceptability: percentage of wins and ties when comparing acceptability score with other submissions

<table>
<thead>
<tr>
<th>System</th>
<th>Pairwise score</th>
<th>Allocation of scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AA</td>
</tr>
<tr>
<td>BBN-1</td>
<td>0.69</td>
<td>81</td>
</tr>
</tbody>
</table>

- Patent examination evaluation: scores in S (perfect), A, B, C, D, and F (worst)

<table>
<thead>
<tr>
<th>System</th>
<th>Allocation of scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td>BBN-1</td>
<td>6</td>
</tr>
</tbody>
</table>
Translation Examples

Source: | MT output: | Reference:
---|---|---
对于每一像素，着色引擎210使用在以上等式(2)-(4)中陈述的边等式来确定所述像素是否在三角形中。 | For each pixel, the **rendering** engine 210 uses the edge equation set forth in equations (2) to (4) above to determine whether the **pixels** in a triangle. | For each pixel, the **shading** engine 210 determines whether the **pixel** is in the triangle using the edge equations set forth in equations (2) - (4) above.

Source: | MT output: | Reference:
---|---|---
上述说明书全面描述了根据本发明原理的改进型可穿透膜片的成分、制造和用途。 | The above description fully describes the composition, manufacture and use of improved penetrable **diaphragm** in accordance with the principles of the present invention. | The above specification provides a complete description of the composition, manufacture and use of the improved penetrable **membrane** in accordance with the principles of the present invention.
Summary

• It was relatively straightforward to port BBN’s MT system to work on patents
  • 4-5 weeks of efforts in NTCIR-9 evaluation
  • 3-4 weeks of efforts in NTCIR-10 evaluation
  • All techniques initially developed for other domains work well on patents

• Special attention to patents helps
  • Better tokenization, special token sharing, optimizing word segmentation
  • Sentence-level LM adaptation
  • Further improvement is possible by exploring special properties of patents

• Lots of potential
  • Patents are easier to translate
  • State-of-the-art accuracies in both automatic and manual evaluations
  • Helpful in real patent examination and possibly other tasks
Related MT Research at BBN

Leading performer in DARPA’s MT programs

- **Text-to-text translation (GALE, BOLT)**
  - Arabic and Chinese to English. newswire, weblogs, web forums, SMS/chat
- **Speech-to-text translation (GALE)**
  - Arabic and Chinese to English. broadcast news and broadcast conversation
- **Speech-to-speech translation (TransTac, BOLT)**
  - English to/from Iraqi Arabic, Farsi, Dari, Pashto, Malay, and Spanish
  - TransTalk: portable (Android), two-way translation device; deployed by US Army
- **Image to text translation (MADCAT)**
  - Foreign text (Arabic, Chinese and Korean) in images (through OCR) to English
- **Multilingual broadcast/web monitoring**
  - Continuous searchable archive of international television broadcasts and web sites
  - Automatic translation to English for deep analysis

Contact: schwartz@bbn.com