Tools-Driven Content Curation & Engine Training

ALEXYANISHEVSKY
Welocalize
October 2014
OVERVIEW

Sparsity of data is bad. Abundance of data is bad. Let’s talk about the proper way to curate data.
ANALYZE

TWO STEPS: CLEAN & ANALYZE

CLEAN
• Remove Markup
• Remove Bullets
• Dedupe
• Remove Misaligned Tus
• Report on Discards (Potential Use as LM)

ANALYZE: Tools Used
• Candadite Scorer — Analysis of Content Suitability & Domain
• Style Scorer — Adhereance to Client’s Style
ANALYZE: CANDIDATE SCORER

How Good is Candidate Text?

- Take Historically “Bad” Text
- POS Tagger on “Bad” Text & Candidate Text
- Exclude List to Reduce False Positives
- Lower Score is Better (Candidate Test Does NOT Match “Bad” Text)

Compare to Historically “Bad” Text
LOWER SCORE IS BETTER
<table>
<thead>
<tr>
<th>Candidate Segment</th>
<th># of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you are viewing regularly acquired print jobs, selecting a period on the left side of the window displays the total costs for that period and the amounts of paper and ink consumed.</td>
<td>7</td>
</tr>
<tr>
<td>You can use this procedure to cancel the regular acquisition of print job logs from printers.</td>
<td>4</td>
</tr>
<tr>
<td>You can also set up automatic email notification of any printer problems or errors to email addresses you designate in advance.</td>
<td>3</td>
</tr>
<tr>
<td>Replace the paper in the printer with the correct paper.</td>
<td>3</td>
</tr>
<tr>
<td>The ink level of every color in the printer.</td>
<td>2</td>
</tr>
<tr>
<td>You can check the status of the hard disk and the documents saved on the hard disk.</td>
<td>1</td>
</tr>
<tr>
<td>If the job has multiple pages, the size of the last page is displayed.</td>
<td>1</td>
</tr>
<tr>
<td>If you do not set the standard prices, you cannot set the price for each roll paper.</td>
<td>1</td>
</tr>
<tr>
<td>Enter the Width and Length of the paper, as well as the Price for the area of the configured width x height.</td>
<td>1</td>
</tr>
<tr>
<td>The job items and details that can be displayed in the job list area are as follows.</td>
<td>1</td>
</tr>
<tr>
<td>This feature collects print job logs from the printer at regular intervals and saves the logs on your computer.</td>
<td>1</td>
</tr>
<tr>
<td>You can use the Job sheet in imagePROGRAF Status Monitor for operations such as pausing and canceling print jobs.</td>
<td>1</td>
</tr>
</tbody>
</table>

| Total | 57 |
Combines PPL Ratio, Dissimilarity Score & Classification Score

- Each Document Receives a Score from 0-4
- Higher Score Indicates Better Match to Style Established by Client’s Documents
- Does NOT Require Parallel Data
- Source Scored for Training/Tuning Suitability
**Tools Used:**
Perplexity Evaluator
Source Content Profiler

**MT Engine Lifecycle**

- **Analyze**
  - source content
- **Select**
  - best output
- **Validate**
  - MT engines
- **Retrain**
  - MT engines
- **Train**
  - MT engines
- **Curate**
  - linguistic assets
CURATE: PERPLEXITY EVALUATOR

- Build Language Model (LM) of Historically “Bad” Text, Historically “Good” Text & Domain-Specific Text
- Is Candidate Text Closer to Historically “Bad” Text, Historically “Good” Text & Domain-Specific Text
- What Domain is Candidate Text Closest To?

LOWER SCORE IS BETTER
FILE LEVEL

<table>
<thead>
<tr>
<th>Document</th>
<th>Client</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>7.78</td>
</tr>
<tr>
<td>bad</td>
<td>2910.85</td>
</tr>
<tr>
<td>client</td>
<td>144.21</td>
</tr>
</tbody>
</table>

TU LEVEL

<tu srclang="EN-US" tuid="75438">  <prop type="x-ppl:train2">208</prop><prop type="x-ppl:techdoc6">191.029</prop><prop type="x-ppl:productLoc">97.073</prop><prop type="x-ppl:sales1">325.983</prop><prop type="x-ppl:productLoc">396.398</prop><prop type="x-ppl:legal1">617.876</prop><prop type="x-ppl:techdoc6">s</prop>Consistent feature set across multiple platforms (Windows, Mac, iOS, Android).<seg>  <prop type="x-ALS:Context">TEXT</prop><prop type="x-ALS:Source File">\DATA\TC\39720\SRC\EN-US\co-02_battle-card_en\co-02_battle-card_en.in</prop><seg>Conjunto de características coherente en varias plataformas (Windows, Mac, iOS, Android)</seg>  </seg>  </tu>
A Collaborative Project Between CNGL & Several Industry Partners

Analyses Source Content & Group Into “Profiles” Based on Set of “Features” — Syntax, Grammar, Readability Score, Terminology, Engineering Characteristics — DNT, Glossary Hits & Tag Ratio
Target Train, Tune, Test Data Selection:

- Random
- Frequency
- Length
- Product or Domain

Check for “Cheating”
Targeted Error Correction of OOVs Based On:

- Ratio Per Sentence
- Token Count
- Ratio Per File
Dashboard for Viewing MT Metrics

✓ Tokenizes Input from Variety of Formats & Runs Several Scoring Algorithms in Parallel

✓ Exports Detailed Analysis to Spreadsheet for Sentence-By-Sentence Review
Visualization of Individual Segments Based on Candidate & Reference

- Precision
- Recall
Correct OOV
Implement Feedback From Production & LQA
Select Best Output: TMT PRIME

- TMT Prime Provides a Commercially Viable Platform that Allows for the Seamless Integration of MT/TM Systems
  - MT/TM System Recommender
- Recommendation Based on TMT Prime Score
  - Confidence Score
  - “Intelligent” Fuzzy Match
- Translation Can Be Obtained from Either In-House or Third Party Systems (Bing or Google)
- A Cloud-Based Framework that Effectively Selects Translation Outputs by Using Different MT/TM Systems

Diagram:
- Analyze source content
- Curate linguistic assets
- Train MT engines
- Validate MT engines
- Retrain MT engines
- Select best output

We localize - doing things differently
Select Best Output: TMT PRIME 2
✓ All TMs for All Languages Have Been Profiled

✓ Implemented Tool to Query Most Relevant TUs by Domain

✓ BLEU Scores Increased by 10 Plus Points on Domain-Specific Engines for Both Asian & European Languages

✓ Productivity Increased by 5-7% on Domain-Specific Engines for Both Asian & European Languages
✓ Develop Domain-Specific Engines for All Languages Localized by the Client (50-60 Engines)

✓ Assess Productivity Increase for All Languages & Engines as Related to BLEU Score Improvement

✓ Use TMT Prime to Recommend & Get Best Engine Translation on a Per Segment Bases as Opposed to Current File Level
THANK YOU

ALEXYANISHEVSKY
Welocalize
October 2014