Machine Translation in Academia and in the Commercial World: a Contrastive Perspective

Alon Lavie
Research Professor – LTI, Carnegie Mellon University
Co-founder, President and CTO – Safaba Translation Solutions

WMT-2014
June 26, 2014
My Two Perspective Views on MT

> Research Professor – Language Technologies Inst., Carnegie Mellon

> Main areas of research:
  > MT evaluation metrics: Meteor
  > Syntax-based MT: syntax-to-syntax models
  > MT System Combination: CMU MEMT System
  > MT into morphologically-rich languages (Arabic)
  > MT for human translation and post-editing

> Co-founder, President and CTO – Safaba Translation Solutions

> Commercial MT technology company focused on solutions and services to global enterprises
Mission Statement: Safaba helps global corporations translate and localize their large volumes of corporate content into the local languages of the markets in which they operate, by dramatically improving translation velocity and reducing translation costs.

Customers: Global corporations, primarily in the hardware, software and IT space, such as Dell, PayPal.

Partners: Select commercial Language Service Providers (LSPs), such as Welocalize, ABBYY-LS.

MT Solutions: Primarily real-time MT services delivered as software-as-a-service (SaaS) using dedicated hosted private-cloud platform.
> **Business Model:**
  > Primary - Full-Service SaaS Model: client delivers data resources, Safaba develops and deploys the MT engines as remote hosted services
  > Secondary – Full-Service with on-site installation
  > Secondary – “Do It Yourself” (DIY) service using Safaba’s EMTGlobal Online platform
  > Clients typically pay us for MT Implementation, Integration and a volume-based annual license

> **Our Largest Deployment:** Dell.com content is translated daily from English into 28 different languages by Safaba's automated translation solutions in collaboration with Welocalize.

> **Volume:** Dell.com translates over 1M words per month through the Safaba EMTGlobal MT platform.
> Enterprise Impact and ROI at Dell of Welocalize + Safaba MT Program:

> Wayne Bourland – Director of Translation, Dell.com

> “Enterprise Language Strategy”, TAUS ILF, June 2014

> Translation cost reduced by nearly 40% on average

> Savings to-date of $2.4M from using MT

> Project delivery times reduced by 40% - 5 days to 3

> Quality has been maintained at the same level as traditional HT

> ROI for MT over 900%
Main MT Technology Stack:
- Predominantly NLP-augmented phrase-based statistical MT technology
- MT runtime decoding platform based on Moses, augmented with Safaba-proprietary pre and post processing modules
- Safaba-proprietary MT development platform based in part on open-source components (Moses, FastAlign, KenLM, etc.)
- DuctTape as a workflow management framework that supports the entire MT development workflow

Main MT Technology Challenges:
- Effective and scalable client-specific adaptation
- Maximizing MT accuracy into many morphologically-rich languages
- Translation of highly-structured content
- Maximizing translator MT post-editing productivity
- Frequent and ongoing adaptation
Talk Objectives

> Provide some deeper insight about the characteristic differences between typical “academic MT systems” (i.e. for WMT and NIST evaluations) and Safaba’s typical commercial systems
> Provide a closer look at some of the main R&D challenges and requirements for delivering advanced hosted real-time Statistical MT services and solutions in commercial settings
> Motivate the broader research community to work more extensively on MT problems and solutions for commercially-relevant content-types and domains
WMT MT Systems vs Safaba MT Systems

> WMT: MT for Assimilation (mostly)
  > Broad-domain systems: News commentary, medical information
  > Training data: Europarl, News commentary, Common Crawl, Gigaword
  > In and out of English and several major European languages

> Safaba: MT for Dissemination (mostly)
  > Client-specific and client-adapted MT engines for enterprise clients
  > Typically domain-focused and consistent content types: product information and documentation, customer support, marketing
  > Training data: Translation Memories and other assets from the client + domain-relevant background data (i.e. TAUS data)
  > Mostly out of English, into 30+ languages (European, Asian, South American variants of ES and PT)
  > Different language variants (FR-France/Canada, PT-Portugal/Brazil, ES-Spain/Latin America, EN-US/GB, etc.)
TAUS: Translation Automation User Society

https://www.taus.net/
TAUS Data

- [https://www.tausdata.org/](https://www.tausdata.org/)
- Data repository consisting of pooled parallel translation data from over 100 contributors (primarily large corporations and LSPs)
- **Total data assets**: about 56 Billion words (including matrix TMAs)
- **Variety of domains**: hardware, software, IT, financial, automotive, medical and bio-pharma, etc.
- Mostly categorized, indexed and word-aligned
- Free online search as a translation memory, terminology DB
- **Coming soon**: freely available for non-commercial academic research!!
- **Data Example:**
  - ENUS-to-ESES: 217.4 M source words
    - Computer Software: 66.9 M words
    - Computer Hardware: 9.0 M words
    - Legal Services: 2.4 M words
    - Other: 138.5M words
Some Contrastive MT System Scores

BLEU Scores of best WMT-2014 MT systems versus Safaba-developed TAUS data generic MT systems

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Best WMT-2014</th>
<th>Safaba TAUS Generic</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-to-FR</td>
<td>35.8</td>
<td>65.4</td>
</tr>
<tr>
<td>EN-to-ES</td>
<td>30.4 *</td>
<td>66.2</td>
</tr>
<tr>
<td>EN-to-RU</td>
<td>29.9</td>
<td>41.6</td>
</tr>
<tr>
<td>EN-to-CS</td>
<td>21.6</td>
<td>43.6</td>
</tr>
<tr>
<td>EN-to-DE</td>
<td>20.6</td>
<td>52.5</td>
</tr>
<tr>
<td>FR-to-EN</td>
<td>35.0</td>
<td>68.0</td>
</tr>
<tr>
<td>RU-to-EN</td>
<td>31.8</td>
<td>---</td>
</tr>
<tr>
<td>ES-to-EN</td>
<td>31.4 *</td>
<td>70.4</td>
</tr>
<tr>
<td>DE-to-EN</td>
<td>29.0</td>
<td>62.4</td>
</tr>
<tr>
<td>CS-to-EN</td>
<td>28.8</td>
<td>---</td>
</tr>
</tbody>
</table>
## Sample Safaba Output

> Unseen test set output, Safaba ES-to-EN TAUS Generic:

<table>
<thead>
<tr>
<th>#</th>
<th>Source</th>
<th>Translation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Utilice el tipo de aprovisionamiento en su factura para indicar quién proporciona los componentes. Por ejemplo, si su proveedor suministra componentes, además de la mano de obra, puede incluirlos en la factura con el tipo de aprovisionamiento Proveedor. Puede cargar una orden de fabricación o un programa repetitivo para estos componentes indicando su coste en el coste estándar del recurso externo de su ruta. O bien, si su proveedor carga estos componentes por separado, puede añadir otro recurso externo específico para estos costes.</td>
<td>Use the supply type on your invoice to indicate who provides components. For example, if your supplier provides components, and the labor, you can include in the invoice with a supply type of Supplier. You can charge a job or repetitive schedule for these components by providing the cost on the standard cost for the outside resource on your routing. Or, if your supplier charges these components separately, you can add another specific external resource costs.</td>
<td>Use the supply type on your bill to indicate who supplies components. For example, if your supplier provides components in addition to labor, you can include the components on your bill with a Supplier supply type. You can charge a job or repetitive schedule for these components by including their cost in the standard cost of the outside resource on your routing. Or, if your supplier charges you separately for these components, you can add another outside resource specifically for these costs.</td>
</tr>
<tr>
<td>2</td>
<td>Uso de Detalles de Selección</td>
<td>Using Selection Details</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Puede probar cambios pendientes si un proceso de modificación se ha interrumpido y determinar si debe utilizar el método Update o CancelUpdate.</td>
<td>You can test for pending changes if an editing process has been interrupted and determine whether you need to use the Update or CancelUpdate method.</td>
<td>You can test for pending changes if an editing process has been interrupted and determine whether you need to use the Update or CancelUpdate method.</td>
</tr>
<tr>
<td>4</td>
<td>Número de años anteriores</td>
<td>Number of past years</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Armas</td>
<td>Arms</td>
<td>4 Number of past years</td>
</tr>
<tr>
<td>6</td>
<td>Controlador IPSEC</td>
<td>IPSEC Driver</td>
<td>6 IPSEC driver</td>
</tr>
</tbody>
</table>
> Used Safaba default EN-to-DE pipeline to develop a WMT-2014 EN-to-DE MT system, as a contrastive reference to our TAUS EN-to-DE system

> Safaba WMT system:
  > Phrase-based system with domain adaptation
  > Constrained WMT-2014 parallel data resources only
  > No extra monolingual data for LM (i.e. GigaWord or CommonCrawl)
  > News Commentary as “in-domain”, everything else as “background”
  > Resulting system scores 17.3 cased BLEU (best system is 20.6)

> Training Statistics:

<table>
<thead>
<tr>
<th></th>
<th>WMT 2014</th>
<th>TAUS Generic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Segments</td>
<td>4,143,962</td>
<td>5,767,915</td>
</tr>
<tr>
<td>Training Tokens (EN)</td>
<td>106,951,743</td>
<td>85,331,463</td>
</tr>
<tr>
<td>Training Tokens (DE)</td>
<td>101,810,648</td>
<td>89,190,947</td>
</tr>
<tr>
<td>Average tokens/segment EN</td>
<td><strong>25.8</strong></td>
<td><strong>14.8</strong></td>
</tr>
<tr>
<td>Average tokens/segment DE</td>
<td>24.6</td>
<td>15.5</td>
</tr>
<tr>
<td>Global length ratio DE/EN</td>
<td>95.2%</td>
<td>104.5%</td>
</tr>
</tbody>
</table>
WMT vs TAUS: EN-to-DE MT Systems

TAUS vs. WMT Input Length Distribution

[Bar chart showing the distribution of input lengths for TAUS and WMT 2014]
WMT vs TAUS: EN-to-DE MT Systems

TAUS vs. WMT Input Length Distribution (cdf)

<table>
<thead>
<tr>
<th></th>
<th>WMT 2014</th>
<th>TAUS Generic</th>
</tr>
</thead>
<tbody>
<tr>
<td># training segments</td>
<td>4,143,962</td>
<td>5,767,915</td>
</tr>
<tr>
<td># training tokens EN</td>
<td>106,951,743</td>
<td>85,331,463</td>
</tr>
<tr>
<td># training tokens DE</td>
<td>101,810,648</td>
<td>89,190,947</td>
</tr>
<tr>
<td># of alignment links, gd</td>
<td>91,519,169</td>
<td>85,607,364</td>
</tr>
<tr>
<td>Average links per token EN</td>
<td>0.856</td>
<td>1.003</td>
</tr>
<tr>
<td>Average links per token DE</td>
<td>0.899</td>
<td>0.960</td>
</tr>
</tbody>
</table>

**TAUS vs. WMT Alignment Link Distribution (EN)**

<table>
<thead>
<tr>
<th>Alignment Links for Token</th>
<th>Fraction of Source Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>1</td>
<td>10.00%</td>
</tr>
<tr>
<td>2</td>
<td>20.00%</td>
</tr>
<tr>
<td>3–10</td>
<td>80.00%</td>
</tr>
</tbody>
</table>

**TAUS vs. WMT Alignment Link Distribution (DE)**

<table>
<thead>
<tr>
<th>Alignment Links for Token</th>
<th>Fraction of Target Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>1</td>
<td>10.00%</td>
</tr>
<tr>
<td>2</td>
<td>20.00%</td>
</tr>
<tr>
<td>3–10</td>
<td>80.00%</td>
</tr>
</tbody>
</table>
## WMT vs TAUS: EN-to-DE MT Systems

> Phrase Extraction Statistics:

<table>
<thead>
<tr>
<th></th>
<th>WMT 2014</th>
<th>TAUS Generic</th>
</tr>
</thead>
<tbody>
<tr>
<td># training tokens EN</td>
<td>106,951,743</td>
<td>85,331,463</td>
</tr>
<tr>
<td># training tokens DE</td>
<td>101,810,648</td>
<td>89,190,947</td>
</tr>
<tr>
<td>Total extracted phrase instances</td>
<td>652,123,624</td>
<td>374,142,109</td>
</tr>
<tr>
<td>Average phrases/token EN</td>
<td>6.10</td>
<td>4.38</td>
</tr>
<tr>
<td>Average phrases/token DE</td>
<td>6.41</td>
<td>4.19</td>
</tr>
<tr>
<td>Unique phrases EN</td>
<td>156,911,242</td>
<td>80,497,425</td>
</tr>
<tr>
<td>Unique phrases DE</td>
<td>168,034,534</td>
<td>97,586,721</td>
</tr>
<tr>
<td>Average instances per phrase EN</td>
<td>4.16</td>
<td>4.65</td>
</tr>
<tr>
<td>Average instances per phrase DE</td>
<td>3.88</td>
<td>3.83</td>
</tr>
<tr>
<td>Total unique phrase pairs</td>
<td>503,220,418</td>
<td>177,760,867</td>
</tr>
<tr>
<td>Average instances per phrase pair</td>
<td>1.30</td>
<td>2.10</td>
</tr>
<tr>
<td>Average translations per phrase EN</td>
<td>3.21</td>
<td>2.21</td>
</tr>
<tr>
<td>Average translations per phrase DE</td>
<td>2.99</td>
<td>1.82</td>
</tr>
</tbody>
</table>
> Phrase Count Distribution Statistics:

<table>
<thead>
<tr>
<th>Phrase Pair Count Histogram:</th>
<th>WMT 2014</th>
<th>TAUS</th>
<th>WMT 2014</th>
<th>TAUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>485,511,302</td>
<td>137,309,184</td>
<td>96.48%</td>
<td>77.24%</td>
</tr>
<tr>
<td>2</td>
<td>10,193,347</td>
<td>26,380,365</td>
<td>2.03%</td>
<td>14.84%</td>
</tr>
<tr>
<td>3</td>
<td>2,710,843</td>
<td>5,760,614</td>
<td>0.54%</td>
<td>3.24%</td>
</tr>
<tr>
<td>4</td>
<td>1,291,623</td>
<td>3,019,769</td>
<td>0.26%</td>
<td>1.70%</td>
</tr>
<tr>
<td>5+</td>
<td>3,513,303</td>
<td>5,290,935</td>
<td>0.70%</td>
<td>2.98%</td>
</tr>
</tbody>
</table>

TAUS vs. WMT Phrase Pair Count Distribution
WMT vs TAUS: EN-to-DE MT Systems

Phrase Translation Ambiguity

TAUS vs. WMT Targets Per Source (cdf)
Test-set Decoding Statistics:

<table>
<thead>
<tr>
<th></th>
<th>WMT 2014 newstest2012</th>
<th>WMT 2014 newstest2014</th>
<th>TAUS Generic test</th>
</tr>
</thead>
<tbody>
<tr>
<td># test set segments</td>
<td>3003</td>
<td>2737</td>
<td>1200</td>
</tr>
<tr>
<td># test set source types</td>
<td>10267</td>
<td>9650</td>
<td>4554</td>
</tr>
<tr>
<td># test set source tokens</td>
<td>73643</td>
<td>62871</td>
<td>19332</td>
</tr>
<tr>
<td>Average test set tokens/segment</td>
<td>24.5</td>
<td>23.0</td>
<td>16.1</td>
</tr>
<tr>
<td># decoder phrases used on test set</td>
<td>39982</td>
<td>34631</td>
<td>8642</td>
</tr>
<tr>
<td>Average decoder source phrase length</td>
<td>1.84</td>
<td>1.82</td>
<td>2.24</td>
</tr>
<tr>
<td># test set OOV types</td>
<td>450</td>
<td>493</td>
<td>82</td>
</tr>
<tr>
<td># test set OOV tokens</td>
<td>720</td>
<td>797</td>
<td>83</td>
</tr>
<tr>
<td>OOV rate (types / type)</td>
<td>4.38%</td>
<td>5.11%</td>
<td>1.80%</td>
</tr>
<tr>
<td>OOV rate (tokens / running token)</td>
<td>0.98%</td>
<td>1.27%</td>
<td>0.43%</td>
</tr>
<tr>
<td>Test set BLEU</td>
<td>15.0</td>
<td>17.1</td>
<td>52.5</td>
</tr>
<tr>
<td>Test set METEOR</td>
<td>34.8</td>
<td>38.8</td>
<td>63.5</td>
</tr>
<tr>
<td>Test set TER</td>
<td>67.9</td>
<td>66.5</td>
<td>38.5</td>
</tr>
<tr>
<td>Test set length ratio (MT/Ref)</td>
<td>97.7</td>
<td>102.8</td>
<td>100.8</td>
</tr>
</tbody>
</table>
WMT vs TAUS: EN-to-DE MT Systems

**TAUS vs. WMT Segment-Level METEOR Score Distribution**

Unstructured text is not necessary for this image.
WMT vs TAUS: EN-to-DE MT Systems

TAUS vs. WMT Decoder Phrase Length Distribution

Fraction of phrase pair instances used on test set

Phrase pair length (source tokens)

TAUS (test)  WMT (nt2014)
What explains the dramatic difference in translation quality between these two setups?

Consistent domain(s) versus broad domain
- Much lower OOV rates for TAUS (0.43% vs. 1.27%)
- Longer phrase matches for TAUS (average 2.24 vs. 1.82)
- Significantly more frequently-occurring phrases for TAUS
- Lower translation ambiguity for TAUS (2.21 vs. 3.21)

Indirect evidence for significantly “cleaner” and more parallel training data
- Denser word alignments for TAUS (1.003 vs. 0.856 links per EN token)
- Significantly fewer unaligned words for TAUS (9.39% vs. 22.27%)
- Significantly more frequently-occurring phrases for TAUS
- Lower translation ambiguity for TAUS (2.21 vs. 3.21)
- TAUS primary data source is highly-QAed commercial TMs

Shorter input segments allow limited-window reordering models to cover a significantly larger fraction of the data

Conclusion: TAUS data is a cleaner, higher-quality and potentially more suitable data source for “clean-lab” experiments with advanced translation models with results having potentially significant commercial relevance.
Multilingual Meteor

- http://www.cs.cmu.edu/~alavie/METEOR/
- We extensively use Meteor at Safaba
  - As an MT evaluation toolkit
  - As a monolingual aligner with flexible matches
Meteor has expanded to cover 17 languages:

### Fully supported languages:

<table>
<thead>
<tr>
<th>Language</th>
<th>Exact Match</th>
<th>Stem Match</th>
<th>Synonym Match</th>
<th>Paraphrase Match</th>
<th>Tuned Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Arabic</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Czech</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>French</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>German</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Spanish</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Partially supported languages:

<table>
<thead>
<tr>
<th>Language</th>
<th>Exact Match</th>
<th>Stem Match</th>
<th>Synonym Match</th>
<th>Paraphrase Match</th>
<th>Tuned Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danish</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Dutch</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Finnish</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Italian</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Norwegian</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Portuguese</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Romanian</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Russian</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Swedish</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
<tr>
<td>Turkish</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>LI</td>
</tr>
</tbody>
</table>
> New support included in Meteor 1.5:
> > Support for **any target language** using only **bi-text** used to build statistical MT systems
> > Learn paraphrases by phrase pivoting (Bannard and Callison-Burch, 2005)
> > Learn function words by relative frequency in monolingual data
> > **Universal parameter set** learned by pooling data from **all WMT languages**
> > **Significantly outperforms** baseline metrics on **unseen languages** with **no development data**.

<table>
<thead>
<tr>
<th>After a sharp <strong>drop</strong> in the morning ...</th>
<th>Después de la rápida <strong>caída</strong> de la mañana ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>... Una <strong>caída</strong> de volumen parecido se registró por última vez ...</td>
<td>... having registered a similarly-ranked <strong>fall</strong> the last time ...</td>
</tr>
<tr>
<td>Learning paraphrase (“drop”, “fall”) by pivoting through “caída”</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>The</strong> weight of one of the world’s longest-running conflicts ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>All <strong>of this is</strong> designed to reinforce one point: <strong>the</strong> Gaza ...</td>
</tr>
<tr>
<td>For <strong>the</strong> source of the problem is neither <strong>the</strong> European ...</td>
</tr>
<tr>
<td>So it <strong>is</strong> surprising that <strong>this</strong> choice <strong>is</strong> not at <strong>the</strong> center of ...</td>
</tr>
<tr>
<td>Learning function words “the”, “of”, “is”, “this” by high frequency</td>
</tr>
</tbody>
</table>
Safaba – MT Architecture Overview

Development Platform

- Development Frontend
- MT Development Workflow
- Deployment

Production Platform

- Safaba MT API and Integration Connectors
- Safaba Server Frontend
- DB
- MT System Production Cloud
Main Alternatives:

- **train-factored-model.perl**
  - For Moses, fossilized 9 steps

- **Experiment.perl**
  - For Moses, customizable

- **LoonyBin [Clark and Lavie, 2009]**
  - General-purpose, customizable

- **DuctTape**
  - Unix-based workflow management system for experimental NLP pipelines
  - General-purpose, customizable, with nice execution properties
  - Open-source, initial development by Jonathan Clark
  - [https://github.com/jhclark/ducttape](https://github.com/jhclark/ducttape)
> Break long pipelines into series of tasks: small block of arbitrary Bash code
> Specify inputs, outputs, configuration parameters, and what tools are required for each task
> Designed to easily test multiple settings via branch points
> DuctTape runs everything in the right order
task align_mkcls_src : mgiza
< corpus=$train_src_for_align
> classes
:: num_classes=50
:: num_runs=2
{
zcat -f $corpus > corpus
$mgiza/bin/mkcls -c num_classes -n num_runs \
   -pcorpus -V classes opt
rm corpus
}

task align_mgiza_direction : mgiza
< src_classes=$classes@align_mkcls_src
< tgt_classes=$classes@align_mkcls_tgt
< ...
> src_tgt_alignments
:: ...
{
   ...
}
DuctTape: Tasks

task **align_mkcls_src** : mgiza
< corpus=${train_src_for_align}
> **classes**
:: num_classes=50
:: num_runs=2
{}
  zcat -f $corpus > corpus
  $mgiza/bin/mkcls -c$num_classes -n$num_runs \\
  -pcorpus -V$classes opt
  rm corpus

  task **align_mgiza_direction** : mgiza
< src_classes=${classes}@align_mkcls_src
< tgt_classes=${classes}@align_mkcls_tgt
< ...
> src_tgt_alignments
:: ...
{}
...
task align_mkcls_src : mgiza
  < corpus=$train_src_for_align
  > classes
    :: num_classes=50
    :: num_runs=2
  {
    zcat -f $corpus > corpus
    $mgiza/bin/mkcls -c$num_classes -n$num_runs -pcorpus -V$classes opt
    rm corpus
  }

task align_mgiza_direction : mgiza
  < src_classes=${classes}@align_mkcls_src
  < tgt_classes=${classes}@align_mkcls_tgt
  < ...
  > src_tgt_alignments
    :: ...
  {
    ...
  }

DuctTape: Tasks
DuctTape: Tasks
task align_mkcls_src : mgiza
  < corpus=$train_src_for_align
  > classes
    :: num_classes=50
    :: num_runs=2
  {
    zcat -f $corpus > corpus
    $mgiza/bin/mkcls -c$num_classes -n$num_runs -pcorpus -V$classes opt
    rm corpus
  }

task align_mgiza_direction : mgiza
  < src_classes=$classes@align_mkcls_src
  < tgt_classes=$classes@align_mkcls_tgt
  < ...
  > src_tgt_alignments
    :: ...
  {
    ...
  }

DuctTape: Branch Points

task align_mkcls_src : mgiza
< corpus=$train_src_for_align>
> classes
:: num_classes=(Classes: small=50 large=1000)
:: num_runs=2
{
zcat -f $corpus > corpus
$mgiza/bin/mkcls -c$num_classes -n$num_runs -pcorpus -V$classes opt
rm corpus
}

num_classes=50
num_classes=1000

task align_mgiza_direction : mgiza
< src_classes=$classes@align_mkcls_src>
< tgt_classes=$classes@align_mkcls_tgt
< ...
> src_tgt_alignments
:: ...
{
...
}
DuctTape: Workflows

DuctTape workflow

ducttape workflow.tape -C myparams.ini
Safaba MT Deployment Process

Deployment involves:

- Packaging a Safaba MT system coming out of the development process
- Staging the system for production
- Migrating the system to our production platform
- Activating the system within production

Packaging:

- Generating a software container with local copies of all data files, software modules and parameter files required to run the MT system in production

Staging:

- The MT system is staged locally as a real-time translator for rigorous functionality and unit-testing

Migration:

- Secure rsync transfer of the staged MT system to the Safaba production platform

Activation:

- Updating of runtime DB and configuration files, and MT engine launch in production
Safaba EMTGlobal™ Online

> Web-based overlay platform and UI that supports remote development, deployment and runtime access and monitoring of Safaba EMTGlobal MT systems

> Provides functionality similar to MS Hub and other cloud-based MT development platforms

> Primary Use Cases:
  > DIY MT Platform for select Safaba clients and partners
  > Monitoring and Testing platform for our end clients
  > Safaba system demonstrations
  > Internal training and development
Welcome user!

Systems

View and access all your systems currently deployed in production. Here you can monitor, restart, redeploy, reconfigure and retrain active systems.

Access the instant translator by selecting a system from the list of available systems.

Development

Develop and deploy new translation systems. Here you can upload training data, set system configuration and manage deployment for each individual translation system.

Note that access to certain screens and functions may be restricted in line with user settings.

Settings

Manage users and permissions.

This area is accessible only to the system administrator.
Select Project / System

Available Translation Systems:

Demo:

<table>
<thead>
<tr>
<th>SYSTEM NAME</th>
<th>TYPE</th>
<th>OWNER</th>
<th>STATUS</th>
<th>ACTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENUS-DEDE, IT Hardware</td>
<td>FULL</td>
<td>User (Read-Only)</td>
<td>Deployed</td>
<td></td>
</tr>
<tr>
<td>Version: 14.02.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version: 14.02.19</td>
<td>FULL</td>
<td>User (Read-Only)</td>
<td>Ready to build</td>
<td>Remove</td>
</tr>
<tr>
<td>Version: 14.02.20</td>
<td>FULL</td>
<td>User (Read-Only)</td>
<td>Configuring initial settings</td>
<td>Remove</td>
</tr>
<tr>
<td>ENUS-FRCA, My system</td>
<td>FULL</td>
<td>User (Read-Only)</td>
<td>Deployed</td>
<td></td>
</tr>
<tr>
<td>Version: 14.03.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version: 14.03.10</td>
<td>FULL</td>
<td>Expert (Read-Only)</td>
<td>Configuring language settings</td>
<td>View</td>
</tr>
<tr>
<td>Version: 14.03.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version: 14.04.21</td>
<td>FULL</td>
<td>Expert (Read-Only)</td>
<td>Uploading data</td>
<td>View</td>
</tr>
<tr>
<td>ENUS-HEIL, Yakov</td>
<td>MPE</td>
<td>Expert (Read-Write)</td>
<td>Uploading data</td>
<td>Remove</td>
</tr>
<tr>
<td>Version: 14.05.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESES-ENUS, Software Support</td>
<td>MPE</td>
<td>User (Read-Only)</td>
<td>Configuring styling settings</td>
<td>Remove</td>
</tr>
<tr>
<td>Version: 14.03.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Safaba EMTGlobal™ Online

DEDE-ENUS, IT Software & Services
Translation System: DEDE-ENUS, IT Software & Services
System ID: 12

Description:
This translation system is provided as a reference for EMTGlobal Online™ users. The system, developed by Safaba's MT experts, can serve as a baseline for individually customized translation systems by selecting it from the 'Reference Systems' drop down menu of the Development workflow's 'Initial Settings' screen.

Build Time: 2018-08-29
Deployed On: boon3.safaba.com
Built By: 

Sentences Translated: 359
Words Translated: 4561
Speed: 1790.35 ms/sentence
Speed: 7.10 words/sec
Last Translation Time: --
Last Input String: --
Managed: Yes

Show credentials

![Graphs and charts showing performance metrics for the translation system.](chart.png)
External Workflow Integrations

Client CMS/TMS Platform

Content Management ➔ Translation Management Workflow ➔ Content Publication

- Translation Memory
- Safaba Connector
- CAT Translation UI

Safaba MT System Production Cloud

- MT
- MT
- MT
- MT
- MT
- MT
- MT
- MT
- MT
- MT

Safaba Production Platform

Safaba MT API and Integration Connectors ➔ Safaba Server Frontend ➔ DB
Translation with MT Post-Editing

Translation Setup:
> Source document is pre-translated by translation memory matches augmented by Safaba MT
> Translation Memory “fuzzy match” threshold typically set at 75-85%
> Pre-translations are presented to human translator as starting point for editing; translators can use or ignore the suggested pre-translations

Training:
> Translation teams typically receive training in MT post-editing

Post-Editing Productivity Assessment:
> Contrastive translation projects that measure and compare translation team productivity with MT post-editing versus translation using just translation memories
> Productivity measured by contrasting translated words per hour under both conditions: MT-PE throughput / HT throughput
MT Post-Editing Productivity Assessment

> Evaluated by Welocalize in the context of our joint Dell MT Program
Commercial enterprise translation data is often in the form of files in structured formats converted for translation into XML-based schemas (i.e. XLIFF and TMX) with tag-annotated segments of source text.

Correctly projecting and placing these segment-internal tags from the source language to the target language is a well-known difficult challenge for MT in general, and statistical MT engines in particular.

Safaba has focused significant effort to developing advanced high-accuracy algorithms for source-to-target tag projection within our EMTGlobal MT solution.

Example:
Source (EN): Click the <g0>Advanced</g0> tab, and click <g1>Change</g1>.
Reference (PT): Clique no separador <g0>Avançado</g0> e em <g1>Alterar</g1>.
Structured Tag Projection Process:

les ordinateurs de bureau <x id="1">les plus populaires</x> pour l'école et la maison
Structured Tag Projection Process:
  > Strip out all internal tags from the input and remember their original contexts.

```
les ordinateurs de bureau <x id="1">les plus populaires</x> pour l’école et la maison
```

```
<x id="1">
les ordinateurs de bureau
les plus populaires
pour l’école et la maison
</x>
```
Structured Tag Projection Process:

> Translate pure text segment and preserve word and phrase alignments.

<table>
<thead>
<tr>
<th>les ordinateurs de bureau</th>
<th>les plus populaires</th>
<th>pour l’école et la maison</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;x id=&quot;1&quot;&gt;</td>
<td></td>
<td>&lt;/x&gt;</td>
</tr>
<tr>
<td>les ordinateurs de bureau</td>
<td>les plus populaires</td>
<td>pour l’école et la maison</td>
</tr>
<tr>
<td>&lt;x id=&quot;1&quot;&gt;</td>
<td></td>
<td>&lt;/x&gt;</td>
</tr>
<tr>
<td>les plus populaires</td>
<td>les ordinateurs de bureau</td>
<td>pour l’école et la maison</td>
</tr>
<tr>
<td>popular</td>
<td>desktops</td>
<td>for school and home</td>
</tr>
</tbody>
</table>
Structured Tag Projection Process:

Reinsert tags with rules based on alignments, contexts and tag types.

- les ordinateurs de bureau <x id="1">les plus populaires</x> pour l’école et la maison
- <x id="1"></x>
- les ordinateurs de bureau les plus populaires pour l’école et la maison
- <x id="1"></x>
- les plus populaires les ordinateurs de bureau pour l’école et la maison
- popular desktops for school and home
- <x id="1">popular</x> desktops for school and home
- <x id="1"></x> desktops for school and home
Goal: Assess tag projection and placement accuracy of EMTGlobal version 1.1 versus 2.1, based on analysis of post-edited MT segments generated by Welocalize for Safaba’s eDell MT engines in production.

Methodology: Estimate accuracy by aligning the target language raw MT output with the post-edited MT version and assess whether each tag is placed between the same target words on both sides.

Example:

Reference: Clique no separador <g0>Avançado</g0> e em <g1>Alterar</g1>.

EMTGlobal v1.1: <g0>Clique na guia Avançado e em</g0> <g1>Alterar</g1>.

EMTGlobal v2.1: Clique na guia <g0>Avançado</g0> e em <g1>Alterar</g1>.
Tag Projection Accuracy Evaluation

[Beregovaya, Lavie and Denkowski, MT Summit 2013]

EMTGlobal version 1.1

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Context Matched</th>
<th>Both</th>
<th>Left</th>
<th>Right</th>
<th>Neither</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td></td>
<td>33.33%</td>
<td>19.44%</td>
<td>11.46%</td>
<td>35.76%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Ending</td>
<td></td>
<td>32.06%</td>
<td>10.10%</td>
<td>8.01%</td>
<td>49.83%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Stand-alone</td>
<td></td>
<td>56.91%</td>
<td>23.98%</td>
<td>18.29%</td>
<td>0.81%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>39.95%</td>
<td>17.54%</td>
<td>12.30%</td>
<td>30.21%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

EMTGlobal version 2.1

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Contexts Matched</th>
<th>Both</th>
<th>Left</th>
<th>Right</th>
<th>Neither</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td></td>
<td>66.67%</td>
<td>12.50%</td>
<td>9.38%</td>
<td>11.46%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Ending</td>
<td></td>
<td>63.41%</td>
<td>10.80%</td>
<td>11.50%</td>
<td>14.29%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Stand-alone</td>
<td></td>
<td>67.89%</td>
<td>18.29%</td>
<td>13.01%</td>
<td>0.81%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>65.90%</td>
<td>13.64%</td>
<td>11.21%</td>
<td>9.26%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

> Fraction of likely **incorrectly placed tags reduced from 30% to 9%**
> Fraction of confirmed **correctly placed tags improved from 40% to 66%**
> Fraction of tags with partially-matched contexts reduced from 30% to 25%

> **Data:** Welocalize post-editing productivity data set
  > 26 target languages, one document per language, 4907 segments
  > For 15 languages (3211 segments), EMTGlobal v1.1 was post-edited
  > For 11 languages (1696 segments), EMTGlobal v2.1 was post-edited
  > Total of 830 tags in PE segments, 821 aligned with MT output (98.9%)
The majority of the MT systems Safaba develops are specifically developed and optimized for specific client content types.

Data Scenario:
- Some amount of client-specific data: translation memories, terminology glossaries and monolingual data resources.
- Additional domain-specific and general background data resources: other client-specific content types, TAUS data, other general parallel and monolingual background data.

Safaba Collection of Adaptation Approaches:
- Data selection, filtering and prioritization methods.
- Data mixture and interpolation methods.
- Model mixture and interpolation methods.
- Client-specific Automated Post-Editing (Language Optimization Engine).
- Styling and Formatting post-processing modules.
- Terminology and DNT runtime overrides.
Challenge: Content Drift

Client-specific systems often degrade in performance over time for two main reasons:

1. Client content, even in controlled-domains, gradually changes over time: new products, new terminology, new content developers
2. The typical integrated setup of MT and translation memories: TMs are updated more frequently, so only “harder” segments are sent to MT

We see strong evidence of “content drift” over time with many of our clients, especially in post-editing setups.

The ongoing generation of new translated content with MT post-editing provides opportunities for generating an MT feedback loop – retrain and/or adapt the MT systems on an ongoing basis.

This motivates our focus on ongoing adaptation approaches.
Challenge: Content Drift

> Evidence from a typical client-specific MT system:
> EN-to-DE MT System - original and retrained systems:
  > February 2013 System: 565K client + 964K background segments
  > March 2014 System: 594K client + 6,795K background segments (including 140K “aged-out” client segments)

> Two test sets:
  > “Original” test set from February 2013 system build (1,200 segments)
  > “Incremental” test set extracted from incremental data (500 segments)

> System Test Scores and Statistics:

<table>
<thead>
<tr>
<th>Lang</th>
<th>System</th>
<th>Gloss Inconsist.</th>
<th>Orig. BLEU</th>
<th>Orig. MET</th>
<th>Orig. TER</th>
<th>Orig. LEN</th>
<th>Orig. OOVs</th>
<th>Incr. BLEU</th>
<th>Incr. MET</th>
<th>Incr. TER</th>
<th>Incr. LEN</th>
<th>Incr. OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>Feb. 2013</td>
<td>55.7 %</td>
<td>51.0</td>
<td>63.4</td>
<td>38.2</td>
<td>101.2</td>
<td>63</td>
<td>41.7</td>
<td>56.6</td>
<td>45.0</td>
<td>101.2</td>
<td>107</td>
</tr>
<tr>
<td>DE</td>
<td>March 2014</td>
<td>24.8 %</td>
<td>52.9</td>
<td>64.2</td>
<td>36.9</td>
<td>100.5</td>
<td>33</td>
<td>60.5</td>
<td>69.9</td>
<td>30.3</td>
<td>99.9</td>
<td>31</td>
</tr>
</tbody>
</table>
Evidence from a typical client-specific MT system:

EN-to-DE MT System - original and retrained systems:

- February 2013 System: 565K client + 964K background segments
- March 2014 System: 594K client + 6,795K background segments (including 140K “aged-out” client segments)

Two test sets:

- “Original” test set from February 2013 system build (1,200 segments)
- “Incremental” test set extracted from incremental data (500 segments)

System Test Scores and Statistics:

<table>
<thead>
<tr>
<th>Lang</th>
<th>System</th>
<th>Gloss Inconsist.</th>
<th>Orig. BLEU</th>
<th>Orig. MET</th>
<th>Orig. TER</th>
<th>Orig. LEN</th>
<th>Orig. OOVs</th>
<th>Incr. BLEU</th>
<th>Incr. MET</th>
<th>Incr. TER</th>
<th>Incr. LEN</th>
<th>Incr. OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>Feb. 2013</td>
<td>55.7 %</td>
<td>51.0</td>
<td>63.4</td>
<td>38.2</td>
<td>101.2</td>
<td>63</td>
<td>41.7</td>
<td>56.6</td>
<td>45.0</td>
<td>101.2</td>
<td>107</td>
</tr>
<tr>
<td>DE</td>
<td>March 2014</td>
<td>24.8 %</td>
<td>52.9</td>
<td>64.2</td>
<td>36.9</td>
<td>100.5</td>
<td>33</td>
<td>60.5</td>
<td>69.9</td>
<td>30.3</td>
<td>99.9</td>
<td>31</td>
</tr>
</tbody>
</table>
Challenge: Content Drift

> Evidence from a typical client-specific MT system:
> EN-to-DE MT System - original and retrained systems:
  > February 2013 System: 565K client + 964K background segments
  > March 2014 System: 594K client + 6,795K background segments (including 140K “aged-out” client segments)

> Two test sets:
  > “Original” test set from February 2013 system build (1,200 segments)
  > “Incremental” test set extracted from incremental data (500 segments)

> System Test Scores and Statistics:

<table>
<thead>
<tr>
<th>Lang</th>
<th>System</th>
<th>Gloss Inconsist.</th>
<th>Orig. BLEU</th>
<th>Orig. MET</th>
<th>Orig. TER</th>
<th>Orig. LEN</th>
<th>Orig. OOVs</th>
<th>Incr. BLEU</th>
<th>Incr. MET</th>
<th>Incr. TER</th>
<th>Incr. LEN</th>
<th>Incr. OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>Feb. 2013</td>
<td>55.7 %</td>
<td>51.0</td>
<td>63.4</td>
<td>38.2</td>
<td>101.2</td>
<td>63</td>
<td>41.7</td>
<td>56.6</td>
<td>45.0</td>
<td>101.2</td>
<td>107</td>
</tr>
<tr>
<td>DE</td>
<td>March 2014</td>
<td>24.8 %</td>
<td>52.9</td>
<td>64.2</td>
<td>36.9</td>
<td>100.5</td>
<td>33</td>
<td>60.5</td>
<td>69.9</td>
<td>30.3</td>
<td>99.9</td>
<td>31</td>
</tr>
</tbody>
</table>
Challenge: Content Drift

- Evidence from a typical client-specific MT system:
  - EN-to-DE MT System - original and retrained systems:
    - February 2013 System: 565K client + 964K background segments
    - March 2014 System: 594K client + 6,795K background segments (including 140K “aged-out” client segments)
  - Two test sets:
    - “Original” test set from February 2013 system build (1,200 segments)
    - “Incremental” test set extracted from incremental data (500 segments)
  - System Test Scores and Statistics:

| Lang | System     | Gloss Inconsist. | Orig. BLEU | Orig. MET | Orig. TER | Orig. LEN | Orig. OOVs | Incr. BLEU | Incr. MET | Incr. TER | Incr. LEN | Incr. OOVs |
|------|------------|------------------|-----------|---------|---------|---------|---------|-----------|---------|---------|---------|---------|---------|
| DE   | Feb. 2013  | 55.7 %           | 51.0      | 63.4    | 38.2    | 101.2   | 63      | 41.7      | 56.6    | 45.0    | 101.2   | 107     |
| DE   | March 2014 | 24.8 %           | 52.9      | 64.2    | 36.9    | 100.5   | 33      | 60.5      | 69.9    | 30.3    | 99.9    | 31      |
Challenge: Content Drift

> Evidence from a typical client-specific MT system:

> EN-to-DE MT System - original and retrained systems:
  > February 2013 System: 565K client + 964K background segments
  > March 2014 System: 594K client + 6,795K background segments (including 140K “aged-out” client segments)

> Two test sets:
  > “Original” test set from February 2013 system build (1,200 segments)
  > “Incremental” test set extracted from incremental data (500 segments)

> System Test Scores and Statistics:

<table>
<thead>
<tr>
<th>Lang</th>
<th>System</th>
<th>Gloss Inconsist.</th>
<th>Orig. BLEU</th>
<th>Orig. MET</th>
<th>Orig. TER</th>
<th>Orig. LEN</th>
<th>Orig. OOVs</th>
<th>Incr. BLEU</th>
<th>Incr. MET</th>
<th>Incr. TER</th>
<th>Incr. LEN</th>
<th>Incr. OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>Feb. 2013</td>
<td>55.7 %</td>
<td>51.0</td>
<td>63.4</td>
<td>38.2</td>
<td>101.2</td>
<td>63</td>
<td>41.7</td>
<td>56.6</td>
<td>45.0</td>
<td>101.2</td>
<td>107</td>
</tr>
<tr>
<td>DE</td>
<td>March 2014</td>
<td>24.8 %</td>
<td>52.9</td>
<td>64.2</td>
<td>36.9</td>
<td>100.5</td>
<td>33</td>
<td>60.5</td>
<td>69.9</td>
<td>30.3</td>
<td>99.9</td>
<td>31</td>
</tr>
</tbody>
</table>
Challenge: Content Drift

> Evidence from a typical client-specific MT system:
> 
> EN-to-DE MT System - original and retrained systems:
> > February 2013 System: 565K client + 964K background segments
> > March 2014 System: 594K client + 6,795K background segments (including 140K “aged-out” client segments)

> Two test sets:
> > “Original” test set from February 2013 system build (1,200 segments)
> > “Incremental” test set extracted from incremental data (500 segments)

> System Test Scores and Statistics:

<table>
<thead>
<tr>
<th>Lang</th>
<th>System</th>
<th>Gloss Inconsist.</th>
<th>Orig. BLEU</th>
<th>Orig. MET</th>
<th>Orig. TER</th>
<th>Orig. LEN</th>
<th>Orig. OOVs</th>
<th>Incr. BLEU</th>
<th>Incr. MET</th>
<th>Incr. TER</th>
<th>Incr. LEN</th>
<th>Incr. OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>Feb. 2013</td>
<td>55.7 %</td>
<td>51.0</td>
<td>63.4</td>
<td>38.2</td>
<td>101.2</td>
<td>63</td>
<td>41.7</td>
<td>56.6</td>
<td>45.0</td>
<td>101.2</td>
<td>107</td>
</tr>
<tr>
<td>DE</td>
<td>March 2014</td>
<td>24.8 %</td>
<td>52.9</td>
<td>64.2</td>
<td>36.9</td>
<td>100.5</td>
<td>33</td>
<td>60.5</td>
<td>69.9</td>
<td>30.3</td>
<td>99.9</td>
<td>31</td>
</tr>
</tbody>
</table>
Objective: Counter “content drift” and help maintain and accelerate post-editing productivity with fast and frequent incremental adaptation retraining

Setting: New additional post-edited client data is deposited and made available for adaptation in small incremental batches

Challenge: Full offline system retraining is slow and computationally intensive and can take several days

Safaba Solution: implement fast “light-weight” adaptations that can be executed, tested and deployed into production within hours (“overnight”)

Suffix-array variant of Moses supports rapid updating of indexed training data

Safaba automated post-editing module supports rapid retraining

KenLM supports rapid rebuilding of language models

Currently in pilot testing with Welocalize and one of our major clients
Real-time Online Adaptation

- **Ultimate Goal**: immediate online feedback loop between MT post-editing and the live MT system in the background

- **Engineering Challenge**: requires a fully integrated online solution where the MT post-editors translation environment is directly connected to the real-time MT engine, and feeds back post-edited segments immediately back to the MT engine for online adaptation

- **MT Challenge**: extend training of all major MT system components to operate in online mode rather than batch mode

- Main focus of Michael Denkowski’s PhD thesis at LTI
- Fully implemented, fully online adapting MT system
- Recently published work:
Real-time Online Adaptation

Static MT System:
- **Grammar**: precompiled corpus level grammar (Chiang, 2005)
- **LM**: kndiscount N-gram model (Chen and Goodman, 1996)
- **Feature Weights**: batch (corpus-level) optimization with MERT (Och, 2003)

Online Adaptive MT System:
- **Grammar**: on-demand sentence level with online learning [Denkowski et al., 2014]
- **LM**: updateable Bayesian N-gram model [Denkowski et al., 2014]
- **Feature Weights**: online learning with MIRA [Chiang, 2012]
- **Online Adaptation**: Update all components immediately after each sentence is post-edited, before MT generated for next sentence
Real-time Online Adaptation
Real-time Online Adaptation

Static

Large LM

Large Bitext

Dynamic

PE Data

Weights

$W_1 \ldots W_n$

Grammar

Input Sentence

Decoder

Post-Editing

$lte$ safaba
Real-time Online Adaptation

- Large LM
- Large Bitext
- Bayesian LM
- PE Data
- Weights
- Grammar
- Input Sentence
- Decoder
- Post-Editing
Real-time Online Adaptation

> **Online Grammar Extraction:**
> Index bi-text with suffix array, extract sentence-level grammars on demand [Lopez, 2008]
> Index bilingual sentences from post-editing data in a separate suffix-array as they become available
> Grammar for each sentence learned using a sample from suffix array (S) and full locally-indexed post-editing data (L)

> **Grammar Rule Features:**

- $C_S(f, e), C_L(f, e)$: counts of $f$ aligning to $e$
- $C_S(f), C_L(f)$: counts of $f$ aligning to anything
- $|S|, |L|$: sample sizes (occurrences of $f$, aligned or not)
### Real-time Online Adaptation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Static</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>coherent $p(e</td>
<td>f)$</td>
<td>$\frac{C_S(f, e)}{</td>
</tr>
<tr>
<td>sample size</td>
<td>$</td>
<td>S</td>
</tr>
<tr>
<td>co-occurrence $\langle f, e \rangle$</td>
<td>$C_S(f, e)$</td>
<td>$C_S(f, e) + C_L(f, e)$</td>
</tr>
<tr>
<td>singleton $f$</td>
<td>$C_S(f) = 1$</td>
<td>$C_S(f) + C_L(f) = 1$</td>
</tr>
<tr>
<td>singleton $\langle f, e \rangle$</td>
<td>$C_S(f, e) = 1$</td>
<td>$C_S(f, e) + C_L(f, e) = 1$</td>
</tr>
<tr>
<td>post-edit support $\langle f, e \rangle$</td>
<td>0</td>
<td>$C_L(f, e) &gt; 0$</td>
</tr>
</tbody>
</table>

Phrase features (rule level)
> Tuning an Online Adaptive System Using Simulated Post-Editing:

> Real post-edited segments are not available during initial system training and tuning

> **Challenge:** How do we learn discriminative weights for our online features?

> **Solution:** Use pre-generated references in place of post-editing [Hardt and Elming, 2010]

<table>
<thead>
<tr>
<th>Source</th>
<th>Target (Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hola contestadora ...</td>
<td>Hello voicemail, my old ...</td>
</tr>
<tr>
<td>He llamado a servicio ...</td>
<td>I’ve called for tech ...</td>
</tr>
<tr>
<td>Ignoré la advertencia ...</td>
<td>I ignored my boss’ ...</td>
</tr>
<tr>
<td>Ahora anochece, y mi ...</td>
<td>Now it’s evening, and ...</td>
</tr>
<tr>
<td>Todavía sigo en espera ...</td>
<td>I’m still on hold. I’m ...</td>
</tr>
<tr>
<td>No creo que me hayas ...</td>
<td>I don’t think you ...</td>
</tr>
<tr>
<td>Ya he presionado cada ...</td>
<td>I punched every touch ...</td>
</tr>
</tbody>
</table>
Real-time Online Adaptation

> Simulated Post-Editing Experiments:
> Baseline MT system (cdec):
>   > Hierarchical phrase-based model with suffix array grammars
>   > Large Modified Kneser-Ney smoothed LM
>   > MIRA optimization
> Online Adaptive Systems:
>   > Update grammars, LM, and weights independently and in combination
> Training Data:
>   > WMT-2012 Spanish–English and NIST 2012 Arabic–English
> Evaluation Data:
>   > WMT News Commentary test sets and out-of-domain TED talks
## Real-time Online Adaptation

### Evaluation Results:

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Dev</th>
<th>In-Dom</th>
<th>Out-of-Dom</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spanish–English</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>29.2</td>
<td>28.0</td>
<td>32.7</td>
</tr>
<tr>
<td>Grammars</td>
<td><strong>29.8</strong></td>
<td><strong>28.3</strong></td>
<td><strong>34.2</strong></td>
</tr>
<tr>
<td>LM</td>
<td>29.2</td>
<td>28.1</td>
<td>33.0</td>
</tr>
<tr>
<td>MIRA</td>
<td>29.2</td>
<td>28.1</td>
<td>33.1</td>
</tr>
<tr>
<td>G+L+M</td>
<td><strong>30.0</strong></td>
<td><strong>28.8</strong></td>
<td><strong>35.2</strong></td>
</tr>
<tr>
<td><strong>Arabic–English</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>21.2</td>
<td>25.9</td>
<td>10.6</td>
</tr>
<tr>
<td>Grammars</td>
<td><strong>21.8</strong></td>
<td><strong>26.2</strong></td>
<td><strong>11.0</strong></td>
</tr>
<tr>
<td>LM</td>
<td>20.6</td>
<td>25.7</td>
<td>10.6</td>
</tr>
<tr>
<td>MIRA</td>
<td><strong>21.3</strong></td>
<td>25.7</td>
<td><strong>10.8</strong></td>
</tr>
<tr>
<td>G+L+M</td>
<td><strong>21.8</strong></td>
<td><strong>26.5</strong></td>
<td><strong>11.4</strong></td>
</tr>
</tbody>
</table>

**Note:** The values represent percentages of some metric (e.g., accuracy or score).
Real-time Online Adaptation

- Evaluation with Live Human Translator Post-Editing:
- Fully integrated adaptive MT system with TransCenter

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 En la pausa, varias personas me preguntaron</td>
<td>At the break, I was asked by several people</td>
<td>4 - Usable</td>
</tr>
<tr>
<td>2 acerca de mis comentarios sobre el debate en torno al envejecimiento.</td>
<td>about my comments about the aging debate.</td>
<td>4 - Usable</td>
</tr>
<tr>
<td>3 Y este será mi único comentario al respecto.</td>
<td>And this will be my only comment on the matter.</td>
<td>5 - Very Good</td>
</tr>
<tr>
<td>4 Y que es que, a mi entender</td>
<td>And that is, I understand</td>
<td>3 - Neutral</td>
</tr>
<tr>
<td>5 los optimistas viven mucho más que los pesimistas.</td>
<td>optimists live much more than the pessimists.</td>
<td>Rate Translation</td>
</tr>
<tr>
<td>6 (Risas)</td>
<td></td>
<td>Rate Translation</td>
</tr>
<tr>
<td>7 Lo que voy a contarles en mis dieciocho minutos es</td>
<td></td>
<td>Rate Translation</td>
</tr>
</tbody>
</table>
Evaluation with Live Human Translator Post-Editing:

Experimental Setup:
- Six translators post-edited 4 talk excerpts totaling 100 MT-generated segments
- Two excerpts translated by static system, two by adaptive system
- Evaluated post-editing effort (HTER) and translator rating of MT suitability

Results:
- Adaptive system significantly outperforms static baseline
- Compared to simulated post-editing with static references
- Small improvement in simulated scenario leads to significant improvement in our live scenario

<table>
<thead>
<tr>
<th></th>
<th>HTER</th>
<th>Rating</th>
<th>SPE BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>19.26</td>
<td>4.19</td>
<td>34.50</td>
</tr>
<tr>
<td>Adaptive</td>
<td>17.01</td>
<td>4.31</td>
<td>34.95</td>
</tr>
</tbody>
</table>
Concluding Remarks

> MT for Dissemination vs. MT for Assimilation: quite different!
> Commercially-relevant data such as TAUS data has some significant advantages for “clean lab” MT modeling research work
> Commercially-useful MT systems have unique requirements and introduce a broad range interesting problems for researchers to focus on:
  > High-accuracy translation of structured content
  > Translation of terminology-heavy content, respecting brand language and style
  > MT adaptation with limited amounts of client-specific data
  > Ongoing adaptation to address content drift
  > Optimizing MT post-editing productivity
  > Real-time online adaptation
> Safaba is doing some cool MT stuff!
Acknowledgements

> **My CMU collaborators, students and contributors:** Chris Dyer, Noah Smith, Michael Denkowski, Greg Hanneman, Austin Matthews, Jonathan Clark, Kenneth Heafield, Wes Feely and the c-lab group

> **My Safaba colleagues, current and former staff members and consultants:** Bob Olszewski, Udi Hershkovich, Jonathan Clark, Michael Denkowski, Greg Hanneman, Sagi Perel, Matt Fiorillo, Austin Matthews, Pradeep Dasigi, Wes Feely, Serena Jeblee, Callie Vaughn, Laura Kieras, Ryan Carlson, Kenton Murray, Chris Dyer.

> **Other Collaborators:** Isabel Lacruz and Greg Shreve @ Kent State; Olga Beregovaya and the team @ Welocalize