MoJo

Bringing Hybrid MT to the Center for Applied Machine Translation

Marianna J. Martindale,
Center for Applied Machine Translation (CAMT)
MT in Research & Industry vs CAMT

• Research is almost entirely StatMT (now Neural)
  • Emphasis on new techniques
  • Most research on high-resource language pairs (except LORELEI & MATERIAL)
  • Not concerned with operational constraints

• In industry StatMT is the norm (for now)
  • Primarily commercially viable language pairs (high-resource)
  • Speed is important, compute resources may or may not be

• CAMT’s GOTS MT is (currently) rule-based
  • Many languages regardless of resource availability
  • Speed is important, compute resources limited (server OR laptop)
  • Fidelity is more important than fluency
Why not StatMT before?

• Technical issues with StatMT
  • Speed
  • Memory
  • Well-engineered systems not readily available
  • Can be tricky to build right
Why not StatMT before?

• Technical issues with StatMT
  • Speed
  • Memory
  • Well-engineered systems not readily available
  • Can be tricky to build right
Why not StatMT before?

- Technical issues with StatMT
  - Speed
  - Memory
  - Well-engineered systems not readily available
  - Can be tricky to build right

- Domain needs
  - Many languages (often low-resource)
Languages Supported in CyberTrans

<table>
<thead>
<tr>
<th>Language</th>
<th>Language</th>
<th>Language</th>
<th>Language</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFRIKAANS</td>
<td>DANISH</td>
<td>JAPANESE</td>
<td>PERSIAN</td>
<td>TAJIK</td>
</tr>
<tr>
<td>ALBANIAN</td>
<td>DARI</td>
<td>JAVANESE</td>
<td>PERSIAN Romanized</td>
<td>TAUSUG</td>
</tr>
<tr>
<td>AMHARIC</td>
<td>DUTCH</td>
<td>KAZAKH</td>
<td>POLISH</td>
<td>TATSUM</td>
</tr>
<tr>
<td>ARABIC Romanized</td>
<td>ESTONIAN</td>
<td>KOREAN</td>
<td>PORTUGUESE</td>
<td>TETUM</td>
</tr>
<tr>
<td>ARMENIAN</td>
<td>FINNISH</td>
<td>KURDIST (Kurmanji)</td>
<td>PUNJABI</td>
<td>THAI</td>
</tr>
<tr>
<td>AZERBAIJANI</td>
<td>FRENCH</td>
<td>KURDIST (Sorani)</td>
<td>ROMANIAN</td>
<td>TOK PISIN</td>
</tr>
<tr>
<td>BALUCHI</td>
<td>GALICIAN</td>
<td>KYRGYZ</td>
<td>RUSSIAN</td>
<td>TURKISH</td>
</tr>
<tr>
<td>BASQUE</td>
<td>GEORGIAN</td>
<td>LAO</td>
<td>RUSSIAN Romanized</td>
<td>TURKMEN</td>
</tr>
<tr>
<td>BELARUSIAN</td>
<td>GEORGIAN Romanized</td>
<td>LATVIAN</td>
<td>SERBIAN</td>
<td>TURKMEN*</td>
</tr>
<tr>
<td>BULGARIAN</td>
<td>GREEK</td>
<td>LINGALA</td>
<td>SERBIAN Cyrillic</td>
<td>UKRAINIAN</td>
</tr>
<tr>
<td>BULGARIAN Romanized</td>
<td>GREEK Romanized</td>
<td>LITHUANIAN</td>
<td>SHONA*</td>
<td>UKRAINIAN Romanized</td>
</tr>
<tr>
<td>CATALAN</td>
<td>HAITIAN CREOLE</td>
<td>MACEDONIAN</td>
<td>SLOVAK</td>
<td>URDU</td>
</tr>
<tr>
<td>CEBUANO</td>
<td>HAUSSA</td>
<td>MACEDONIAN Romanized</td>
<td>SLOVENE</td>
<td>URDU Romanized</td>
</tr>
<tr>
<td>CHAVACANO</td>
<td>HEBREW</td>
<td>MAGUINDANAON</td>
<td>SOMALI</td>
<td>UYGHUR</td>
</tr>
<tr>
<td>CHECCHEN</td>
<td>HINDI</td>
<td>MALAGASY</td>
<td>SPANISH</td>
<td>UZBEK</td>
</tr>
<tr>
<td>CHINESE Simplified</td>
<td>Hmong</td>
<td>MALAYSIAN</td>
<td>SRANAN</td>
<td>UZBEK Romanized</td>
</tr>
<tr>
<td>CHINESE Traditional</td>
<td>HUNGARIAN</td>
<td>NORWEGIAN</td>
<td>SUNDANESE</td>
<td>VIETNAMESE</td>
</tr>
<tr>
<td>CROATIAN</td>
<td>INDONESIAN</td>
<td>PAPIAMENTO</td>
<td>SWAHILI</td>
<td>WOLOF</td>
</tr>
<tr>
<td>CZECH</td>
<td>ITALIAN</td>
<td>PASHTO</td>
<td>SWEDISH</td>
<td>YAKAN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PASHTO Romanized</td>
<td>TAGALOG</td>
<td>YORUBA</td>
</tr>
</tbody>
</table>
Why not StatMT before?

• Technical issues with StatMT
  • Speed
  • Memory
  • Well-engineered systems not readily available
  • Can be tricky to build right

• Domain needs
  • Many languages (often low-resource)
  • Little or no in-domain parallel text
  • Frequent sometimes urgent updates
  • Fidelity as priority (accurate, traceable)
Why not StatMT before?

• Technical issues with StatMT
  • Speed
  • Memory
  • Well-engineered systems not readily available
  • Can be tricky to build right

• Domain needs
  • Many languages (often low-resource)
  • Little or no in-domain parallel text
  • Frequent sometimes urgent updates
  • Fidelity as priority (accurate, traceable)

MoTrans
  • Human instead of bitext
  • Updated based on actual text submitted
  • Easy to trace input to output
  • Caveat: Sacrifice fluency for fidelity
Features of Rule-based and StatMT

**Rule-based**
- Rules are composed by language experts
- Performs a deep source language analysis
- Easy to update, adapt to new domains
- Easy to trace input to output
- Very fast

**StatMT**
- Learns automatically from example translations
- Doesn’t require language-specific knowledge
- Leverages Big Data
- More fluent translations
- Recent engineering advances make adoption easier

**Best of both worlds?**
Best of both worlds

MoTrans
Human constructed
Domain focused
Knowledge-rich
CAMT linguistic and
technical investment

Statistical
Learned automatically
Generic
Language-agnostic
Commercially dominant
Open source

Hybrid
Example (Russian)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Motrans</td>
<td>He noted, that presidential pre-election campaign provoked “discrepant and often frequently vulgar rhetoric,” eating away democracy and society.</td>
<td></td>
</tr>
<tr>
<td>StatMT</td>
<td>He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” разъедающую democracy and society .</td>
<td></td>
</tr>
<tr>
<td>Hybrid</td>
<td>He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” eating away democracy and society.</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>He said the presidential campaign has brought “divisive and often vulgar rhetoric” that corrodes democracy and society.</td>
<td></td>
</tr>
</tbody>
</table>
Example (Russian)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motrans</td>
<td>From Moscow to Sochi on the train about two days! Really? Is it possible? You want to lead two days in the uncomfortable train?</td>
</tr>
<tr>
<td>StatMT</td>
<td>From Moscow to Sochi to train about two days! Do you want to spend two days in awkward train?</td>
</tr>
<tr>
<td>Hybrid</td>
<td>From Moscow to Sochi on the train about two days! Do you want to spend two days in uncomfortable train?</td>
</tr>
<tr>
<td>Human</td>
<td>From Moscow to Sochi by train is close to 2 days! Do you really want to spend two days in an uncomfortable train?</td>
</tr>
</tbody>
</table>
## Example (Swahili)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motrans</td>
<td>LABLA America in/at what region? America is big.</td>
</tr>
<tr>
<td>StatMT</td>
<td>“Maybe America in what state? The United States is the greatest.</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Maybe America in what region? The United States is big.</td>
</tr>
<tr>
<td>Human</td>
<td>To be more precise, which state in America? America is vast.</td>
</tr>
</tbody>
</table>
## Example (Swahili)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motrans</td>
<td>Pat: say (you/it) succeeded to get children and MUMEO Mohammed HAMIS?</td>
</tr>
<tr>
<td>StatMT</td>
<td>Pat: is ulifanikiwa to children and mumeo Mohamed Hamis?</td>
</tr>
<tr>
<td>Hybrid</td>
<td>Pat: <em>Have you succeeded in</em> getting children mumeo Mohamed Hamis?</td>
</tr>
<tr>
<td>Human</td>
<td>Pat: Were you successful at having children by that husband Mohamed Hamis?</td>
</tr>
</tbody>
</table>
Best of both worlds

MoTrans

Joshua

MoJo
Apache Joshua

- Open-source Java statistical machine translation system
- Apache project currently in Incubation Stage
- Provides both phrase-based and hiero StatMT
- Pre-built language packs available for download
- **Written in Java** (like CyberTrans)
- **Development lead conveniently located at JHU HLTCOE in nearby Baltimore**

http://joshua.apache.org/
Best of both worlds

MoTrans

Joshua

MoJo
MoTrans Translator

- Morphological Translator
- Fast
- Deep morphological analysis
- Expressive lexicon and grammar
- Continually updated by lexicographers
- Quick “better than nothing” for Low Resource Languages
- Currently over 40 languages
- Many users, positive feedback
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR|sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla | N | lace
  • pas/ar | V.AR | pass
  • de | DE | of
  • pas de puntillas | V.AR | sidestep | S--

Source

Usually lemma
But not always

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla\(N\) lace
  • pas/ar \(V.AR\) pass
  • de \(DE\) of
  • pas de puntillas\(V.AR\) sidestep\(S--\)

• Rules
  PROCLITIC:\(V.CONJ\):ha:has \%s
  SUFFIX:\(V.AR\):ado:%s:+pastp:\(V.CONJ\)
  SUFFIX_PATTERN:\(N\):(\*V)(X-0\{s\}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-∅{s}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar|V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules

  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-∅{s}):%s:+plural
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC: V.CONJ: ha: has %s
  SUFFIX: V.AR: ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN: N: (*V)(X-0{s}):%s:+plural

Part-of-Speech
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC:V.CONJ: ha: has %s
  SUFFIX:V.AR: ado: %s:+pastp:V.CONJ
  SUFFIX_PATTERN:N: (*V)(X-0{s}): %s:+plural

Source Transformation
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha: has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-0{0}:%s:+plural

Target Transformation
MoTrans Lexicon Example

• Lexical entries
  • puntilla | N | lace
  • pas/ar | V.AR | pass
  • de | DE | of
  • pas de puntillas | V.AR | sidestep | S--

• Rules
  PROCLITIC: V.CONJ: ha: has %s
  SUFFIX: V.AR: ado: %s: + pastp: V.CONJ
  SUFFIX_PATTERN: N: (*V)(X-0{s}): %s: + plural

Target Conjugation
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR| sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(^V)(X-$0{s}):%s:+plural

New Part-of-Speech
MoTrans Lexicon Example

• Lexical entries
  • puntilla | N | lace
  • pas/ar | V.AR | pass
  • de | DE | of
  • pas de puntillas | V.AR | sidestep | S--

• Rules
  PROCLITIC: V.CONJ: ha: has %s
  SUFFIX: V.AR: ado: %s: + pastp: V.CONJ
  SUFFIX_PATTERN: N: (*V) (X-∅{s}): %s: + plural

For a noun that ends in a vowel
Add an ‘s’
Pluralize the English

puntillas -> lace
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas|V.AR|sidestep|S--

• Rules
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-0{s}):%s:+plural
MoTrans Lexicon Example

- **Lexical entries**
  - puntilla\textit{N} | lace
  - pas\textit{ar} |V.AR| pass
  - de |DE| of
  - pas de puntillas |V.AR| sidestep |S--

- **Rules**
  PROCLITIC:V.CONJ:ha:has %s
  SUFFIX:V.AR:ado:%s:+pastp:V.CONJ
  SUFFIX_PATTERN:N:(*V)(X-\theta\{s\}):%s:+plural

- puntillas ->
  lace

- ha pasado ->
  has passed

- ha pasado de puntillas ->
  has sidestepped
MoTrans Lexicon Example

• Lexical entries
  • puntilla|N| lace
  • pas/ar |V.AR| pass
  • de |DE| of
  • pas de puntillas |V.AR| sidestep |S--

• Rules
  PROCLITIC: V.CONJ: ha: has %s
  SUFFIX: V.AR: ado: %s: + pastp: V.CONJ
  SUFFIX_PATTERN: N: (*V)(X-\{S\}): %s: + plural

puntillas ->
  PUNTILLA

ha pasado ->
  PAS (PROCLITIC: V.CONJ: HA: has %s: + pastp)
  (SUFFIX: V.AR: ADO: %s: + pastp):
  V.CONJ: has passed

ha pasado de puntillas ->
  PAS DE PUNTILLAS
  (PROCLITIC: V.CONJ: HA: has %s: + pastp)
  (SUFFIX: V.AR: ADO: %s: + pastp):
  V.CONJ: has sidestepped
Best of both worlds

MoTrans

Joshua

MoJo
Building the Hybrid: Base StatMT

Parallel sentences (bitext)

English text

Translation Model

Language Model

New document

MT system

Translation
Building the Hybrid: Black Box

Parallel sentences (bitext)

Translation Model

Rule-based knowledge applied to domain data to feed StatMT

Language Model

English text

MT system

New document

Translation
Building the Hybrid: Black Box

Parallel sentences (bitext)

- English
- Foreign

English text

Translation Model

Language Model

Parallel “sentences” (pseudo-bitext)

Domain Data

“English”

MoTrans

New document

MT system

Translation
Building the Hybrid: Black Box

Parallel sentences (bitext)

- Foreign
- English

English text

- English text

Translation Model

- Language Model

Parallel “sentences” (pseudo-bitext)

- Domain Data
- “English”

MoTrans

New document

MT system

Translation

How do we update?


Austin, Oct 28 - Nov 1, 2016 | p. 692
Building the Hybrid: On Demand

Parallel sentences (bitext)

English text

Parallel “sentences” (pseudo-bitext)

New document

MoTrans

MT system

Translation

Translation Model

Language Model

MoTrans

Domain Data

“English”


Austin, Oct 28 - Nov 1, 2016 | p. 693
Building the Hybrid: Live Updates

Parallel sentences (bitext)

English text

Translation Model

Language Model

Parallel “sentences” (pseudo-bitext)

New document

MoTrans

Translation

MT system

Live Updates

Building the Hybrid: Live Updates

Parallel sentences (bitext)

English text

Translation Model

Language Model

Parallel "sentences" (pseudo-bitext)

New document

MoTrans

Live Updates

MT system

Translation

"English"

Domain Data

MoTrans

Live Updates

"English"

Translation
User View
Workflow

MoTrans → USTs → Translation

Lexicographer

Expert User
Workflow

MoTrans -> USTs

Joshua

MoJo

OOVs

Translation

Lexicographer

Expert User

Austin, Oct 28 - Nov 1, 2016 | p. 698
Conclusions

• MoJo online in CyberTrans “soon”
  • Productization in progress
    • Starting with Spanish
    • Other languages will follow
  • Shortly thereafter will be available as add-on for CyberTrans distributions

Questions?
Backup Slides
## Example (Portuguese)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motrans</td>
<td>however, the balloting already finished but not yet there are results</td>
</tr>
<tr>
<td></td>
<td>ends.</td>
</tr>
<tr>
<td>StatMT</td>
<td>Meanwhile, the ballot finished but there is still no final results.</td>
</tr>
<tr>
<td>Hybrid</td>
<td>However, the ballot has finished but there is still no final results.</td>
</tr>
<tr>
<td>Human</td>
<td>However, the audit is over but there still are no final results.</td>
</tr>
</tbody>
</table>
Example (Arabic)

<table>
<thead>
<tr>
<th>System</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motrans</td>
<td>And adds Dr. Syrian/Suri that this is reason viewing it to some the patients/al-Murdi who come to him after the passing of the time of their treatment, in addition to the resistance of their bodies to drugs specific/Mu'inah.</td>
</tr>
<tr>
<td>StatMT</td>
<td>He said d. Sorry, this is a reason to see some patients who come to him after it was too late for treatment, as well as to resist their bodies certain drugs.</td>
</tr>
<tr>
<td>Hybrid</td>
<td>D. Syrian adds that this is the reason see some patients who come to him after the passing of time their treatment, as well as to resist their bodies specific drugs.</td>
</tr>
<tr>
<td>Human</td>
<td>Dr. Sory also stated that is why he sometimes had diseased people come in when it was too late for treatment and why there was resistance to certain drugs.</td>
</tr>
</tbody>
</table>
Center for Applied Machine Translation

- DoD-recognized Center of Excellence for Machine Translation
  - Serving the US Government for over 14 years.

- Flagship product: CYBERTRANS
  - Integrated suite of automated tools for MT, language and encoding identification, spelling and text enhancement, and encoding conversion.

中华人民共和国和阿拉伯联合酋长国建立外交关系的联合公报

The Peoples Republic of China and the United Arab Emirates established diplomatic relations with the joint communique
CyberTrans Usage

• Primary customers don’t know language
  • Triage, filtering, selection
  • Free translators from spending time on low value material

• Secondary customers do know language
  • Gisting
  • Seed translation
Hybrid Approaches
Human Comprehension Results

Mean Comprehension Score

- Arabic
- Farsi
- Portuguese
- Russian
- Swahili
BLEU Scores (In-domain)
### Three Hybrid Approaches

<table>
<thead>
<tr>
<th><strong>Black Box</strong></th>
<th>Run Motrans on large text, build a regular StatMT model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct Conversion</strong></td>
<td>Convert Motrans rules directly to StatMT phrase pairs</td>
</tr>
<tr>
<td><strong>On Demand</strong></td>
<td>StatMT system queries Motrans, incorporates its suggestions</td>
</tr>
</tbody>
</table>
Black Box results (in-domain)

<table>
<thead>
<tr>
<th>Language</th>
<th>MoTrans</th>
<th>StatMT Baseline</th>
<th>Black Box</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pashto</td>
<td>+2.0</td>
<td>+1.9</td>
<td>+2.0</td>
</tr>
<tr>
<td>Swahili</td>
<td>+3.3</td>
<td>+3.9</td>
<td>+3.3</td>
</tr>
<tr>
<td>Urdu</td>
<td>+3.9</td>
<td>+4.4</td>
<td>+3.9</td>
</tr>
<tr>
<td>Somali</td>
<td>+4.4</td>
<td>+9.8</td>
<td>+4.4</td>
</tr>
<tr>
<td>Arabic</td>
<td>+9.8</td>
<td>+1.1</td>
<td>+9.8</td>
</tr>
<tr>
<td>Farsi</td>
<td>+2.9</td>
<td>+2.9</td>
<td>+2.9</td>
</tr>
<tr>
<td>Russian</td>
<td>+1.2</td>
<td>+1.2</td>
<td>+1.2</td>
</tr>
<tr>
<td>Spanish</td>
<td>+1.2</td>
<td>+1.2</td>
<td>+1.2</td>
</tr>
<tr>
<td>French</td>
<td>+1.2</td>
<td>+1.2</td>
<td>+1.2</td>
</tr>
<tr>
<td>Portuguese</td>
<td>+1.2</td>
<td>+1.2</td>
<td>+1.2</td>
</tr>
</tbody>
</table>
Direct Conversion Approach

Parallel sentences (bitext)

English text

Translation Model

Converted Translation Model

Language Model

New document

MT system

Translation

MoTrans Lexicon
Direct Conversion Results

![Bar chart showing blue and red bars for Swahili, Urdu, and Farsi in StatMT Baseline and Direct Conversion.

- Swahili: StatMT Baseline, Direct Conversion
- Urdu: StatMT Baseline, Direct Conversion
- Farsi: StatMT Baseline, Direct Conversion]
Challenges of the Conversion Approach

- MoTrans rules have complex, unique syntax
  - Parsing requires in-depth knowledge of MoTrans
  - Some rule types not yet handled
- Not feasible to expand all rules
  - Some rules apply to full sentence, not phrases
  - Rule chaining creates exponential possibilities
- Cybertrans does pre- and post-processing that’s not described in lexicon
On Demand Results

<table>
<thead>
<tr>
<th>Language</th>
<th>StatMT Baseline</th>
<th>On Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pashto</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Swahili</td>
<td>0.9</td>
<td>-0.1</td>
</tr>
<tr>
<td>Urdu</td>
<td>-0.1</td>
<td>6.8</td>
</tr>
<tr>
<td>Farsi</td>
<td>6.8</td>
<td></td>
</tr>
</tbody>
</table>
Combined Hybrid Results

![Bar graph showing performance improvements in different languages. The chart compares StatMT Baseline, Black Box, and Black Box + On Demand for Pashto, Swahili, Urdu, Arabic, Farsi, and Portuguese. Portuguese shows the highest improvement with +7.1.]