You got your StatMT on my rules!

You got your rules in my StatMT!

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

OPEN SOURCE

Why not StatMT before?

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

OPEN SOURCE

- Domain needs
  - Many languages (often low-resource)

Languages Supported in CyberTrans

AFRIKAANS
ALBANIAN
AMHARIC
ARABIC
ARABIC Romanized
ARMENIAN
AZERBAIJANI
BALUCHI
BASQUE
BELARUSIAN
BULGARIAN
BULGARIAN Romanized
CATALAN
CEBUANO
CHAVACANO
CHECHEN
CHINESE Simplified
CHINESE Traditional
CROATIAN
CZECH
DANISH
DARI
DUTCH
ESTONIAN
FINNISH
FRENCH
GALICIAN
GEORGIAN
GEORGIAN Romanized
GERMAN
GREEK
GREEK Romanized
HAITIAN CREOLE
HAUSA
HEBREW
HINDI
HMONG
HUNGARIAN
INDONESIAN
ITALIAN
JAPANESE
JAVANESE
KAZAKH
KOREAN
KURDISH (Kurmanji)
KURDISH (Sorani)
KYRGYZ
LAO
LATVIAN
LINGALA
LITHUANIAN
MACEDONIAN
MACEDONIAN Romanized
MAGUINDANAON
MALAGASY
MALAYSIAN
NORWEGIAN
PAPIAMENTO
PASHTO
PASHTO Romanized
PERSIAN
PERSIAN Romanized
POLISH
PORTUGUESE
PUNJABI
ROMANIAN
RUSSIAN
RUSSIAN Romanized
SERBIAN
SERBIAN Cyrillic
*SHONA**
SLOVAK
SLOVENE
SOMALI
SPANISH
SRANAN
SUNDANESE
SWAHILI
SWEDISH
TAGALOG
TAJIK
TAJIK Romanized
TAUSUG
TETUM
THAI
TOK PISIN
TURKISH
TURKMEN
TWI*
UKRAINIAN
UKRAINIAN Romanized
URDU
URDU Romanized
UYGHUR
UZBEK Cyrillic
UZBEK Romanized
VIETNAMESE
WOLOF
YAKAN
YORUBA

Coming soon

Why not StatMT before?

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

OPEN SOURCE

- 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

OPEN SOURCE

- 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

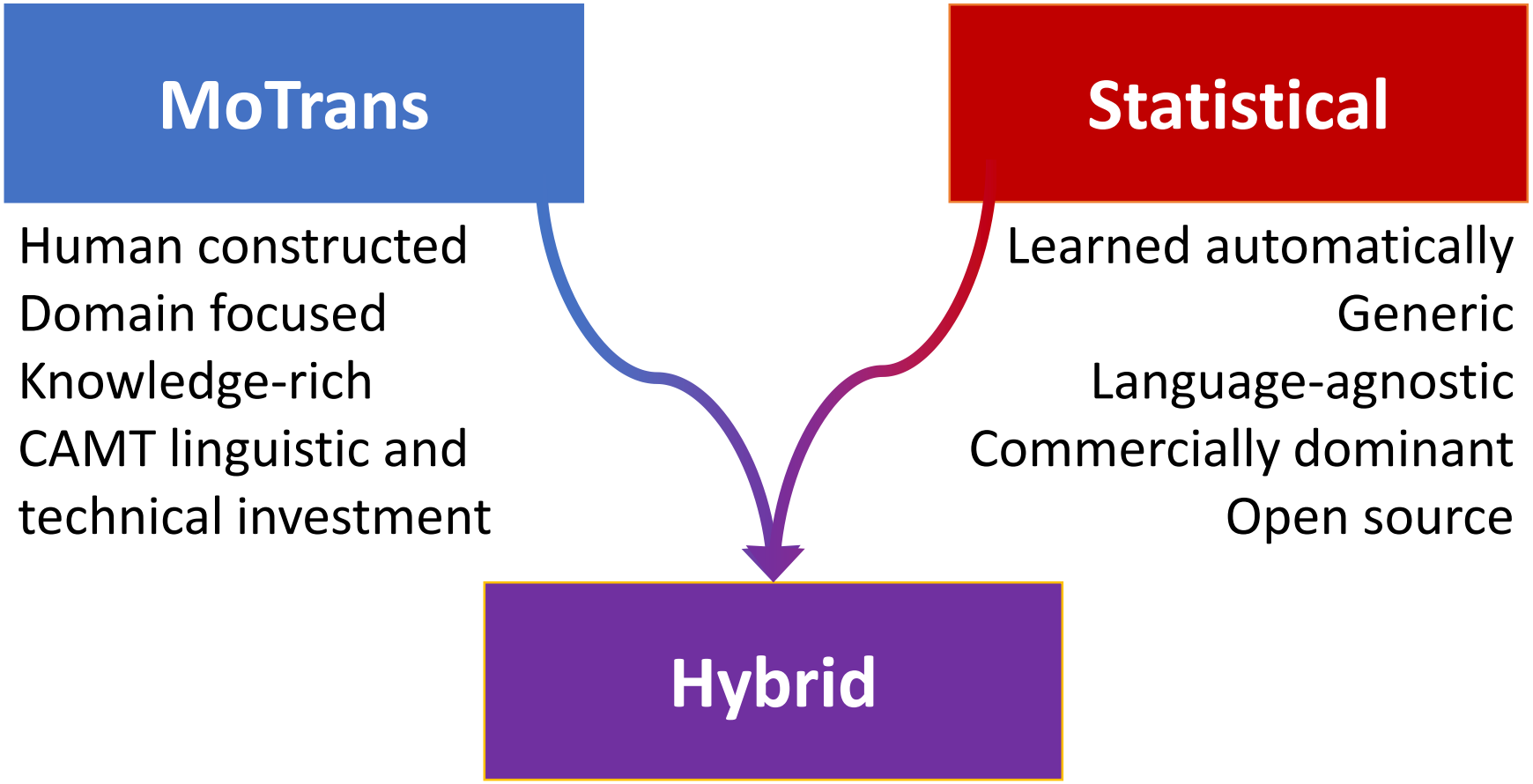
- Rules are composed by language experts
- Performs a deep source language analysis
- Easy to update, adapt to new domains
- Easy to trace input to output
- Very fast

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



Example (Russian)

System	Output
Motrans	He noted, that presidential pre-election campaign provoked “discrepant and often frequently vulgar rhetoric,” eating away democracy and society.
StatMT	He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” разъедающую democracy and society .
Hybrid	He noted that the presidential electoral campaign has provoked “inconsistent and often vulgar rhetoric,” eating away democracy and society.
Human	He said the presidential campaign has brought “divisive and often vulgar rhetoric” that corrodes democracy and society.

Example (Russian)

System	Output
Motrans	From Moscow to Sochi on the train about two days! Really? Is it possible? You want to lead two days in the uncomfortable train?
StatMT	From Moscow to Sochi to train about two days! Do you want to spend two days in awkward train?
Hybrid	From Moscow to Sochi on the train about two days! Do you want to spend two days in uncomfortable train?
Human	From Moscow to Sochi by train is close to 2 days! Do you really want to spend two days in an uncomfortable train?

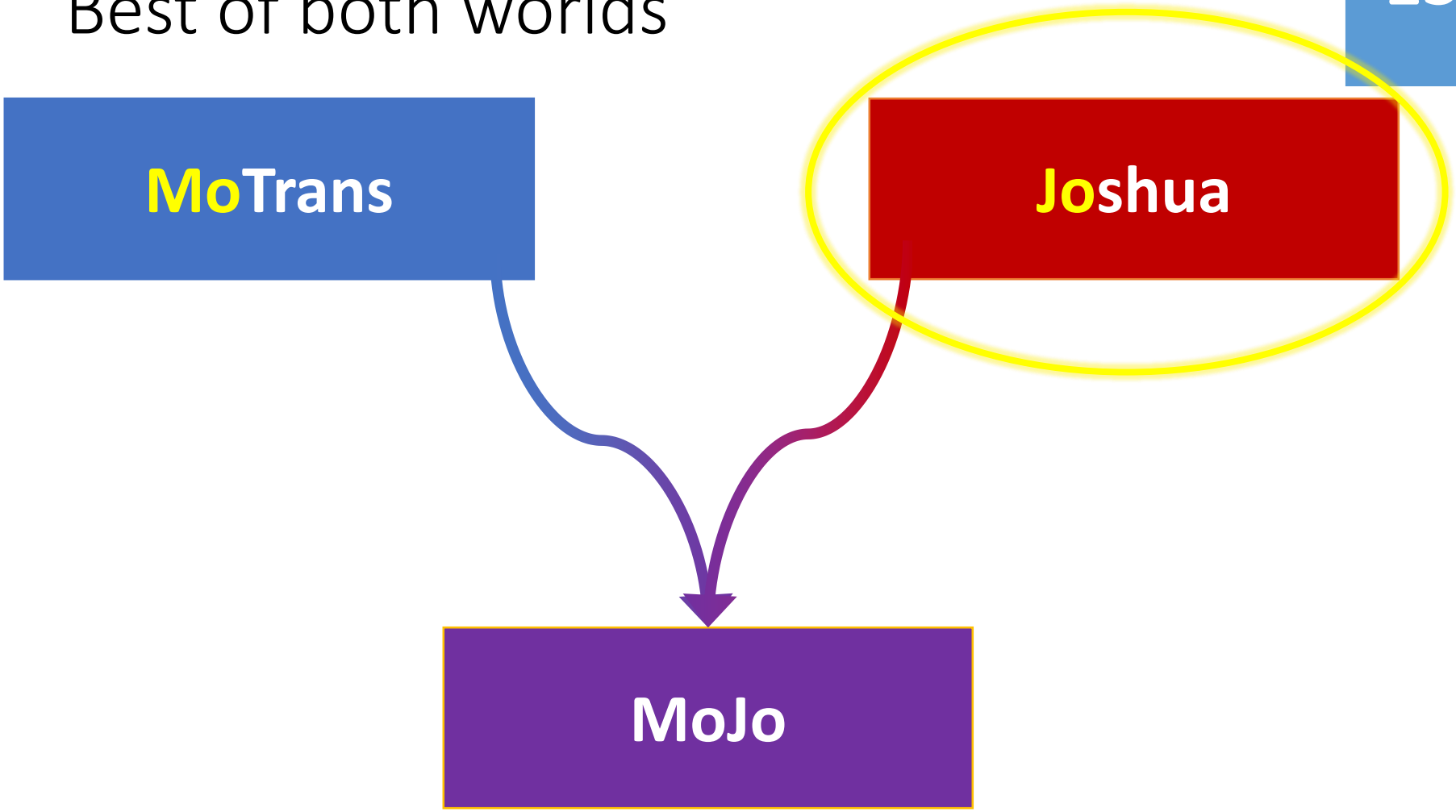
Example (Swahili)

System	Output
Motrans	LABLA America in/at what region? America is big.
StatMT	“Maybe America in what state? The United States is the greatest.
Hybrid	Maybe America in what region? The United States is big.
Human	To be more precise, which state in America? America is vast.

Example (Swahili)

System	Output
Motrans	Pat: say (you/it) succeeded to get children and MUMEO Mohammed HAMIS?
StatMT	Pat: is ulifanikiwa to children and mumeo Mohamed Hamis?
Hybrid	Pat: *Have you* succeeded in *getting* children mumeo Mohamed Hamis?
Human	Pat: Were you successful at having children by that husband Mohamed Hamis?

Best of both worlds

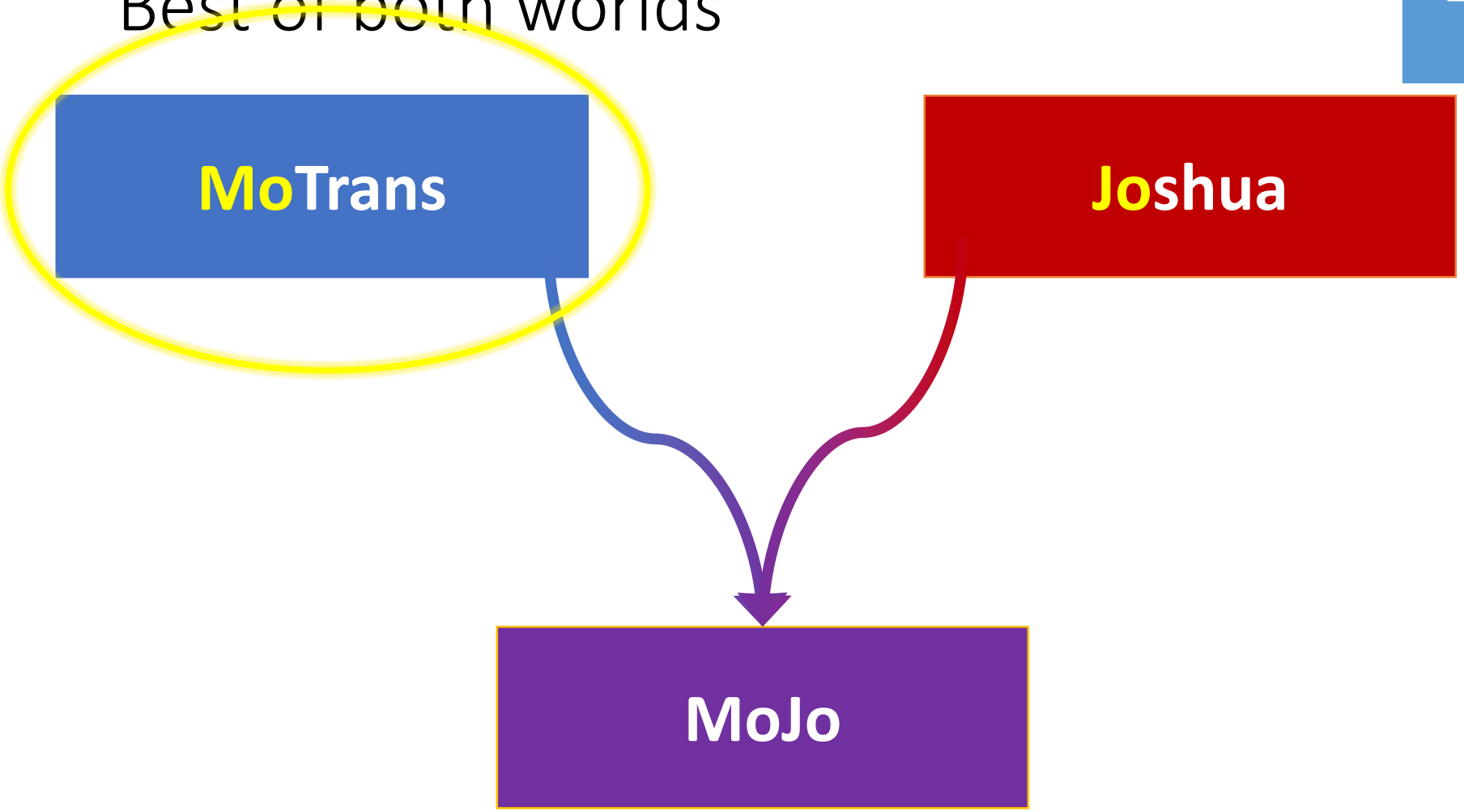


Apache Joshua

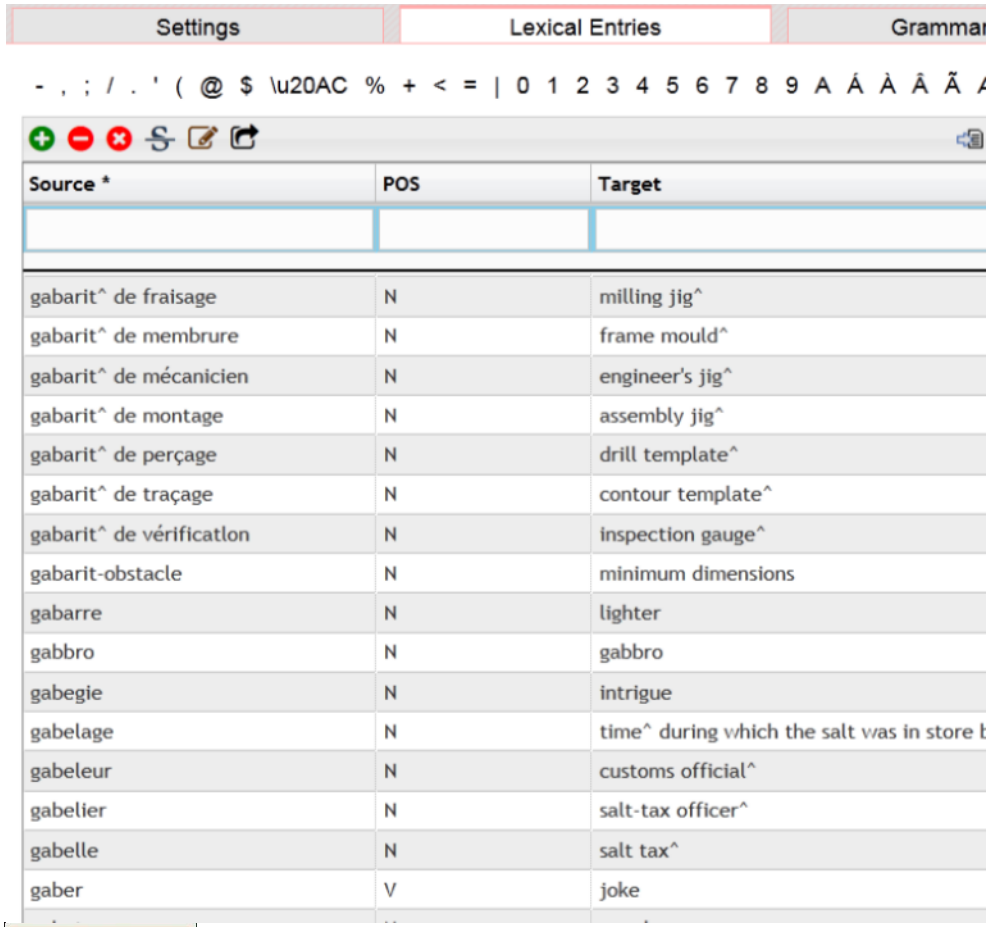
- Open-source Java statistical machine translation system
- Apache project currently in Incubation Stage
- Provides both phrase-based and hiero StatMT
- Pre-built language packs available for download
- **Written in Java** (like CyberTrans)
- **Development lead conveniently located at JHU HLTCOE in nearby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

http://joshua.apache.org/

Best of both worlds



MoTrans 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--

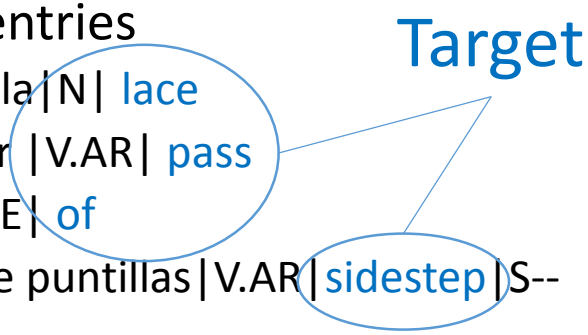
Part of Speech

- 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--

Stem indicator

- 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--

Affix Location Pattern

- 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
```

Type

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{s}):%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-0{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
```

puntillas ->

lace

ha pasado ->

has passed

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
```

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-0{s}):%s:+plural

puntillas ->

PUNTILLA (SUFFIX_PATTERN:N:(*V)(X-0{S}):%s:+plural):N: lace

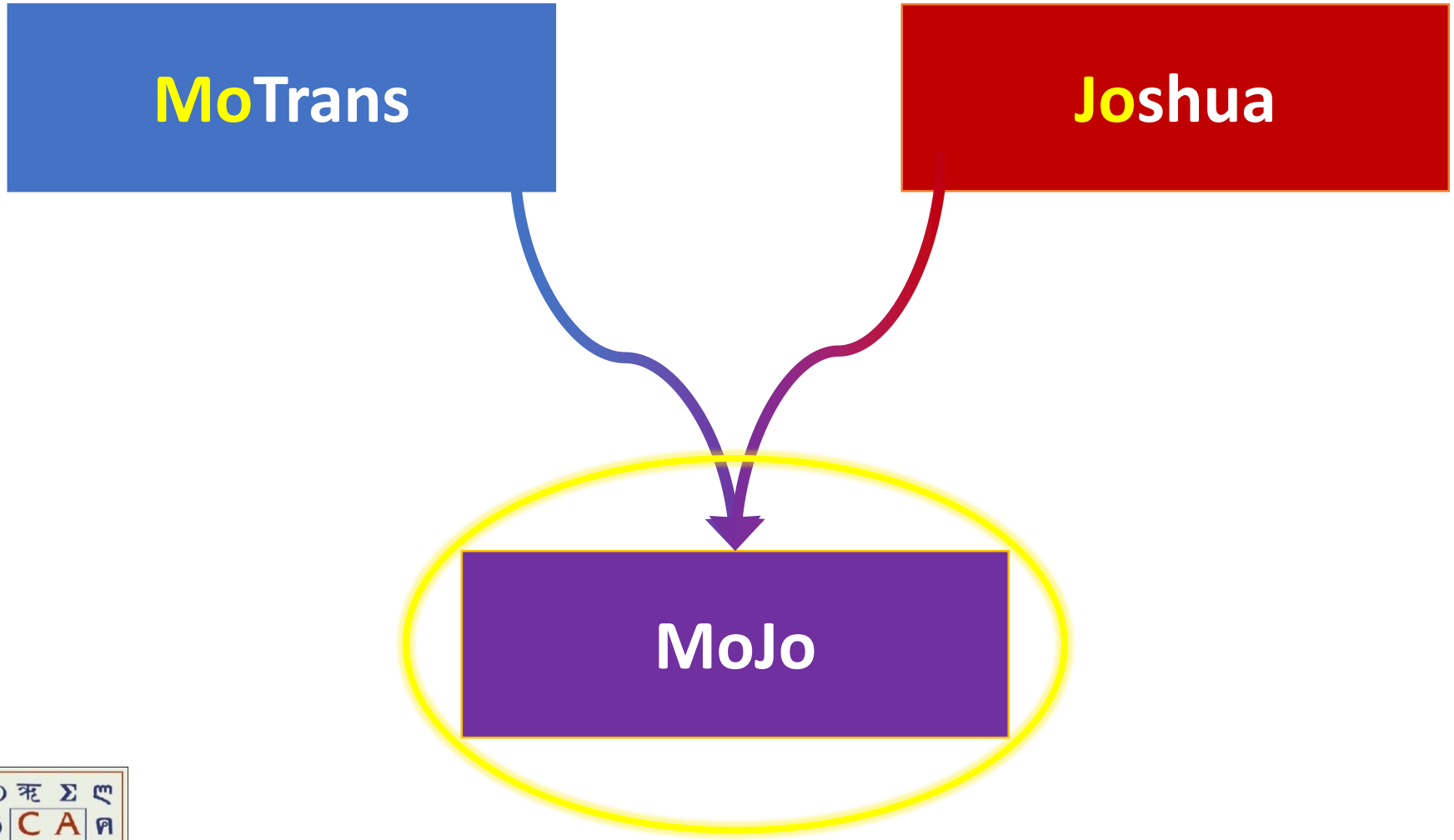
ha pasado ->

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

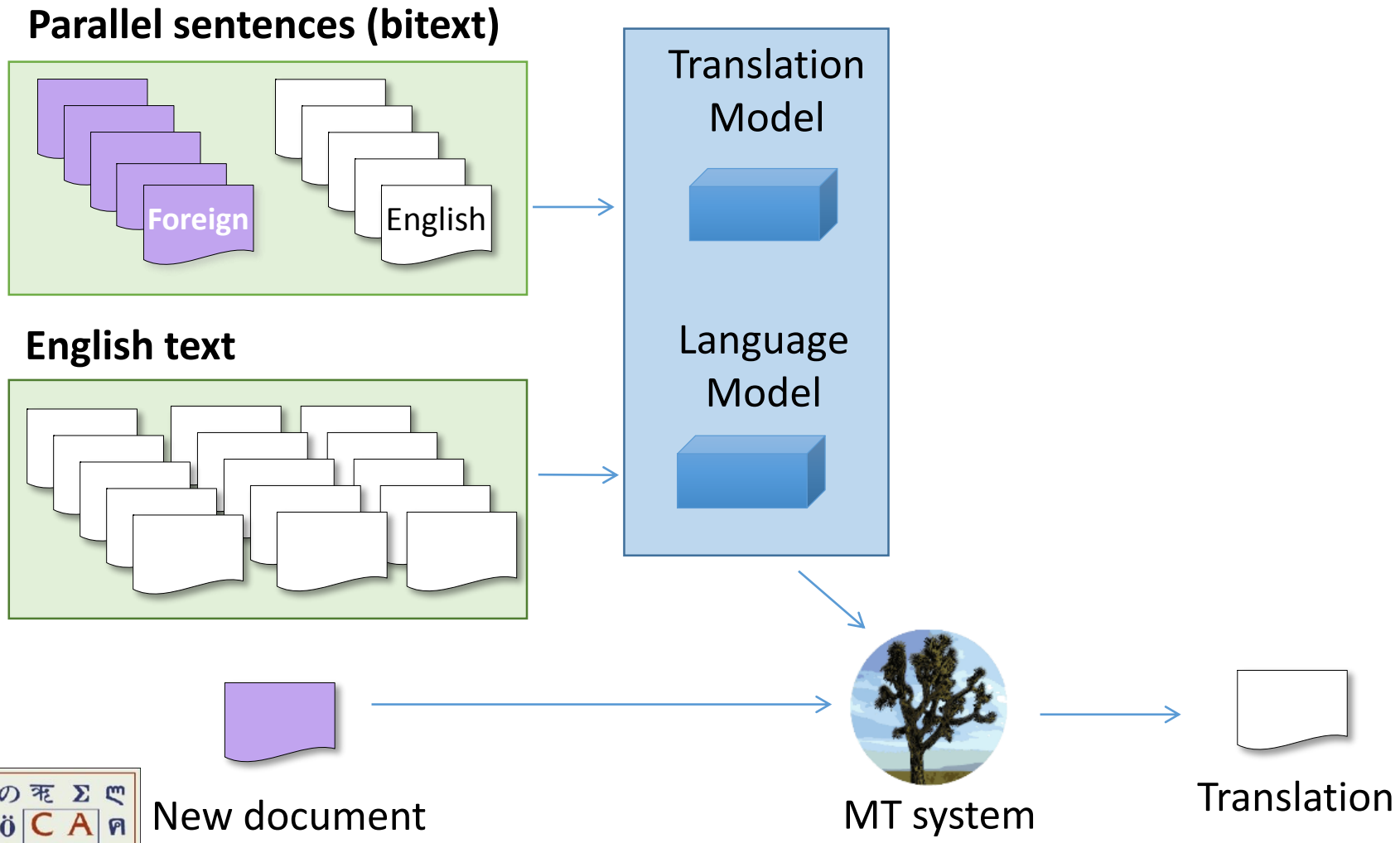
ha pasado de puntillas ->

PAS DE PUNTILLAS (PROCLITIC:V.CONJ:HA:has %s:+pastp) (SUFFIX:V.AR:ADO:%s:+pastp):V.CONJ: has sidestepped

Best of both worlds

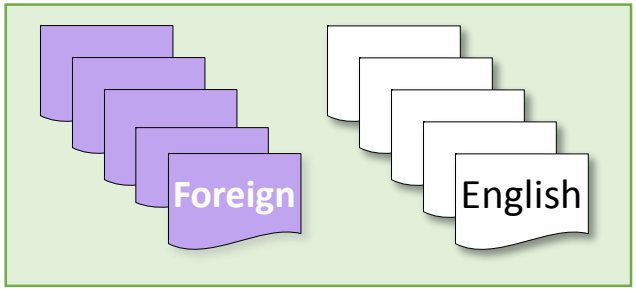


Building the Hybrid: Base StatMT

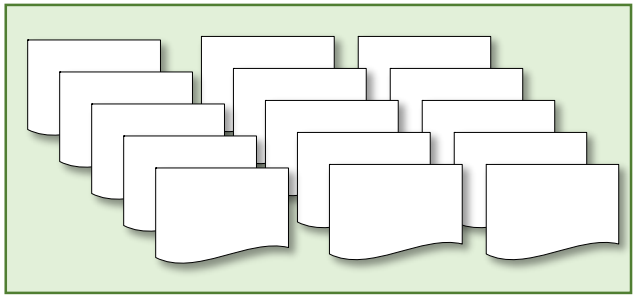


Building the Hybrid: Black Box

Parallel sentences (bitext)



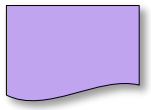
English text



Translation Model

Language Model

Rule-based knowledge applied to domain data to feed StatMT

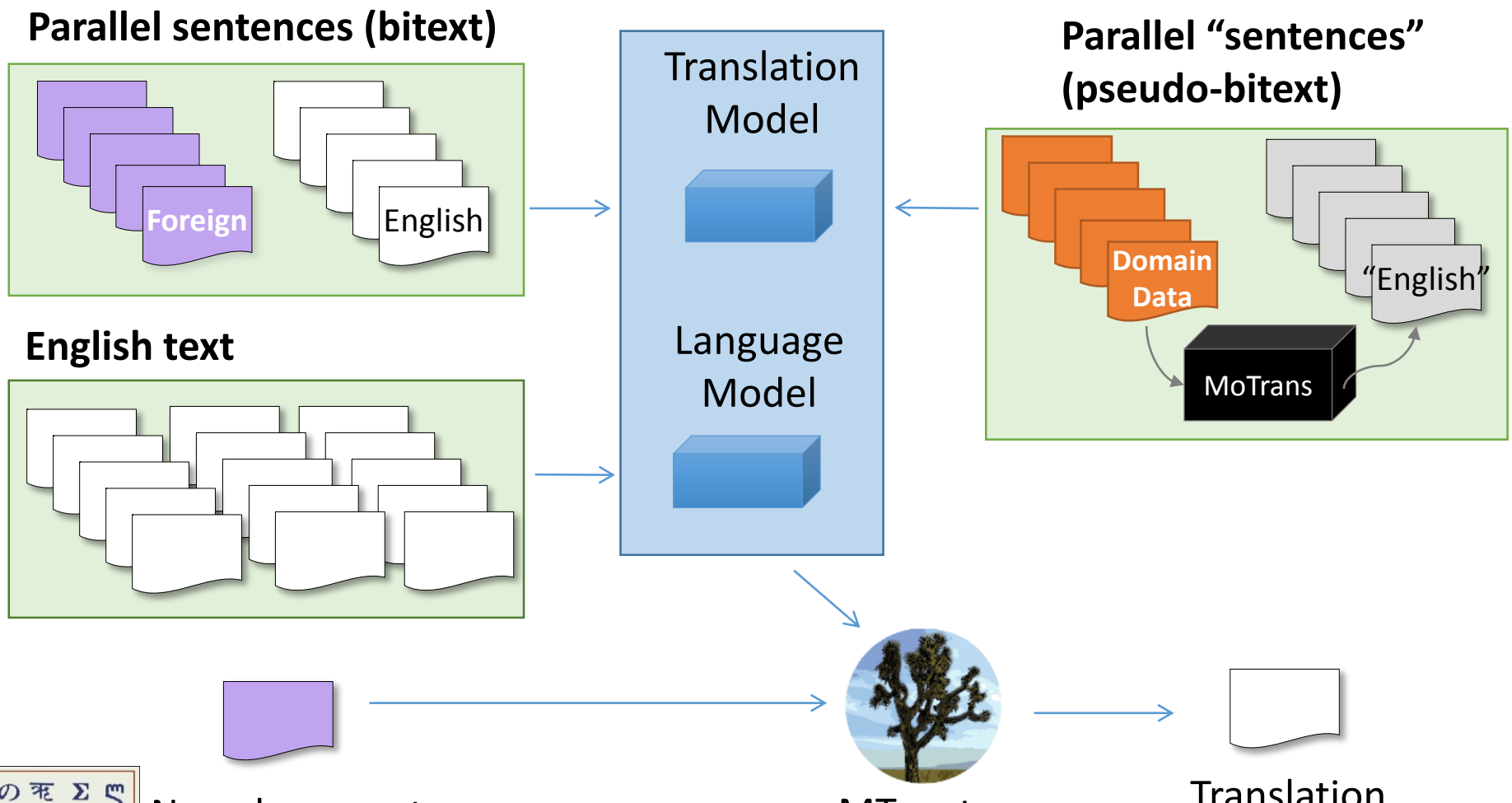


New document

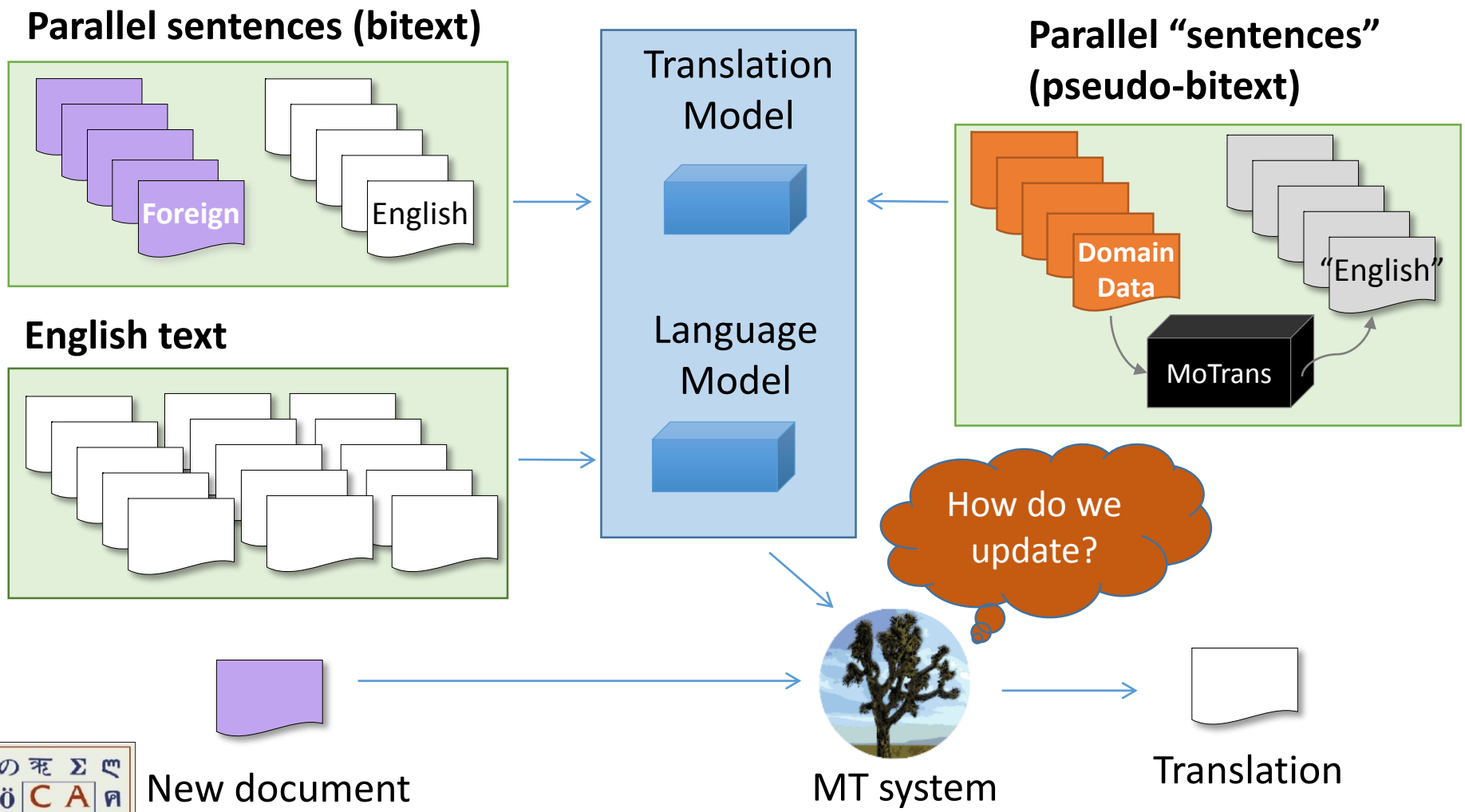
MT system

Translation

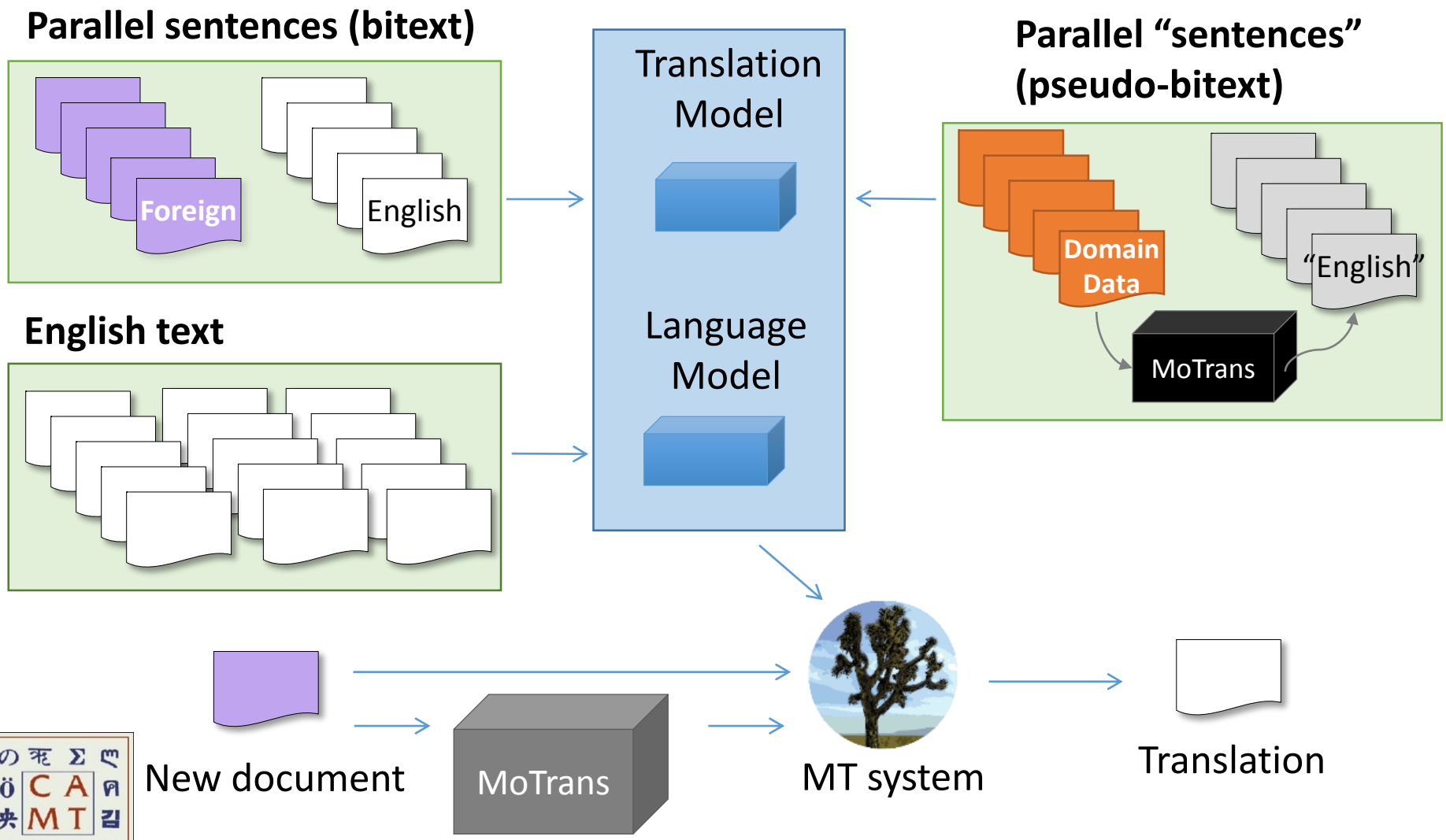
Building the Hybrid: Black Box



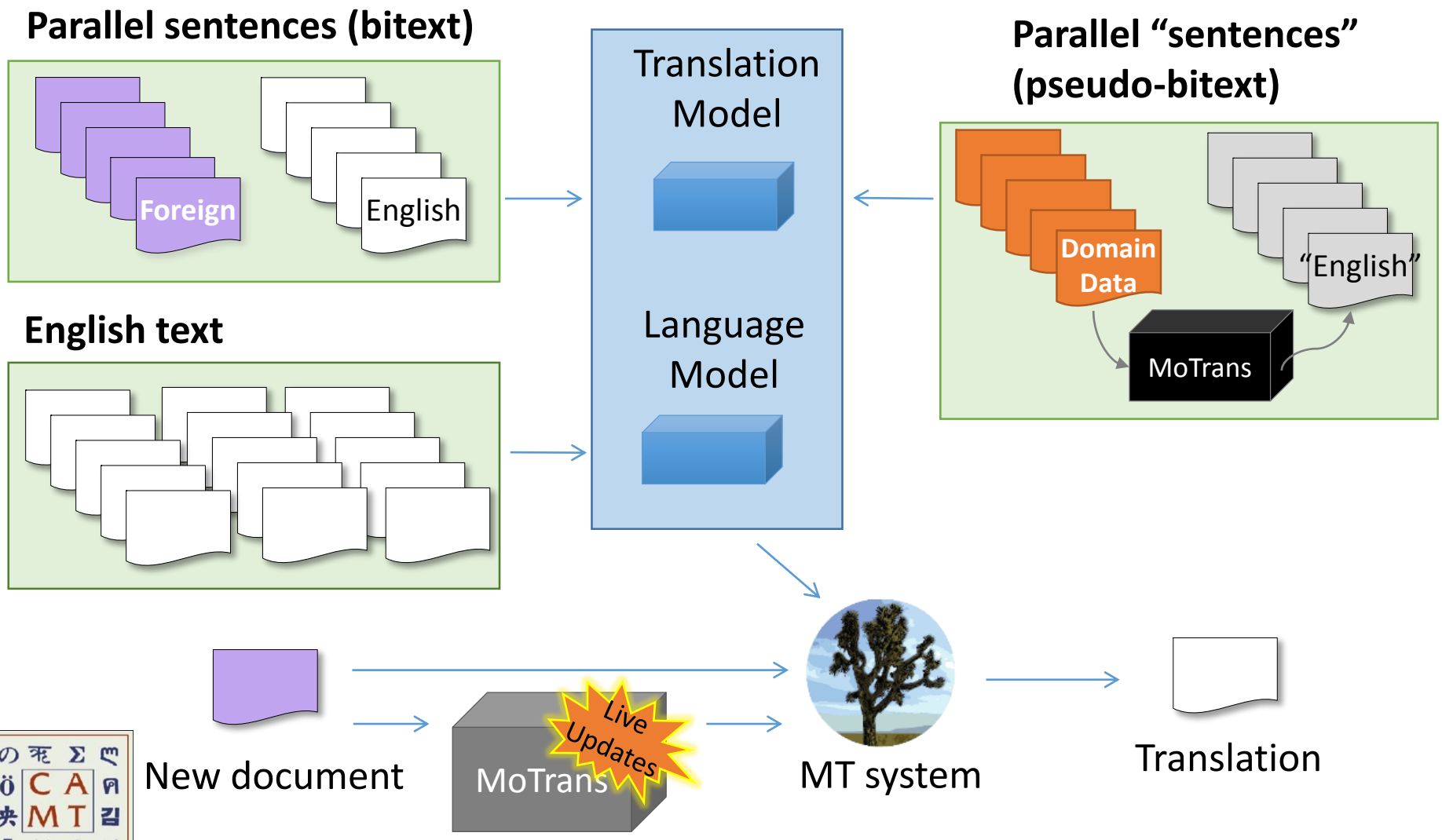
Building the Hybrid: Black Box



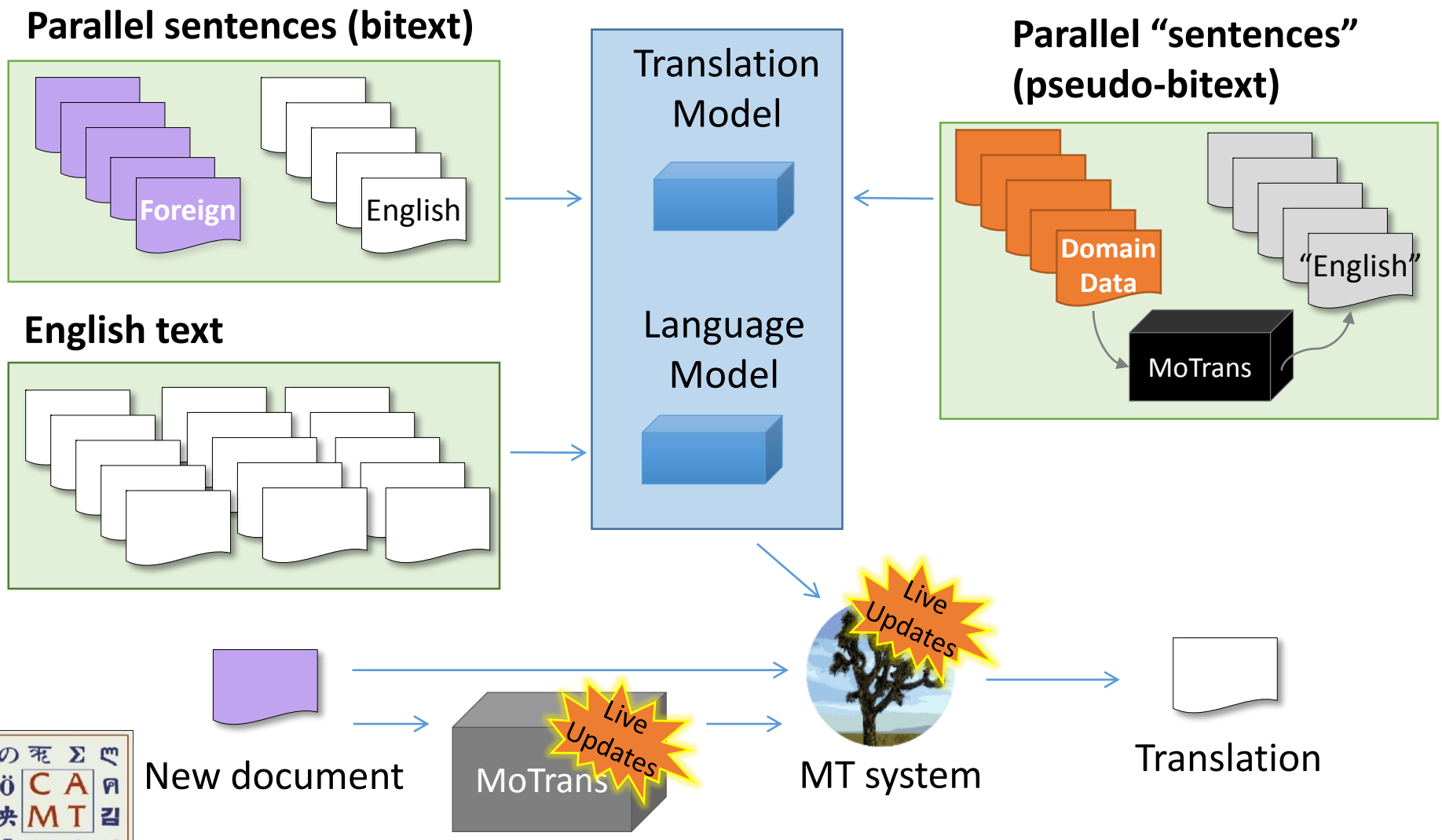
Building the Hybrid: On Demand



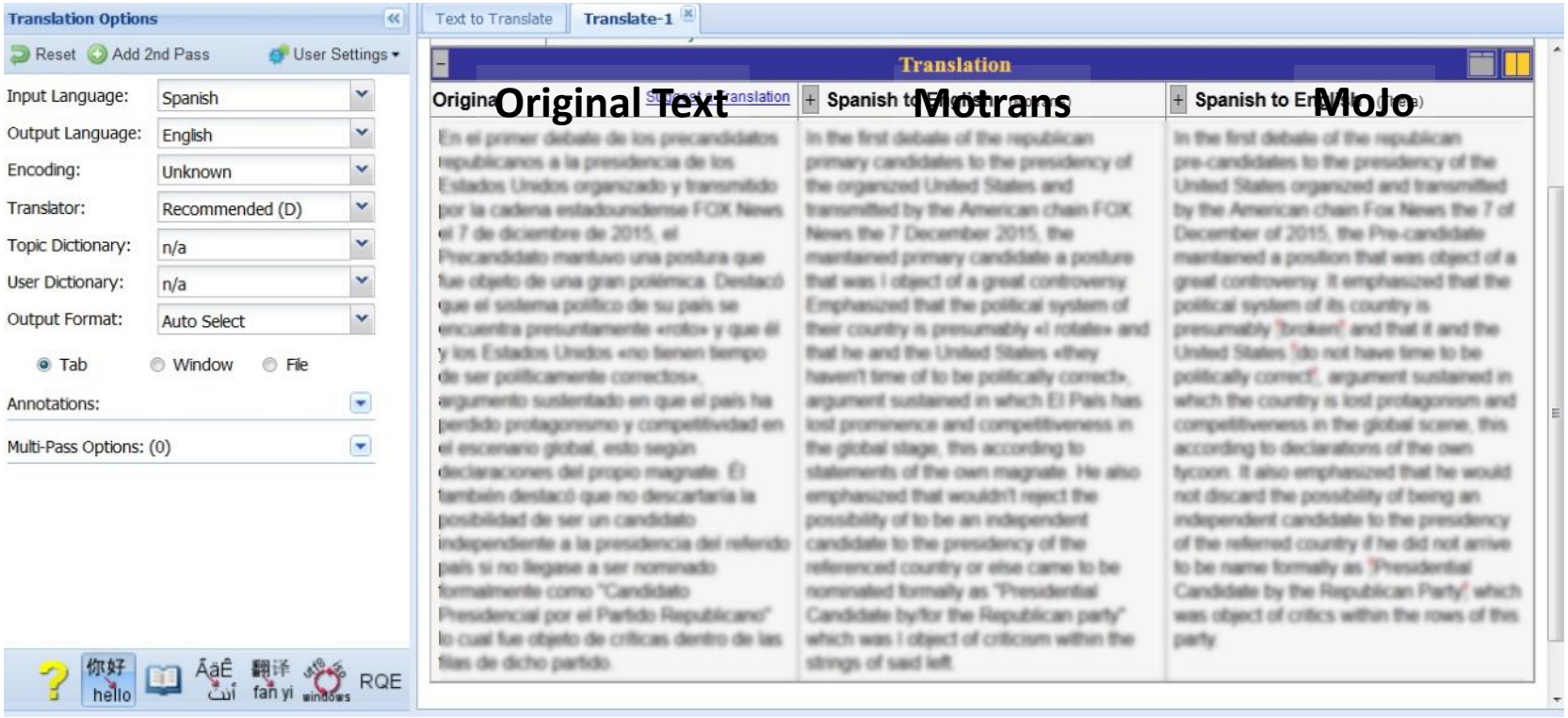
Building the Hybrid: Live Updates



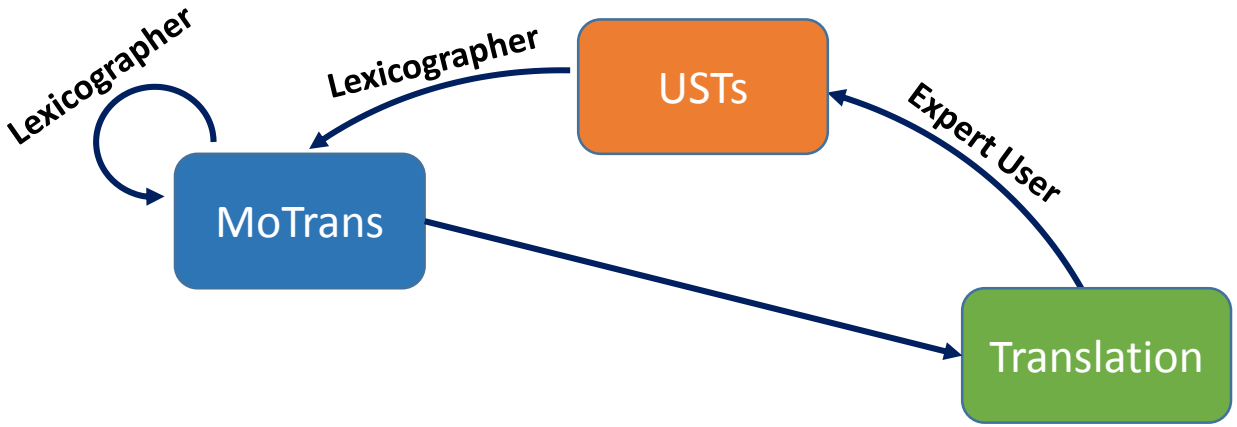
Building the Hybrid: Live Updates



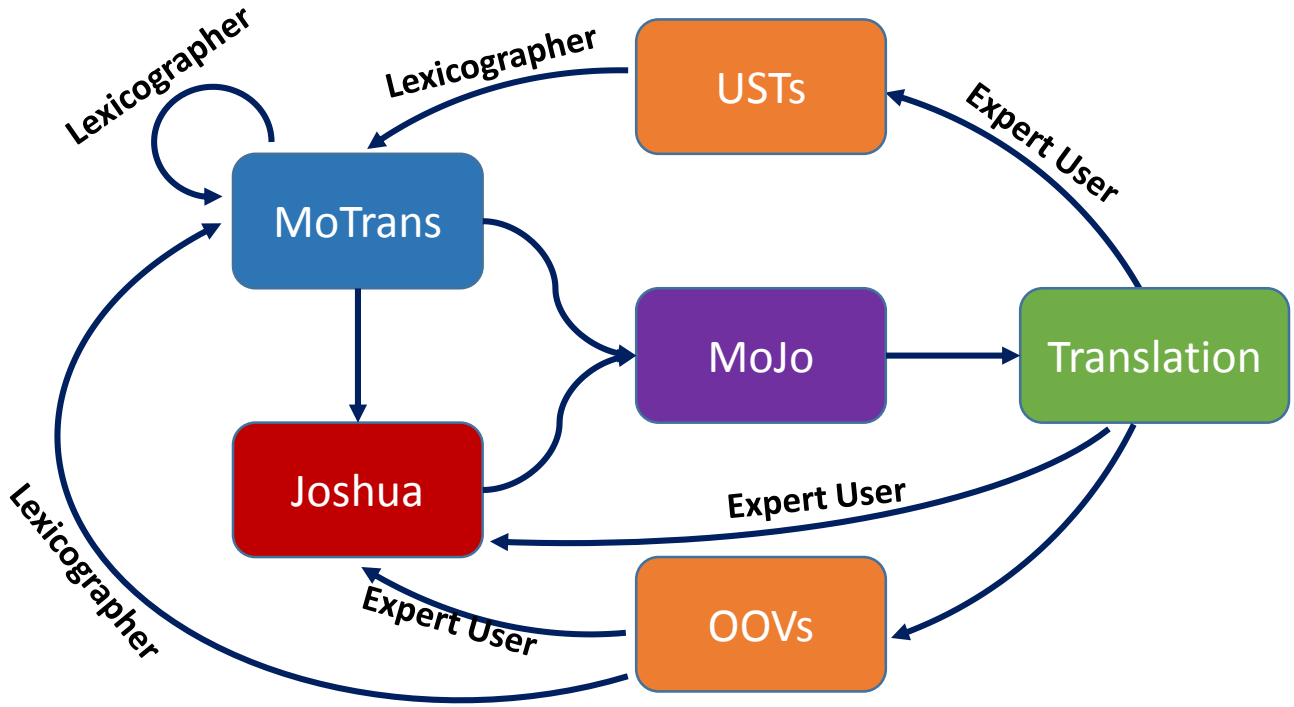
User View


Translation Options
Reset
Add 2nd Pass
User Settings
Input Language: Spanish
Output Language: English
Encoding: Unknown
Translator: Recommended (D)
Topic Dictionary: n/a
User Dictionary: n/a
Output Format: Auto Select
Tab
Window
File
Annotations:
Multi-Pass Options: (0)
Text to Translate
Translate-1
Translation
Original Text
Motrans
MoJo
RQE

Workflow


Lexicographer
Lexicographer
USTs
Expert User
MoTrans
Translation

Workflow


Lexicographer
Lexicographer
USTs
Expert User
MoTrans
MoJo
Translation
Joshua
Expert User
Lexicographer
Expert User
OOVs

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)

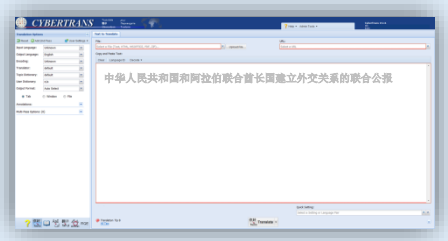
System	Output
Motrans	however, the balloting already finished but not yet there are results ends.
StatMT	Meanwhile, the ballot finished but there is still no final results.
Hybrid	However, the ballot has finished but there is still no final results.
Human	However, the audit is over but there still are no final results.

Example (Arabic)

System	Output
Motrans	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.
StatMT	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.
Hybrid	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.
Human	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.

Center for Applied Machine Translation

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

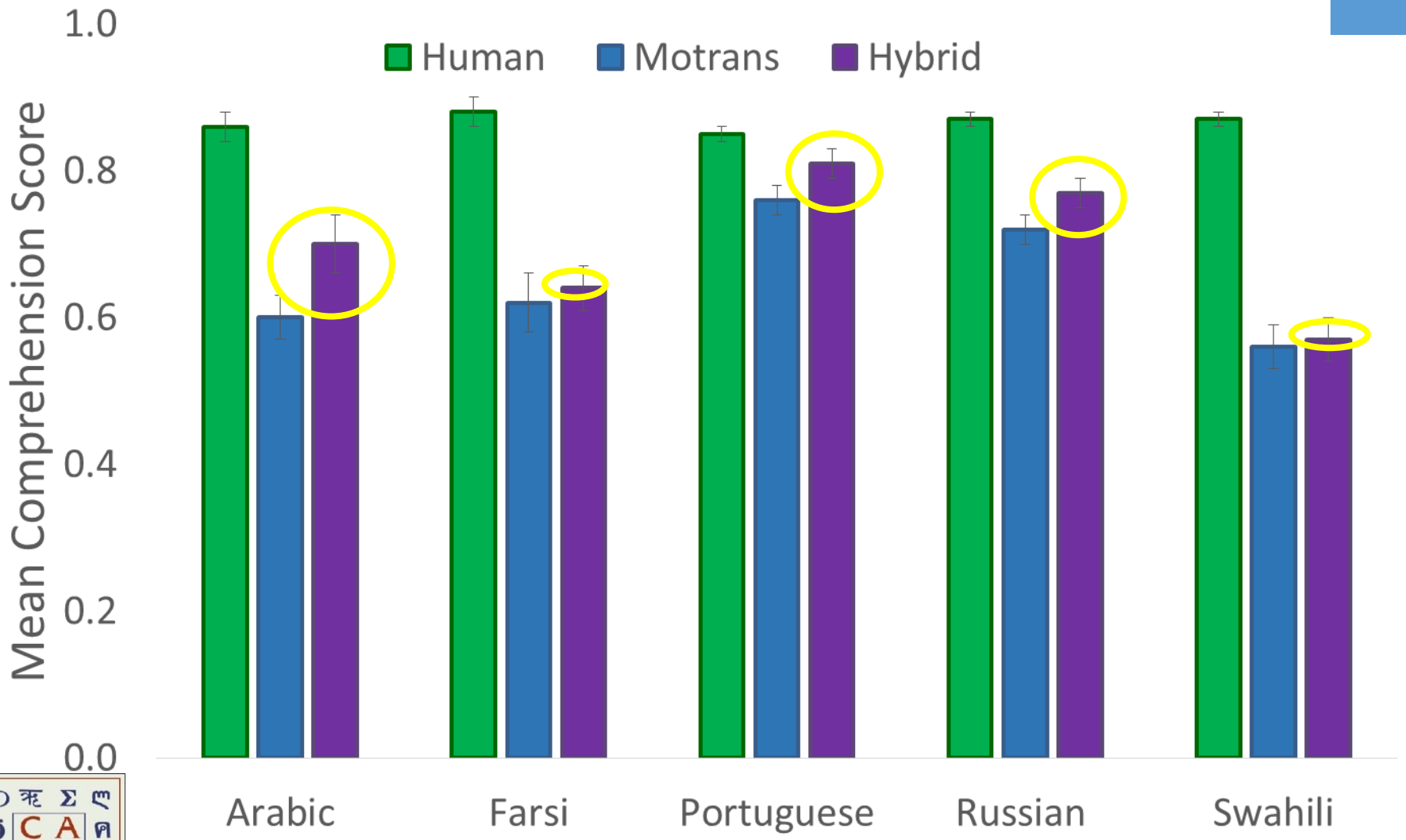
- DoD-recognized Center of Excellence for Machine Translation
  - Serving the US Government for over 14 years.
- Flagship product: **CYBERTRANS**
  - Integrated suite of automated tools for MT, language and encoding identification, spelling and text enhancement, and encoding conversion.

CyberTrans Usage

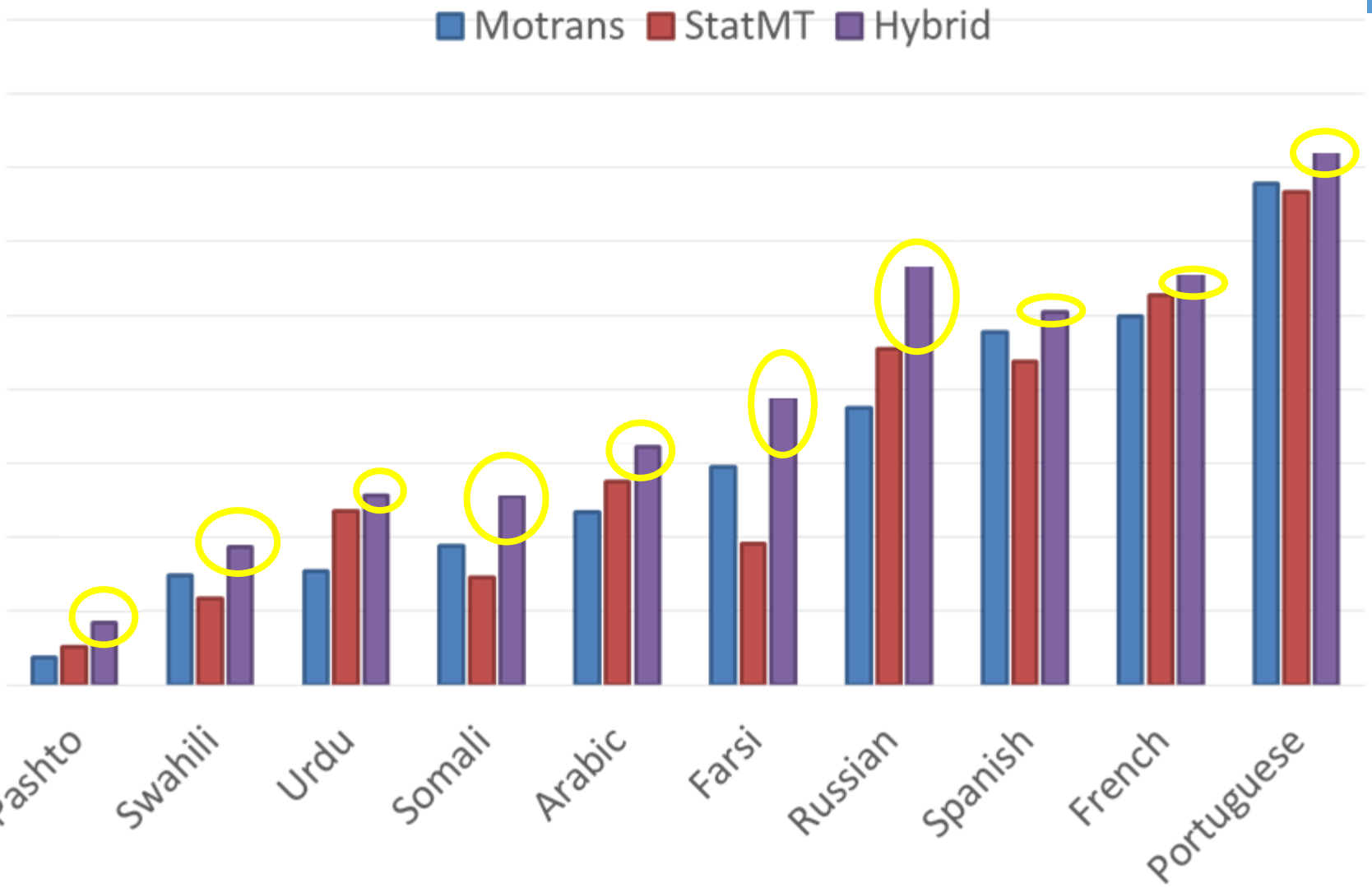
- Primary customers **don't** know language
  - Triage, filtering, selection
  - Free translators from spending time on low value material
- Secondary customers **do** know language
  - Gisting
  - Seed translation

Hybrid Approaches

Human Comprehension Results



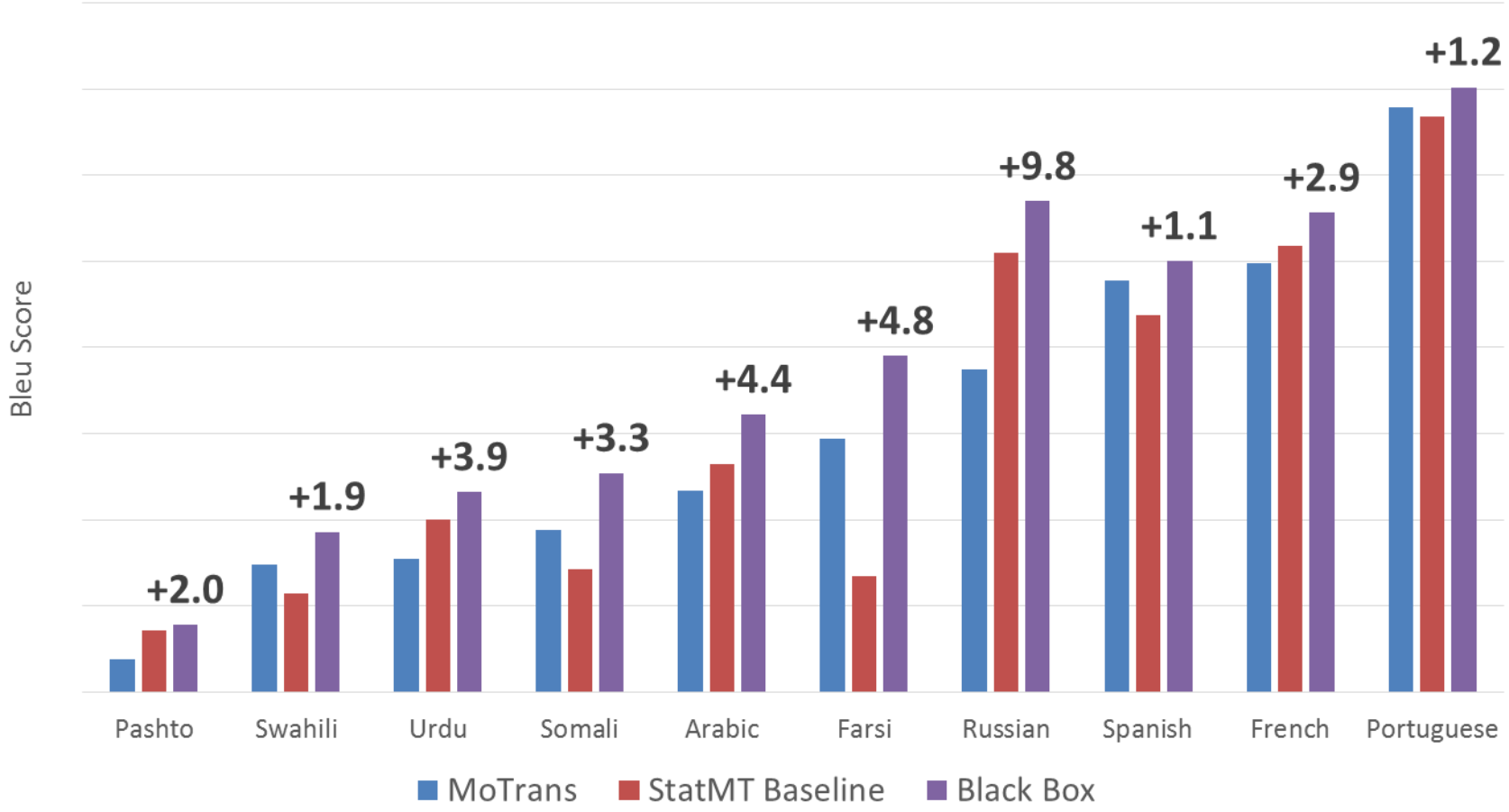
BLEU Scores (In-domain)



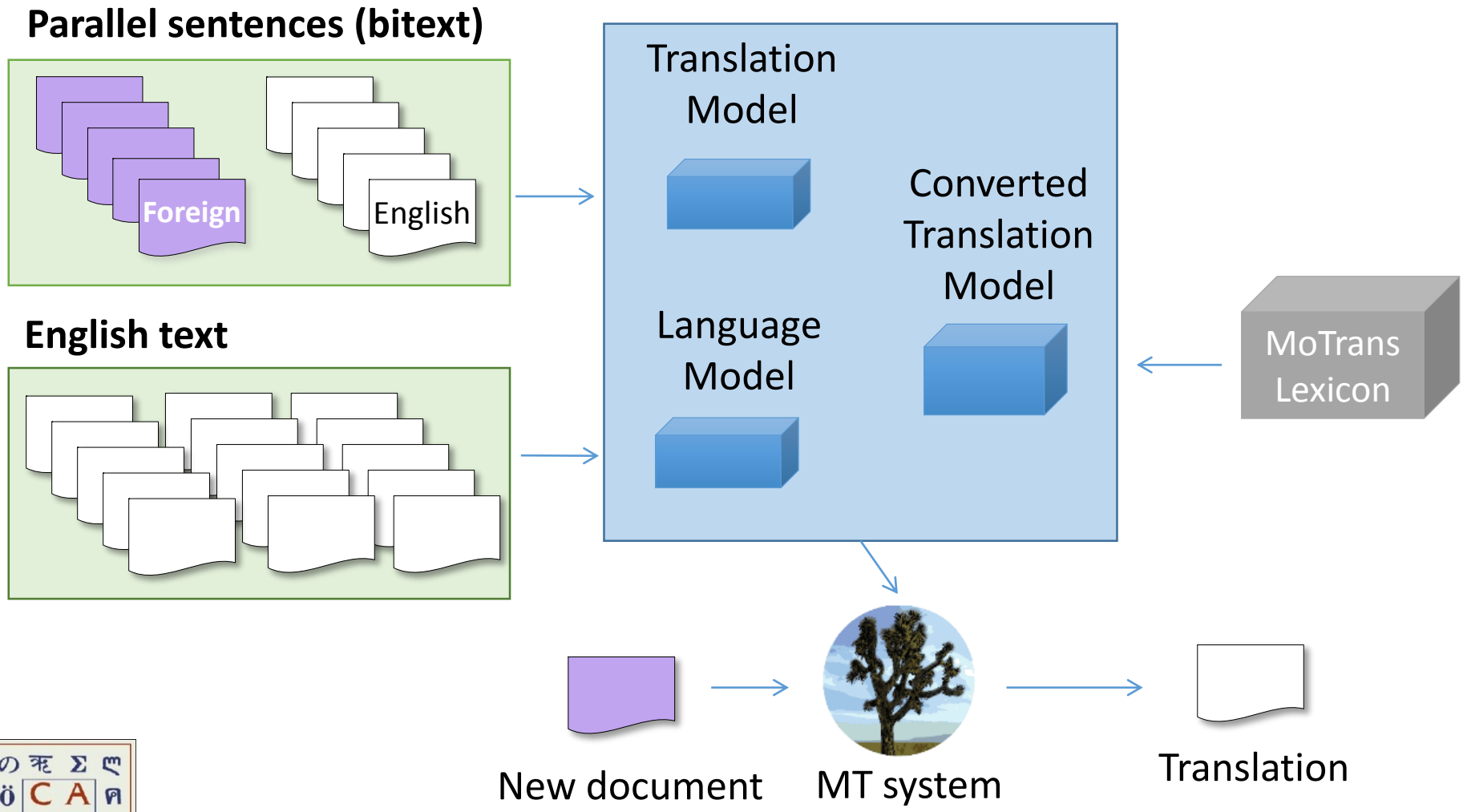
Three Hybrid Approaches

Black Box	Run Motrans on large text, build a regular StatMT model
Direct Conversion	Convert Motrans rules directly to StatMT phrase pairs
On Demand	StatMT system queries Motrans, incorporates its suggestions

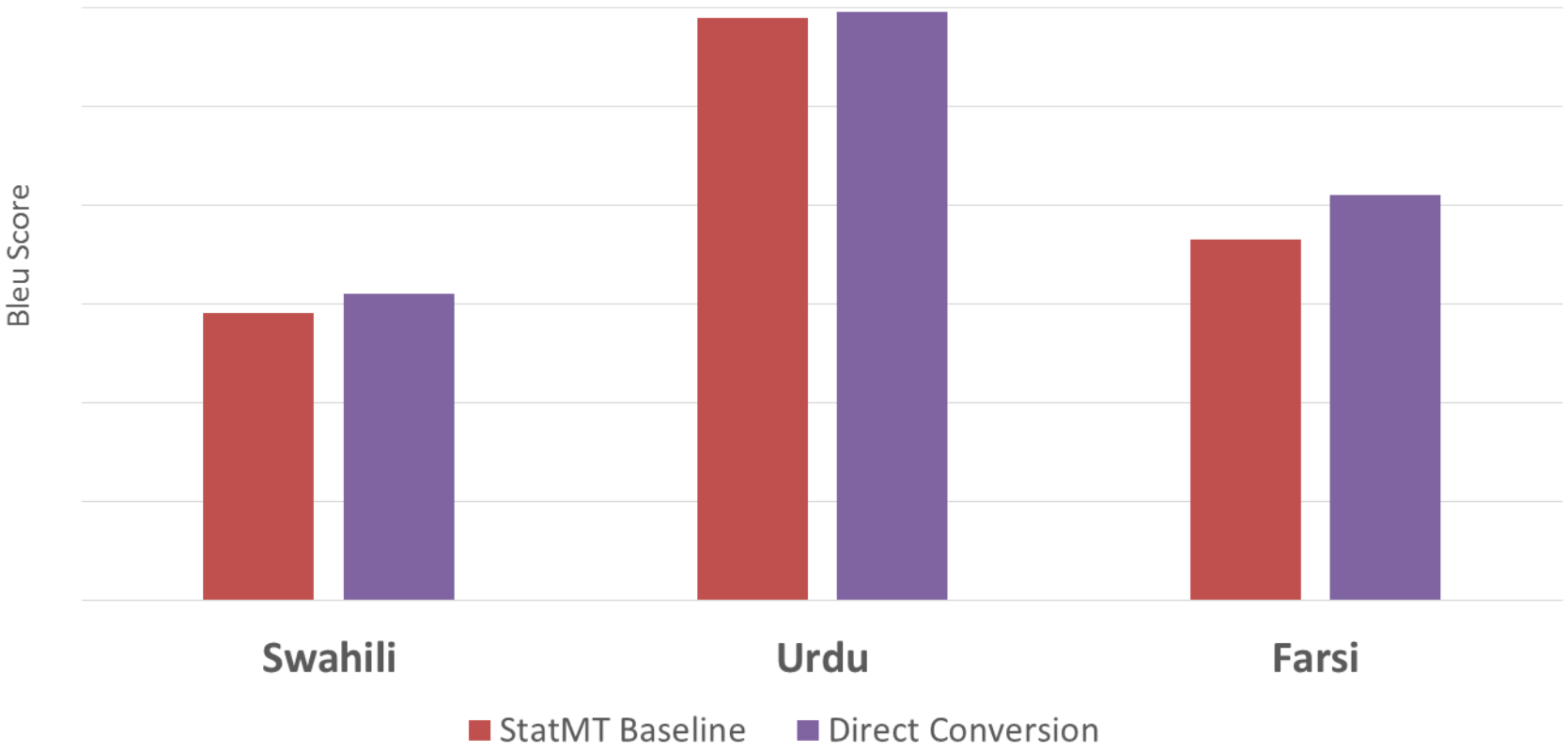
Black Box results (in-domain)



Direct Conversion Approach



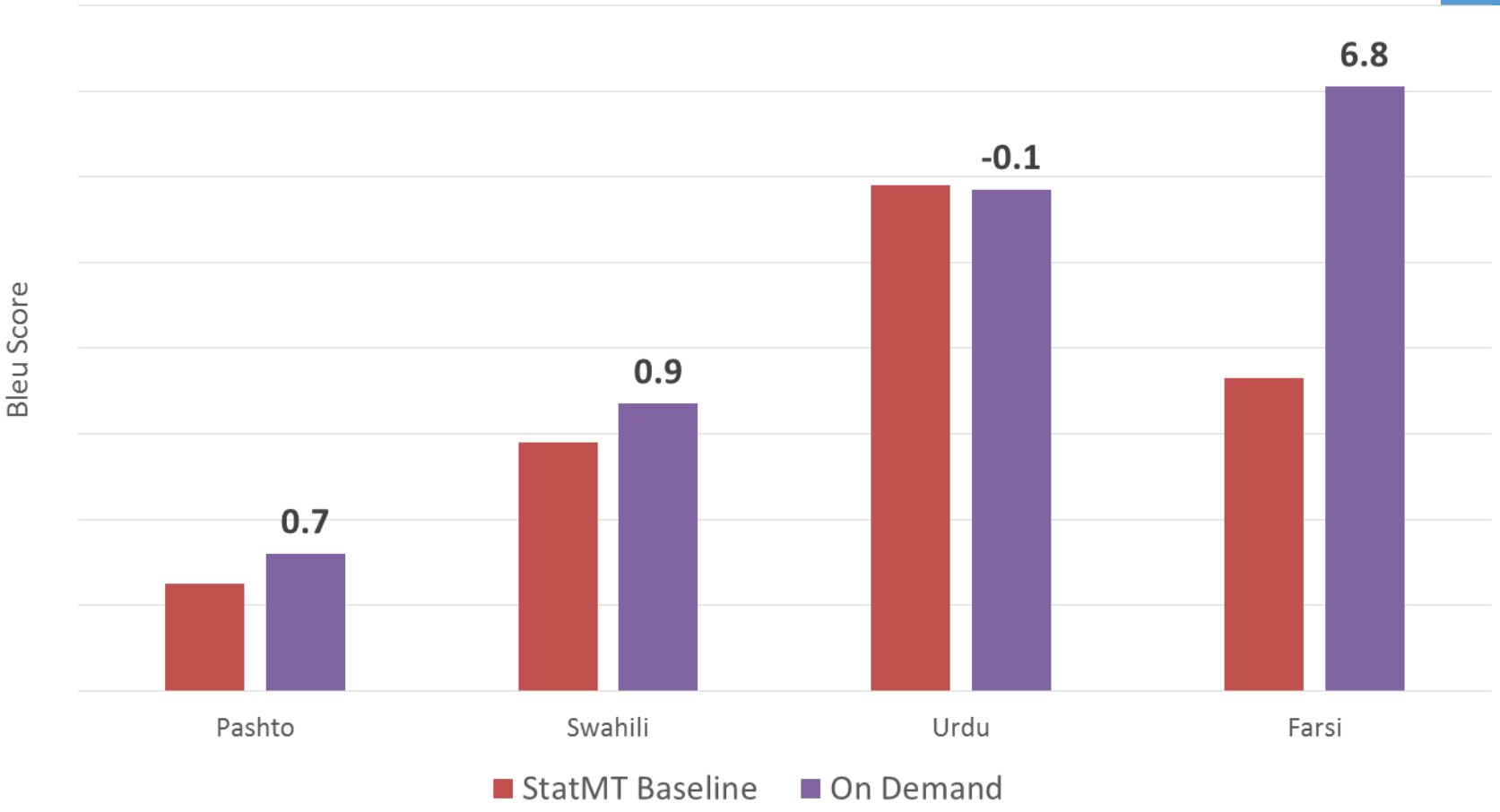
Direct Conversion Results



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



Combined Hybrid Results