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

Safaba Translation Solutions

Safaba Translation Solutions

Safaba Translation Solutions

- **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

Safaba Translation Solutions

- **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.

Safaba Translation Solutions

- **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%**

Safaba – MT Technology Overview

- **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

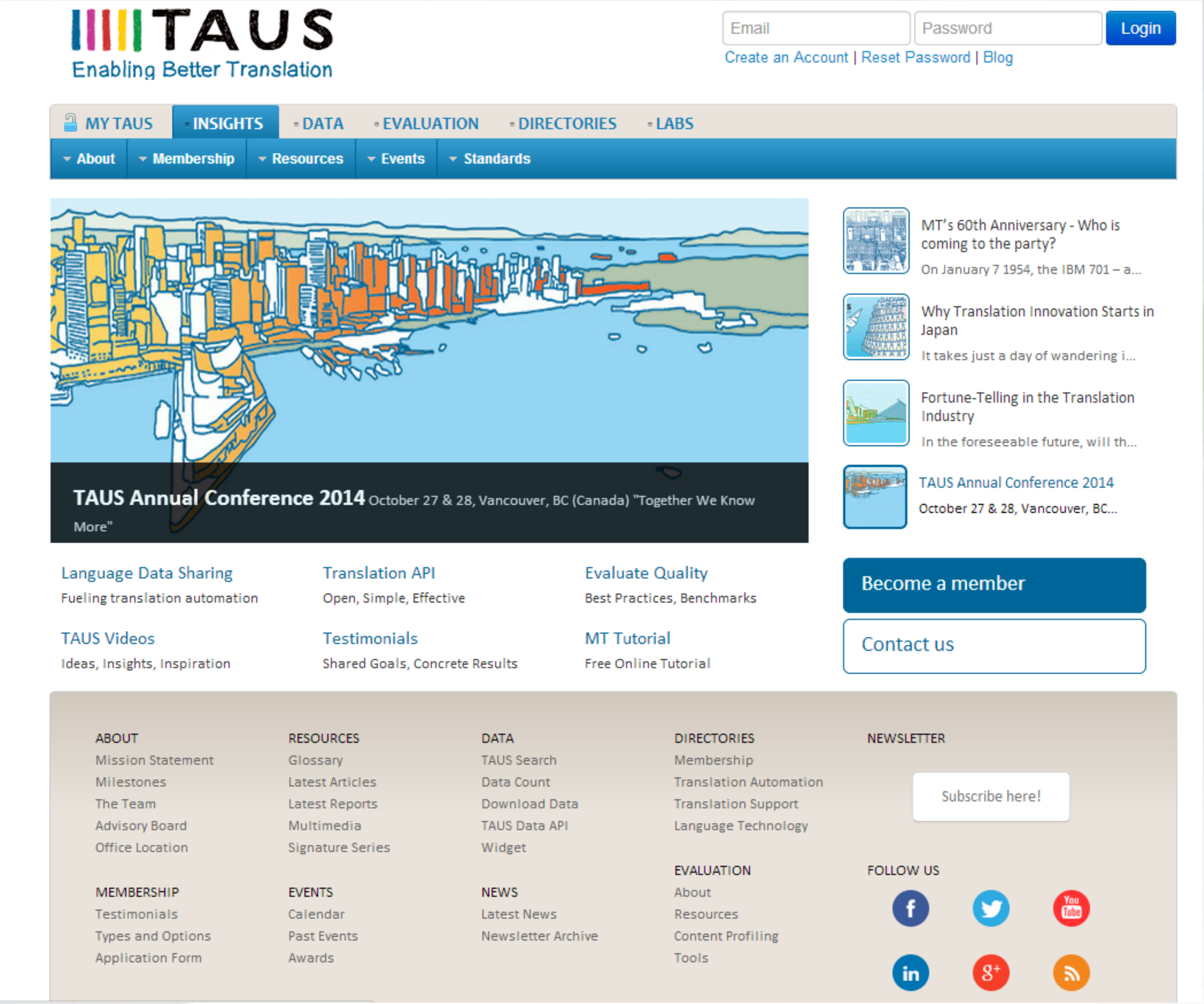
- Provide some deeper insight about the characteristic differences between typical “academic MT systems” (i.e. for WMT and NIST evaluations) and Safaba’s typical commercial systems
- Provide a closer look at some of the main R&D challenges and requirements for delivering advanced hosted real-time Statistical 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/
- 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 TMs)
- **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

Language Pair	Best WMT-2014	Safaba TAUS Generic
EN-to-FR	35.8	65.4
EN-to-ES	30.4 *	66.2
EN-to-RU	29.9	41.6
EN-to-CS	21.6	43.6
EN-to-DE	20.6	52.5
FR-to-EN	35.0	68.0
RU-to-EN	31.8	---
ES-to-EN	31.4 *	70.4
DE-to-EN	29.0	62.4
CS-to-EN	28.8	---

Sample Safaba Output

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

#	Source	Translation
1	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.	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.
2	Uso de Detalles de Selección	Using Selection Details
3	Puede probar cambios pendientes si un proceso de modificación se ha interrumpido y determinar si debe utilizar el método Update o CancelUpdate .	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.
4	Número de años anteriores	Number of past years
5	Armas	Arms
6	Controlador IPSEC	IPSEC Driver

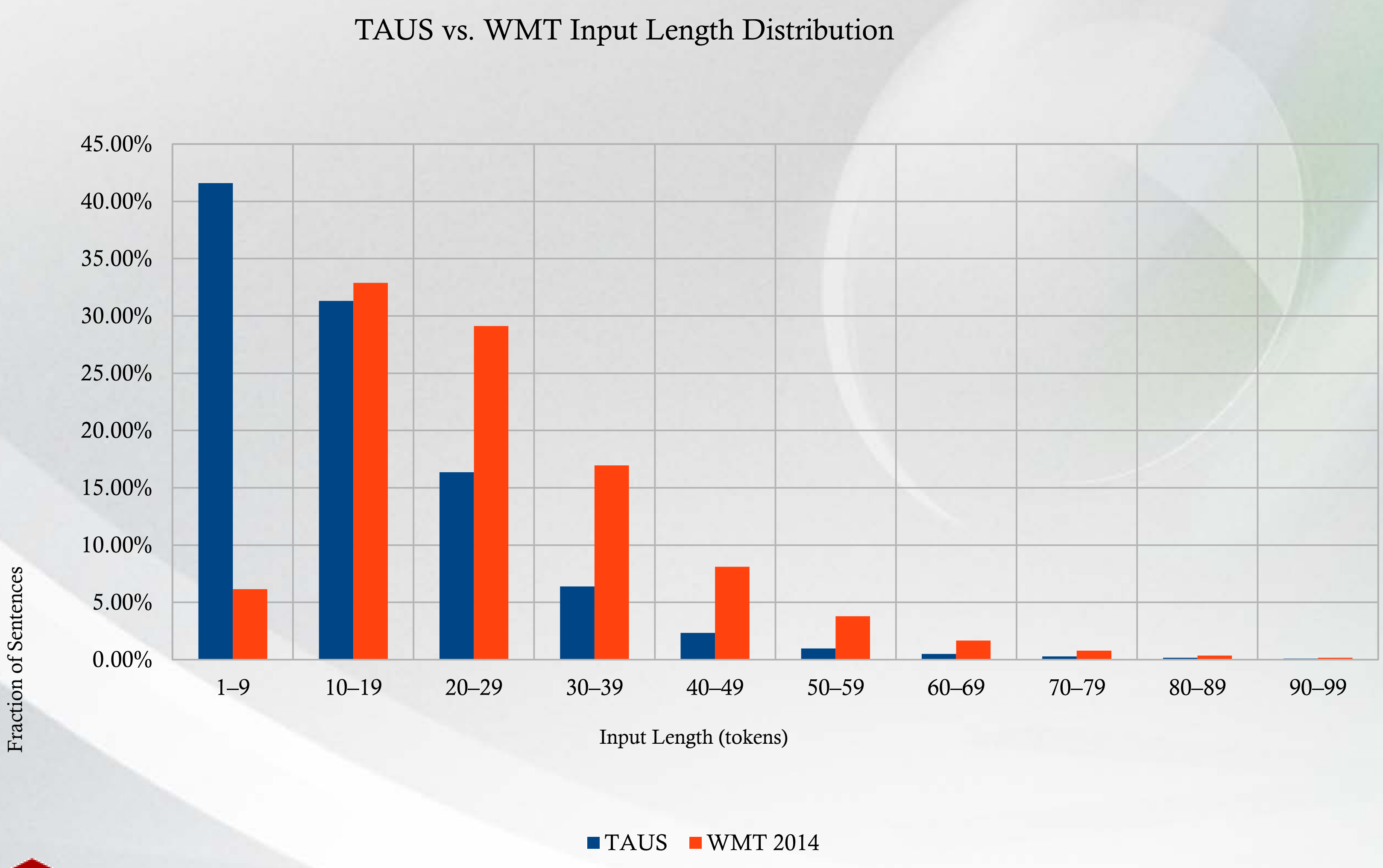
	Reference
1	Use the supply type on your bill to indicate who supplies components. For example, if your supplier provides components in addition to labor, you can include the components on your bill with a Supplier supply type. You can charge a job or repetitive schedule for these components by including 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.
2	Using Selection Details
3	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.
4	Number of past years
5	Weapons
6	IPSEC driver

WMT vs TAUS: EN-to-DE MT Systems

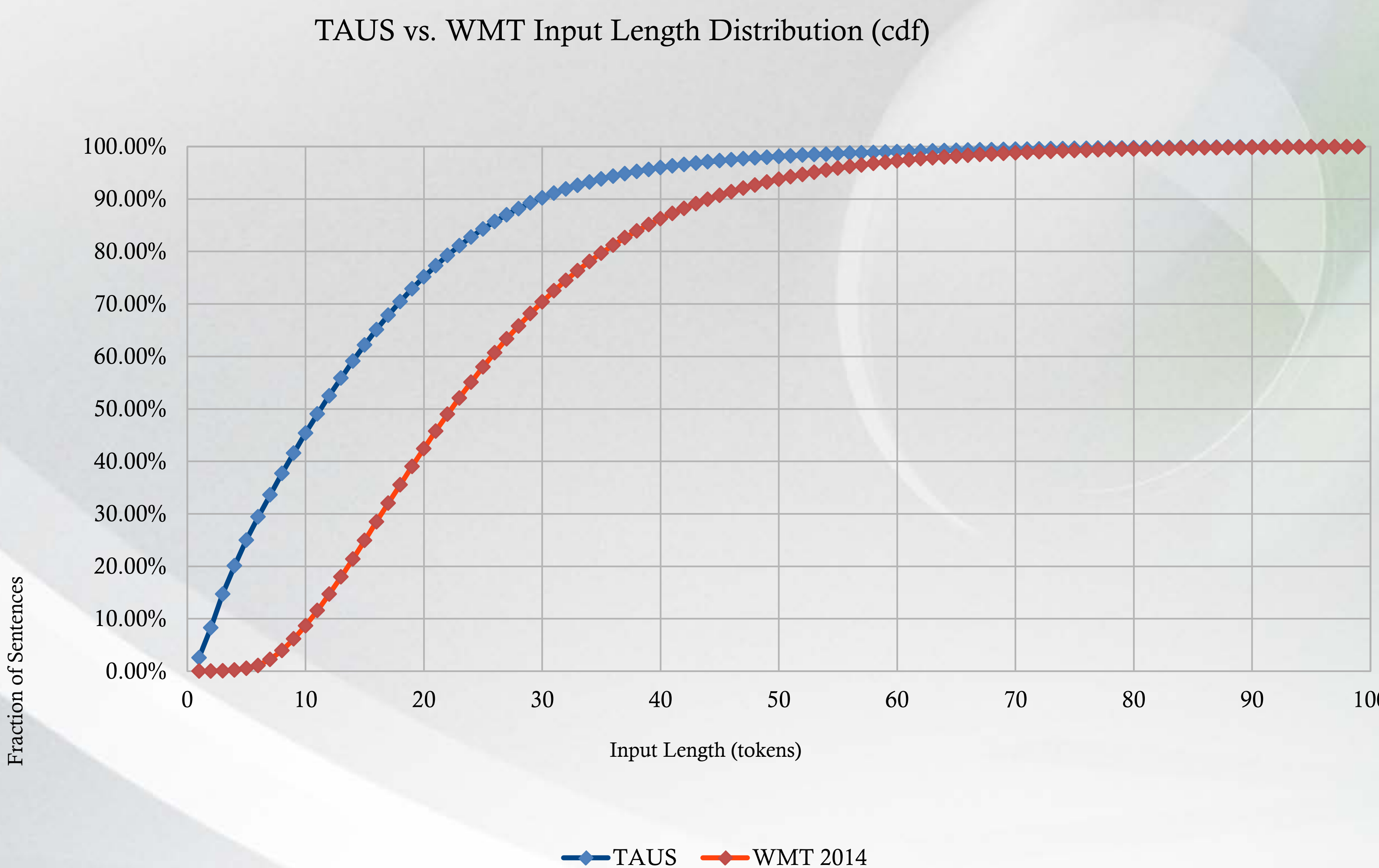
- Used Safaba default EN-to-DE pipeline to develop a WMT-2014 EN-to-DE MT system, as a contrastive reference to our TAUS EN-to-DE system
- Safaba WMT system:
  - Phrase-based system with domain adaptation
  - Constrained WMT-2014 parallel data resources only
  - No extra monolingual data for LM (i.e. 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:

	WMT 2014	TAUS Generic
Training Segments	4,143,962	5,767,915
Training Tokens (EN)	106,951,743	85,331,463
Training Tokens (DE)	101,810,648	89,190,947
Average tokens/segment EN	25.8	14.8
Average tokens/segment DE	24.6	15.5
Global length ratio DE/EN	95.2%	104.5%

WMT vs TAUS: EN-to-DE MT Systems



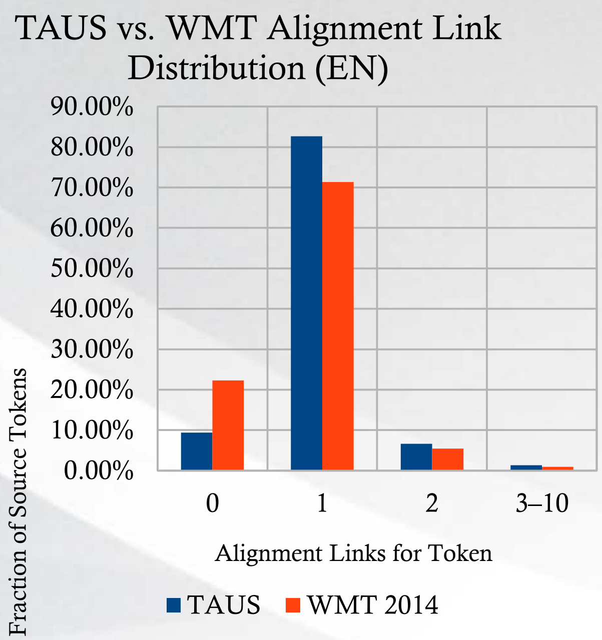
WMT vs TAUS: EN-to-DE MT Systems

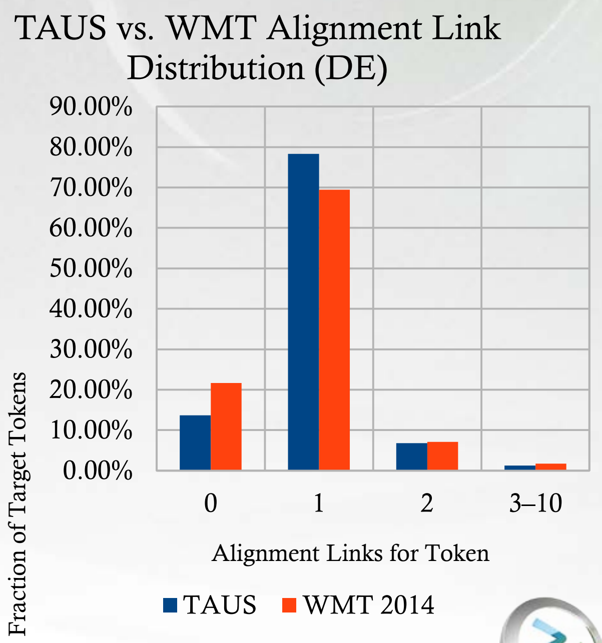


WMT vs TAUS: EN-to-DE MT Systems

> Word Alignment Statistics: Mgiza++, Grow-diag sym.

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





WMT vs TAUS: EN-to-DE MT Systems

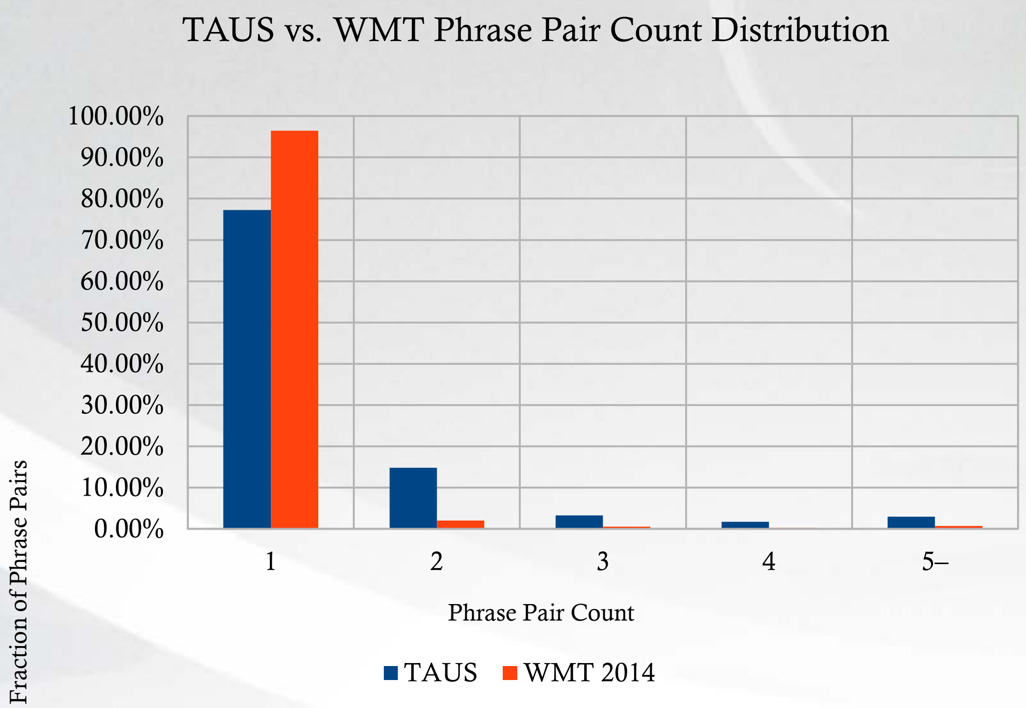
> Phrase Extraction Statistics:

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

WMT vs TAUS: EN-to-DE MT Systems

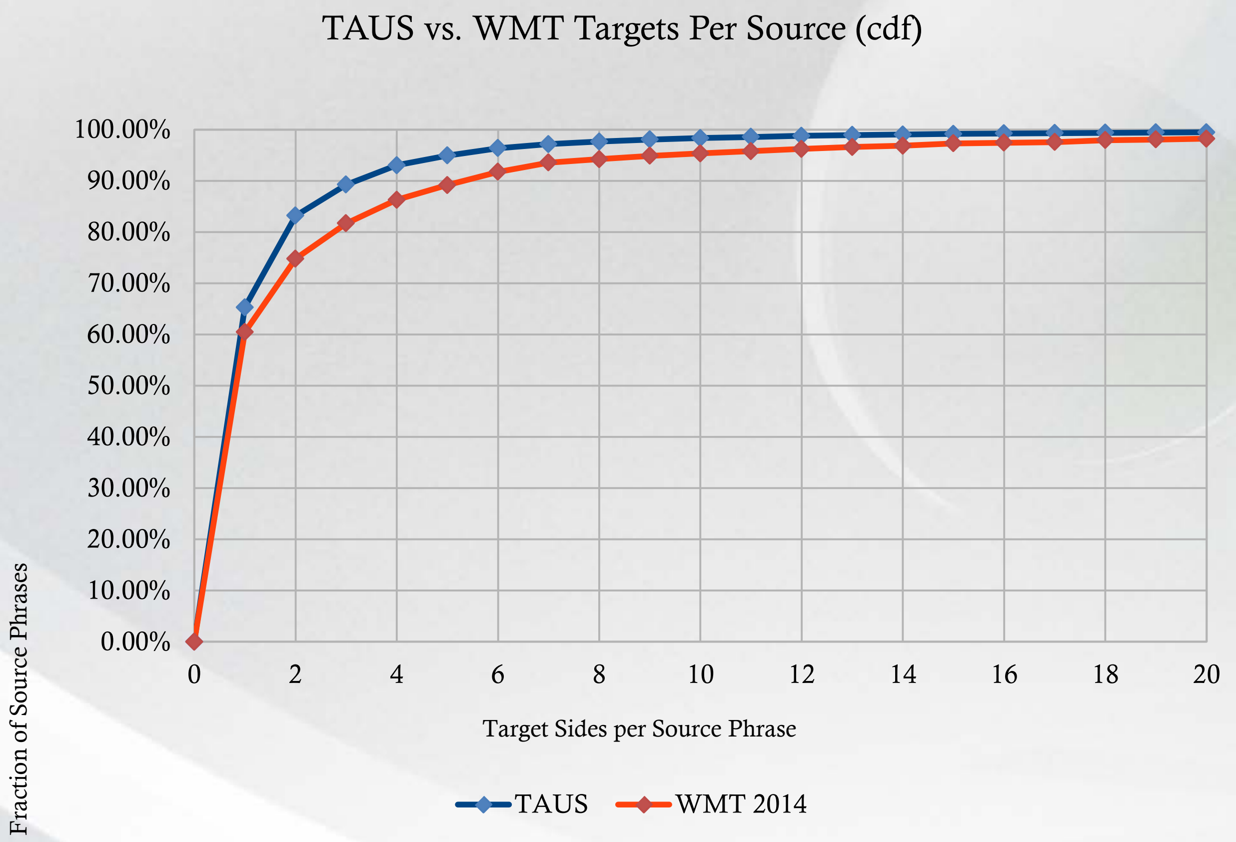
> Phrase Count Distribution Statistics:

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



WMT vs TAUS: EN-to-DE MT Systems

> Phrase Translation Ambiguity

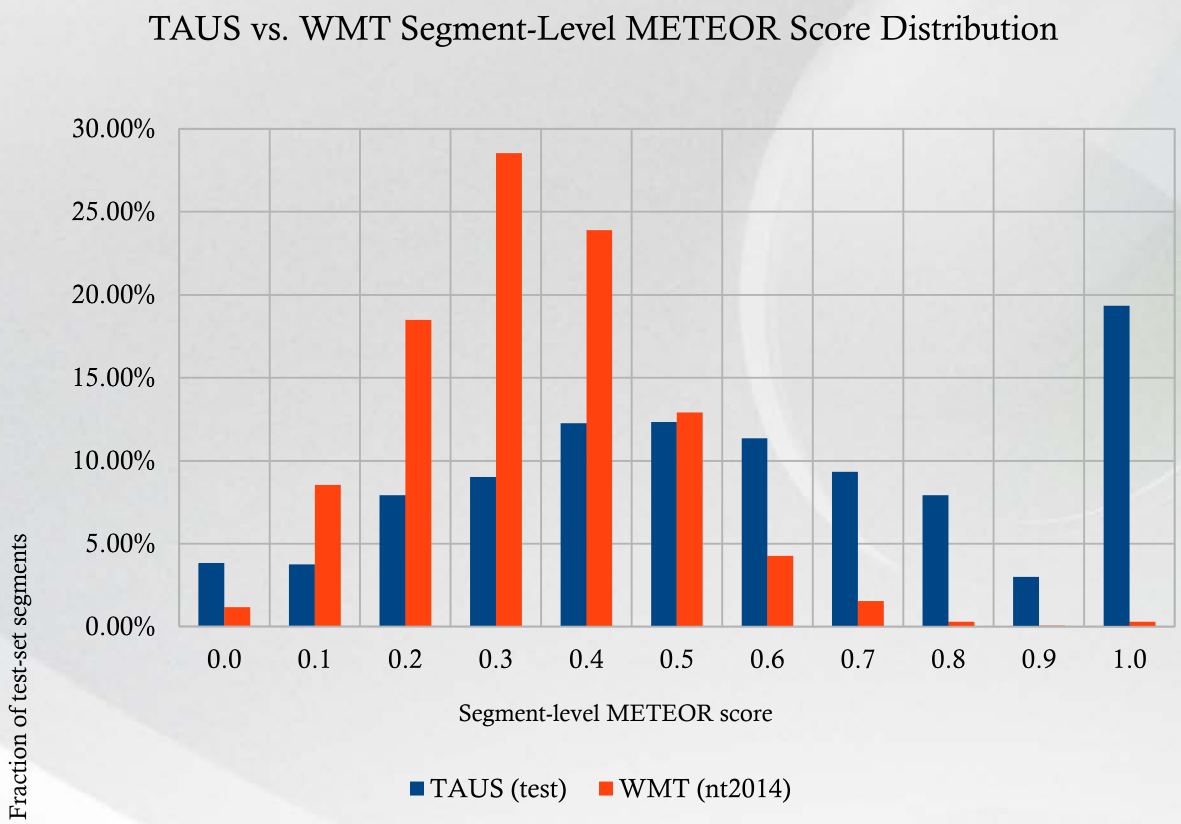


WMT vs TAUS: EN-to-DE MT Systems

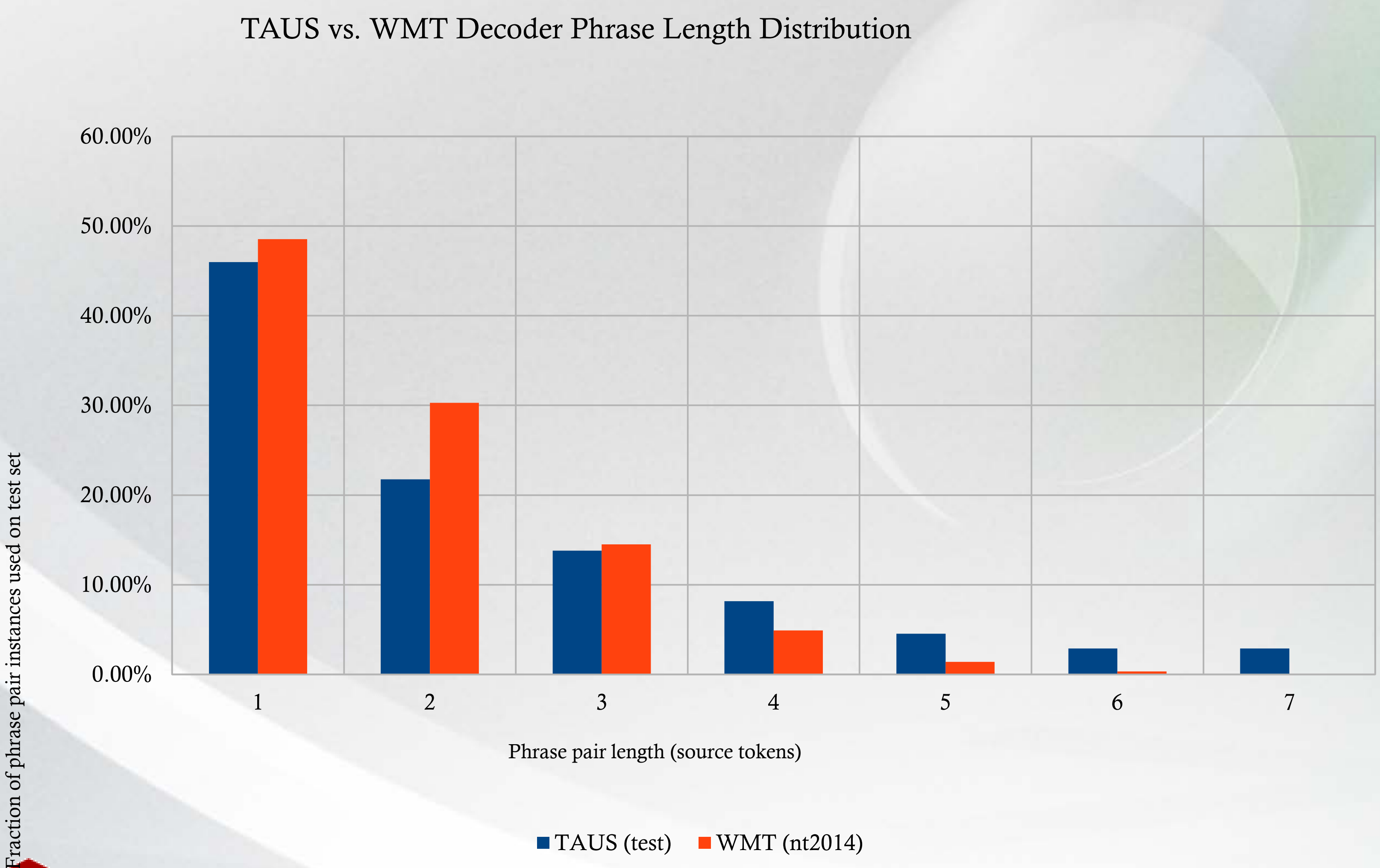
> Test-set Decoding Statistics:

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

WMT vs TAUS: EN-to-DE MT Systems



WMT vs TAUS: EN-to-DE MT Systems

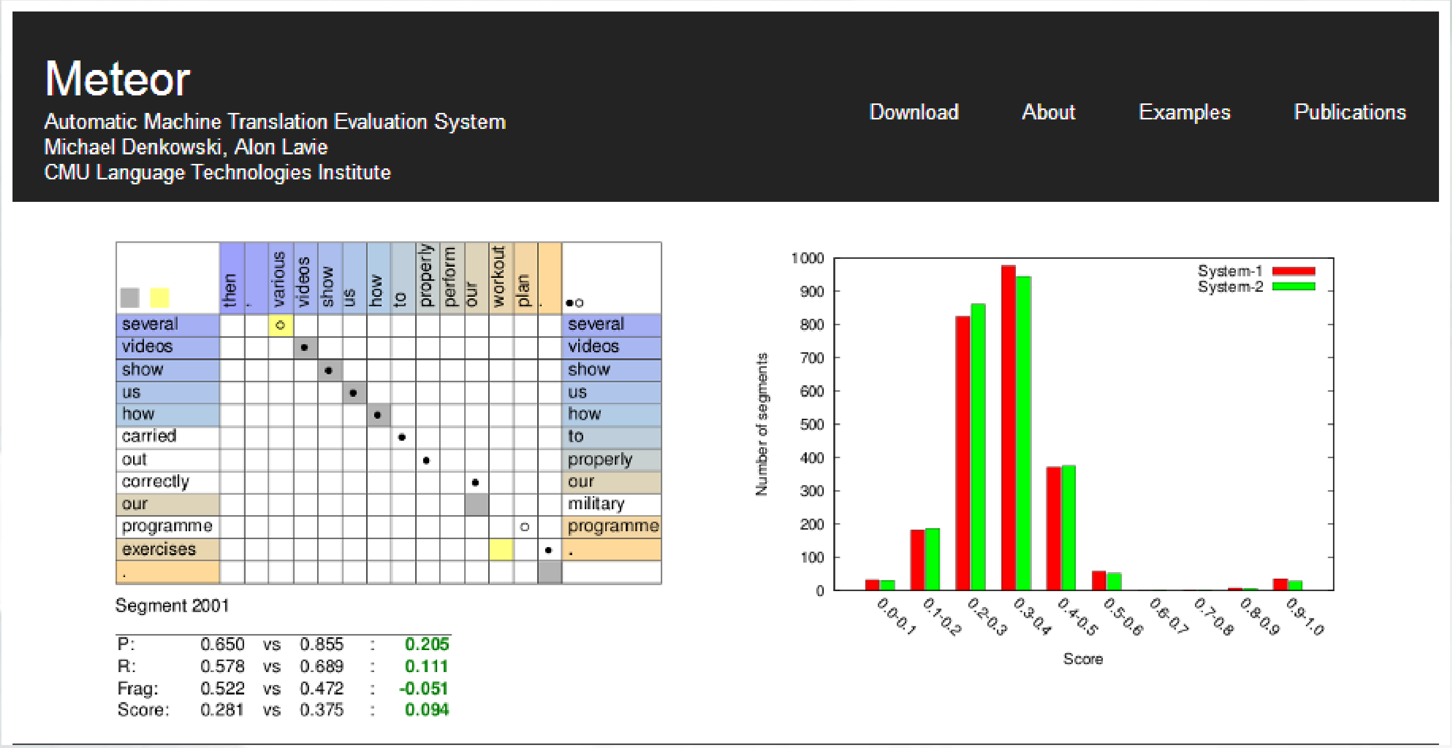


WMT vs TAUS MT Systems - Insights

- 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



Multilingual Meteor

> Meteor has expanded to cover 17 languages:

Fully supported languages:

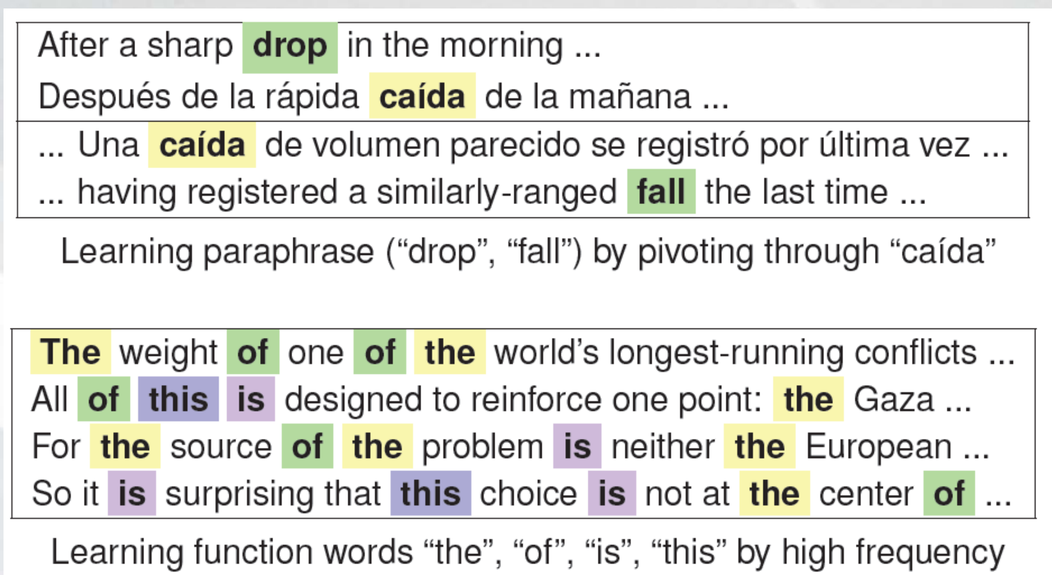
Language	Exact Match	Stem Match	Synonym Match	Paraphrase Match	Tuned Parameters
English	Yes	Yes	Yes	Yes	Yes
Arabic	Yes	No	No	Yes	Yes
Czech	Yes	No	No	Yes	Yes
French	Yes	Yes	No	Yes	Yes
German	Yes	Yes	No	Yes	Yes
Spanish	Yes	Yes	No	Yes	Yes

Partially supported languages:

Language	Exact Match	Stem Match	Synonym Match	Paraphrase Match	Tuned Parameters
Danish	Yes	Yes	No	No	LI
Dutch	Yes	Yes	No	No	LI
Finnish	Yes	Yes	No	No	LI
Hungarian	Yes	Yes	No	No	LI
Italian	Yes	Yes	No	No	LI
Norwegian	Yes	Yes	No	No	LI
Portuguese	Yes	Yes	No	No	LI
Romanian	Yes	Yes	No	No	LI
Russian	Yes	Yes	No	No	LI
Swedish	Yes	Yes	No	No	LI
Turkish	Yes	Yes	No	No	LI

Meteor Universal

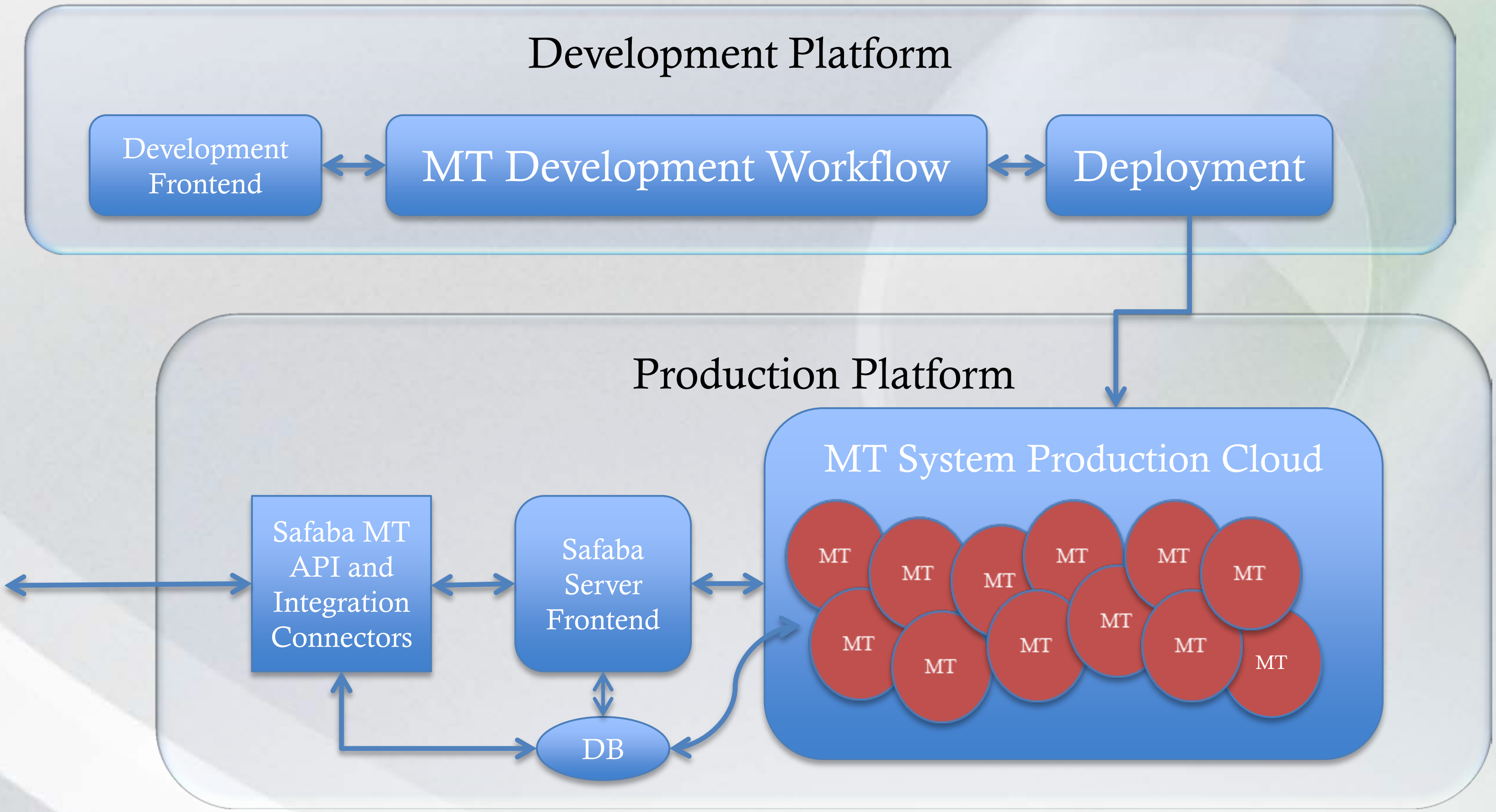
- [Denkowski and Lavie, 2014] WMT-2014 Metrics Task
- New support included in Meteor 1.5:
  - Support for **any target language** using only **bi-text** used to build statistical MT systems
  - Learn paraphrases by phrase pivoting (Bannard and Callison-Burch, 2005)
  - Learn function words by relative 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**.



Learning paraphrase ("drop", "fall") by pivoting through "caída"

Learning function words "the", "of", "is", "this" by high frequency

Safaba – MT Architecture Overview



MT Development Workflow Management

- 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**

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

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: 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
}
```

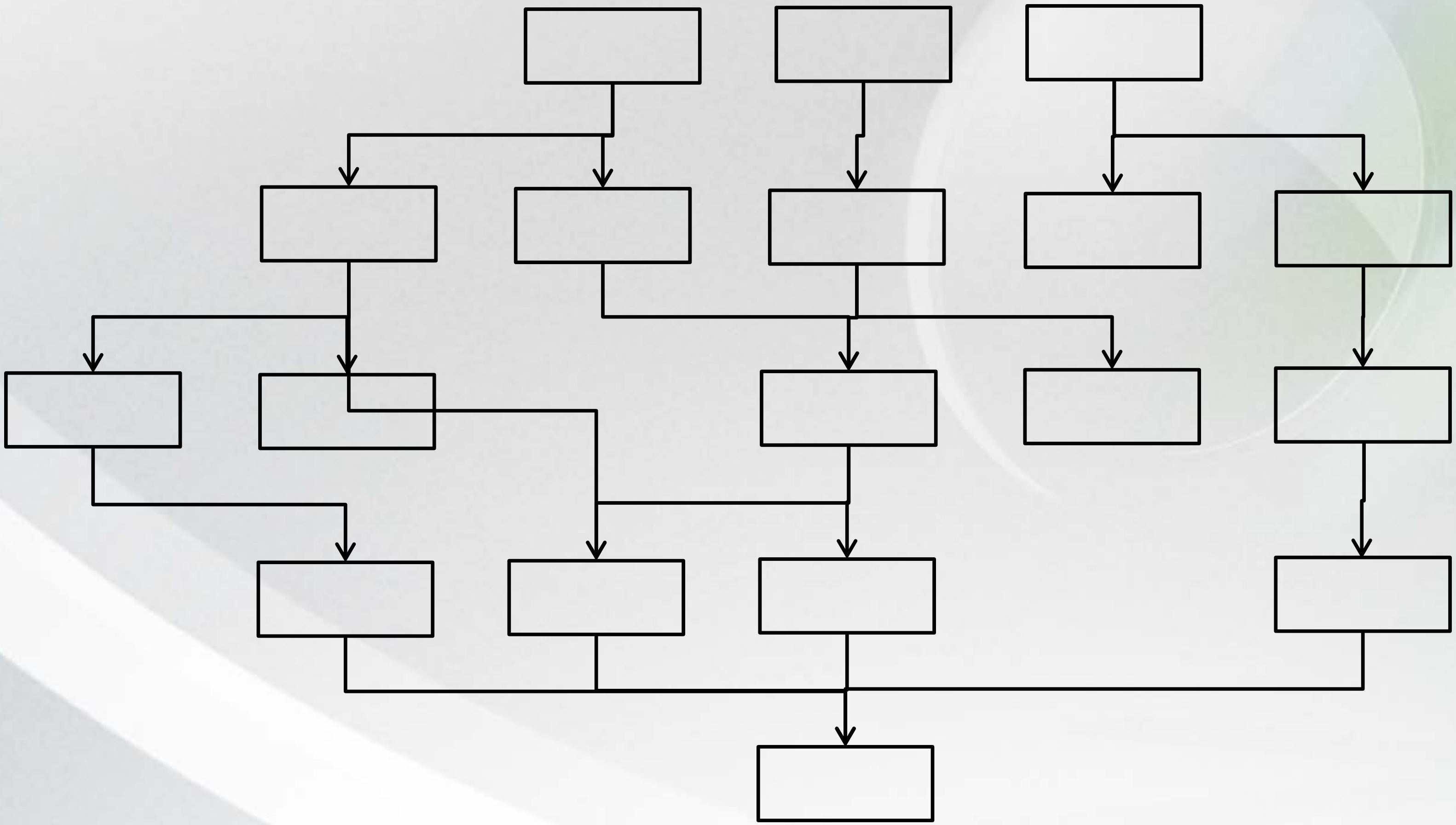
```
task align_mgiza_direction : mgiza
    < src_classes=$classes@align_mkcls_src
    < tgt_classes=$classes@align_mkcls_tgt
    < ...
    > src_tgt_alignments
    :: ...
{
    ...
}
```

DuctTape: 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: Tasks



DuctTape: Branch Points

```
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

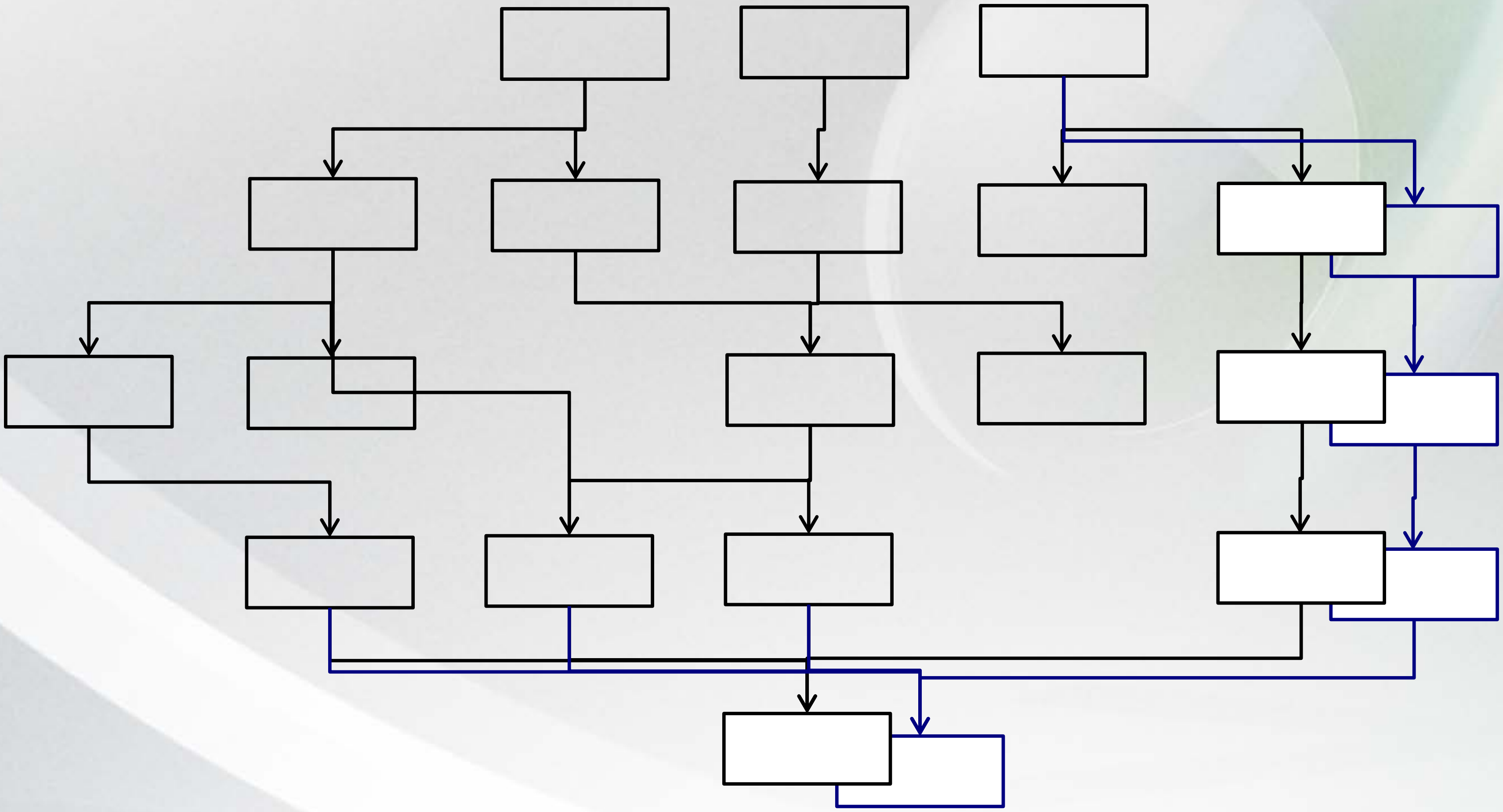
num_classes=1000

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

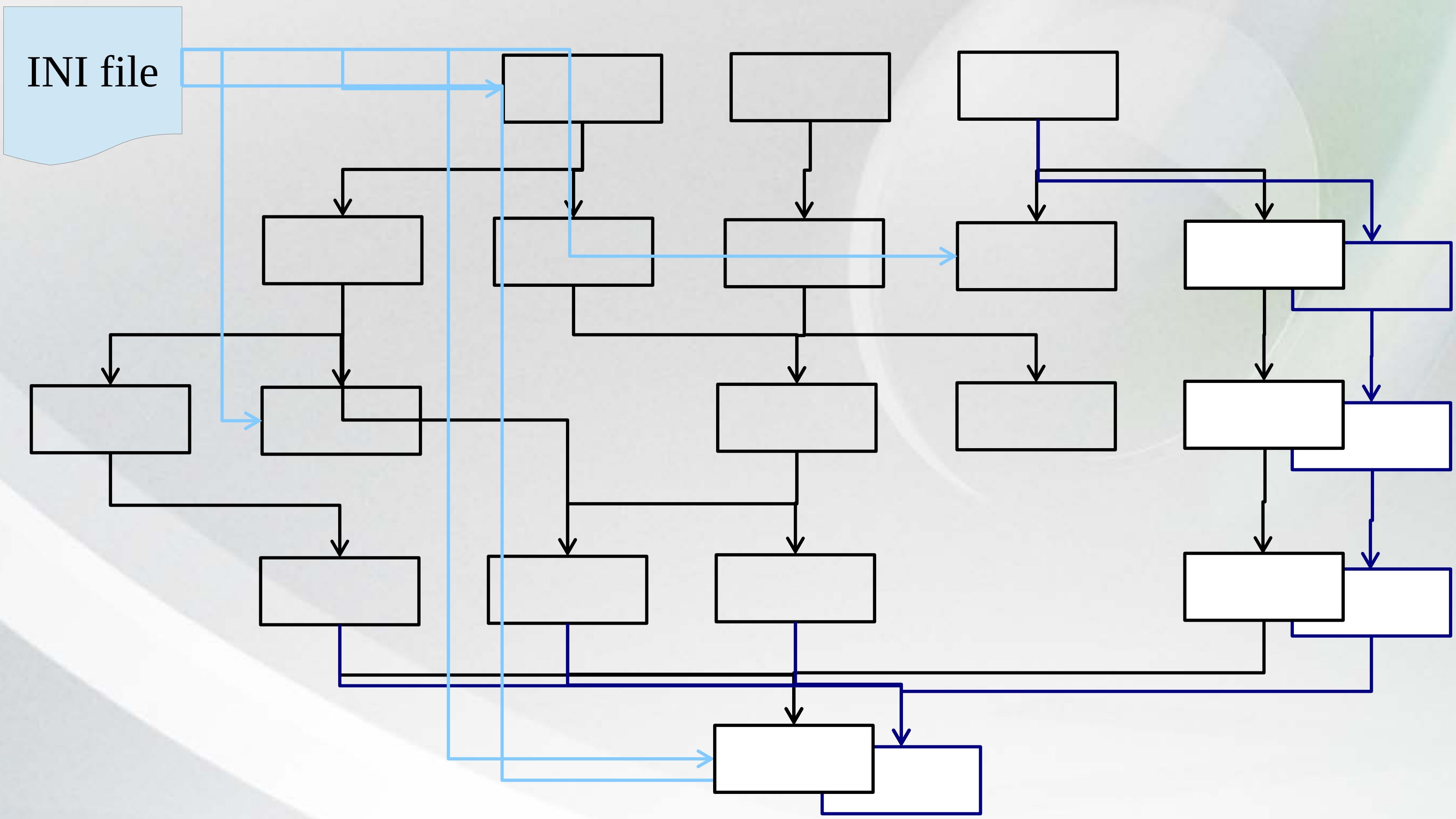

```
giza_direction : mgiza
asses=$classes@align_mkcls_src
asses=$classes@align_mkcls_tgt

_alignments
```

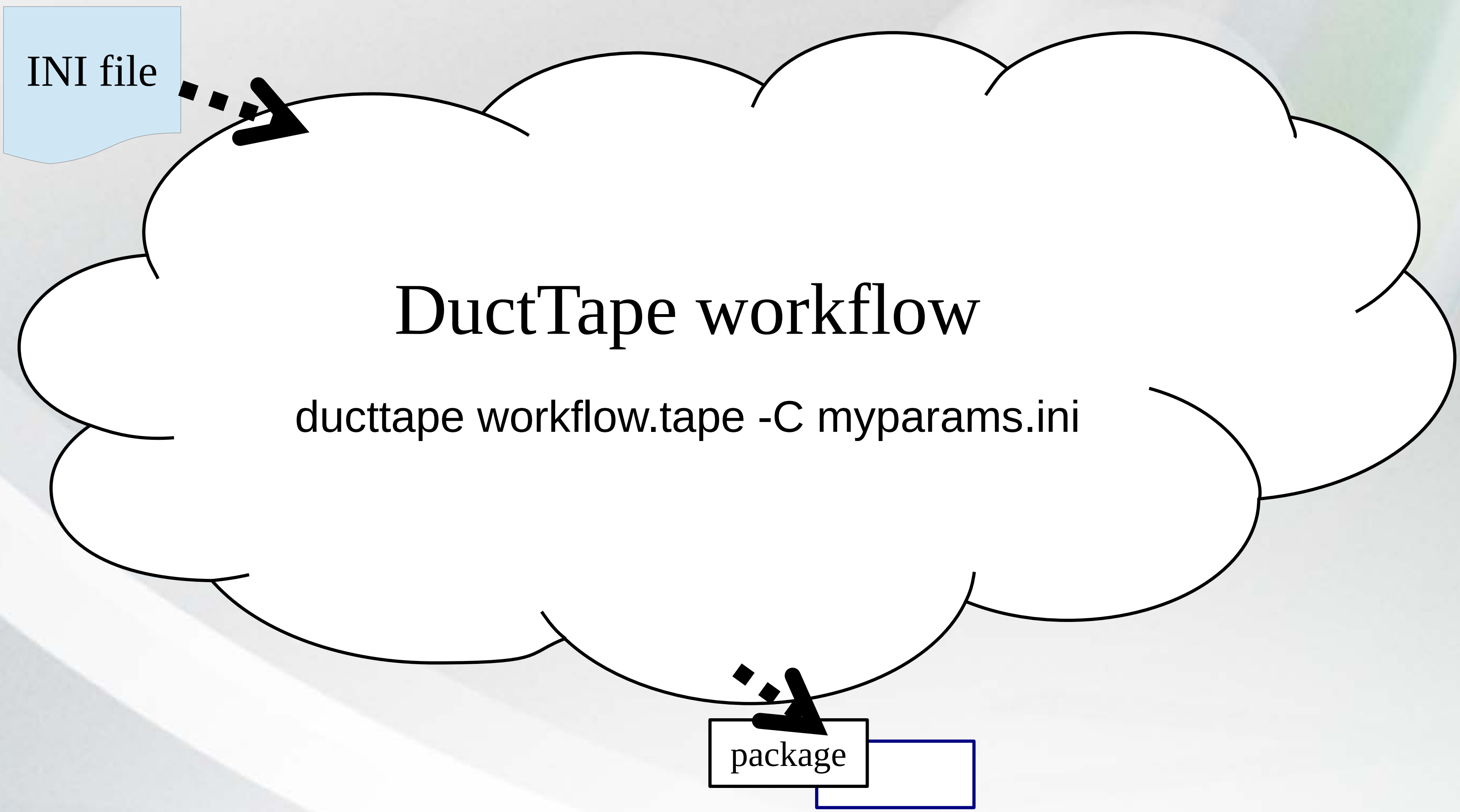
DuctTape: Branch Points



DuctTape: Workflows



DuctTape: Workflows



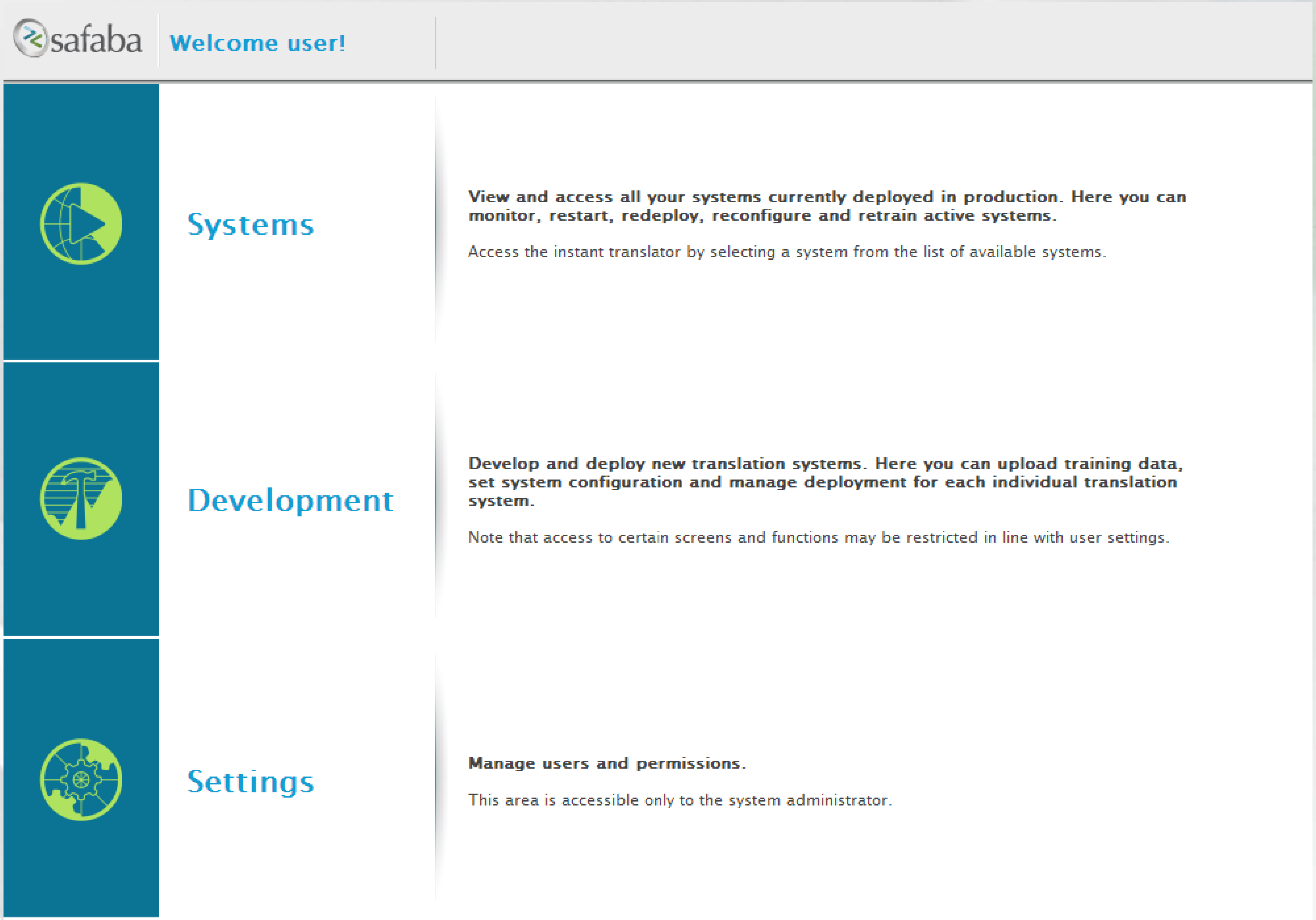
Safaba MT Deployment Process

- 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

Safaba EMTGlobal™ Online



Safaba EMTGlobal™ Online

safaba Welcome user!

Home> Development

EMTGlobal™ ONLINE

BUILD A NEW TRANSLATION SYSTEM

SYSTEMS | DEVELOPMENT | SETTINGS

Demo

- ENUS-DEDE, IT Hardware
- ENUS-FRCA, My system
- ENUS-HEIL, Yakov
- ESES-ENUS, Software Support

Select Project / System

Available Translation Systems :

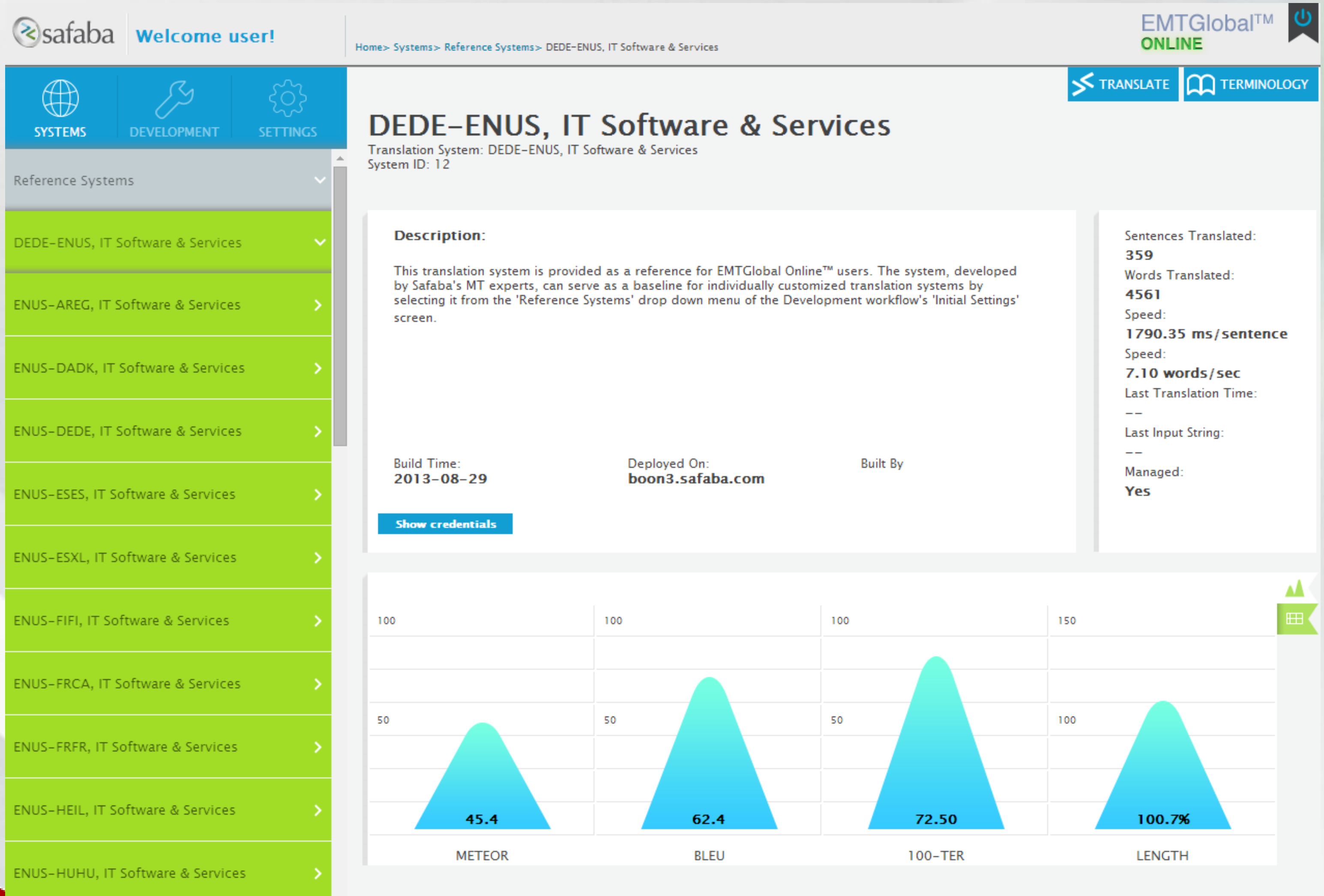
Demo:

SYSTEM NAME	TYPE	OWNER	STATUS	ACTIONS
ENUS-DEDE, IT Hardware				
Version : 14.02.04	FULL	User (Read-Only)	Deployed	
Version : 14.02.19	FULL	User (Read-Only)	Ready to build	Remove
Version : 14.03.30	FULL	User (Read-Only)	Configuring initial settings	Remove
ENUS-FRCA, My system				
Version : 14.03.05	FULL	User (Read-Only)	Deployed	
Version : 14.03.30	FULL	Expert (Read-Only)	Configuring language optimization settings	View
Version : 14.03.31	FULL	Expert (Read-Only)	Uploading data	View
ENUS-HEIL, Yakov				
Version : 14.05.28	MPE	Expert (Read-Write)	Uploading data	Remove
ESES-ENUS, Software Support				
Version : 14.03.05	MPE	User (Read-Only)	Configuring styling settings	Remove

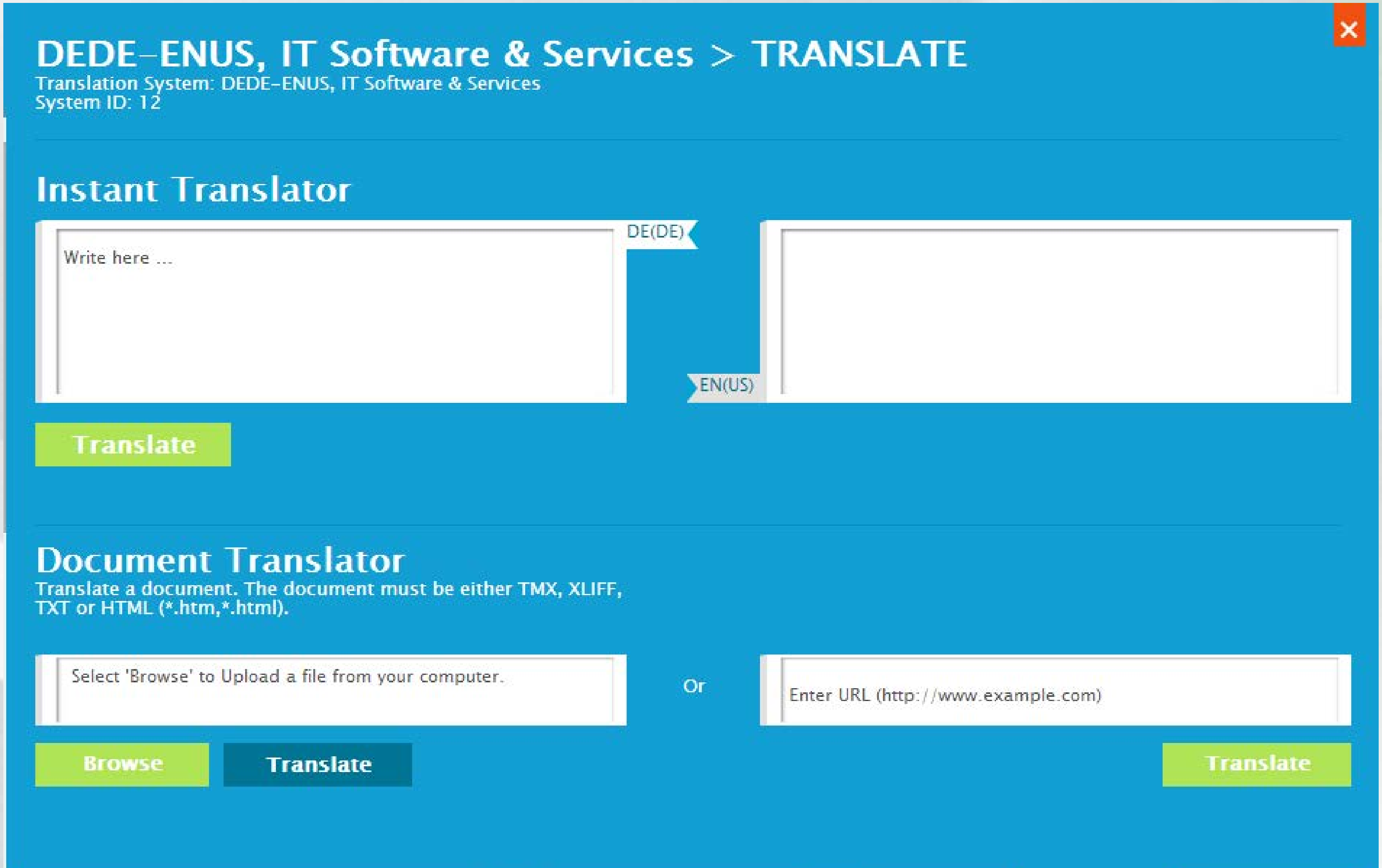
lti

safaba

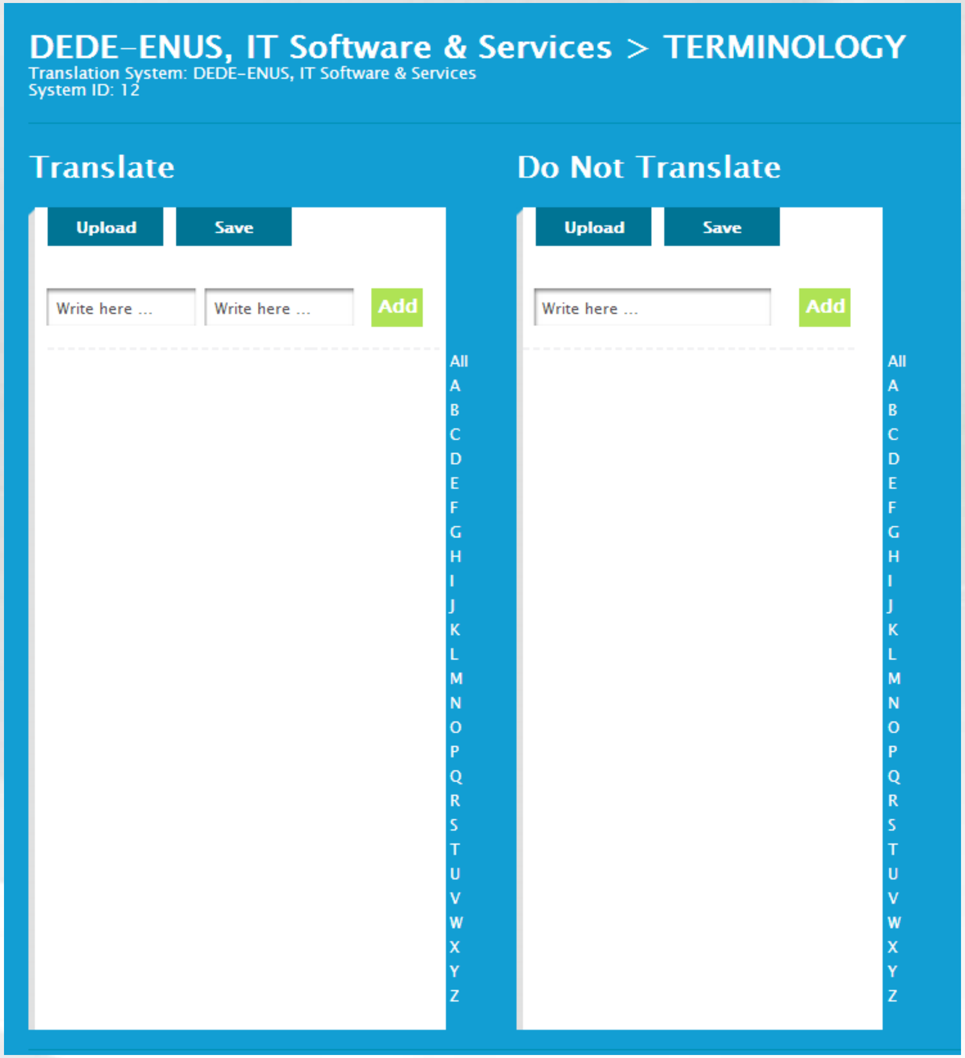
Safaba EMTGlobal™ Online



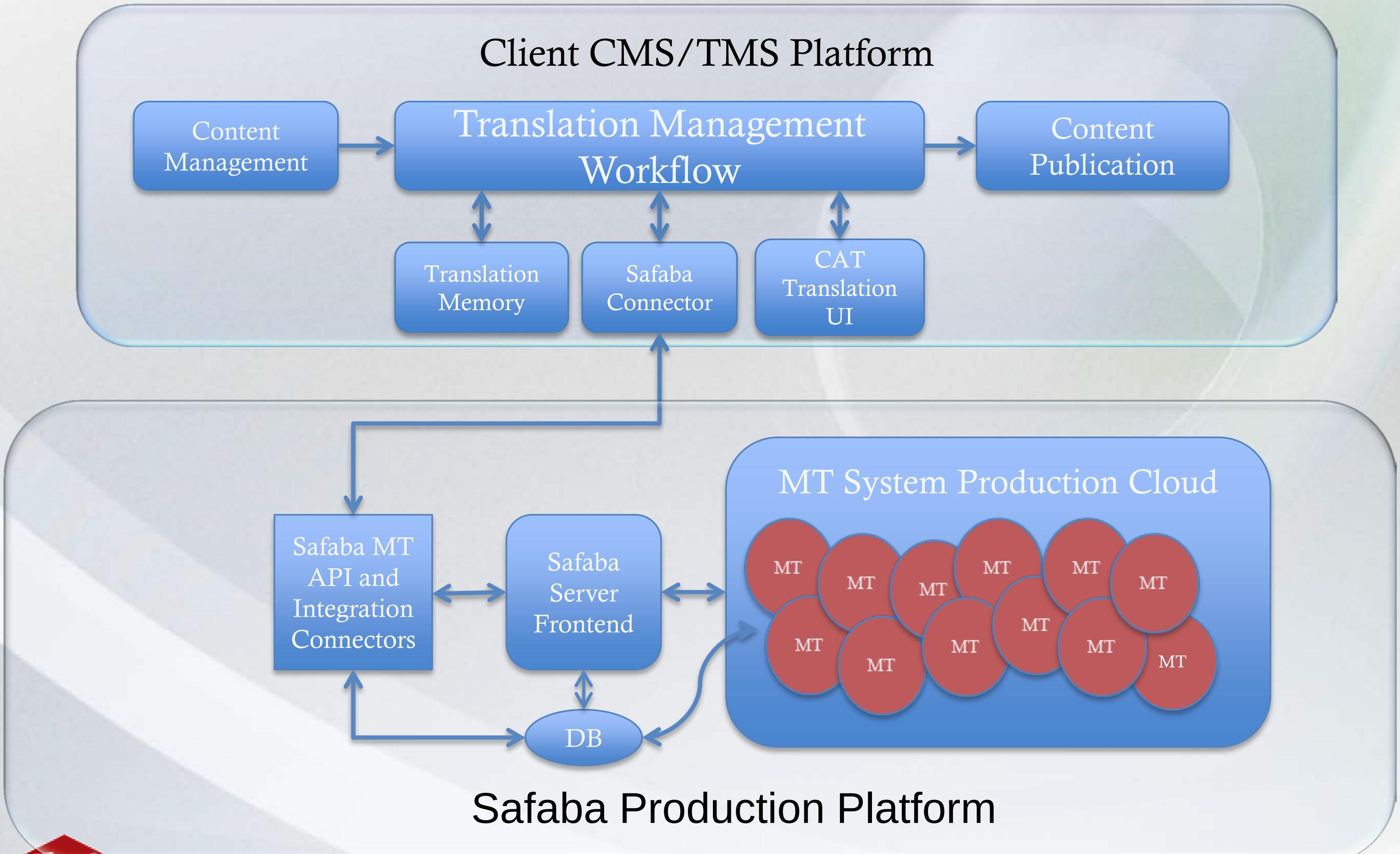
Safaba EMTGlobal™ Online



Safaba EMTGlobal™ Online



External Workflow Integrations

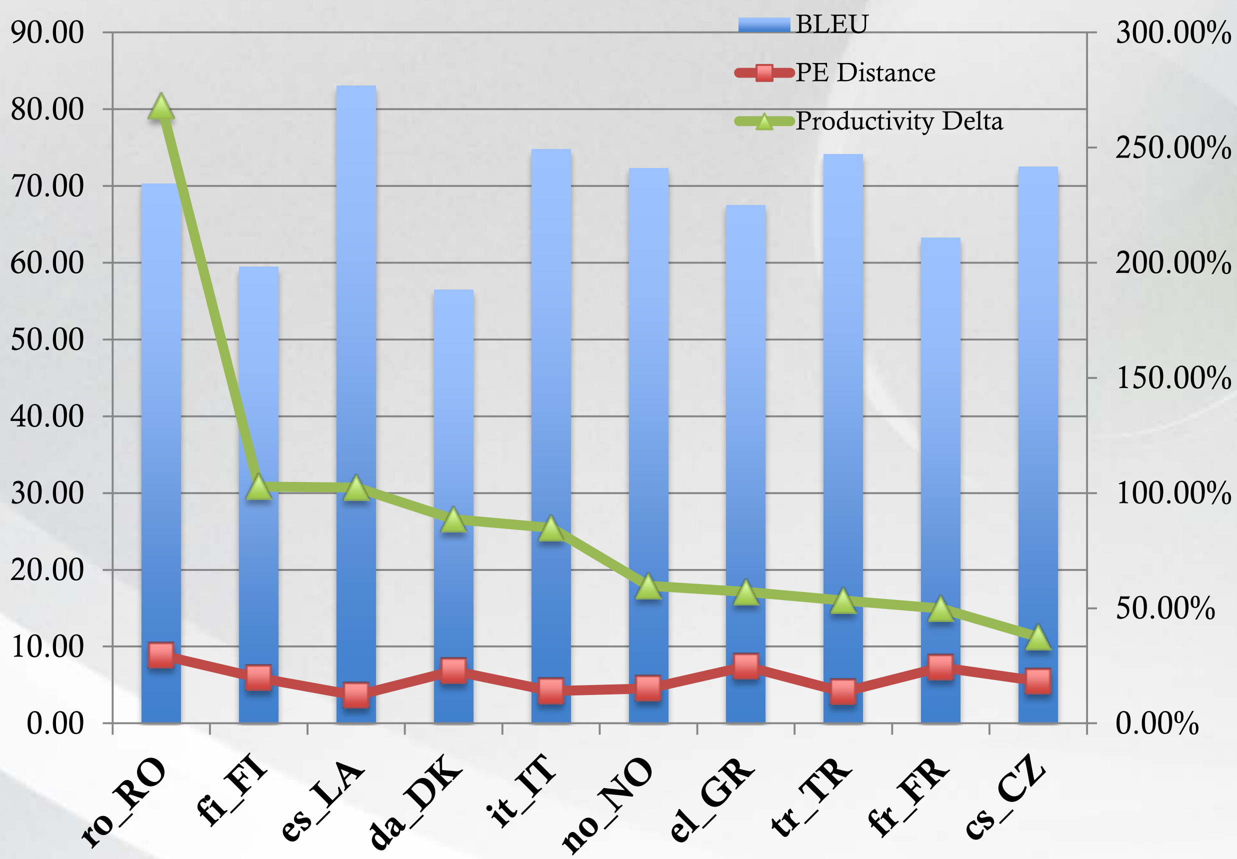


Translation with MT Post-Editing

- 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



Challenge: Structured Content Translation

- 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>.

Challenge: Structured Content Translation

- Structured Tag Projection Process:

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

Challenge: Structured Content Translation

- 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">		</x>	
les ordinateurs de bureau		*les plus populaires*		*pour l'école et la maison*

Challenge: Structured Content Translation

- Structured Tag Projection Process:
  - Translate pure text segment and preserve word and phrase alignments.

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*

Challenge: Structured Content Translation

- 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">*popular*</x> *desktops for school and home*

Tag Projection Accuracy Evaluation

- **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

	Context Matched				
Tag Type	**Both**	**Left**	**Right**	**Neither**	**Total**
Beginning	33.33%	19.44%	11.46%	35.76%	100.00%
Ending	32.06%	10.10%	8.01%	49.83%	100.00%
Stand-alone	56.91%	23.98%	18.29%	0.81%	100.00%
Total	**39.95%**	17.54%	12.30%	**30.21%**	**100.00%**

EMTGlobal version 2.1

	Contexts Matched				
Tag Type	**Both**	**Left**	**Right**	**Neither**	**Total**
Beginning	66.67%	12.50%	9.38%	11.46%	100.00%
Ending	63.41%	10.80%	11.50%	14.29%	100.00%
Stand-alone	67.89%	18.29%	13.01%	0.81%	100.00%
Total	**65.90%**	13.64%	11.21%	**9.26%**	**100.00%**

- 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%)

Client-Specific Adaptation

- 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:

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:

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:

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:

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:

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:

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

“Overnight” Incremental Adaptation

- **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

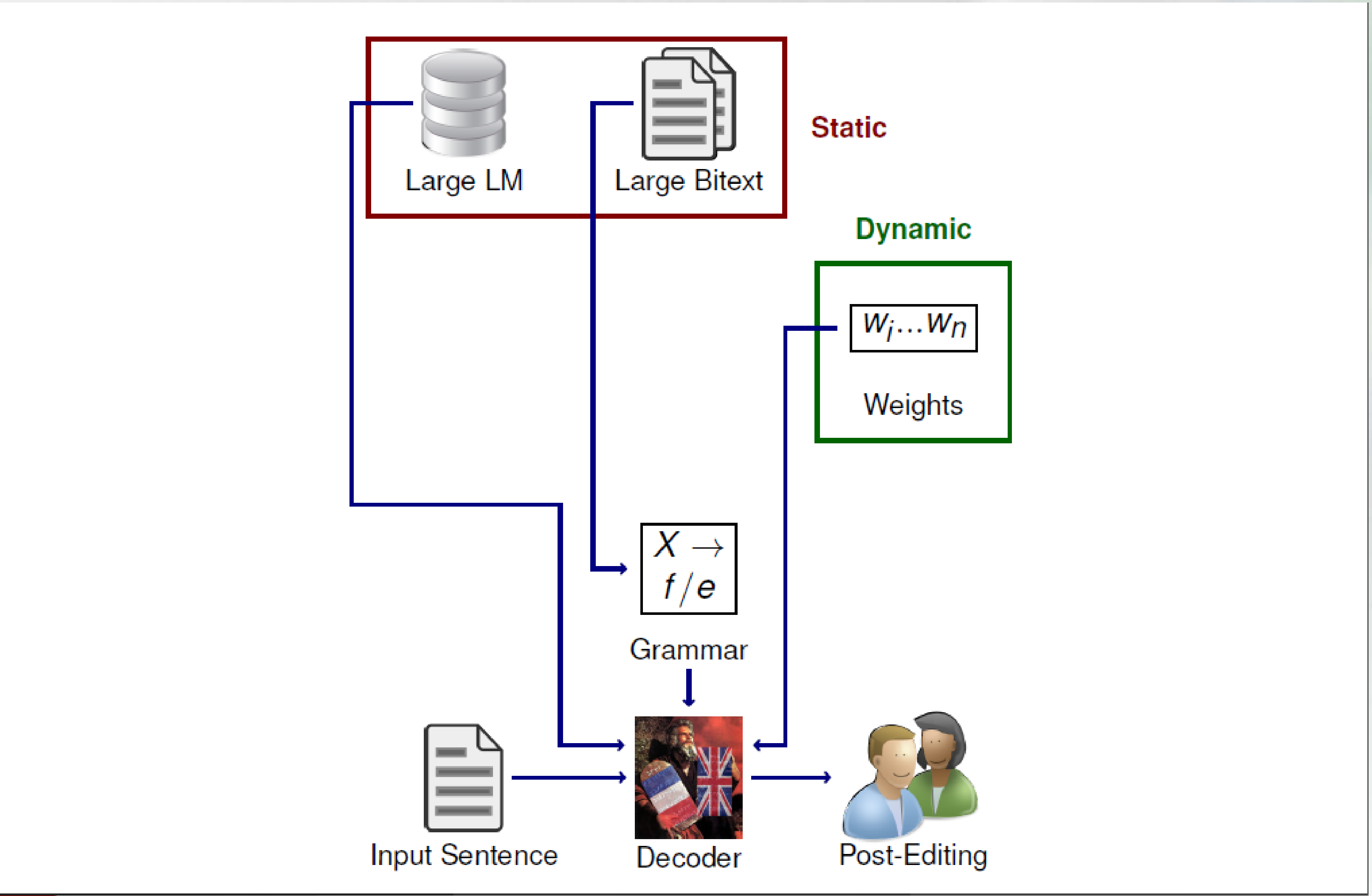
- **Ultimate Goal:** immediate online feedback loop between MT post-editing and the live MT system in the background
- **Engineering Challenge:** requires a fully integrated online solution where the MT post-editors translation environment is directly connected to the real-time MT engine, and feeds back 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:
  - [Denkowski, Dyer and Lavie, 2014] EACL 2014
  - [Denkowski, Lavie, Lacruz and Dyer, 2014] EACL 2014 Workshop on Humans and Computer-assisted Translation

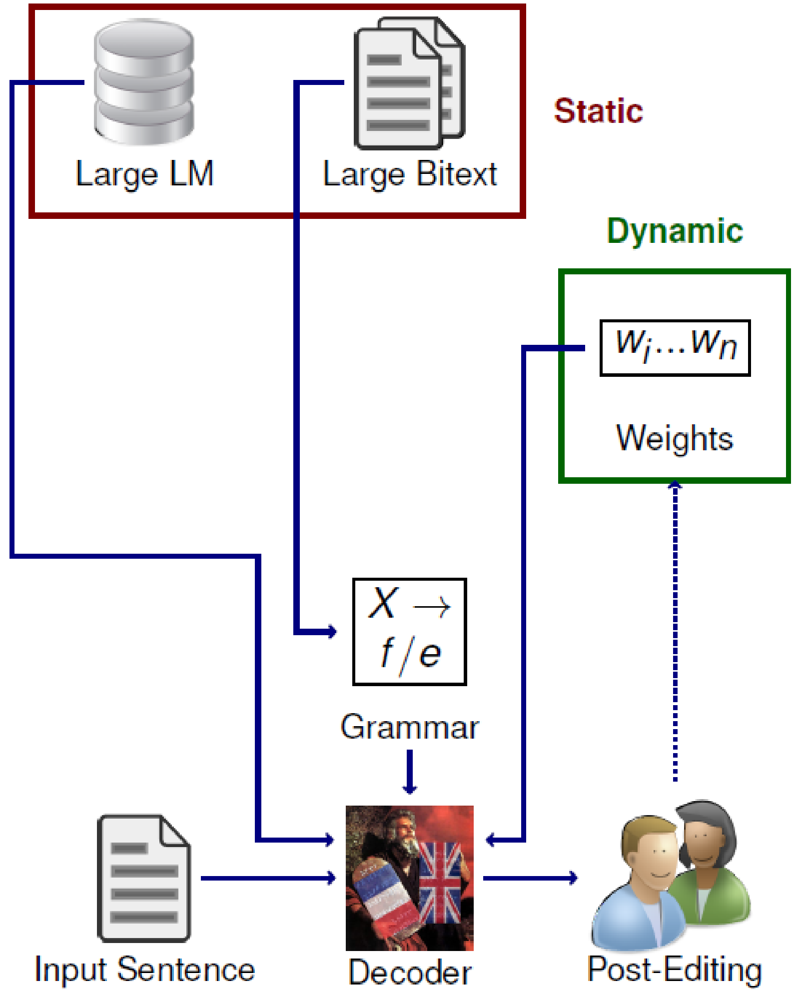
Real-time Online Adaptation

- **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

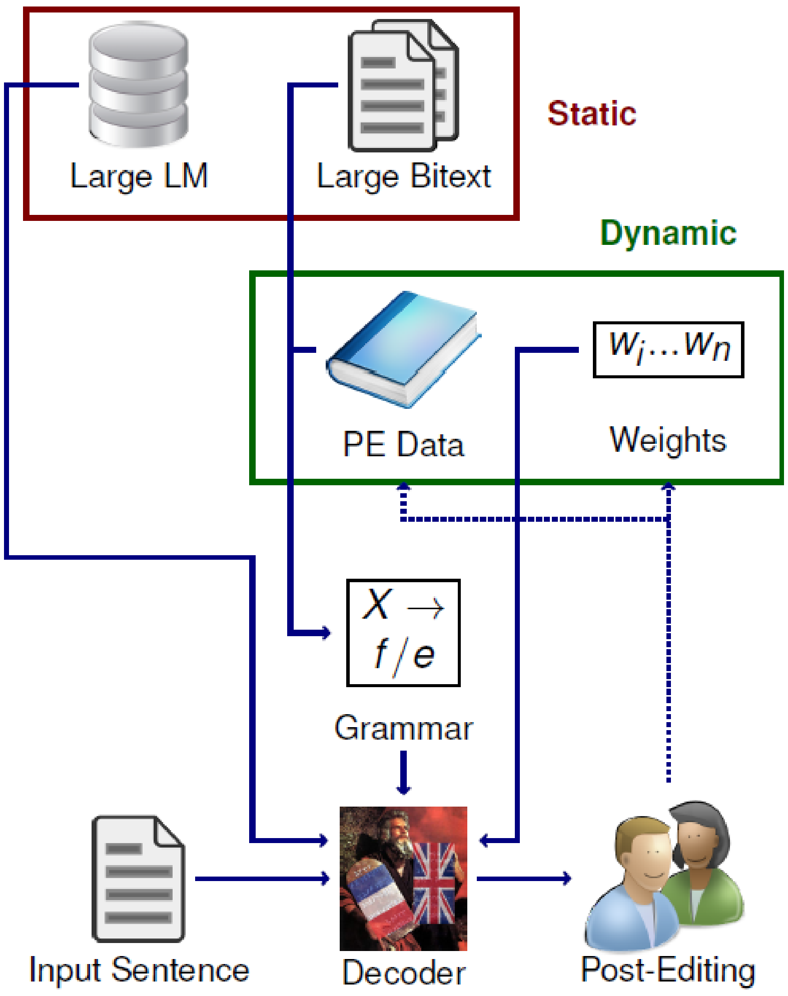
Real-time Online Adaptation



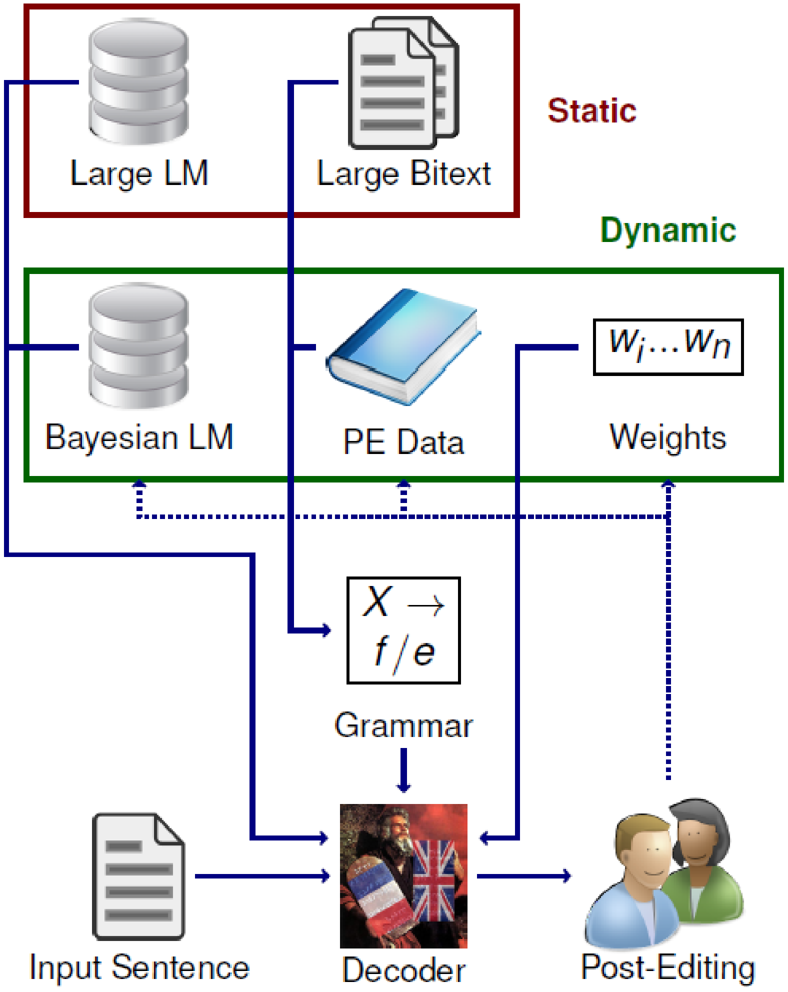
Real-time Online Adaptation



Real-time Online Adaptation



Real-time Online Adaptation



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:**
  - $c_{\mathcal{S}}(f,e)$, $c_{\mathcal{L}}(f,e)$: counts of f aligning to e
  - $c_{\mathcal{S}}(f)$, $c_{\mathcal{L}}(f)$: counts of f aligning to anything
  - $|\mathcal{S}|$, $|\mathcal{L}|$: sample sizes (occurrences of f, aligned or not)

Real-time Online Adaptation

Feature	Static	Adaptive
coherent $p(e\|f)$	$\frac{C_{\mathcal{S}}(f,e)}{\|\mathcal{S}\|}$	$\frac{C_{\mathcal{S}}(f,e)+C_{\mathcal{L}}(f,e)}{\|\mathcal{S}\|+\|\mathcal{L}\|}$
sample size	$\|\mathcal{S}\|$	$\|\mathcal{S}\|+\|\mathcal{L}\|$
co-occurrence $\langle f,e\rangle$	$C_{\mathcal{S}}(f,e)$	$C_{\mathcal{S}}(f,e)+C_{\mathcal{L}}(f,e)$
singleton f	$C_{\mathcal{S}}(f)=1$	$C_{\mathcal{S}}(f)+C_{\mathcal{L}}(f)=1$
singleton $\langle f,e\rangle$	$C_{\mathcal{S}}(f,e)=1$	$C_{\mathcal{S}}(f,e)+C_{\mathcal{L}}(f,e)=1$
post-edit support $\langle f,e\rangle$	0	$C_{\mathcal{L}}(f,e)>0$

Phrase features (rule level)

Real-time Online Adaptation

- **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]

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:**

	Dev	In-Dom	Out-of-Dom	
Spanish–English	WMT10	WMT11	TED1	TED2
Baseline	29.2	28.0	32.7	29.7
Grammars	**29.8**	**28.3**	**34.2**	**30.7**
LM	29.2	**28.1**	**33.0**	**29.8**
MIRA	29.2	**28.1**	**33.1**	**29.8**
G+L+M	**30.0**	**28.8**	**35.2**	**31.3**
Arabic–English	MT08	MT09	TED1	TED2
Baseline	21.2	25.9	10.6	10.9
Grammars	**21.8**	**26.2**	**11.0**	**11.7**
LM	20.6	25.7	10.6	10.9
MIRA	**21.3**	25.7	**10.8**	**11.0**
G+L+M	**21.8**	**26.5**	**11.4**	**11.8**

Real-time Online Adaptation

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

Talk 5 — Pause | Submit | ?

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

Real-time Online Adaptation

- 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

	HTER	Rating	SPE BLEU
Baseline	19.26	4.19	34.50
Adaptive	**17.01**	**4.31**	**34.95**

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

Thank You!