I Ate Too Much Cake: Beyond Domain-Specific MT Engines

Alex Yanishevsky

welocalize
doing things differently

How Much Cake Is Too Much Cake?

What is the Tipping Point?

welocalize
doing things differently

AGENDA

Recap of Previous Experiments

Challenges for Mature MT Programs

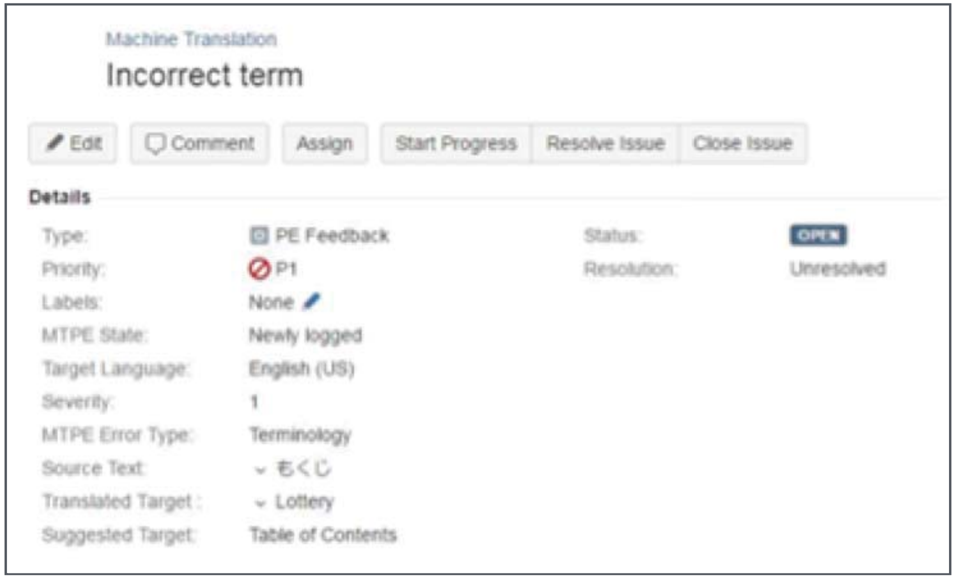
Opportunities for Mature MT Programs

Recap of Previous Experiments // Part 1

- Criteria for training domain-specific engines
  - ✓ Environment: elegant deployment?
  - ✓ Cost
  - ✓ How different are they from each other
  - ✓ Maintenance (engineering and linguistic feedback implementation)
- Trained and deployed over 50 engines in 13 languages (to and from English)
- Corrected over 300 linguistic issues

Recap of Previous Experiments // Part 2

Implementing Linguistic Feedback



Key	Summary & Description	Target Language	Source Text	Translated Target	Suggested Target	Translation from New Engine	Comments
MTR-136	term (Portable devices) Portable device = terminal mobile	French	Portable devices	Appareils portables	Terminaux mobiles	Terminaux mobiles	Term translation correct do NOT add to tune or UD: checked with lead translator, and both translations are correct depending on context
MTR-143	term & gender (appliance) appliance = appliance (masculin)	French	appliance	dispositif	appliance	appliance	add to UD.

Recap of Previous Experiments // Part 3

Savings on MT per quarter

Recap of Previous Experiments // Part 4

MT Usage Per Month

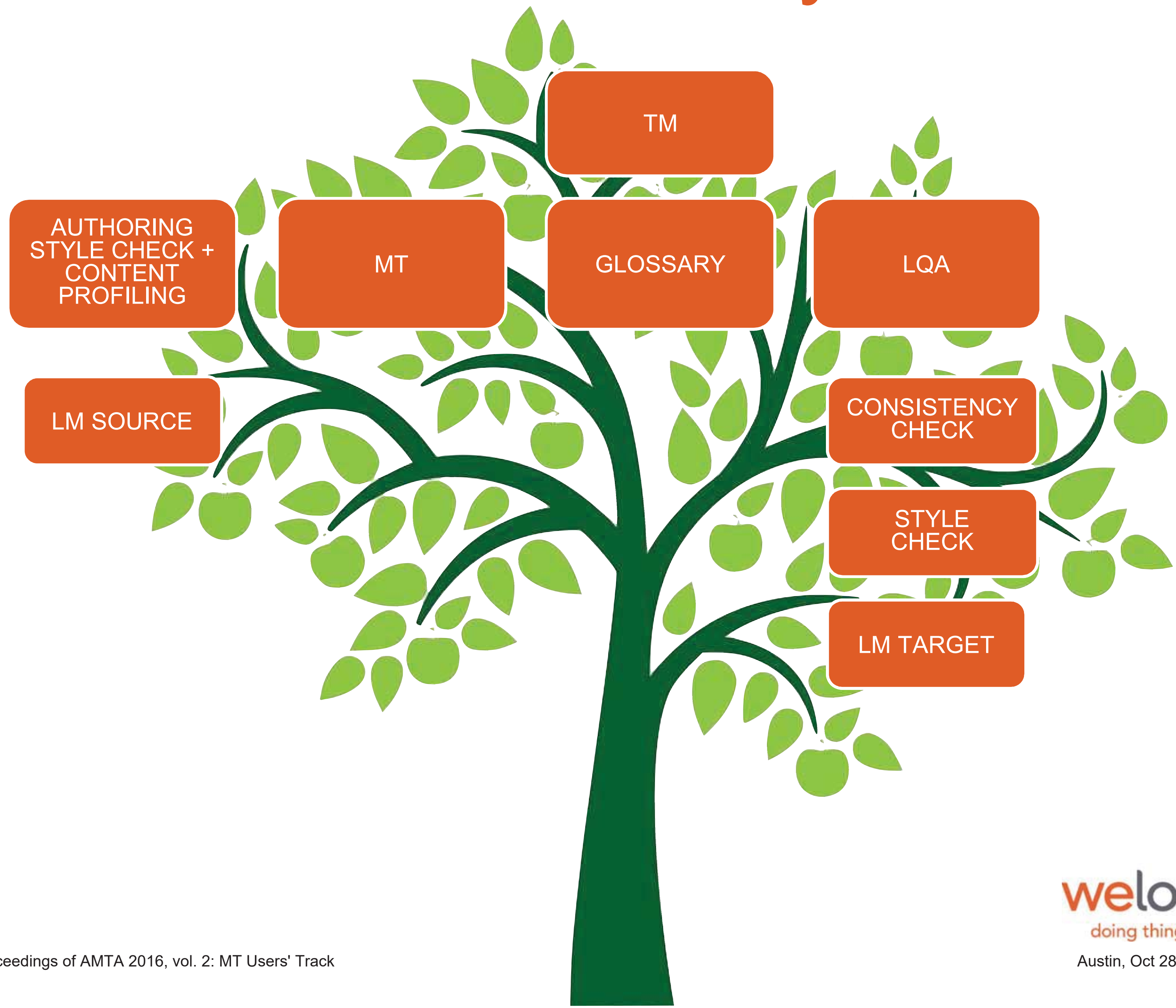
Challenges for Mature MT Programs

- Engage in only those activities that can have an objective, measurable value and/or ROI to the program
- Be wary of not making the engines worse - a threshold beyond which re-training may not be optimal
- Less concerned with automatic scoring as the overriding benchmark for quality since the engines are already at a high quality level
- Stress should be diverted to greater lexical coverage and to fixing high priority and/or high severity linguistic issues that occur numerous times in a corpus

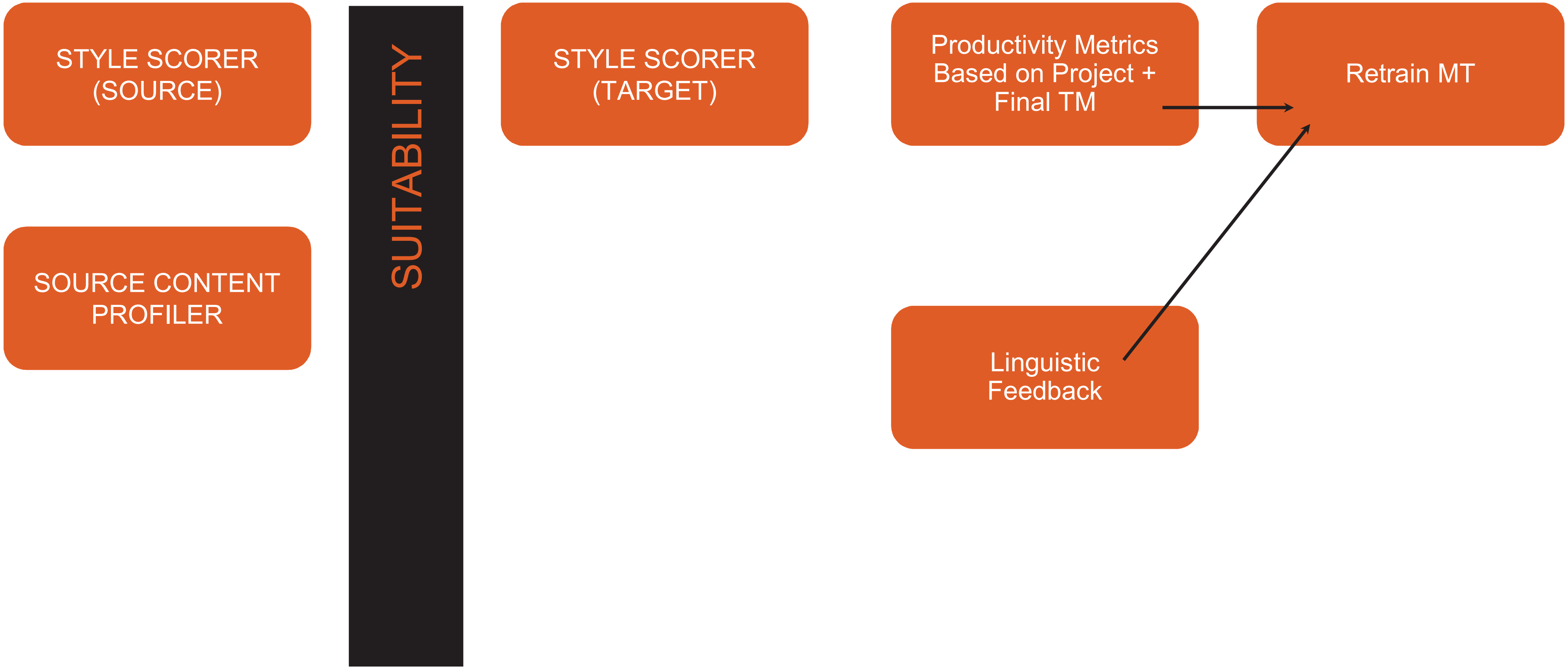
Opportunities for Mature MT Programs

- Pushing the MT engagement upstream
- Analyzing the source content for suitability
- A correlation between the quality of source and the quality and efficacy of MT
- Forecast an MT program, including expected productivity and discounts and make data-driven decisions about the source and its impact before any MT even takes place

The TM Family Tree



Workflow


STYLE SCORER (SOURCE)
SOURCE CONTENT PROFILER
SUITABILITY
STYLE SCORER (TARGET)
Productivity Metrics Based on Project + Final TM
Linguistic Feedback
Retrain MT

What is Style?

- *Style Can be Formally Defined in a Style Guide That Authors & Translators are Requested to Adhere to (But it Doesn't Have to Be)*
- *Style is a Consistency of Voice Across Multiple Documents*
- *Style Tells us Something About the Target Audience*
- *Style Tends to Reflect Patterns of Conscious Grammatical Decisions*
- *For the Purposes of Style Scorer, the Documents Define the Style, Rather than the Style Defining the Documents*

Source: TAUS 2016, Dave Landan, Welocalize

welocalize
doing things differently

Style Scorer Overview

Combines PPL Ratios, Dissimilarity Score + Classification Score

EN-US	OLH	SCORE
DOCUMENT	SM_MANAGER_TAILORING	3.98
DOCUMENT	_Marketing_whitepaper_4aa5-7132enw	0.74

Style Scorer: Under the Hood

- Score between 0 and 4, with higher score indicating better style match.

- Dissimilarity
  Use character n-gram frequency to generate dissimilarity scores. For each document in the Gold Standard, find its maximum dissimilarity compared with all other documents in the Gold Standard. Let G be the set of Gold Standard documents, and g be a document in G. For each g_i in G, calculate $D_{max}(g_i, G)$. For a document t in the set of Test Documents, calculate $D(g_i, t)$ for all g_i. We want to find the average of the ratio of $D(g_i, t) / D_{max}(g_i, G)$ across all g_i. That average is the dissimilarity score component.

- Classification
  Using a one-class classifier, return 1 if the Test Document is in the Gold Standard class; otherwise return -1.

- Perplexity
  Build a language model from the Gold Standard, and get perplexity score for each document within the Gold Standard to establish PPL_{min}, the theoretical floor for perplexity. For each document in Test Documents, calculate PPL. PPL_{min}/PPL_{Test} will be in the range (0,1].

welocalize
doing things differently

Why Use Style Scorer?

Source

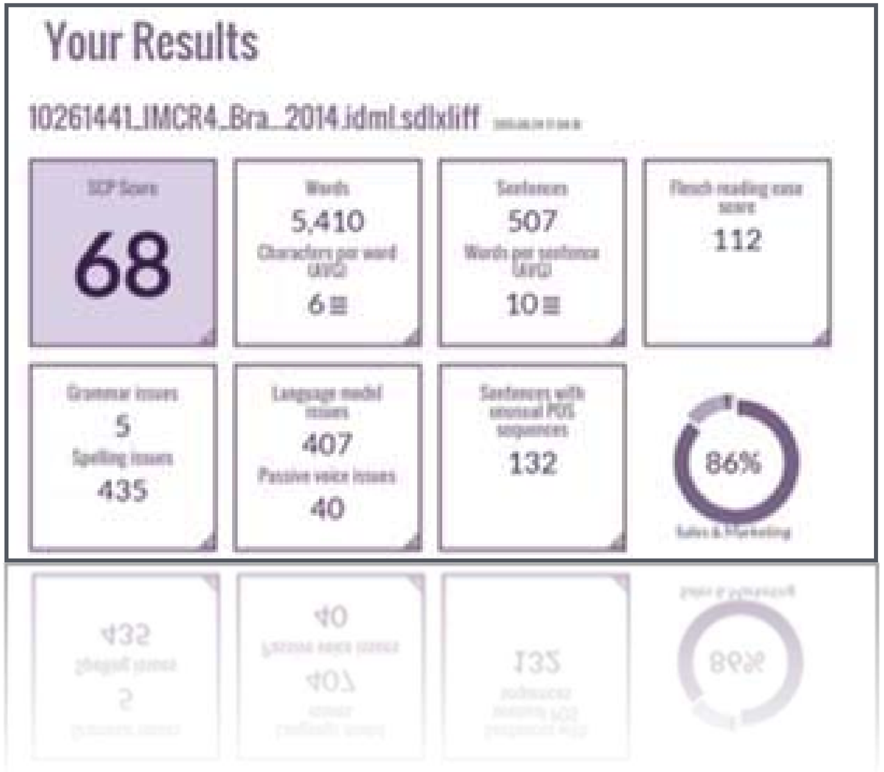
- ✓ Is this really a support document? To what degree is it similar to other support documents, tech doc documents, etc.?
- ✓ Dissimilarity can point to worse quality for raw MT and/or reduced post-editing productivity
- ✓ Find supplemental training data

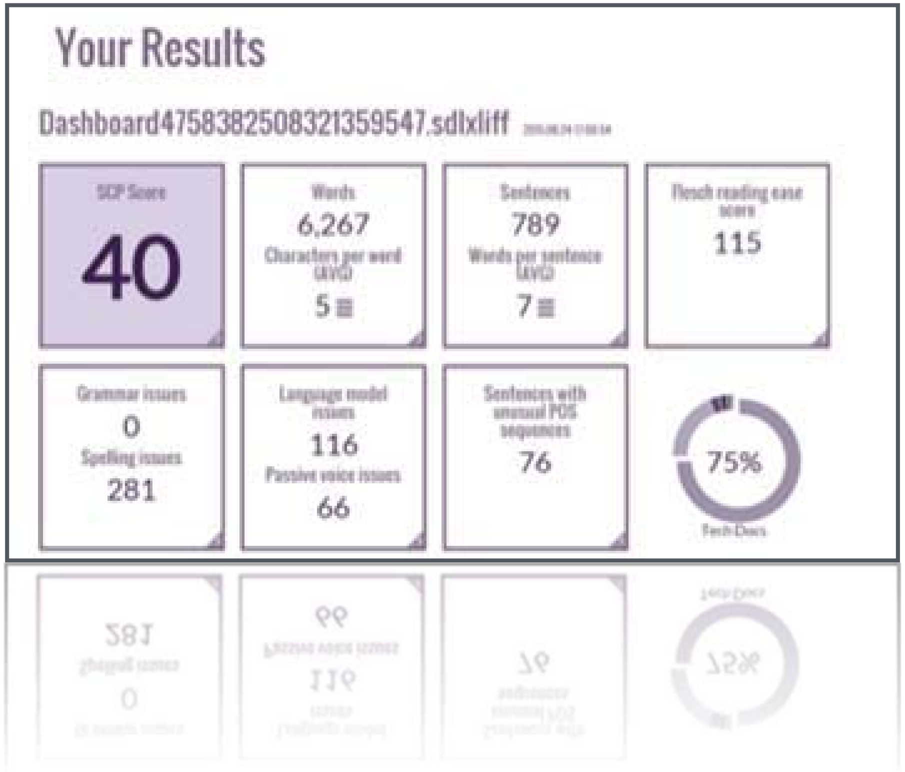
Target

- ✓ Does this target match the style that the client found to be acceptable in the past?
- ✓ Dissimilarity can point to worse quality and reduced post-editing productivity

welocalize
doing things differently

Source Content Profiler // Part 1





welocalize
doing things differently

Source Content Profiler: Part 2

- SCP helps you classify a document
- SCP only works on English source

Words per Sentence

Words per sentence	Occurrence
1	33
2	30
3	31
10+	3
20+	0

Source Content Profiler: Part 3

SCP Highlights Source Issues on a Segment Level

- ✓ *Difficult constructions (e.g. noun phrases)*
- ✓ *Very short or very long sentences*
- ✓ *Passive constructions*

Productivity Metrics

Segment Level

- Source
- Pre-edit Target
- Post-edit Target
- Time to edit (overall, keystroke, pause)
- Number of visits
- Source word count
- Target word count
- Total character inserted
- PE Distance as %

src	preeditText	segTargetFinal	totalVisitCount	totalPeTimeMilliseconds	totalInsertCharCount	srcWordcount
Optional storage bays (used with 3.5" and mixed configuration arrays) house up to 12 DAEs.	Las bahías de almacenamiento opcionales (utilizadas con 3,5 " y arreglos combinados de configuración) contienen hasta 12 DAE.	Las bahías de almacenamiento opcionales (utilizadas con arreglos de 3,5" y de configuración combinada) contienen hasta 12 DAE.	4	372413	25	13

Source	Hypothesis	Reference	Lev. Dist.	PE Dist. (% of ref. le	Words
Contact support	Contactez l'assistance technique	Contacter l'assistance	11	47.83%	2

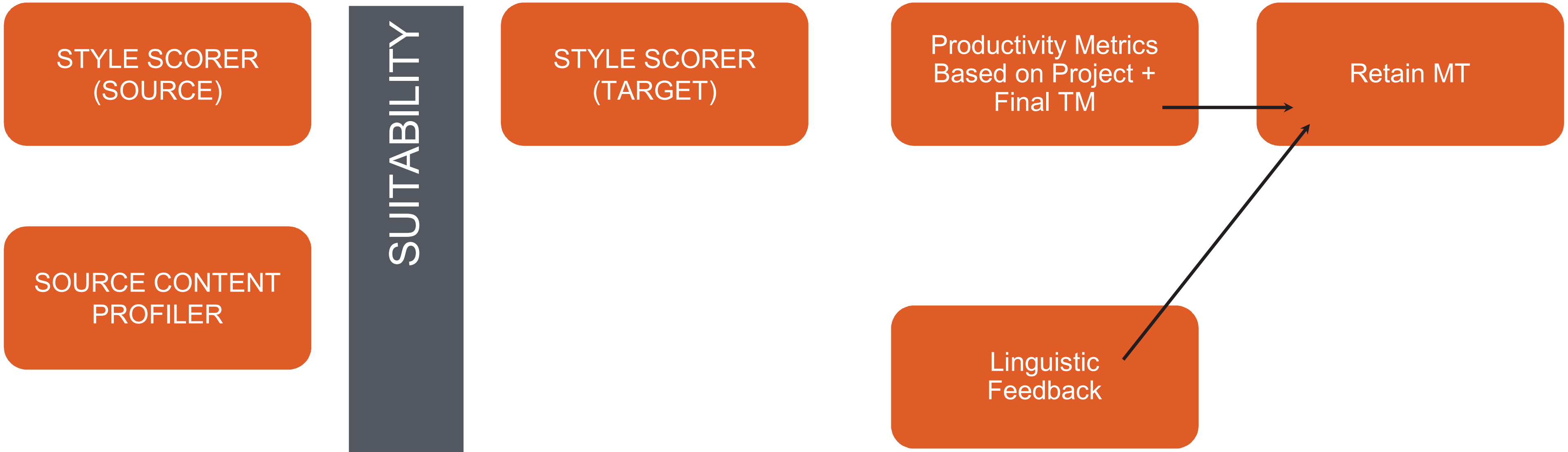
Automatic Scoring

File or Project Level

- BLEU
- Meteor
- GTM
- Precision
- Recall
- TER
- PE Distance as %

BLEU	NIST	METEOR	GTM	Avg. PE	TER	Precision	Recall	Length (Hyp./Ref.)	Segs.	Words
46.90	9.86	62.51	68.86	31.80%	40.71	0.69	0.69	1.01	13797	148862

Goal



Goal is to find correlation between source, LQA effort required for target and productivity metrics and use data to evangelize changes at the beginning of the content creation cycle.

Next Steps

- Build LMs per domain for English source for Style Scorer
- Build LMs per domain for target languages (tier 1 and subsequently tier 2) for Style Scorer
- Build LMs per domain for English source for Source Content Profiler
- Calculate auto-scoring including PE distance for before_PE, after_PE, after_client_review
- Find how strong the correlation is between all the metrics above
- What can be done to prevent some of the issues?

welocalize
doing things differently


Thank You!
welocalize
doing things differently