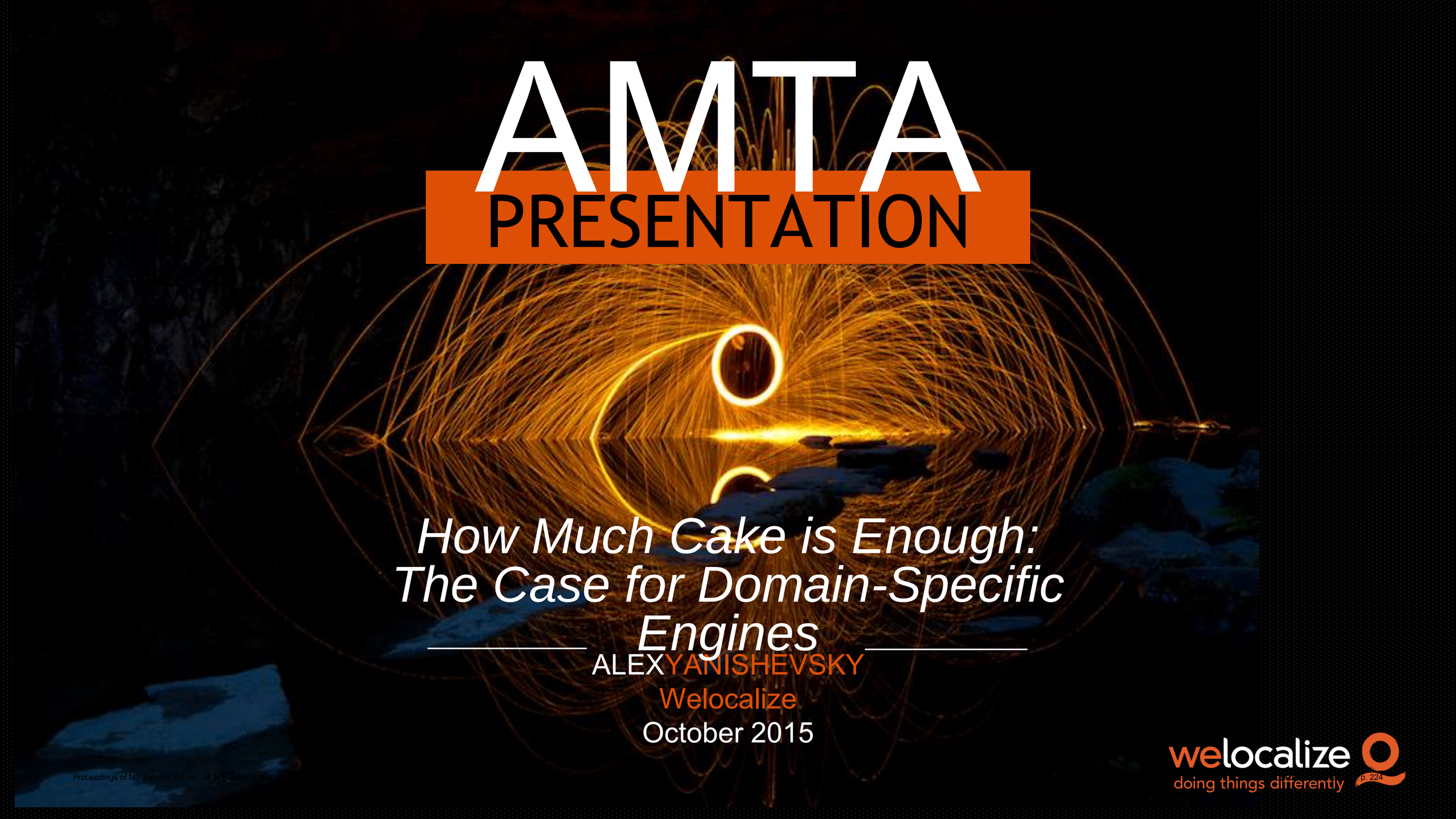


AMTA PRESENTATION

How Much Cake is Enough: The Case for Domain-Specific Engines

ALEXYANISHEVSKY
Welocalize

October 2015

welocalize
doing things differently

How Much Cake Is Too Much Cake?

What is the Tipping Point?

AGENDA

- How Many Engines
- How to Split Domains
- How to Measure Success
- How to Improve

welocalize
doing things differently

HOW MANY ENGINES: CRITERIA

- Environment: Elegant Deployment?
- Cost
- How Different are They From Each Other?
- Maintenance: Engineering + Linguistic Feedback Implementation

HOW TO SPLIT DOMAINS: CRITERIA

- Content Owner Feedback
- Historical Experience Based On Business Unit or Portfolio
- Naming Convention
- Style Analysis: Difference in Characteristics Based on Lexical Diversity, Sentence Length + Syntactic Complexity

HOW TO SPLIT DOMAINS: TOOLS

HOLISTIC APPROACH BASED ON SEVERAL TOOLS:

- Build Domain-Specific Language Models + Select TUs for Domain by PPL
- Source Content Profiler – Helps Identify Domain Based on Language Models, as well as Other Stylistic Characteristics
- Style Scorer – Higher Score Indicates Better Match to Style Established by Client's Documents

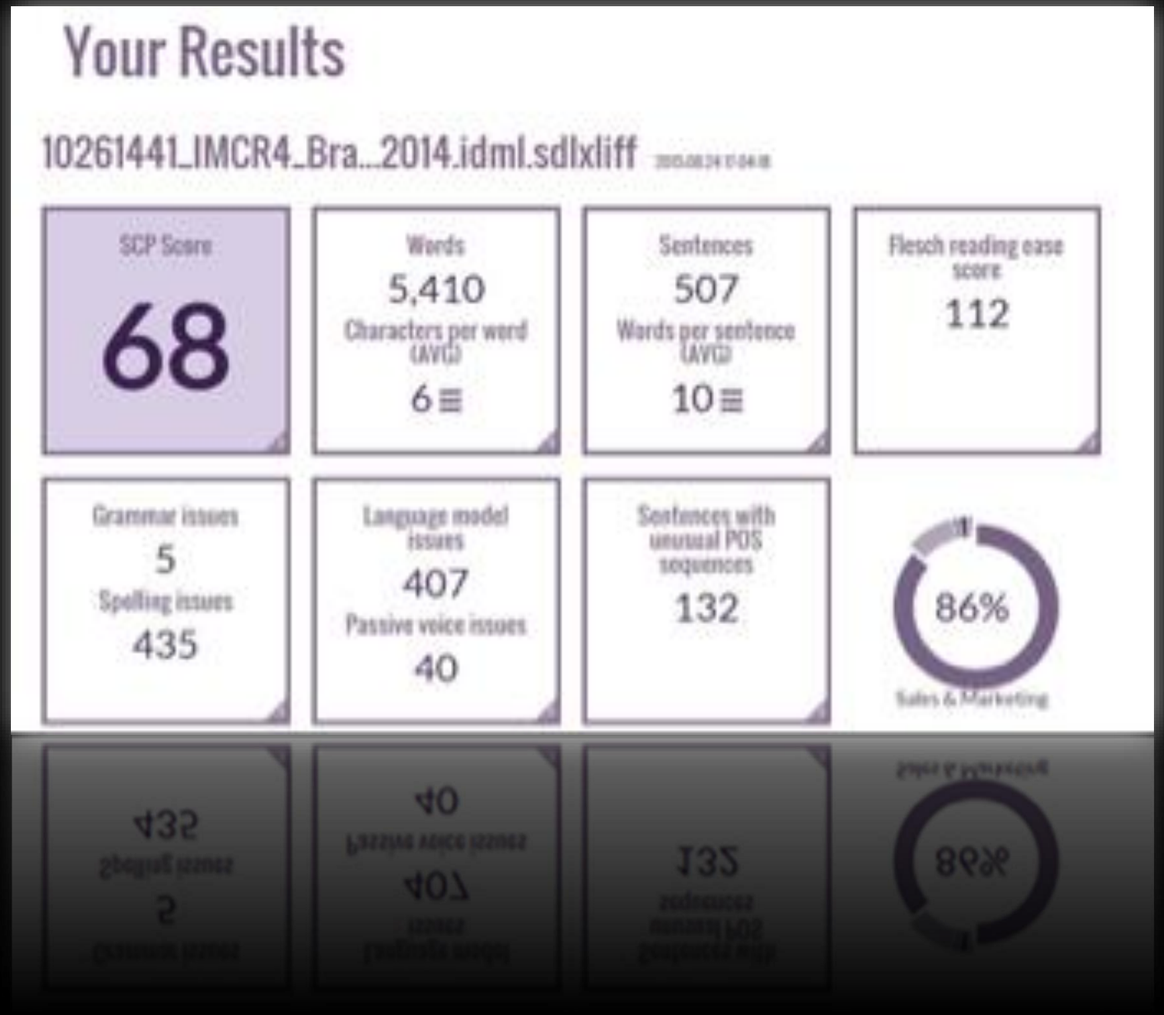
TOOLS: PERPLEXITY EVALUATOR

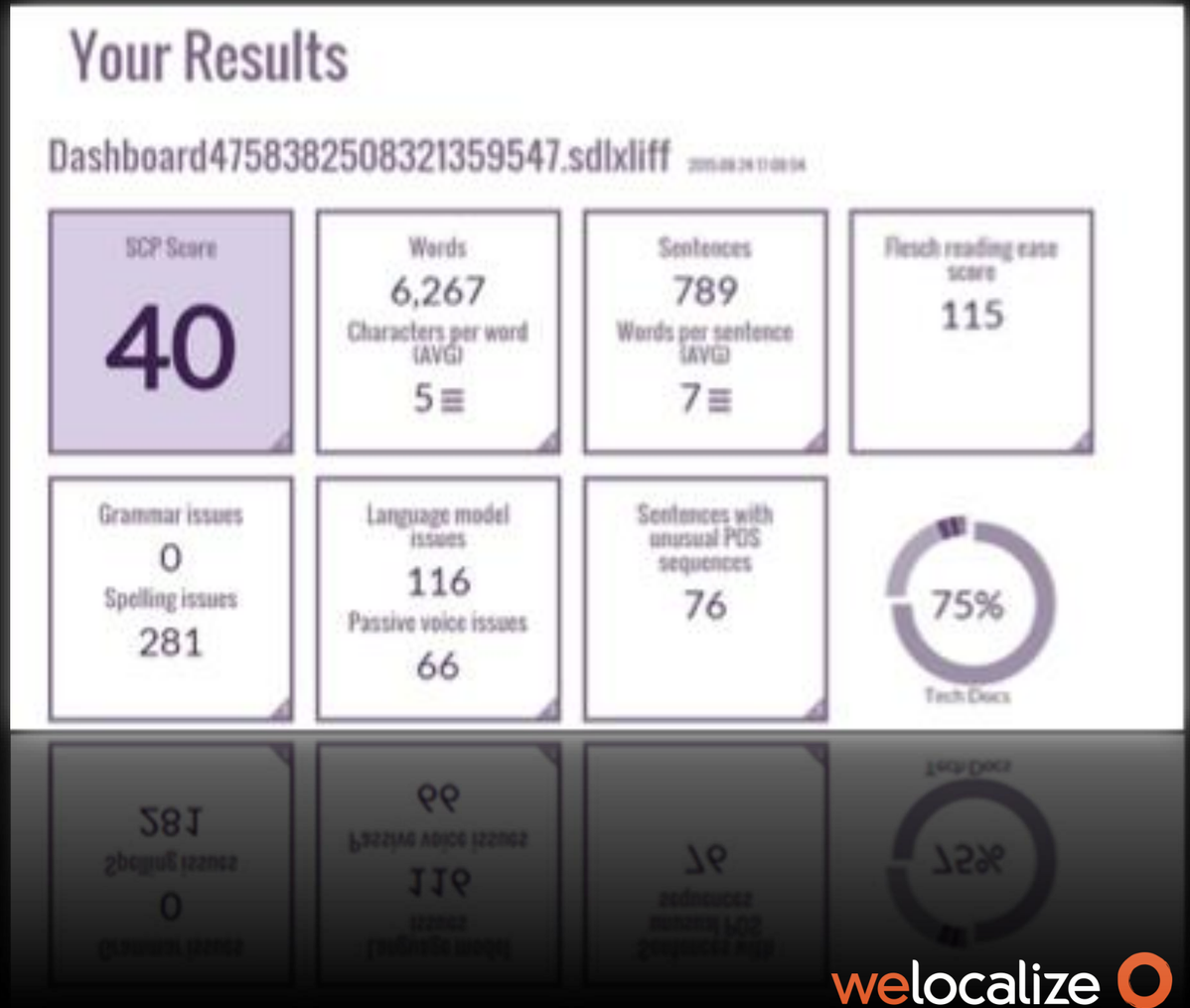
TU LEVEL

<tu srclang="EN-US" tuid="75438"> <prop type="x-ppl:train2">208</prop><prop type="x-ppl:techdoc6">191.025</prop><prop type="x-ppl:support2">325.983</prop><prop type="x-ppl:sales1">97.0736</prop><prop type="x-ppl:productLoc1">396.398</prop><prop type="x-ppl:legal1">617.876</prop><tuv xml:lang="EN-US"> <seg>Consistent feature set across multiple platforms (Windows, Mac, iOS, Android).</seg> </tuv> <tuv changedate="20140325T122530Z" changeid="serviceaaa" creationdate="20140325T122530Z" creationid="serviceaaa" lastusagedate="20140325T122530Z" usagecount="0" xml:lang="ES-XL"> <prop type="x-ALS:Context">TEXT</prop> <prop type="x-ALS:Source File">\\DATA\TC\39720\SRC\EN-US\co-02__battle-card_en\co-02__battle-card_en.inx</prop> <seg>Conjunto de características coherente en varias plataformas (Windows, Mac, iOS, Android)</seg> </tuv> </tu>

welocalize
doing things differently

TOOLS: SOURCE CONTENT PROFILER

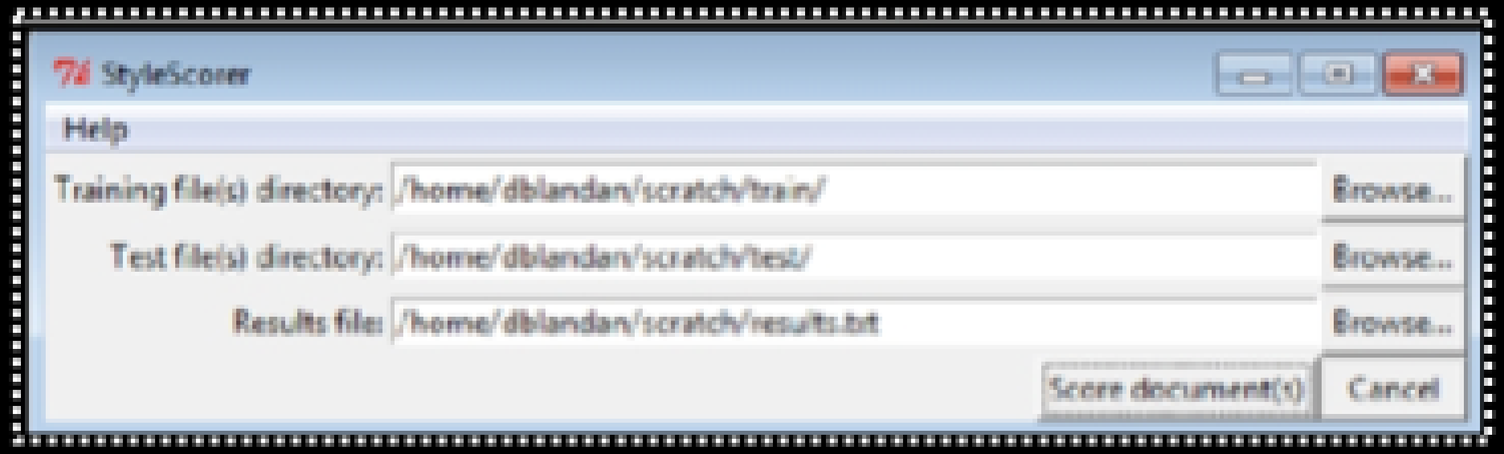




welocalize
doing things differently

TOOLS: STYLE SCORER

COMBINES PPL RATIOS, DISSIMILARITY SCORE + CLASSIFICATION SCORE



TEST CATEGORY	TRAINING CATEGORY	SCORE
SUPPORT	TECH DOC	3.16
TECH DOC	TECH DOC	2.94
TECH DOC	LEGAL	,02

welocalize
doing things differently

WHY USE STYLE SCORER?

- Identify similarity of source document to “gold standard” documents from that domain and other domains
- Identify similarity of target document to “gold standard” documents from that domain and other domains
- Example: 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

STYLE SCORER + SCP

- SCP Helps Classify a Document
- Style Scorer Tells You How Good a Match a Document is to a Profile
- SCP Only Works on English Source
- Style Scorer Works on English Source + Non-English Target

CASE STUDY
ONE DOMAIN?

One ring
to rule them all

Three Rings for the Elven-kings under the sky,
Seven for the Dwarf-lords in their halls of stone,
Nine for Mortal Men doomed to die,
One for the Dark Lord on his dark throne
In the Land of Mordor where the Shadows lie.
One Ring to rule them all, One Ring to find them,
One Ring to bring them all and in the darkness bind them
In the Land of Mordor where the Shadows lie.

welocalize
doing things differently

CASE STUDY: HOW MANY DOMAINS?

- Started With 6 Domains: Technical Documentation, Legal, Support, Training, Product UI, Sales/Marketing
- Found that Technical Documentation, Support + Training Were Very Similar Based on LMs Scores Against Each Other, Length of Sentences, Similar Grammatical Structures
- Found that Product UI was Close Enough to Above 3 That Making a Separate Engine was Not Warranted
- Found that Legal + Sales/Marketing Were Different Enough from Above Domains and From Each Other Based on LMs Scores Against Each Other + Length of Sentences

CASE STUDY: GATHERING ASSETS

TMs

- Old
- Somewhat Recent
- Current

- Termbases in MultiTerm
- Existing User Dictionaries + Normalization Dictionaries
- New User Dictionaries Based on Term Extractions + Auto-Import for Some Languages

CASE STUDY: CURATING ASSETS

- Cleaned TMs
- Based on LM Perplexity
- Kept the UDs + Normalization Dictionaries As Is
- Additional term extraction for weak languages or languages with insufficient assets

CASE STUDY: ENGINE ITERATIONS

Based on options in Systran:

- RBMT only
- Hybrid with Stemming, LM Order, Distortion, etc.
- SMT only

HOW TO MEASURE SUCCESS

- Automatic scores
- Human evaluations
- Decrease in PE distance
- Decrease in linguistic issues reported

CASE STUDY: AUTOMATIC SCORES SALES/MARKETING1

	ar-SA		de-DE		es-ES		fr-CA		fr-FR		it-IT	
	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc
BLEU:	29.41	26.82	50.17	41.01	65.06	60.57	49.36	47.10	53.40	54.47	56.57	56.55
NIST:	6.89	6.53	9.23	7.87	11.10	10.64	9.37	9.11	9.76	9.83	10.17	10.10
METEOR:	14.69	13.98	60.07	53.43	79.66	76.44	64.80	62.98	67.41	68.11	70.06	70.04
GTM:	57.37	54.67	72.06	64.52	83.75	81.27	73.22	71.50	75.14	75.50	78.14	77.63
Avg. PE Dist.	29.27%	32.32%	29.07%	38.89%	22.56%	25.68%	27.51%	29.37%	28.87%	28.82%	24.85%	24.91%
TER:	54.09	57.68	38.61	49.53	24.08	27.26	36.53	38.76	34.51	34.17	30.34	30.99
Precision:	0.59	0.56	0.76	0.65	0.86	0.84	0.78	0.76	0.78	0.78	0.81	0.80
Recall:	0.56	0.53	0.69	0.64	0.81	0.79	0.69	0.68	0.72	0.73	0.75	0.75
Length (Mean Ref./Cand. Len.)	0.95	0.95	0.91	0.98	0.94	0.94	0.89	0.90	0.92	0.93	0.92	0.95
Sample size (Segments):	999	999	999	999	999	999	999	999	999	999	999	999
(Target Words):	9808	9808	12670	12670	12063	12063	11836	11836	11956	11956	11927	11927

CASE STUDY: AUTOMATIC SCORES SALES/MARKETING2

	ja-JP		ko-KR		pl-PL		pt-BR		ru-RU		zh-CN	
	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc
BLEU:	62.04	45.96	59.05	44.23	52.38	30.98	64.38	53.56	53.42	43.86	55.01	49.39
NIST:	10.10	7.92	9.84	7.95	9.37	6.68	10.97	9.81	9.52	8.30	10.01	9.33
METEOR:	71.79	58.09	70.63	58.28	65.22	44.20	76.55	68.62	66.69	58.16	69.23	64.54
GTM:	78.99	68.00	77.60	67.42	74.07	56.02	83.30	77.21	74.04	66.07	77.99	74.23
Avg. PE Dist.	40.48%	56.25%	39.48%	53.61%	26.97%	45.44%	18.20%	24.82%	27.62%	35.72%	36.09%	42.32%
TER:	33.64	48.17	34.79	48.76	35.33	54.26	23.31	31.40	35.13	43.67	33.85	38.73
Precision:	0.84	0.73	0.81	0.69	0.79	0.59	0.86	0.79	0.78	0.69	0.83	0.79
Recall:	0.75	0.64	0.74	0.66	0.70	0.53	0.81	0.76	0.71	0.63	0.74	0.70
Length (Mean Ref./Cand. Len.)	0.89	0.88	0.91	0.94	0.89	0.91	0.94	0.96	0.91	0.92	0.89	0.90
Sample size (Segments):	991	991	931	931	999	999	999	999	999	999	999	999
(Target Words):	11951	11951	10834	10834	11153	11153	11776	11776	11770	11770	11542	11542

CASE STUDY: AUTOMATIC SCORES LEGAL1

	ar-SA		de-DE		es-ES		fr-CA		fr-FR		it-IT	
	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc
BLEU:	46.90	30.73	48.57	32.59	62.58	46.10	56.79	38.45	61.24	38.04	56.76	48.12
NIST:	8.94	7.03	9.04	6.90	10.77	9.03	10.15	8.07	10.49	7.99	10.09	9.10
METEOR:	60.90	45.67	59.12	46.80	77.41	65.67	70.19	56.27	72.95	55.52	69.88	62.75
GTM:	70.81	59.47	71.04	58.53	81.97	72.50	77.28	65.59	79.07	65.14	77.53	71.65
Avg. PE Dist.	28.00%	42.63%	30.90%	46.05%	20.36%	30.69%	25.42%	37.86%	27.87%	42.46%	24.64%	31.90%
TER:	40.69	54.53	41.48	57.82	26.04	37.57	31.60	45.96	29.39	46.69	30.68	38.11
Precision:	0.74	0.63	0.74	0.59	0.86	0.75	0.81	0.69	0.82	0.67	0.80	0.73
Recall:	0.68	0.57	0.68	0.58	0.79	0.70	0.74	0.63	0.77	0.63	0.75	0.70
Length (Mean Ref./Cand. Len.)	0.92	0.91	0.93	0.99	0.92	0.94	0.91	0.91	0.94	0.93	0.93	0.95
Sample size (Segments):	1000	1000	999	999	999	999	999	999	999	999	999	999
(Target Words):	9412	9412	10648	10648	10845	10845	11893	11893	10835	10835	10665	10665

CASE STUDY: AUTOMATIC SCORES LEGAL2

	ja-JP		ko-KR		pl-PL		pt-BR		ru-RU		zh-CN	
	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc
BLEU:	55.99	41.51	57.62	37.40	50.78	25.04	59.06	45.89	46.77	29.94	65.13	43.77
NIST:	9.26	7.38	9.60	7.03	9.21	5.93	10.35	8.99	8.66	6.55	10.92	8.64
METEOR:	67.02	54.00	69.28	52.16	64.30	38.05	72.31	62.13	60.51	45.18	76.15	59.33
GTM:	75.52	64.79	76.39	62.03	73.61	51.58	79.94	72.31	68.78	55.18	82.09	69.72
Avg. PE Dist.	43.46%	61.59%	38.14%	60.56%	26.83%	50.97%	22.10%	30.00%	35.02%	49.37%	28.73%	44.97%
TER:	38.14	52.43	36.09	55.43	36.15	59.89	27.62	37.03	41.44	56.23	28.22	44.44
Precision:	0.82	0.69	0.81	0.63	0.78	0.54	0.82	0.74	0.73	0.58	0.84	0.73
Recall:	0.70	0.61	0.73	0.61	0.70	0.49	0.78	0.71	0.65	0.53	0.80	0.67
Length (Mean Ref./Cand. Len.)	0.85	0.88	0.90	0.96	0.89	0.90	0.95	0.96	0.89	0.90	0.95	0.91
Sample size (Segments):	966	966	999	999	999	999	999	999	999	999	998	998
(Target Words):	10469	10469	10119	10119	9485	9485	10601	10601	10805	10805	9073	9073

HOW TO IMPROVE

OPPORTUNITIES FOR RESEARCH

- Eradicate High-Frequency Inconsistencies Between TMs, Termbases + User Dictionaries (UDs)
- Create Domain-Specific UDs
- Pre-MT Source Check: Was This Content Properly Categorized?
- Send Best Reply: TMT Prime, Send Best Translation Irrespective of Domain

SUMMARY

- Domain-specific Engines Yield Better Results as Evidenced by Auto Scores, Human Evaluations and Reduced PE Distance
- Group Closely-related Content into One Domain
- Determine How Many Engines Your Infrastructure Can Support

ALEXYANISHEVSKY
Welocalize
October 2015