The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

Introduction to MT

Mike Dillinger

Ebay

Jay Marciano

Lionbridge
Technologies

SAN DIEGO, CA
OCTOBER 28- NOVEMBER 1, 2012

This tutorial is for people who are beginning to evaluate how well machine translation will fit their needs or who are curious to know more about how it is used. We assume no previous knowledge of machine translation. We focus on background knowledge that will help you both get more out of the rest of AMTA2010 and to make better decisions about how to invest in machine translation.

Past participants have ranged from tech writers and freelance translators who want to keep up to date to VPs and CEOs who are evaluating technology strategies for their organizations.

The main topics for discussion are common FAQs about MT (Can machines really translate? Can we fire our translators now?) and limitations (Why is the output so bad? What is MT good for?), workflow (Why buy MT if it's free on the internet? What other kinds of translation automation are there? How do we use it?), return on investment (How much does MT cost? How can we convince our bosses to buy MT?), and steps to deployment (Which MT system should we buy? What do we do next?).

Presenters

- Mike Dillinger, PhD, Principal of Translation Optimization Partners, an independent consulting group that helps clients optimize communication in global markets.
- Jay Marciano, Director of Real-time Translation Development at Lionbridge Technologies, a leading translation and localization company.

Introduction to
Machine Translation
Tutorial at AMTA 2012, San Diego

Mike Dillinger (eBay)

Jay Marciano (Lionbridge)

2012
AMTA
20 Years

Overview

- The Problem: Too much to translate
- FAQs: What's MT?
- ROI: Benefits of MT
- Which MT system should I choose?
- Evaluation: What's good enough?
  - Automated Metrics
  - Human Evaluation
- Kinds of MT Systems

2012
AMTA
20 Years

About Us

Mike Dillinger

- – President-elect of AMTA
- – Industry Consultant; Computational Linguist @ eBay
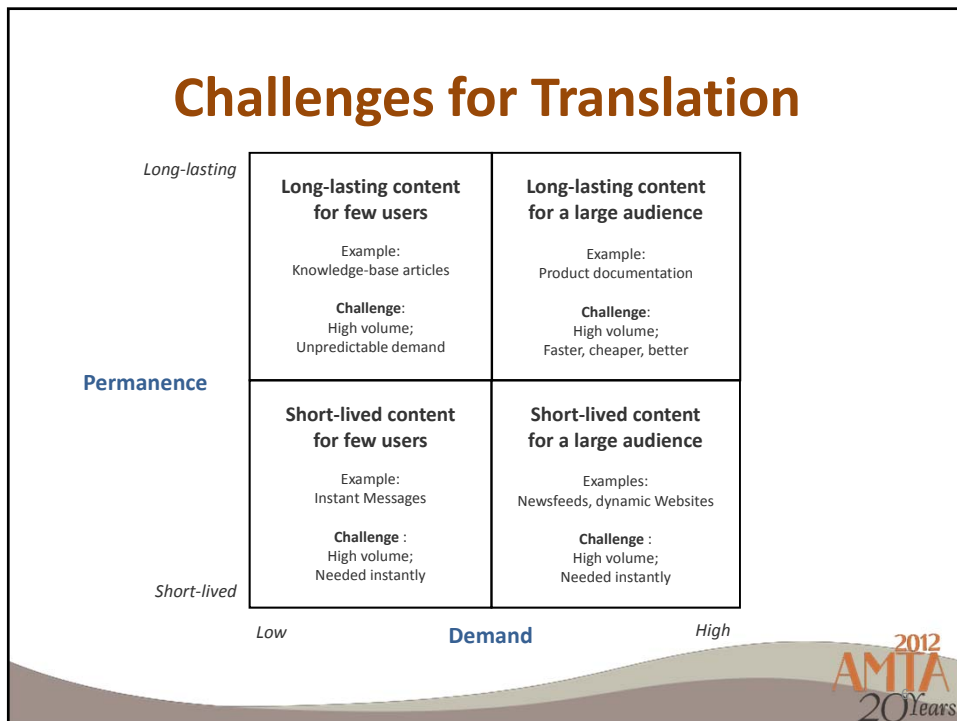- – 15 years' experience in the development & deployment of MT systems

Jay Marciano

- – Director of Real-time Translation @ Lionbridge
- – 15 years' experience in the development & application of MT systems

The Problem:
Too much to translate; too little time

What's going on here?

Translators

Content

We work right here!

Our

Are you starting to feel the pressure yet?

© Philip Wallick/Corbis

5

Challenges for Translation

	Low — Demand — High	
Long-lasting	**Long-lasting content for few users** Example: Knowledge-base articles **Challenge**: High volume; Unpredictable demand	**Long-lasting content for a large audience** Example: Product documentation **Challenge**: High volume; Faster, cheaper, better
Permanence	**Short-lived content for few users** Example: Instant Messages **Challenge** : High volume; Needed instantly	**Short-lived content for a large audience** Examples: Newsfeeds, dynamic Websites **Challenge** : High volume; Needed instantly
Short-lived		

AMTA 2012 20 Years

| More | Better | Faster | Cheaper |

More   **Better**   **Faster**   **Cheaper**

More   **Better**   **Faster**   **Cheaper**

More Better Faster Cheaper

More Better Faster Cheaper

7

The Problems	The Requirement

Scalability, Cost, Time
•Human translation is (usually) wonderful but **it doesn't scale well**
  •Bigger projects = more costs
  •Bigger projects = more issues
  •More languages = more costs + more issues
•Human translation is **expensive**
•Human translation is **slow**	

Technology Options for Translation

	Volume of Translations		
	Small	Medium	Large
	<50 pgs/day ~3 HT ~1,000 pgs/mo	~100 pgs/day ~10 HT ~2,000 pgs/mo	>300 pgs/day ~30 HT ~6,000 pgs/mo
Months **Weeks**	HT + TM	HT + TM	HT + TM
Days **Hours**	HT + TM	HT + TM or MT/TM + Post-Editing	MT/TM + Post-Editing
Minutes **Seconds**	Simultaneous Interpreting	Simultaneous Interpreting or MT/TM, no Post-Editing	MT/TM, no Post-Editing

Response time for final Translations

Real-time Translation

9

---

10 ways to reach a global audience

**1,000,000 words
into 10 languages**

$6,000,000	1.	Re-create in each target language
$2,000,000	2.	Translate by hand with the very best quality
$1,600,000	3.	Use customized MT + extensive post-editing
$1,200,000	4.	Use customized MT + minimal post-editing
$1,000,000	5.	Translate by hand as cheaply as possible
$1,000,000	6.	Use Global English
$ 300,000	7.	Use customized MT alone
$ 100,000	8.	Use raw MT
$ 0	9.	Use "technical" English
$ 0	10.	Use magic, telepathy, or prayer

The good news: There are *lots* of options.

10

FAQs: What's MT?
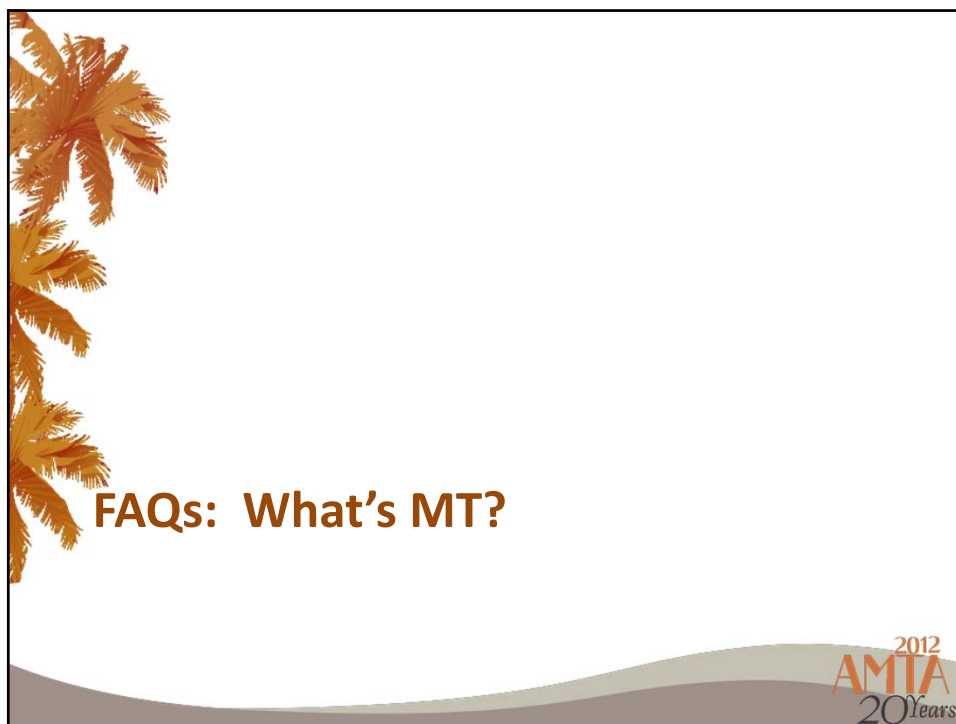
What's MT?

Machine Translation systems are software products that translate electronic texts (and speech) into other languages automatically.

- **Electronic Dictionaries?**
  - o One term at a time; do not use context; do not combine terms into sentences

- **Translation Memories?**
  - o Retrieve known segments from "memory"
  - o Few suggestions for unknown segments

- **Machine Translation Systems**
  - o Calculate best-guess translations for all segments, based on known translations for parts

12

What's MT?

• **Do you mean systems like Google Translate and Babelfish?**

•Yes, the basic technology is the same, **but** for companies we adapt the system extensively, to meet your needs. The result is very different!

• **We already use machine translation from Trados, right?**

•Trados is one good use of very *old* machine translation technology – it's called "translation memory". It doesn't work well with new sentences or new topics. Modern machine translation technology can do a much better job with new input; think of MT as "translation *reasoning*".

13

What's MT?

• **Can we fire our human translators?**

No. In most situations, MT *requires* human translators. Their job just changes so they can do more translation faster. Many translation agencies already use MT for draft translations because it saves them time and money. Remember that there are many situations where there is too much to translate for humans.

• **You guys really hate translators, don't you?**

Not at all! Some overly enthusiastic MT researchers in the old days talked about replacing humans, which scared the pants off the translators. It also embarrassed us to death. Nowadays, we even invite translators to our conferences :)

14

What's MT?

• **The whole idea of MT is ridiculous. Only humans can *really* translate. You have to understand the subtleties of language and culture.**

It turns out that very, very many kinds of sentences are routine enough that machines can do a great job without subtle understanding. Most useful texts are neither poetic nor sophisticated.

• **I've seen MT on the web. It's laughable junk.**

Millions of people use MT every day and very few complain. Besides, it's free. What would a free Mercedes look like? Brand new Enterprise MT, customized to your organization's needs is very, very different from what you see on the web.
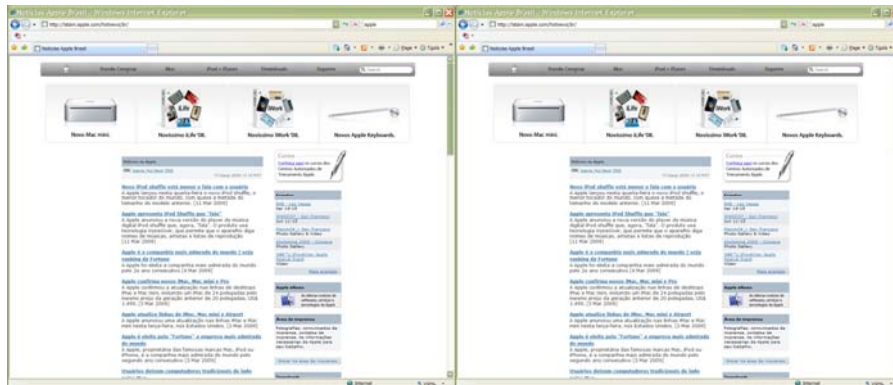
15

"MT will have a negative impact on my brand."

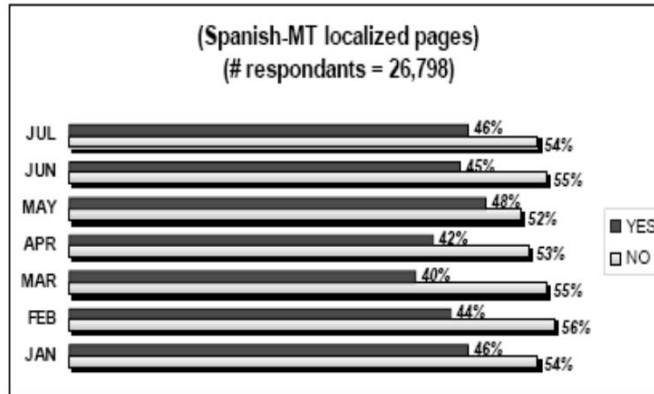**Your site translated without MT**   **Your site translated *with* MT**

Faster; Cheaper; More consistent

16

But is MT good enough?

Did this information help answer your question?
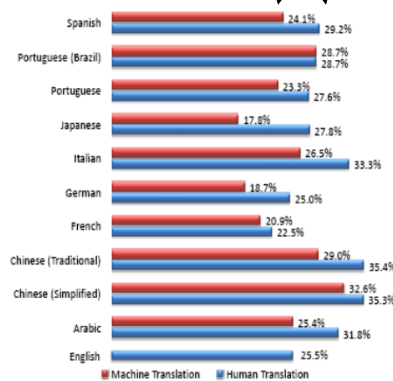
Intel

Burgett & Chang, 2008

Microsoft

Wendt, 2008

Human or machine?

Adobe InDesign CS 5.5

Engage, inform and attract readers with documents created in Adobe ® InDesign ® CS5.5, incorporating interactivity, video and sound. Develop strong and persuasive messages, and with advanced interactive documents.

Tools Folio Producer
Create, view and collect documents to interactive digital devices such as Apple iPad tablet and a variety of tablets Android ™. Load the documents to Adobe Digital Publishing Suite1para more production and distribution.

Presentations and interactive documents
Add interactivity, motion, audio and video documents and presentations, exporting directly to SWF for playback in Adobe Flash ® Player.

Export to Flash Professional
Export InDesign documents to Flash Professional to add sophisticated interactivity, animation and navigation to complex layouts.

Draw documents more engaging, rich media and images are automatically resized to fit any screen. Have more control over the content and typography to export documents to the EPUB format.

2012
AMTA
20 Years

Use Cases for MT

Use Cases
• Pre-translation of printed Product Information (**Localization**)
• Translation of Dynamic Web Content
• Monitoring Information from Foreign Sources
• Translation of Subtitles, TV, Radio Broadcasts
• Translation of eCommerce content
• Translation of Chat, SMS, Tweets
• Translation of User Generated Content about Products
• Speech to Speech Translation

20

2012
AMTA
20 Years

Who's really using MT?

- **"In-bound" Translation** (from other languages to yours)
  - Global Public Health Information Network (Public Health, Canada)
  - Many military and business organizations
  - Internet users around the world

- **"Out-bound" Translation** (from yours to other languages)
  - eBay, Symantec, Adobe, Cisco, Microsoft, Intel, European Community, etc.
  - Internet users around the world

- **"Real-time" Translation** (between two languages)
  - Translated subtitles (news, Jay Leno), translated TV and radio broadcasts
  - Internet users around the world: translated chat, translated SMS

21

Who uses MT for real-time applications?

FreeTranslation.com Year-over-Year Usage

Legend: Sep2001, Sep2002, Sep2003, Sep2004, Sep2005, Sep2006

Y-axis: Transactions per Minute (0 to 5000)
X-axis: Time of Day (Time, 12:59, 13:59, 14:59, 15:59, 16:59, 17:59, 18:59, 19:59, 20:59, 21:59, 22:59, 23:59, 0:59, 1:59, 2:59, 3:59, 4:59, 5:59, 6:59, 7:59, 8:59, 9:59, 10:59)
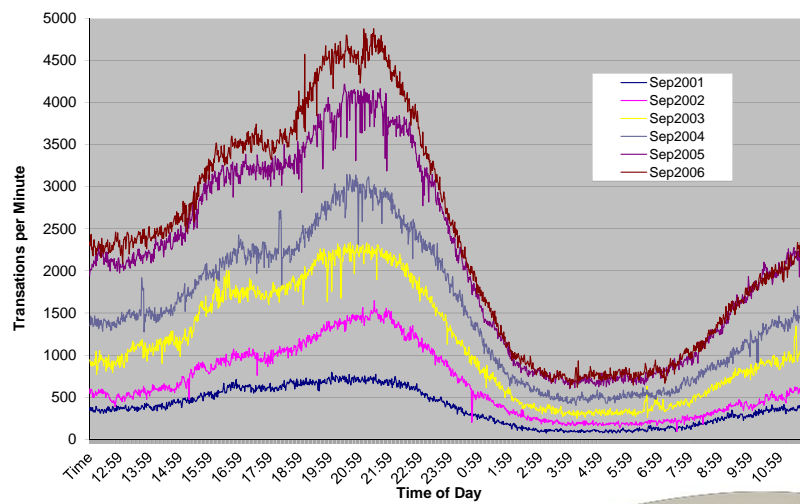
Who uses real-time MT?

Today we have more than **200 million monthly active users** on translate.google.com (and even more in other places where you can use Translate, such as Chrome, mobile apps, YouTube, etc.).

And our users are truly global: more than 92 percent of our traffic comes from outside the United States.

In a given day we translate roughly as much text as you'd find in **1 million books.**

To put it another way: what all the professional human translators in the world produce in a year, our system translates in roughly a single day.

Franz Och, Google Translate
26 April 2012

(http://googleblog.blogspot.com/2012/04/breaking-down-language-barriersix-years.html)

From the Translator's point of view…

Machine Translation
is an extension of Translation Memory

"Translation technologies" like TM and machine translation **don't translate.**

- They **suggest** the closest segments that they can find (TM) or assemble (machine translation).
- They help translators re-use the work that they've done before.
- They're translator "accelerators",
   not translator "replacements"

Machine translation products do the same things that TM products do, and more.

Good Matches

TM products

- Check new segment against existing segments that we put in memory
- Return the "good" matches

Machine Translation products

- Check new segment against existing segments that we put in memory
- Return the "good" matches

You can simply plug your existing TMs into many machine translation products to re-use them as usual.
If…

Fuzzy Matches

TM products
- Give an approximate suggested translation
- Assemble a suggested translation from fragments

We get approximate or "assembled" translations.

We ignore or fix the suggestions.

Machine Translation products
- Give an approximate suggested translation
- Assemble a suggested translation from fragments *in much more sophisticated ways*

We get more, better "assembled" translations.

We ignore or fix the suggestions.

Non-Matches

TM products
- Leave non-matches blank
- Provide no suggestions

We translate from scratch.

Machine Translation products
- Give an approximate suggested translation
- Assemble a suggested translation from fragments *in much more sophisticated ways*

We get "assembled" translations instead of nothing.

We ignore or fix the suggestions.

Translating from TM output

"Post-editing" from TM + MT output

ROI: The Benefits of MT

Benefits of Real-Time Translation

- In very large-scale and real-time scenarios, there is a stark choice:
  - Machine Translation with variable quality, or
  - NO Translation at all

  Period.

- Some access to web content, emails, chat, SMS, knowledge bases, eCommerce, etc. **or none at all**
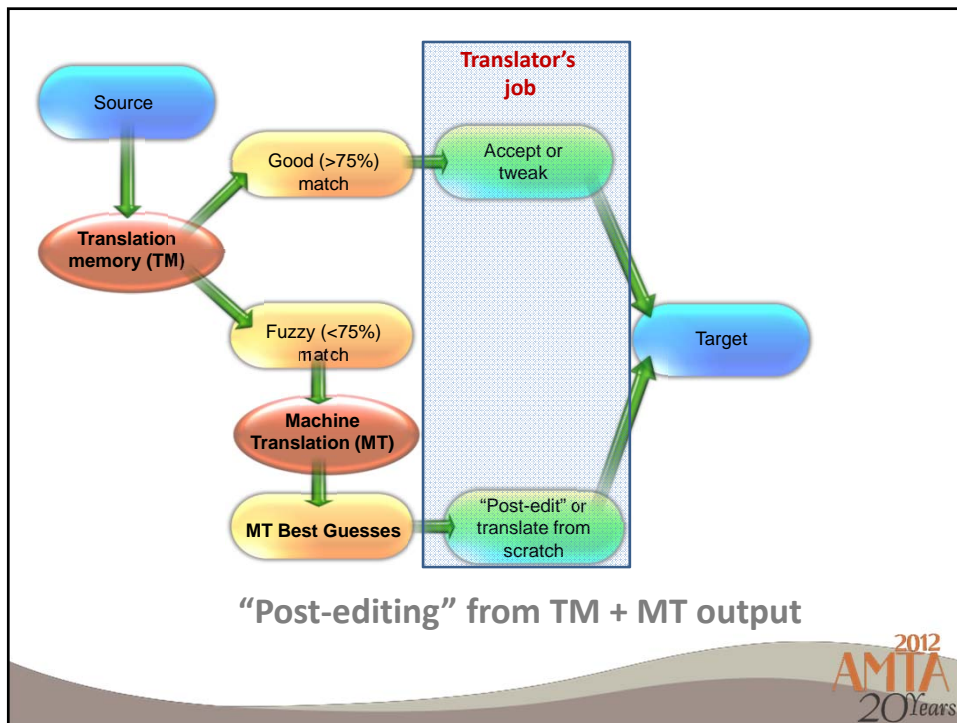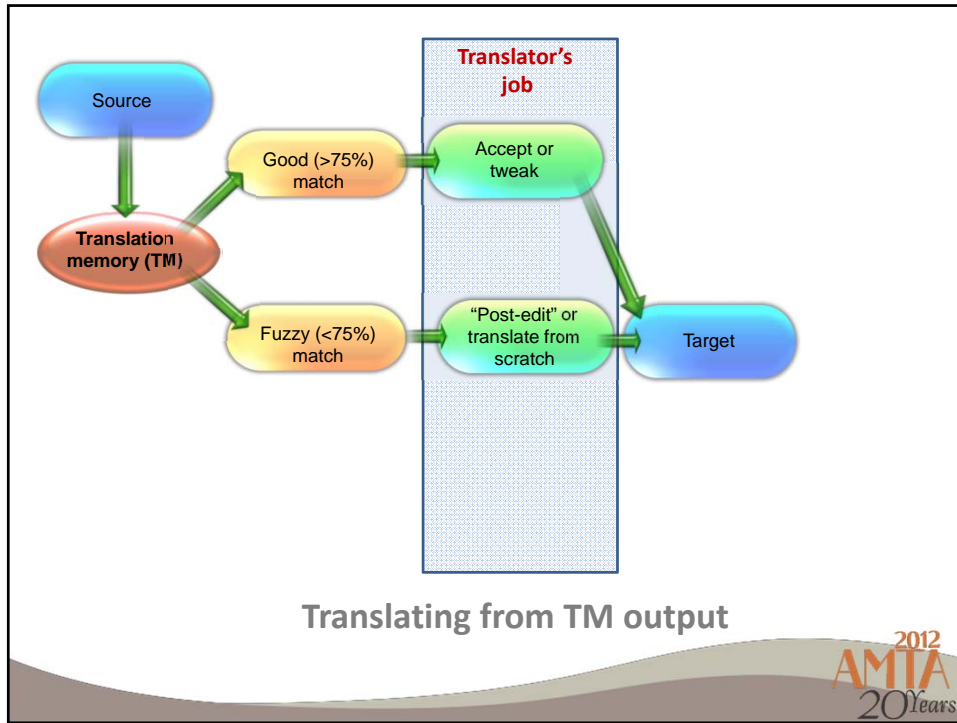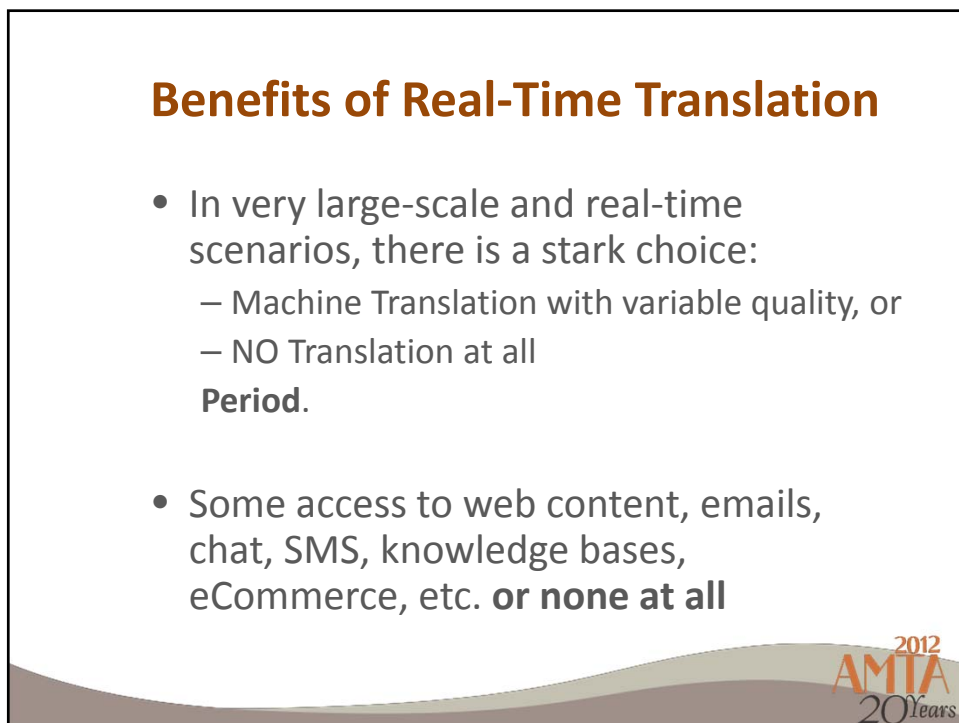
Benefits of MT for Localization

Machine Translation systems provide **faster and cheaper** translations than humans with translation memory tools alone.

• MT captures translator knowledge and effort in additional ways (memory vs. *reasoning*)

•MT requires more disciplined writing, which leads to additional efficiency and savings

•MT shifts the translator's workload from slower, more complex tasks (translation) to faster, simpler tasks (revising)

There are a range of different scenarios for translation automation, with and without MT.

33

2012
AMTA
20 Years

Total *effort* **(time * cost)** for the same 1,000,000-word project

Area = amount of effort

Cost (USD)

$ 250,000
$ 240,000
$ 230,000
$ 220,000
$ 210,000
$ 200,000
$ 190,000
$ 180,000
$ 170,000
$ 160,000
$ 150,000
$ 140,000
$ 130,000
$ 120,000
$ 110,000
$ 100,000
$ 90,000
$ 80,000
$ 70,000
$ 60,000
$ 50,000
$ 40,000
$ 30,000
$ 20,000
$ 10,000
$ -

No tools

TM only

MT+

TM+MT

TM+customized MT

TM+MT
+better source

TM+MT+much better source

MT

1 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500

Delivery time (person-days)

2012
AMTA
20 Years

Benefits of MT for Localization

Time	Internal benefits	Market benefits
Delivery time 4 times faster or more	Much more flexibility Shorter launch schedule	More sales opportunities Better user experience
Volume		
4 (or more) times more content in the same period	Scalability	Better user experience
Consistency		
More consistent terminology use	Better indexing and search	Better user experience
More consistent writing	Better indexing and search	Better user experience
Lower operating costs		
Generally 50% lower	More funds for product improvements	More sales opportunities Better user experience
More languages		
Less translation effort per language	Scalability	More sales opportunities Better user experience

35

How does MT save time and money in Localization?

Task analysis
Translation includes different activities, each with different **speeds** and **costs**

Translators:

- **Translate** from scratch other sentences (non-matches)
- **Revise** translations that are worth fixing. ("fuzzy" matches)
- **Approve** translations that are correct. ("perfect" matches)
- **Skip** sentences that have already been translated. ("ICE" matches)

Different tools divide "translation" into these activities in different ways

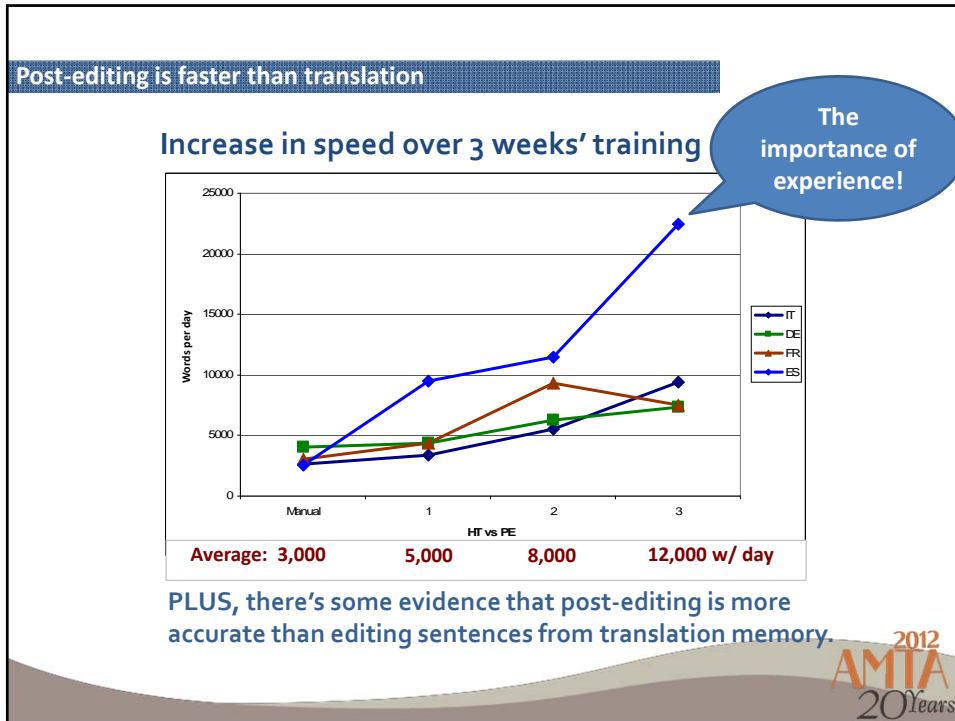**Important assumption:**
Output quality is *the same* in all scenarios.

Sample project
1,000,000 words

wds/day	Activity	%	cost
2,000 wds/day	Translate	50%	17¢/wd
8,000 wds/day	Revise	25%	10¢/wd
12,000 wds/day	Approve	15%	3¢/wd
50,000 wds/day	Skip	10%	1¢/wd

Total time: 293 person-days
Total cost: USD $115,963

37

Translation Workflow Scenarios
Progressive automation

Words
1,000,000

	Translate with no tools	TM only New content using existing TMs	Customized MT only	TM + MT	TM + MT with extra Customization	TM + MT Better source	TM + MT Much better source	Raw MT
Total cost (USD)	$ 170,000	$ 115,963	$ 93,670	$ 81,453	$ 69,935	$ 61,235	$ 43,090	$ 5,000
Cost per word (USD)	0.17	0.12	0.09	0.08	0.07	0.06	0.04	0.01
% change		-32%	-19%	-13%	-14%	-12%	-30%	

38

19

Post-editing is faster than translation

Increase in speed over 3 weeks' training

The importance of experience!

Average: 3,000   5,000   8,000   12,000 w/ day

PLUS, there's some evidence that post-editing is more accurate than editing sentences from translation memory.

OK, you convinced me. I want one. Now.

Which MT system should I choose?

How *not* to choose MT

The usual (mis)steps:
- Hear salesperson say how great product X is
- Ask lots of questions about "quality", speed, interfaces, cost
- Get evaluation versions of product X and others
- Translate some of your documents
- Ask translators about "quality" of translation
- **Get puzzled about poor output quality**
- **Decide not to use MT**

Outcomes
- Wasted time, effort, money
- Little understanding, little learning
- Negative reputation for MT
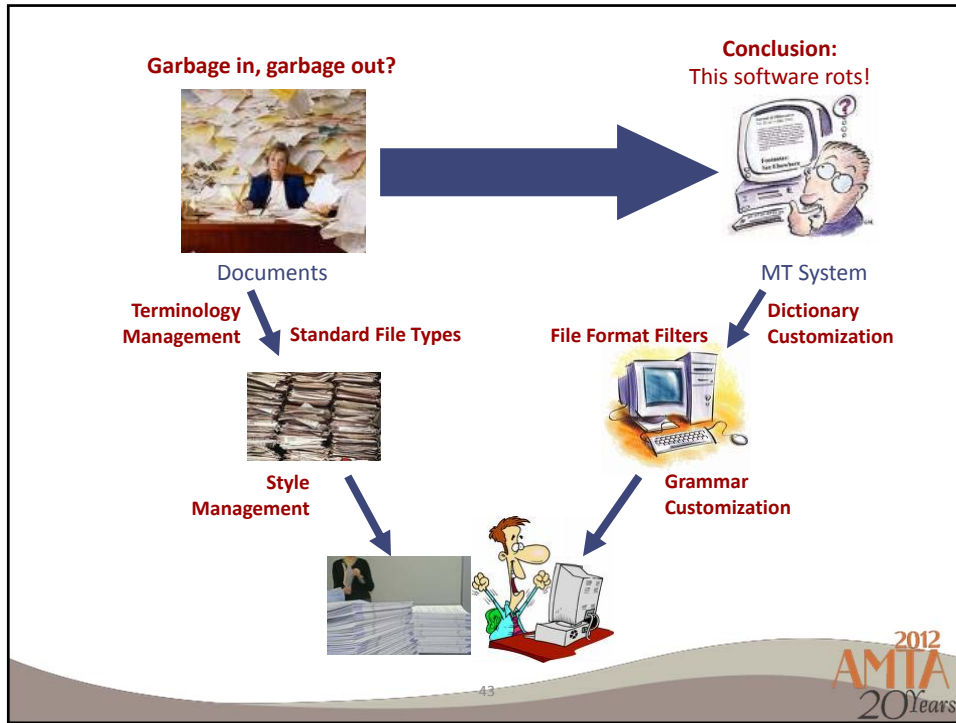

That's like evaluating a car…

Outcomes

I'll take it for a drive…

Conclusion:
"These cars are junk!"

Garbage in, garbage out?

Documents

Conclusion:
This software rots!

MT System

Terminology Management   **Standard File Types**   **File Format Filters**   **Dictionary Customization**

Style Management   **Grammar Customization**

MT, prepared poorly

What certification FSC?

The FSC - Forest Stewardship Council is a not governmental agency with headquarters in Germany, with world-wide performance, that it all regulates practical of handling of impact reduced in the forests of the world.

Forest Handling according to FSC is based on the three pillars of sustentation: correct, beneficial and ambiently socially economically viable. What it allows the withdrawal of the wood of a less impactante form for the environment, but income-producing for the society.

The certification is a voluntary process, in which the forest is evaluated by an independent agency, the certifier, who verifies the fulfilment of ambient questions, economic and social that is part of the Principles and Criteria of the FSC.

In January of 2007, the Brazil Log conquered certification FSC. Our furniture that possesss stamp FSC comes from areas of forest handling certifyd FSC that carry through the forest inventory, identifying to the forest species gifts, its dimensions and its geographic reference.

If to desire to get more information on the FSC, has access the site www.fsc.org

MT, prepared better

What FSC certification?

The FSC - Forest Stewardship Council is a non-governmental agency with headquarters in Germany, with world-wide activity, that all regulates the practices of low-impact management in the forests of the world.

Forest Management according to FSC is based on the three pillars of sustainability: environmentally correct, beneficial for society and economically viable. What it allows the harvesting of the lumber of a more sustainable way for the environment, but income-producing for the society.

The certification is a voluntary process, in which the forest is evaluated by an independent agency, the certifier, who verifies the fulfilment of environmental, economic, and social requirements that they are part of the Principles and Criteria of the FSC.

In January of 2007, Tora Brasil earned FSC certification. Our furniture that possesss FSC approval comes from FSC-certified managed forests that carry out the forest inventory, identifying ocurring forest species, its dimensions and its geographic reference.

If you want more information on the FSC, to go to the site www.fsc.org

Options for MT

MT Services Google, Bing, SDL/Language Weaver	MT SDKs Systran, PROMT, Open Source
PROs:	**CONs:**
• Ready Now (except SDL/Language Weaver) • Good Coverage of Languages • Good Translation Usability • Proven, Large-scale Deployment • Halo Effect	• Require Development • Less Coverage of Languages • Good Translation Usability after Training • Large-scale Deployments only with Systran • No Halo Effect • Open Source has fewer tools; requires more development
CONs:	**PROs:**
• Crucial Business Dependency • Issues with Data Privacy • High Cost per MB in Production; Bing is cheaper • Google: No Control over Translation Usability; Bing is better • Google: No Control of Coverage of your Content; Bing is better	• No Crucial Business Dependencies • No Issues with Data Privacy • Low Cost per Language in Production • Good Control over Translation Usability • Good Control of Coverage of your Content

Wait a second!
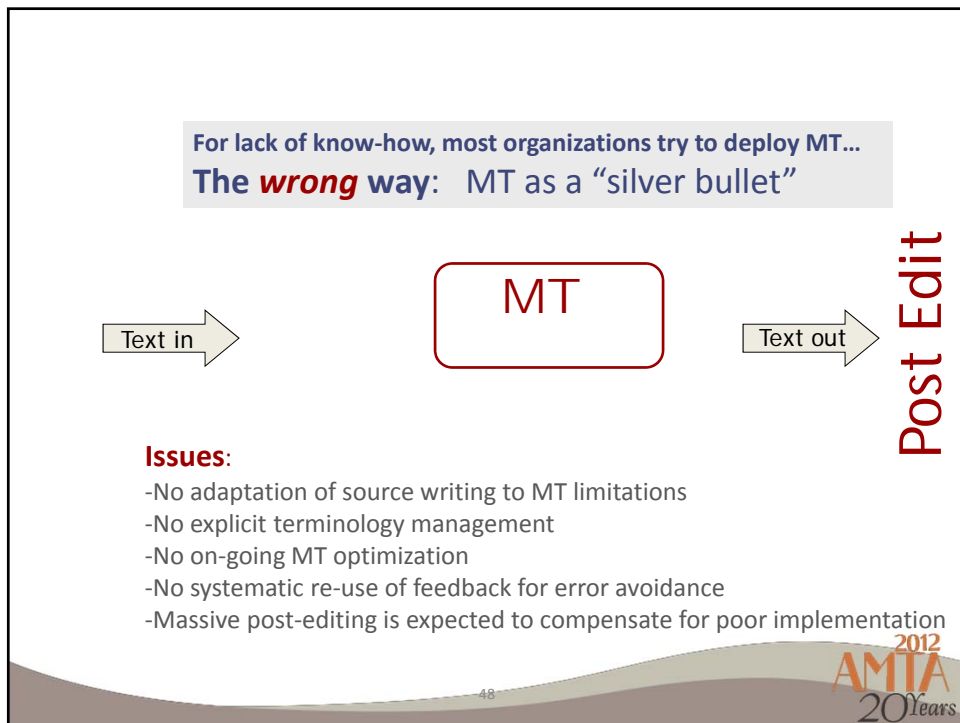**I can just use free MT from Google Translate**
(or Babblefish, Bing Translator, etc.).
That'll save me lots of money!
:) :)
Right?

Determinants of Translation "Quality"

Input Sentence

Pre-Processing

Dictionary Coverage and Quality

Corpora Coverage and Quality

①

②

MT

Translation Memory

Translator Quality

Reviewer Quality

Coverage

Sub-string Match | Full-string Match
Unknown sentences | Known sentences
Variable Quality | Human Quality

Translated Sentence

③ **User**

For lack of know-how, most organizations try to deploy MT...
The *wrong* way: MT as a "silver bullet"

Text in

MT

Text out

Post Edit

Issues:
-No adaptation of source writing to MT limitations
-No explicit terminology management
-No on-going MT optimization
-No systematic re-use of feedback for error avoidance
-Massive post-editing is expected to compensate for poor implementation

48

The *right* way:
Step 1. Optimize processes *without* MT

Write for readability

Text in

Pre-processing

Normalize1
Filter with TM
Normalize2
Dic Prep

Translate by hand

Post-processing

Normalize3

Text out

Analyze errors

Approach:
-Create infrastructure for on-going optimization
-Accumulate know-how
-Use feedback and communication to prevent future errors

49

2012
AMTA
20 Years

The *right* way:
Step 2. Add MT

Write for readability

Text in

Pre-processing

Normalize1
Filter with TM
Normalize2
Dic Prep

MT

Translate by hand

Post-processing

Normalize3

Text out

Post Edit
Analyze errors

Approach:
-MT accelerates existing effective processes
-MT does not make up for lack of effective processes
-Optimization know-how is the competitive advantage
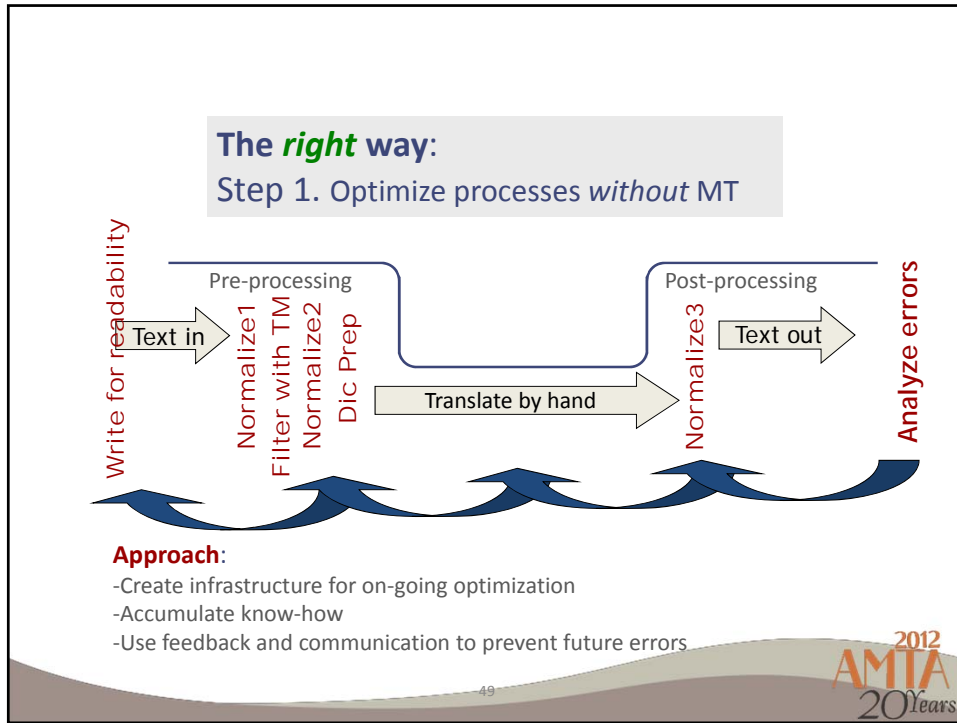
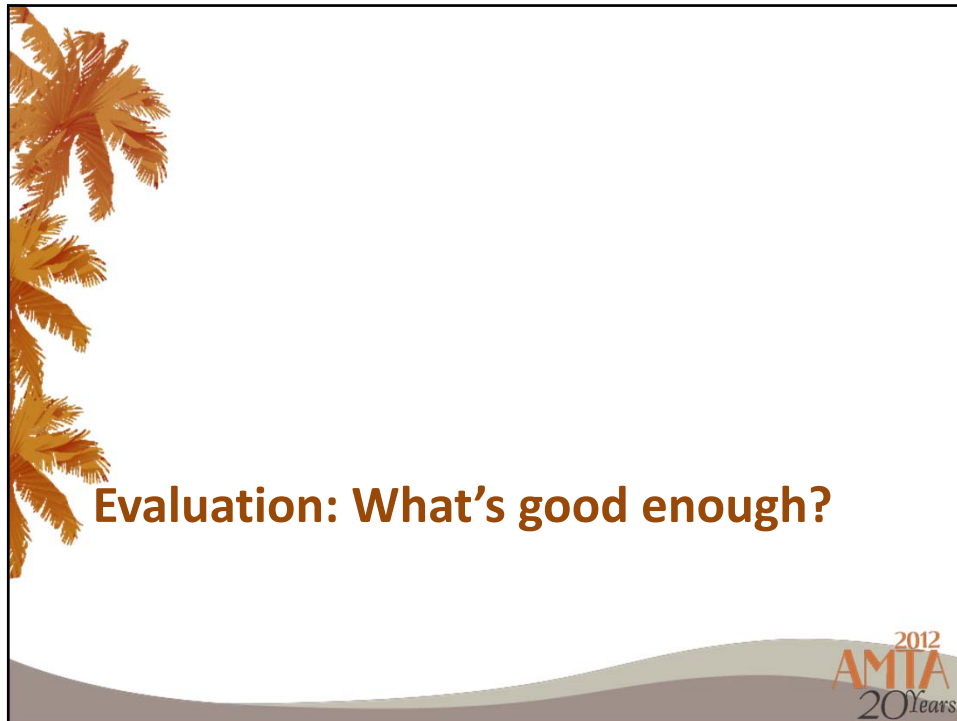50

2012
AMTA
20 Years

Evaluation: What's good enough?

Can *humans* really translate?

In Tambaba no neckness will be punished

A northeastern beach full of misteries, exciting views and no human touch forests, is an obliged topic on brasilians shoreline News. Tambaba, at Conde city is known all over the world because it was the first regional beach purposed to neckness. In 1989 the beach were titled as one of the ten more beautiful beaches. I'ts one of the few northeastern ecologics places, where the humam absense contributed for the conservation. The beach is rounded dor big trees and in all area – a bit more than one kilometer – many, naturals polls, large palm trees and big "FALÉSIAS".

From a human-translated tourist brochure
about the Brazilian state of Paraíba

52
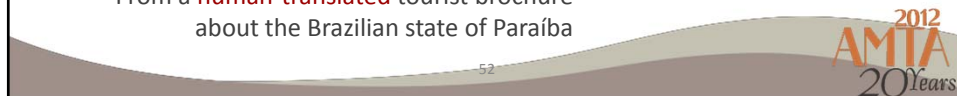
Can *humans* really translate?

- **The flattening of underwear with pleasure is the job of the chambermaid.** (In a Yugoslavian hotel)

- **You are invited to take advantage of the chambermaid.** (In a Japanese hotel)

- **Ladies, leave your clothes here and spend the afternoon having a good time.** (In a Rome laundry)

- **We offer special cocktails for the ladies with nuts.** (In a Tokyo bar)

- **Ladies are requested not to have children in the bar.** (In a Norwegian cocktail lounge)

- **Specialist in women and other diseases.** (In the office of a Roman doctor)

- **The manager has personally passed all the water served here.** (In an Acapulco hotel)

- **Daily plate -- shrimp in spit.** (In a Brazilian restaurant)

Can *humans* really translate?

- **Visitors are expected to complain at the office between the hours of 9 and 11 A.M. daily.** (In a hotel in Athens)

- **Take one of our horse driven city tours - we guarantee no miscarriages.** (In a Czechoslovakian tourist agency)

- **Please do not feed the animals. If you have any suitable food, give it to the guard on duty.** (at a Budapest zoo)

- **Cooles and Heates: If you want just condition of warm in your room, please control yourself.** (from a Japanese information booklet about using a hotel air conditioner)

- **To stop the drip, turn cock to right.** (In a Finnish washroom)

Conclusion:
Human translators are useless. (?!)

2: faucet consisting of a rotating device for regulating flow of a liquid
[syn: stopcock, turncock]

Translation "Quality"

So far, we've only used *translators'* criteria for quality.

What do *end users* notice and not notice? **What are *their* criteria for quality?**

- Should translators "fix" errors in the source documents?
- Should they reorganize source documents?

Correct source/target equivalence is still a question of *trust*, not of measurement.

Quality of target document **=**
f (quality of *source* document +
quality of target sentences)

Dimensions of Information Quality

- Content quality (relevant, complete, accurate information)
- Design quality (easy to find and maintain information)
- Linguistic quality (easy to understand information)
  - Term consistency; Stylistic simplicity
- Process quality (cost, consistency, reliability, etc.)

Translation "Quality" depends on ALL of these...

2012 AMTA 20 Years

Why is MT output so bad?

-Source issues-
- Poor writing in the source text
- Formatting issues in the source text

-Mismatch issues-
- Terms and expressions of the source that are not in the MT dictionary
- Sentence types of the source that are not covered by the MT system

-MT issues-
- Incorrect word sense chosen
- Incorrect sentence structure analysis

2012 AMTA 20 Years

Why is MT output so bad?

Technology is designed and built for optimal performance *in specific conditions* (ex: paved road, competent driver, correct fuel)

> *Even a wonderful, brand-new BMW looks like junk when it's tested outside the design specs.*

Using technology outside of its "comfort zone" requires adaptation.

Action Items
- Understand MT's "comfort zone"
- Assess which adaptations are necessary:
    - Input
    - People
    - Process
    - Technology

57

MT's "comfort zone"

Adaptations by writers	MT is designed for:	Adaptations of MT
Manage terminology and vocabulary explicitly	**Familiar (to the system) words and phrases**	Add words and phrases to dictionary
Manage writing style explicitly	**Familiar (to the system) sentence types**	Extend grammatical coverage; couple translation memory
Manage writing style explicitly	**Literal, predictable meanings**	Extend semantic coverage; couple translation memory
Use standard file formats	**Standardized file formats**	Add filters and converters
Write to minimize post-editing	**Post-editing**	Extend system performance to minimize post-editing
		The adaptations converge
	Very fast processing	
	Very large volumes	
	Good quality	

58

Assess the Adaptations
that you need

Input
- Make writing more translatable
- Standardize file formats

People
- Train writers
- Train/hire post-editors
- Train/hire MT operator(s)

Action Items
- Assess adaptations in more detail
- Estimate deployment effort

Process
- Develop pre- and post-processing tools
- Develop metrics

Technology
- Customize/train MT

59

Evaluation

How can I tell if MT is working?

Evaluation of translation – even traditional translation – is a complicated process!

There are MANY ways to evaluate translation:
- Automated metrics
- Human evaluation metrics
- Task-based metrics

Choose a metric that actually tells you what you need to know!

Automated Metrics

Evaluation

Automated Metrics
(BLEU, Edit Distance, NIST, Meteor, PER, etc.)

Advantages:
- Quick and inexpensive (if you have what you need to run them)
- Objective and repeatable
- Easy to show how improvement over baseline
- Very helpful when used appropriately

Disadvantages:
- Require reference translations
- False positives and negatives are possible
- Do not directly consider the rendering of source meaning in target
- Don't provide a "warm and fuzzy" feeling for the translation
- Everyone wants them to be more than they are

Evaluation: BLEU Score

BLEU (Bilingual Evaluation Understudy) is an algorithm for calculating the n-gram correspondence between the MT output for a given source sentence and one or more human translations of that sentence.

- Scores ranges from 0 to 1
- 1 indicates that the MT corresponds perfectly with the reference translations
- Test bed should be more than 1,000 sentences (and their translations) that were NOT used in the training of the system
- BLEU was designed to speed up the development of SMT.

We believe that BLEU will accelerate the MT R&D cycle by allowing researchers to rapidly home in on effective modeling ideas....

[BLEU] correlates highly with human judgments by averaging out individual sentence judgment errors over a test corpus *rather than attempting to divine the exact human judgment for every sentence: quantity leads to quality.*

From "BLEU: a Method for Automatic Evaluation of Machine Translation" (2002)
K. Papineni, S. Roukos, T. Ward, and W. Zhu

Evaluation: BLEU Score (Appropriate use)

BLEU Score

	MT System 1 (Baseline)	MT System 1 (Customized)
EN>DE (Chat)	0.1929	0.3006
EN>ES (Chat)	0.3945	0.474
EN>ZHCN (Chat)	0.4439	0.5401
EN>DE (Doc)	0.3321	0.3549
EN>ES (Doc)	0.4842	0.5387
EN>ZHCN (Doc)	0.4853	0.5235

BLEU is valuable for comparing iterations of the same MT system with the same test set.

## Evaluation: BLEU Score (Inappropriate use)

BLEU Score

	MT System 1	MT System 2
EN>DE (Chat)	0.3312	0.3006
EN>ES (Chat)	0.5493	0.474
EN>ZHCN (Chat)	0.2469	0.5401
EN>DE (Doc)	0.4359	0.3549
EN>ES (Doc)	0.5417	0.5387
EN>ZHCN (Doc)	0.4344	0.5235

BLEU is *not* particularly good for comparing two different MT systems (even with the same test set) and is *not at all* appropriate for comparing the quality of different language directions.

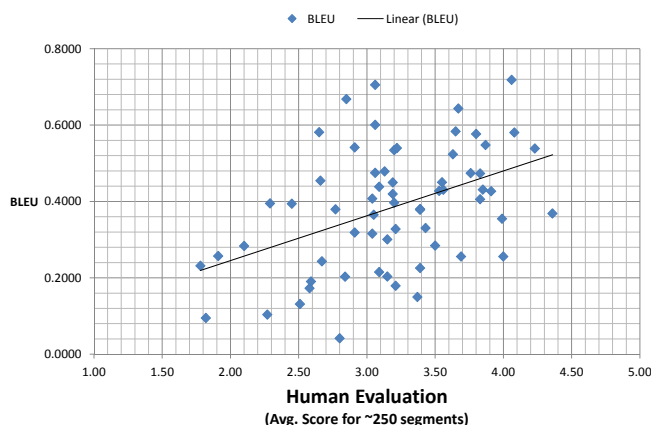## Automated Evaluations: BLEU Scores

BLEU Scores
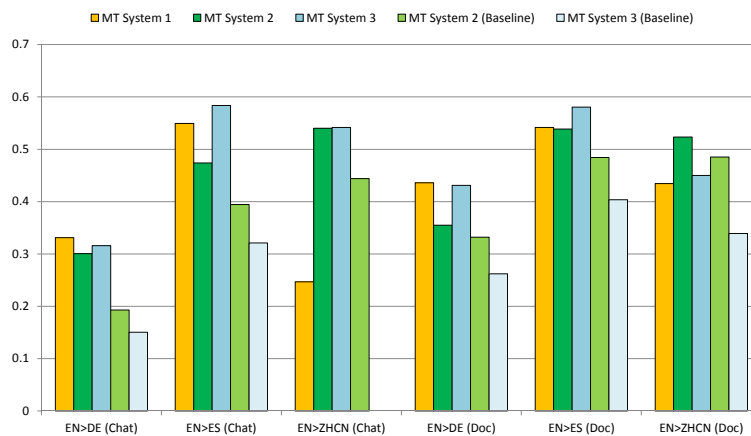
(BLEU test bed: 2500 segments not included in training material)

Automated Evaluations: BLEU Scores

Correlation of BLEU Score to Human Evaluation

BLEU test bed: 2500 segments not included in training material
Human evaluation: 10% random sampling of BLEU test bed

Evaluation: Edit Distance

Edit Distance (also called Translation Error Rate) is an algorithm for calculating the minimum number of edits that have to be made to a string (the MT output) to make it match another string (the reference translation).

- The lower a score is, the better.
- A score of 0 indicates that no changes have to be made to match one string to the other.
- Edit Distance is particularly helpful in predicting post-editing efficiency.
- But does not offer information on what kinds of changes are required.

Evaluation: Edit Distance
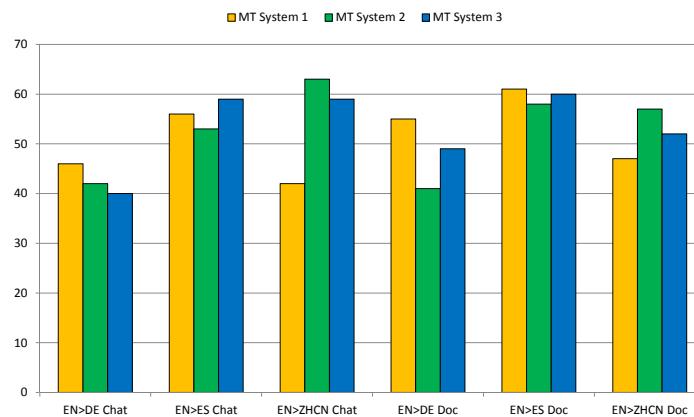
Language	MT System 1	MT System 2	MT System 3
EN>DE Chat	54%	58%	60%
EN>ES Chat	44%	47%	41%
EN>ZHCH Chat	58%	37%	41%
EN>DE Doc	45%	59%	51%
EN>ES Doc	39%	39%	40%
EN>ZHCH Doc	53%	40%	48%

Automated Evaluations: Edit Distance

Edit Distance
(lower score is better)

(Edit Distance test bed: 2500 segments not included in training material)

Human Evaluation

Evaluation

Human Evaluation

Advantages:

- Measures the rendering of source meaning in target
- Can provide information on type, severity, and frequency of error
- Does not require a reference translation
- Provides an invaluable "reality check" on a human level

Disadvantages:

- Time-consuming and expensive
- Existing processes vary from company to company
- Often developed for traditional translation
- Subjective, despite efforts to the contrary

Evaluation: Human Assessment

Human Assessment is still the gold standard for testing the accuracy, correctness, and comprehensibility of a translation.

After all, translations are produced for people, so asking a person what they think is the appropriate "user acceptance test".

But tests have to be designed carefully!
- To avoid bias
- To generate valuable information

Points to consider:
- How much data should be tested?
- How should the data be collected?
- How should the evaluators judge the translations?
  - On a point scale?
  - Pass/Fail?
  - With clarification of types of errors?
- What kind of skills does an evaluator need?
- How many evaluators should you use?
- How should the data be presented?

Human Evaluation Scale

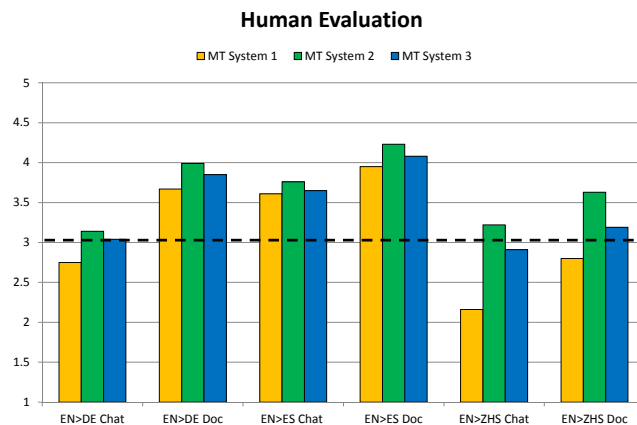| Ratings	Descriptions of Ratings
5 – Excellent	The information was translated clearly and with appropriate grammar, vocabulary, and style.
4 – Very Good	There may be minor errors in the translation, but the meaning of the original is very clear.
3 – Good Enough	There are errors in the translation, but the meaning of the original is reasonably clear.
2 – Not good enough	Errors in grammar, vocabulary and style make the meaning of the original difficult to understand.
1 – Poor	Fundamental errors in grammar and vocabulary prevent conveyance of the meaning of the original.
0 – System Failure	This score is for those cases when the system produces output that cannot be judged on the 1-5 scale. For example: the output is in Chinese characters although it is supposed to be in French; or an entire sentence is inappropriately left untranslated. If the translation is recognizable as the target language, it should not receive this rating.

Evaluation: Human Assessment

ItemNo	Segments for Evaluation	Rating
148	**May I know which navisphere software you are able to view?**	
	¿ Me podría indicar que el software Navisphere, podrá ver ?	3
	¿Puedo saber qué software Navisphere puede ver?	4
149	**because few of them are not customer accessible.**	
	Porque algunos de ellos no son accesibles a los clientes.	5
	porque algunos de ellos no son productos que el Cliente puede acceder a él.	4
150	**For inq utility: please navigate through Home > Support > Product and Diagnostic Tools > INQ Utility.**	
	Para utilidad INQ: Navegar por favor hogar > Soporte > herramientas de diagnóstico y productos > utilidad de INQ.	4
	Para utilidad inq : por favor navegar a través de Inicio > Soporte > Herramientas de productos y diagnóstico > de programa de utilidad.	3
151	**inq download requires ftp site access**	
	INQ requiere acceso al sitio FTP	4
	descargar inq requiere de acceso al sitio FTP	4
152	**Is this not accesible?**	
	¿Esto es no accesible?	5
	¿ No es accesible?	5

Evaluation: Human Assessment

(Human Assessment test bed: 250 randomly selected segments from BLEU/ED Test Suite)

Evaluation: Human Assessment

Evaluator Analysis: En>De Document Content

Score	MT System 1			MT System 2			MT System 3		
	Eval 1	Eval 2	Eval 3	Eval 1	Eval 2	Eval 3	Eval 1	Eval 2	Eval 3
0	0	0	0	0	0	0	0	0	0
1	4	3	3	2	2	2	1	2	1
2	13	29	71	6	11	31	5	20	44
3	60	85	60	34	71	66	44	75	72
4	99	52	54	111	70	68	129	65	62
5	74	81	62	97	96	83	71	88	71

Score Type	MT System 1			MT System 2			MT System 3		
	Eval 1	Eval 2	Eval 3	Eval 1	Eval 2	Eval 3	Eval 1	Eval 2	Eval 3
Average	3.90	3.72	3.40	4.18	3.99	3.80	4.06	3.87	3.63
Median	4.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	4.0
Mode	4.0	3.0	2.0	4.0	5.0	5.0	4.0	5.0	3.0

Evaluation: Sometimes a confusing picture!

BLEU Score	MT System 1	MT System 2	MT System 3
EN>DE Chat	1	3	2
EN>ES Chat	2	3	1
EN>ZHCN Chat	3	2	1
EN>DE Doc	1	3	2
EN>ES Doc	2	3	1
EN>ZHCN Doc	3	1	3

Edit Distance	MT System 1	MT System 2	MT System 3
EN>DE Chat	1	2	3
EN>ES Chat	2	3	1
EN>ZHCN Chat	3	1	2
EN>DE Doc	1	3	2
EN>ES Doc	1	3	2
EN>ZHCN Doc	3	1	2

Human Evaluation	MT System 1	MT System 2	MT System 3
EN>DE Chat	3	1	2
EN>ES Chat	3	1	2
EN>ZHCN Chat	3	1	2
EN>DE Doc	3	1	2
EN>ES Doc	3	1	2
EN>ZHCN Doc	3	1	2

Kinds of MT Systems

Methods of Machine Translation

Building a translation based on what is known about language

MT

Finding a translation based on analysis of translated material

Rules-based MT

Data-driven MT

Direct — Transfer — Interlingua

EBMT — SMT — TM

Syntactic — Semantic

Word-based — Phrase-based

Hybrid

Parallel — Sequential — Who knows?

Rules-based Machine Translation

Source Language Input

Analysis → Transfer → Synthesis

Source Language Lexicon & Grammars

Source >Target Language Lexical &Structural Rules

Target Language Lexicon & Grammars

Target Language Output

S
NP₁ NP₂
SL representation

S
NP₁ PP
représentation P NP₂
de TL

Rules-based Machine Translation

Cons	Pros
• Relatively costly to develop	• More predictable output
• Costly to add new languages	• Modest hardware requirements
• Manual customization	• Precise grammatical customization
	• Does well with most grammatical aspects of language (tense, aspect, number, case, agreement)
• Often less fluent translation	• Few errors of omission

Statistical Machine Translation

1. Calculate all of the possible moves.

2. Throw away all but the best one.

Statistical Machine Translation

$$\hat{e} = \arg\max_{e} \ p(e)p(f|e)$$

Given source string (*e*), the translation (*ê*) is the string that is produced by the Translation Model that is most likely to occur in the target language, according to the Language Model.

- *p*(f|e) is the **Translation Model**
  - A bilingual table of source words and collocations with many associated translations and their frequency
  - Assigns higher probabilities to target language strings that occur frequently in translations of sentences that contain the source string.
  - **The Translation Model proposes translation candidates.**

- *p*(e) is the **Language Model**
  - A monolingual table of target language words and collocations and their frequency of occurence
  - Provides the "likeliness" or probability of a given **phrase** (or n-grams) in a target corpus.
  - Provides the likeliness of target language strings (or n-grams) occurring in juxtaposition.
  - **The Language model selects best options.**

Statistical Machine Translation

Pros	Cons
• Relatively inexpensive to develop	• Less predictable output
• Easy to add new languages*	• Massive hardware requirements
• Automated customization *	• Imprecise customization
	• Has trouble with some grammatical aspects of language (tense, aspect, number, case, agreement)
• Often more fluent output	• Errors of omission

* Given the availability of adequate and appropriate training material.

Hybrid MT (we hope!)

Rule-based MT	Hybrid MT	Statistical MT
~600 words per second		~200 words per second
Better with word order		Issues with word order
Better with sentence structure		Issues with sentence structure
Issues choosing phrasing and stylistics		Better choosing phrasing and stylistics
Targeted customization		Global customization
Many tools for targeted customization		Few tools for targeted customization
More complex customization from existing translations		Simple, efficient training from existing translations
All plug into different content management systems		
Hard to build for new languages		Easy to build for new languages
Generally less expensive		For the moment, more expensive

86

Wrap-up

Wrap up

Many factors pressure us to localize more, better, faster, and cheaper

- Use MT to leverage one translator action into many, many changes in the translated output

- Use MT to emphasize cheaper, faster translator activities

- Use MT for cost reduction and increased throughput

- Use editing feedback to improve MT

- Every investment in more consistent, more readable source documents yields huge returns for localization

- **Don't go it alone – get help to choose and deploy MT**

- **Educate *all* your stakeholders about MT, continuously**

88

Four Components of High-quality MT Services

1. Normalization of Input
2. Training and Customization of Translation Engines
3. Good Translation Engines   *The Naïve View*
4. On-going Monitoring and Improvement of Output Quality

2012
AMTA
20 Years

Thanks for your attention!

Mike Dillinger
- President-elect of AMTA
- Computational Linguist @ eBay
-

Jay Marciano
- Director of Real-Time Translation, Lionbridge

- **Mike.Dillinger @translationOptimization.com**
- **Jay.Marciano @lionbridge.com**

2012
AMTA
20 Years