An introduction to Machine Translation

Mike Dillinger, PhD
Principal, Translation Optimization Partners
mike@translationOptimization.com

Translation Optimization Partners

Overview

- The problem: industrial-scale translation
- FAQs: what's MT?
  - Can machines really translate?!
  - Can we fire our translators now?
- Limitations: what MT can't do
  - Why is the output so bad? What is MT good for?
- Return on investment: Benefits of MT
  - How much does MT cost?
  - How can we convince our bosses to buy MT?
- Workflow: MT in action
  - Why buy MT if it's free on the internet?
  - What other kinds of translation automation are there?
  - How do we use it?
- Kinds of MT Systems
  - Rule-based, statistical, and hybrid

The Problem: Industrial-scale translation

Houston, we have a problem.

Communication is the lifeblood of business.
Without communication, business can't happen.

- Communication with clients
- Internal communication

But now our clients and operations are **global**

It's too hard, too time consuming, and too expensive to re-write everything from scratch in each language.

So, translation is *inescapable*.

And it's not just 20 or 30 pages, eh?

Why MT?

We need *industrial-scale* translation, part 1

There are more products = ***more*** **content**

The products are more complex = ***more*** **content**

The products have more uses = ***more*** **content**

Products change more rapidly = ***more*** **content**

Operations are more complex = ***more*** **content**

Content is used in more places = ***more*** **translation**

Why MT?

We need *industrial-scale* translation, part 2

Product development is faster = ***faster*** translation
Time to market is faster = ***faster*** translation
Support has to be faster = ***faster*** translation
Organizations have to be flexible = ***faster*** translation

We need *industrial-scale* translation, part 3

Oh, and by the way,

You can't spend any more money to get all of this done!
=***cheaper*** translation

"Industrial scale" means…

More Better Faster Cheaper

More Better Faster Cheaper

More Better Faster Cheaper

More Better Faster Cheaper

A Perfect Storm

The practice of global communication is falling apart:

a) Goals **Publish** documents	**60% or more** are not used
b) Scale Easier access to more **documents**; language *pairs*	Information **overload**; crisis of confidence; a **few** languages
c) Process *write + translate + use + support* are all **independent**	Each **interferes** with the other

The problems

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

Goals

We need **industrial-scale** translation processes: more, better, faster, cheaper

FAQs: What's MT?

What's MT?

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

- **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!
- **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.
- **We already use machine translation from Trados, right?**
  - Trados is one good use of *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*".

What's MT?

- **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 :)
- **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 needs is very, very different from what you see on the web.

"MT will have a negative impact on my brand"

Your site translated without MT

Your site translated *with* MT

Faster
Cheaper
More consistent

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)
  - 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

Introduction to MT

Limitations: What MT can't do

Translation Optimization Partners

Translation "quality"

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

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

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

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

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

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.)

How good?

Which MT system should I choose?

How *not* to evaluate MT

The usual (mis)steps:

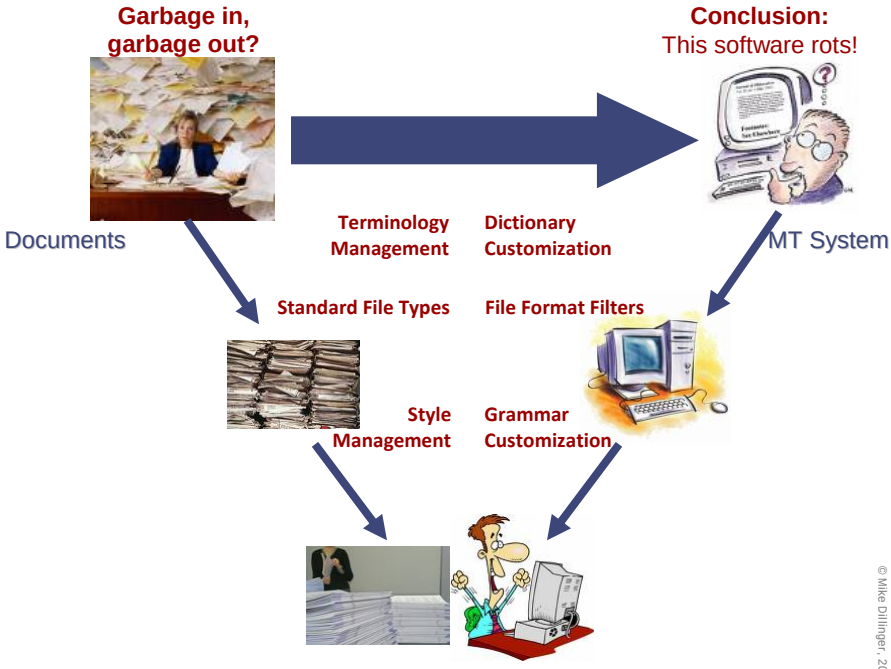
- Hear salesperson say how great product X is
- Ask lots of questions about "quality", speed, interfaces, cost
- Get an evaluation versions of product X and others
- Translate some of your documents
- Ask translators about "quality" of translation
- **Get puzzled about poor output quality**
- **Decide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

Outcomes

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

Conclusion:

"These MT products are junk!"



Original source

Emerging Markets Take Record Share of World Equity (Update1)

By Michael Patterson and Laura Cochrane

July 3 (Bloomberg) -- Developing countries' share of worldwide equity value climbed to a record as the fastest-growing economies lured investors amid the first global recession since World War II.

The 22 nations <that were> classified as "emerging" by index provider MSCI Inc. comprised 24 percent of world market capitalization, up from 18 percent at the start of this year, the highest proportion since Bloomberg began compiling the data in 2003.

China shares surpassed $3 trillion yesterday for the first time since August, from $1.8 trillion at the end of 2008.

The increase signals growing confidence in developing countries as equity investors, spurred by interest-rate cuts and stimulus plans, redeploy cash after the worst U.S. losses last since the Great Depression.

The MSCI Emerging Markets Index rose 35 percent, beating a 2.9 percent advance in the MSCI World Index of developed economies and lifting the value of stocks to $8.6 trillion from $5.1 trillion in 2008

"Everyone is trying to jump on that bandwagon," said Nicholas Field, who helps manage about $11 billion in emerging- market stocks at Schroders Plc in London.

Raw Google (customization not possible), original source

Emerging Markets Take Record Part du World Equity (Update1)

By Michael Patterson et Laura Cochrane

3 juillet (Bloomberg) - Développer la part des pays dans le monde entier la valeur a atteint un record de la plus forte croissance économique a attiré les investisseurs au milieu de la première récession mondiale depuis la Seconde Guerre mondiale.

Les 22 pays classés comme "émergents" de l'indice MSCI Inc fournisseur comprend 24 pour cent de la capitalisation boursière mondiale, contre 18 pour cent au début de cette année, la proportion la plus élevée depuis Bloomberg a commencé la compilation des données en 2003.

La Chine partage dépassé $ 3 billion, hier, pour la première fois depuis le mois d'août, à partir de $ 1,8 billions à la fin de 2008.

L'augmentation des signaux de plus en plus confiance dans les pays en développement que les investisseurs, stimulée par des taux d'intérêt des coupures et des plans d'incitation, de redéployer en espèces après la pire des États-Unis pertes dernier depuis la Grande Dépression.

L'indice MSCI Emerging Markets Index a augmenté de 35 pour cent, en battant l'avance de 2,9 pour cent dans l'indice mondial MSCI des pays développés et la levée de la valeur des stocks à 8,6 billions de $ 5.1 billions de $ en 2008

Tout le monde tente de sauter sur cette aventure», a déclaré Nicholas Field, qui permet de gérer environ 11 milliards de dollars en actions des marchés émergents à Schroders Plc à Londres.

Adapted source	Improved MT with adapted source and dic customization
Emerging Markets Take Record Share of World Equity (Update1)	Les marchés émergent prennent partie record de situation nette modiale (Update1)
By Michael Patterson and Laura Cochrane	Par Michael Patterson et Laura Cochrane
July 3 (Bloomberg). Developing countries now have a record-breaking part of world equity. Their economies are growing quickly and they lured investors, even amid the first global recession since World War II.	Le 3 (Bloomberg) juillet. Les pays en voie de développement ont maintenant une partie record de situation nette modiale. Leurs économies croissent rapidement et ils ont leurré des investisseurs, même entre la première récession globale depuis seconde guerre mondiale.
The 22 nations that were classified as "emerging countries" by index provider MSCI Inc. comprised 24 percent of world market capitalization. This increased from 18 percent at the start of this year. This was the highest proportion since Bloomberg began compiling the data in 2003.	Les 22 nations qui ont été classifiées comme "pays émergents" par fournisseur de l'indice MSCI Inc. compris 24 pour cent de capitalisation de marché mondial. Cela a augmenté de 18 pour cent au début de cette année. C'était la plus haute proportion depuis que Bloomberg a commencé à compiler les données dans 2003.
China's part surpassed $3 trillion yesterday for the first time since August, from $1.8 trillion at the end of 2008.	La partie de Chine a dépassé $3 billion hier pour la première fois depuis août, de $1.8 billion à la fin de 2008.
The increase signals growing confidence in developing countries as equity investors. Spurred by interest-rate cuts and stimulus plans, they are redeploying cash after the worst U.S. losses since the Great Depression.	L'augmentation signale la confiance croissante au pays en voie de développement comme investisseurs des capitaux propres. Talonnés par les réductions de l'intérêt-taux et les plans du signal, ils redistribuent la liquidité après le plus mauvais U.S. pertes depuis la Grande Baisse.
The MSCI Emerging Markets Index rose 35 percent. This beat a 2.9 percent advance in the MSCI World Index of developed economies and lifted the value of stocks to $8.6 trillion from $5.1 trillion in 2008.	L'Indice MSCI des marchés émergents a augmenté 35 pour cent. Cela a battu un 2.9 pour cent avance dans l'Indice Mondiale MSCI d'économies développées et a soulevé la valeur de titres à $8.6 billion de $5.1 billion dans 2008.
"Everyone is trying to jump on that bandwagon," said Nicholas Field, who helps to manage about $11 billion in emerging-market stocks at Schroders Plc in London.	Tout le monde essaie de prendre ce train en marche, a dit Nicolas Field qui aide pour diriger approximativement $11 milliard dans les actions des marchés émergents à Schroder Plc à Londres.

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 that are not in the MT dictionary
- Sentence types that are not covered by the MT system

-MT issues-

- Incorrect word sense chosen
- Incorrect sentence structure analysis

What's MT good for?

- Draft translations of clean source texts
- Fast translations of low-value information
- Translation of very high-volume information
- Information extraction and decision support

Lessons learned

- 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

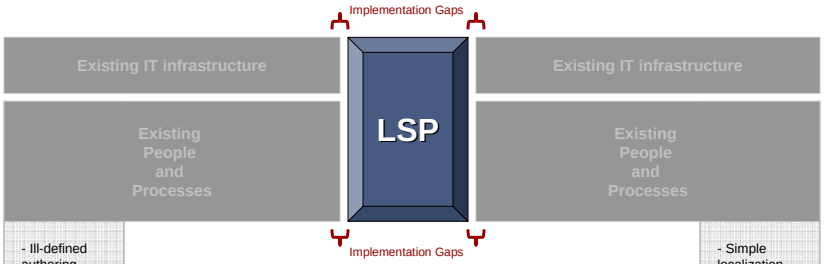
- Understand MT's "comfort zone"
- Assess which adaptations are necessary:
  - Input
  - People
  - Process
  - Technology

The "comfort zone" for MT

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
	Very fast processing	
	Very large volumes	
	Good quality	

The adaptations converge

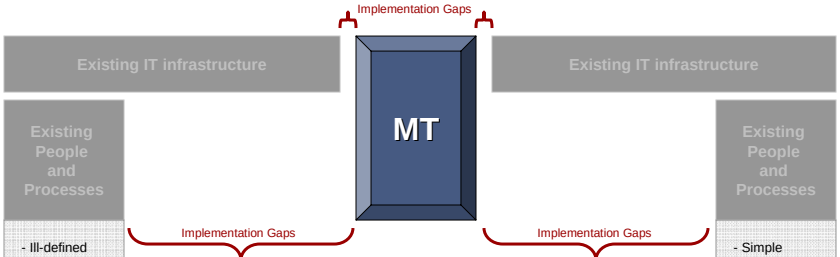
The *status quo*



- Ill-defined authoring processes
- Simple localization management processes
- Informal work flow definition
- Few quantitative metrics in use

- Simple localization management processes
- Informal quality control of target
- No error analysis

Why deploying MT seems difficult



- Ill-defined authoring processes
- Simple localization management processes
- Informal work flow definition
- Few quantitative metrics in use

- **term detection and management**
- **well-defined authoring processes**
- **systematic quality control of source**
- **document pre-processing**

- **post-editing**
- **systematic quality control of target**
- **error analysis for feedback**
- **document post-processing**

- Simple localization management processes
- Informal quality control of target
- No error analysis

Who can do this?
- in-house staff?
- new hires?
- outsourced?

ROI: Benefits of MT

ROI: Benefits of MT

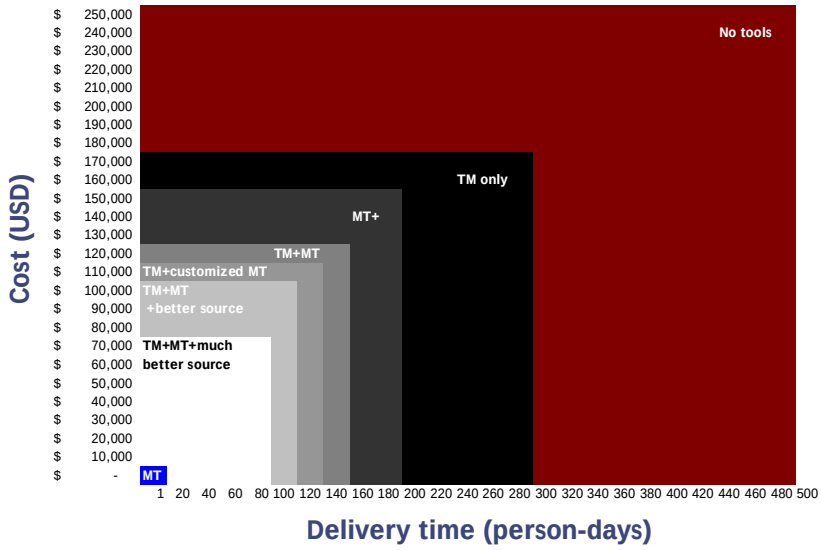
Machine Translation systems provide **faster and cheaper** translations than humans with translation memory tools alone.

- 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.

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

Area = amount of effort



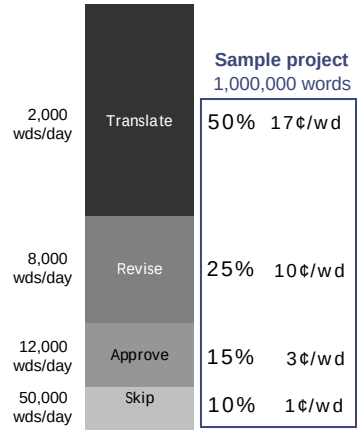
Benefits of MT

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 improvements	More sales opportunities Better user experience
More languages		
Less translation effort per language	Scalability	More sales opportunities Better user experience

How?

HOW does Translation Automation save time and money?

Scenario 1



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

Task analysis

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

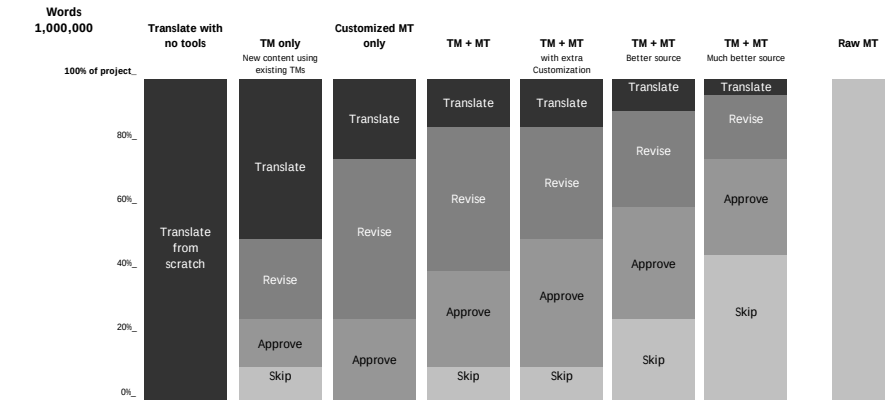
Translators:

- **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.

Translation Workflow Scenarios — Progressive automation



	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%**	
Lost Sales for Client (total)	$ 2,550,000	$ 1,377,634	$ 736,366	$ 572,472	$ 525,031	$ 394,181	$ 226,338	
Savings for Client (over TM only)			$ 641,269	$ 805,163	$ 852,603	$ 983,453	$ 1,151,297	
Lost Sales (per day)	$ 30,000							
Lost Sales (per month)	$ 900,000							
Sales per year	$ 10,800,000							

Translation Optimization Partners

A focus on price?

1,000,000 words

Cost per word	Total cost
0.30	$ 300,000
0.25	$ 250,000
0.20	$ 200,000
0.15	$ 150,000
0.10	$ 100,000
0.05	$ 50,000
0.00	:)
0.01	**$ 10,000**

$ 10,000,000 **Sales per year**
$ 27,397 **Sales per day**

Delivery time (days)	Lost sales
10	$ 273,973
30	$ 821,918
60	$ 1,643,836
90	$ 2,465,753
120	$ 3,287,671
150	$ 4,109,589
180	$ 4,931,507
7	**$ 191,781**

Or on delivery time?

Focusing on a lower vendor price per word is a recipe for *disaster*

How much does MT cost?

Evaluation system

Direct costs

Trial installation: US$1,000 for desktop version (or loaner from vendor)

Consultant for planning and training personnel

Indirect costs

Personnel time

Production system

Direct costs (initial installation)

Server: US$30,000 - $200,000 *per language pair*

Vendor or consultant services for MT customization/training

Consultant for planning and training personnel

(on-going costs)

Maintenance fee: ~20% of server price, per year

Indirect costs

Personnel time

What are these costs for? (1)

Input		
Monitor translatability	**Beneficial without MT**	
Convert to standard file formats	**Beneficial without MT**	
People		
Writers: Manage terminology and vocabulary explicitly	**Beneficial without MT**	
Writers: Manage writing style explicitly	**Beneficial without MT**	
Writers: Use standard file formats	**Beneficial without MT**	
Editors: Train to edit for MT input	**Beneficial without MT**	
Project Managers: Train to minimize post-editing and delivery time	**Beneficial without MT**	
MT operator: Train or hire	***	

What are these costs for? (2)

Process		
Develop pre- and post-processing tools and procedures	Beneficial without MT	
Develop evaluation metrics	Beneficial without MT	
Technology		
Add words and phrases to dictionary	Beneficial without MT	
Extend grammatical coverage; couple translation memory	***	
Extend semantic coverage; couple translation memory	***	
Add filters and converters	***	
Extend system performance to minimize post-editing	***	
Integration with existing systems	***	

Benefits of MT

Faster, cheaper translation:

- Better scalability
- Better capture and re-use of translator knowledge and effort
  - Complements TM
- Shifts translator workload to simpler tasks
- Promotes better writing

ROI calculator

Discussion

Questions?

Introduction to MT

Workflow: MT in action

Translation Optimization Partners

More output from each translator

Investment in tools: none

Word processor

Investment in tools: small

Translation Memory

Investment in tools: large

TM + MT

Industrial-scale translation

More output from each translator

One translator action yields:

- translation of 1 sentence, in one place, in one document, into one language
  1:1

- translation of *some* source sentences, in different places, in many documents, into many languages
  1:1,000

Translation Memory

- translation of *all* source sentences, in many documents, into many languages
  1:10,000

TM + MT

Leveraging one action into many.

More output from each translator

Next place?
Next document?

Translator's action	Result	Changes in work
Translate one segment in one document	**Manual Translation** One translated segment, in one place, in one document	-None
We need to capture and multiply translator effort = create "leverage"		
Translate one segment in one document	**With Translation Memory** The **whole** translated segment *in different places, in many documents* -**Not "portable"** to different domains **Some leverage, some revision**	-Skip some sentences -**Revise more** -Translate from scratch less
Translate one segment in one document	**With TM + Machine Translation** ***Pieces*** of the same translated segment *in different places, in many documents* -**Now "portable"** to different domains **More leverage, more revision**	-Skip more sentences -**Revise more** -Translate from scratch less

Focus

More from each translator

Focus on multiplying the results of translators' effort:

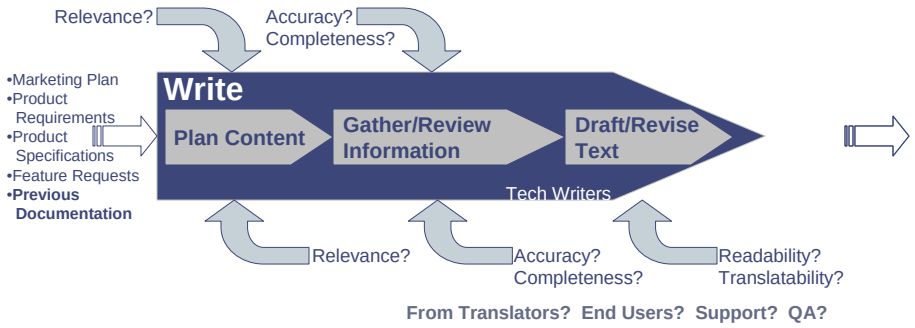
- Capturing effort is important
- Re-using effort is important
- Different tools have different limitations

- **Translation tools capture and re-use translator effort to improve scalability**
  - In different ways

Discussion

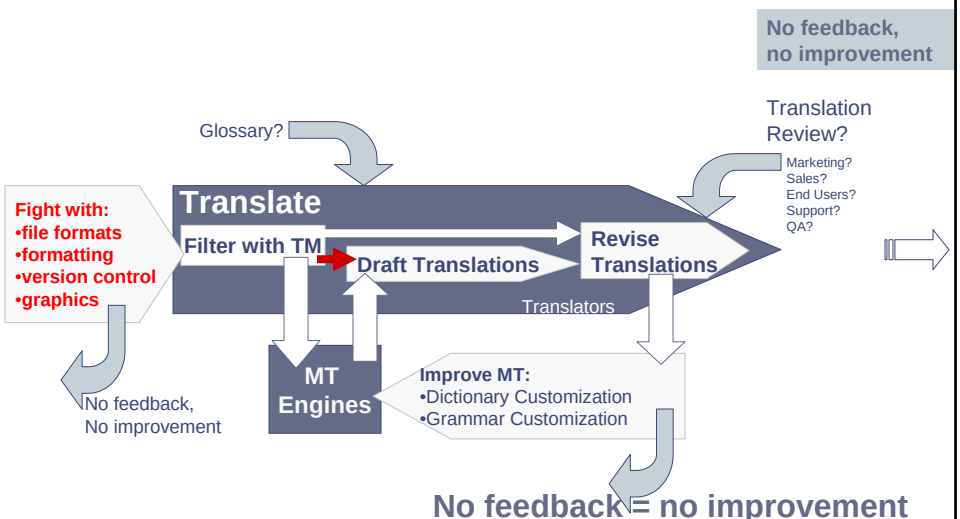
Questions so far?

The Content Supply Chain

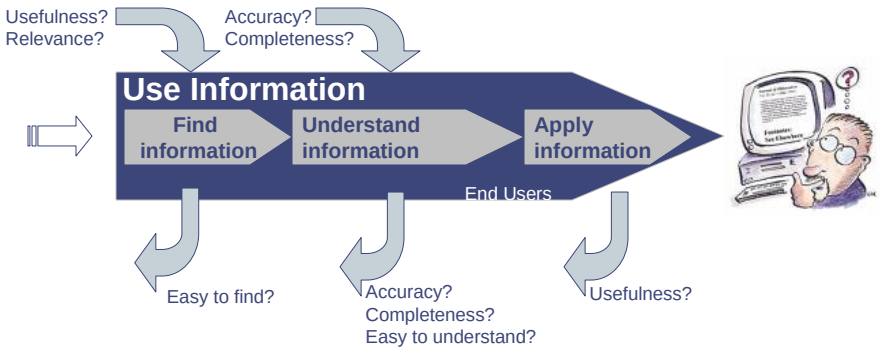


No feedback = no improvement

The Content Supply Chain



The Content Supply Chain

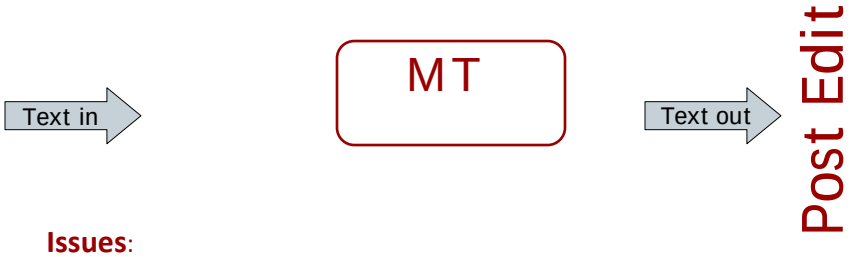


Wait!

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

For lack of know-how, most organizations try to deploy MT…

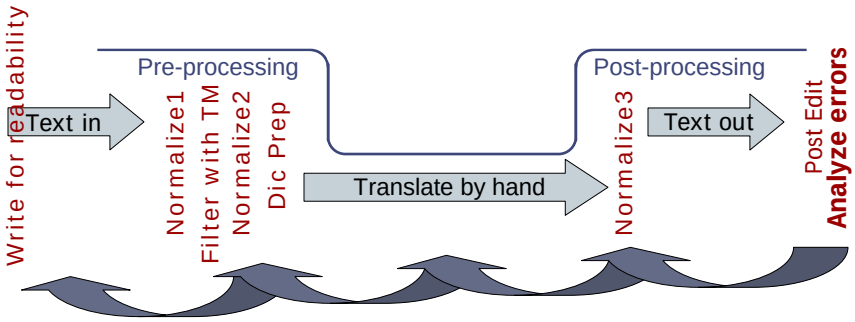
The *wrong* way: MT as a "silver bullet"



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

Translation Optimization Partners

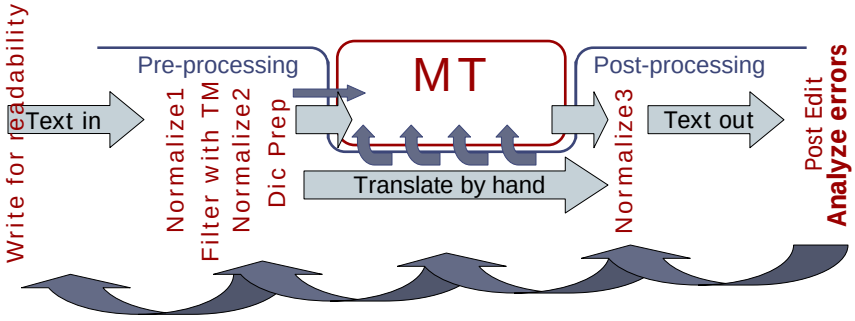
The *right* way: Step 1. Optimize processes *without* MT



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

Translation Optimization Partners

The *right* way: Step 2. Add MT



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

Translation Optimization Partners

Assess adaptations that are needed

Input

- Make writing more translatable
- Standardize file formats

People

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

Process

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

Technology

- Customize/train MT

Action items

- Assess adaptations in more detail
- Estimate deployment effort

Kinds of MT systems: rule-based, statistical, and hybrid

Kinds of MT systems — Rule-based MT

Rule-based MT		Why you should care
~600 words per second		Usually not a factor for localization
Better with word order	☑	Fewer complex edits
Better with sentence structure	☑	Fewer complex edits
Issues choosing phrasing and stylistics	☒	More edits about word choice
Targeted customization	☑	Can fix very specific errors and prepare the system for specific projects
Many tools for targeted customization	☑	Can fix very specific errors and prepare the system for specific projects
More complex to customize from existing translations	☒	On-going investment in system improvement
Hard to build for new languages	☒	May not be available for a language that you need, ex: long-tail strategies
Generally less expensive	☑	

Kinds of MT systems — Statistical MT

Statistical MT		Why you should care
~200 words per second		Usually not a factor for localization
Issues with word order	☒	More complex edits
Issues with sentence structure	☒	More complex edits
Better at choosing phrasing and stylistics	☑	Fewer edits about word choice
Global customization		Very convenient but
Few tools for targeted customization	☒	Hard to make specific changes
Simple, efficient training from existing translations	☑	Very convenient built-in feedback to reuse human translations
Easy to build for new languages	☑	But only if you have many existing translations
For the moment, more expensive	☒	

Kinds of MT systems

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

Action items

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

- Use tools to leverage **one** translator action **into many, many** changes in the translated output
- Use tools to emphasize cheaper, faster activities
- Use tools for cost reduction and increased throughput
- Use editing feedback to improve tools
- Every investment in more consistent, more readable source documents yields huge returns for localization

Wrap up

Learn more about MT

- Don't go it alone – **hire a consultant** to help choose and deploy MT
- Educate *all* your stakeholders about MT, continuously

We are independent translation automation consultants who help you to:

- Improve your strategic decision making and planning to get it right
- Understand your current situation and effective paths to reach your goals
- Troubleshoot your existing processes and tools to solve your immediate problems

About us

Translation Optimization Partners

Principal:
Mike Dillinger
mike@translationoptimization.com

Thanks for your attention.

Questions?

An Introduction to Machine Translation

Mike Dillinger, PhD
Principal, Translation Optimization Partners
mike@translationOptimization.com

AMTA 2010

Translation Optimization Partners